INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Multi-attribute Factorized Hidden Layer Adaptation for DNN Acoustic Models

Lahiru Samarakoon, Khe Chai Sim

National University of Singapore
lahiruts@comp.nus.edu.sg, simkc@comp.nus.edu.sg

Abstract

Recently, the Factorized Hidden Layer (FHL) adaptation is proposed for speaker adaptation of deep neural network (DNN) based acoustic models. In addition to the standard affine transformation, an FHL contains a speaker-dependent (SD) transformation matrix using a linear combination of rank-1 matrices and an SD bias using a linear combination of vectors. In this work, we extend the FHL based adaptation to multiple variabilities of the speech signal. Experimental results on Aurora4 task show 26.0% relative improvement over the baseline when standard FHL adaptation is used for speaker adaptation. The Multi-attribute FHL adaptation shows gains over the standard FHL adaptation where improvements reach up to 29.0% relative to the baseline.

Index Terms: Automatic Speech Recognition, Adaptation, Factorized Hidden Layers.

1. Introduction

DNNs, like all other machine learning techniques, are susceptible to performance degradations due to the mismatches between the training and testing conditions. The adaptation techniques are used to normalize these mismatch causing variabilities by transforming the model to match testing conditions or by augmenting the runtime features to match the model. Maximum a posteriori (MAP) [1] and maximum likelihood linear regression (MLLR) [2, 3] are commonly used to adapt GMM-hidden Markov model (HMM) systems. In addition, speaker adaptive training (SAT) has been applied to GMM-HMM systems [4, 5]. However, it is not possible to directly use GMM adaptation techniques for DNN adaptation due to the generative nature of GMMs and the difficulty in interpreting DNNs as meaningful structures as it is possible for GMMs. Therefore, DNN adaptation is challenging, especially when performed with a small amount of data in an unsupervised fashion. Adaptation for DNNs is important as it reduces error rates significantly [6, 7, 8, 9, 10, 11, 12].

In this paper, we extend our FHL based speaker adaptation method [13] to multiple variabilities of the speech signal. In addition to the standard affine transformation, an FHL contains an SD transformation matrix using a linear combination of rank-1 matrices and an SD bias using a linear combination of vectors. Therefore, the FHL adaptation learns a set of bases for speaker variability during training and interpolation weights are learned during adaptation. In Multi-attribute FHL adaptation, we learn these bases for noise, utterance and channel variabilities in addition to the speaker variability. We have evaluated the proposed method in the Aurora4 [14] noisy speech recognition task.

The rest of the paper is organized as follows. Section 2 describes the related work while in Section 3 we briefly review the FHL adaptation framework. In Section 4 we describe the proposed Multi-attribute FHL adaptation. The results are reported in Section 5 and we conclude our work in Section 6.

2. DNN Adaptation

Adaptation techniques for DNNs can be categorized into two broad approaches: test-only adaptation (simply refers to as adaptation), and adaptive training. Adaptation techniques can be categorized into 3 classes: linear transformation based adaptation, subspace or subset adaptation and regularized adaptation. Linear transformation based adaptation methods augment the original DNN model with a linear layer [15, 16, 17, 18, 19, 20]. In subspace or subset methods, the adaptation is performed to a subset of model parameters or on a pruned model [21, 22, 23, 24, 25, 26, 27, 28]. Regularization based methods perform adaptation conservatively by employing regularization into the adaptation criterion [8, 29]. Adaptive training for DNNs can be categorized into 3 classes: cluster adaptive training [30, 31], feature normalization [3, 32] and speaker-aware training.

In speaker-aware training (SaT), speaker features are provided during DNN training. Techniques like i-vectors [9, 10, 33, 6] and bottleneck features [34, 35] are commonly used to extract speaker features. The standard approach in SaT is to concatenate the acoustic features with the speaker features before DNN training. In that case, speaker information can be considered as a bias to the layer above as given below.

$$\mathbf{h}^l = \sigma(\mathbf{W}^l\mathbf{h}^{l-1} + \mathbf{b}^l + \mathbf{U}^l\mathbf{v}_s), \tag{1}$$

$\mathbf{v}_s$ is the speaker representation and $\mathbf{U}^l$ is the speaker representation transformation weight matrix for layer l, respectively.

3. Factorized Hidden Layers Adaptation

In this section, we briefly review the general formulation of the FHL adaptation. In the standard SaT where only an SD bias is used, all the phonemes of a speaker are adapted with a fixed bias which is not optimal. Therefore, in an FHL, in addition to the SD bias, an SD transformation is used as given below:

$$\mathbf{h}^l = \sigma(\mathbf{W}_s^l\mathbf{h}^{l-1} + \mathbf{b}_s^l) \tag{2}$$

where the SD transformation matrix, $\mathbf{W}_s^l$ is given by:

$$\mathbf{W}_s^l = \mathbf{W}^l + \sum_{i=1}^{|\mathbf{d}_s^l|} \mathbf{d}_s^l(i)\mathbf{B}^l(i) \tag{3}$$

where $\{\mathbf{B}^l(1), \mathbf{B}^l(2)..,\mathbf{B}^l(|\mathbf{d}_s^l|)\}$ is the set of basis for the SD transformation and $\mathbf{d}_s^l \in \mathbb{R}^{|\mathbf{d}_s^l|\times 1}$ is the SD interpolation vector. Similarly, the SD bias vector, $\mathbf{b}_s^l$, for hidden layer l is given by:

http://dx.doi.org/10.21437/Interspeech.2016-1233

$$\mathbf{b}_s^l = \mathbf{b}^l + \sum_{i=1}^{|\mathbf{v}_s^l|} \mathbf{v}_s^l(i)\mathbf{u}^l(i) \tag{4}$$

where $\{\mathbf{u}^l(1), \mathbf{u}^l(2)..,\mathbf{u}^l(|\mathbf{v}_s^l|)\}$ is the set of basis for the SD bias and $\mathbf{v}_s^l \in \mathbb{R}^{|\mathbf{v}_s^l|\times 1}$ is the SD interpolation vector.

Furthermore, in our model $\mathbf{B}^l(i)$ weight bases are constrained to be rank-1 matrices. This allows us to formulate the SD transformation as given below:

$$\begin{aligned}\mathbf{W}_s^l &= \mathbf{W}^l + \sum_{i=1}^{|\mathbf{d}_s^l|} \mathbf{d}_s^l(i)\mathbf{B}^l(i) \\ &= \mathbf{W}^l + \sum_{i=1}^{|\mathbf{d}_s^l|} \mathbf{d}_s(i)^l\mathbf{a}^l(i)\mathbf{b}^{l\top}(i) \\ &= \mathbf{W}^l + \mathbf{\Gamma}^l\mathbf{D}_s^l\mathbf{\Psi}^{l\top}\end{aligned} \tag{5}$$

where $\mathbf{B}^l(i) = \mathbf{a}^l(i)\mathbf{b}^{l\top}(i)$ and $\mathbf{D}_s^l \in \mathbb{R}^{|\mathbf{d}_s^l|\times|\mathbf{d}_s^l|}$ is a diagonal matrix ($\mathbf{D}_s^l = \text{diag}(\mathbf{d}_s^l)$) and $\mathbf{a}^l(i)$, $\mathbf{b}^l(i)$ are i-th column vectors for $\mathbf{\Gamma}^l$, $\mathbf{\Psi}^l$ respectively.

3.1. Training

The diagonality of $\mathbf{D}_s^l$ matrix allows us to initialize $\mathbf{d}_s^l$ with the speaker i-vector and train the system in the form given in equation 5. We refer to this step of FHL adaptation as the "Initialized" step.

3.2. Adaptation

During the adaptation, we estimate the SD coefficients for both training and testing speakers while keeping all other model parameters fixed. Therefore, for training speakers, the adaptation is performed in supervised fashion while for test speakers unsupervised adaptation is used. Furthermore, it is worth noting that we also adapt the speaker representations ($\mathbf{v}_s^l$) for SD biases. In this paper, the scenario where the speaker adaptation is performed only for testing speakers is refered as "Initialized + Adapt" step.

3.3. Optimization of the Bases

During this stage, the new representations that are learned for both training and testing speakers are used as the interpolation coefficients ($\mathbf{d}_s^l$) to learn a new set of bases. Then, for test speakers, additional adaptation step is performed in an unsupervised fashion. In this paper, we refer to this stage as "Optimized Bases (OB) + Adapt".

4. Multi-attribute FHL

In Multi-attribute FHL, in addition to the speaker variability, we extend the FHL adaptation to multiple variabilities of the speech signal like noise, channel and utterance. Multi-attribute FHL can be represented as below:

$$\mathbf{h}^l = \sigma(\mathbf{W}_{c_{1..n}}^l\mathbf{h}^{l-1} + \mathbf{b}_{c_{1..n}}^l) \tag{6}$$

where $c_1, c_2, ..., c_n$ are the different variabilities modelled by the Multi-attribute FHL and the multi-attribute transformation matrix, $\mathbf{W}_{c_{1..n}}^l$ is given by:

$$\mathbf{W}_{c_{1..n}}^l = \mathbf{W}^l + \sum_{j=1}^{n}\sum_{i=1}^{|\mathbf{d}_{c_j}^l|} \mathbf{d}_{c_j}^l(i)\mathbf{B}_{c_j}^l(i). \tag{7}$$

Similarly, the multi-attribute bias vector, $\mathbf{b}_{c_{1..n}}^l$, for hidden layer l is given by:

$$\mathbf{b}_{c_{1..n}}^l = \mathbf{b}^l + \sum_{j=1}^{n}\sum_{i=1}^{|\mathbf{v}_{c_j}^l|} \mathbf{v}_{c_j}^l(i)\mathbf{u}_{c_j}^l(i). \tag{8}$$

Since the bases are rank-1, the equation 7 can be formulated as given below:

$$\begin{aligned}\mathbf{W}_{c_{1..n}}^l &= \mathbf{W}^l + \sum_{j=1}^{n}\sum_{i=1}^{|\mathbf{d}_{c_j}^l|} \mathbf{d}_{c_j}^l(i)\mathbf{B}_{c_j}^l(i) \\ &= \mathbf{W}^l + \sum_{j=1}^{n}\sum_{i=1}^{|\mathbf{d}_{c_j}^l|} \mathbf{d}_{c_j}(i)^l\mathbf{a}_{c_j}^l(i)\mathbf{b}_{c_j}^{l\top}(i) \\ &= \mathbf{W}^l + \mathbf{\Gamma}^l\mathbf{D}_{c_{1..n}}^l\mathbf{\Psi}^{l\top}\end{aligned} \tag{9}$$

Therefore, in Multi-attribute FHL, attribute-specific bases are learned. As in FHL adaptation for speaker, we use the attribute-specific i-vector for training of the Initialized step. Then, we learn new representations only for the speaker variability by performing speaker adaptation for both training and testing speakers as discussed in Section 3.2. During the "optimized bases (OB)" step, we use this new speaker representations and i-vectors for other variabilities to learn a new set of optimized bases. Finally, only speaker representations are adapted for test speakers for "OB + Adapt" step.

4.1. Orthogonality Analysis

As shown in equation 9, in Multi-attribute FHL, we learn set of bases for each attribute. In this section, we investigate the variation of the orthogonality between different attribute bases at initialization and OB stages. To achieve this, we calculate the absolute cosine similarity between bases for attribute p and attribute q as given below:

$$Similarity = \frac{\sum_{i=1}^{|\mathbf{d}_{c_p}^l|}\sum_{j=1}^{|\mathbf{d}_{c_q}^l|}\frac{|c_p^l(i).c_q^l(j)|}{||c_p^l(i)||||c_q^l(j)||}}{|\mathbf{d}_{c_p}^l||\mathbf{d}_{c_q}^l|} \tag{10}$$

where $c_p^l(i) = \begin{bmatrix}\mathbf{a}_p^l(i)\\ \mathbf{b}_p^l(i)\end{bmatrix}$ and $c_q^l(j) = \begin{bmatrix}\mathbf{a}_q^l(j)\\ \mathbf{b}_q^l(j)\end{bmatrix}$. These $\mathbf{a}_p^l(i)$, $\mathbf{b}_p^l(i)$, $\mathbf{a}_q^l(j)$ and $\mathbf{b}_q^l(j)$ are as defined in equation 9. The orthogonality between bases of different attributes increases when the absolute cosine similarity defined in the equation 10 decreases.

5. Experiments

5.1. Experimental Setup

We use the Aurora4 multi-condition training set with 83 speakers for training and the development set with 10 speakers for validation. Aurora4 consists of recordings from 2 different

Table 1: *Word Error Rates (WER %) for FHL adaptation for speaker variability. The dimensionality of speaker representation is 100.*

Stage	WER (%)				
	A	B	C	D	Avg.
None	3.2	7.4	6.7	18.4	11.9
Initialized	2.6	6.7	5.4	16.7	10.6
Initialized + Adapt	2.4	5.8	4.2	14.0	9.0
OB + Adapt	2.5	5.6	4.3	13.9	8.8

channels, namely Channel-1 (single microphone) and Channel-2 (18 different microphones). The test set is divided into four subsets: A (clean speech + Channel-1), B (noisy speech + Channel-1), C (clean speech + Channel-2) and D (noisy speech + Channel-2). The results are reported on the test set with 8 speakers.

First, we extracted the MFCC features from the speech using a 25-ms window and a 10-ms frame-shift. We obtain the LDA features by first splicing 7 frames of 13-dimensional MFCCs and then projecting downwards to 40 dimensions using LDA. A single semi-tied covariance (STC) transformation [36] is applied on top of the LDA features. The GMM-HMM system for generating the alignments for DNN training is trained on top of these 40 dimensional LDA+STC features.

The DNN-HMM baseline is trained on the LDA+STC features that span a context of 11 neighboring frames. Before being presented to the DNN, CMVN is performed on the features globally. To train the network, layer-wise discriminative pretraining is used. The initial DNN has 7 sigmoid hidden layers with 2048 units per layer, and 2031 senones as the outputs. All the DNNs are trained to optimize the cross-entropy criterion with a mini-batch size of 256. We use CNTK [37] to train the DNNs. The Kaldi toolkit [38] is used to built the GMM-HMM systems and for the i-vector extraction. The i-vectors are trained on top of the same 40 dimensional LDA+LTC features. The UBM consist of 128 Gaussians. All the decodings are performed with the pruned 5K tigram language model of WSJ0. We do not change the training alignment during the process. We use the alignments from the GMM-HMM system. This allows us to compare the gains.

5.2. Results

The Table 1 reports the performance of the FHL adaptation for speaker variability. The stage "None" stands for the baseline system (11.9% WER). We only present the results for the best combination of FHL layers in this paper, where bottom 6 layers are FHLs. In addition, only the first FHL has a SD bias connected. As it can be seen clearly, the average performance improves considerably with every stage of the FHL adaptation. The best performance of 8.8% is achieved when the optimized bases model is adapted for test speakers which is a 26% relative improvement over the baseline.

The Table 2 presents the results for Multi-attribute FHL adaptation for speaker and utterance attributes. For fair comparison with the FHL adaptation which uses 100 dimensional speaker i-vector, we estimate speaker i-vectors of 85 dimensions and utterance-level i-vectors of 15 dimension. This allows us to keep the same number of model parameters between these two methods. In addition, this also has a similar configuation to FHL model where bottom 6 layers are Multi-attribute FHLs. Similarly, the bias component is only present in the first

Table 2: *WER % for Multi-attribute FHL with speaker and utterance level information. The dimensionality of speaker and utterance representations are 85 and 15 respectively.*

Stage	WER (%)				
	A	B	C	D	Avg.
Initialized	2.6	6.5	5.8	16.5	10.5
Initialized + Adapt	2.4	5.6	4.3	13.7	8.8
OB + Adapt	2.5	5.5	4.0	12.9	8.4

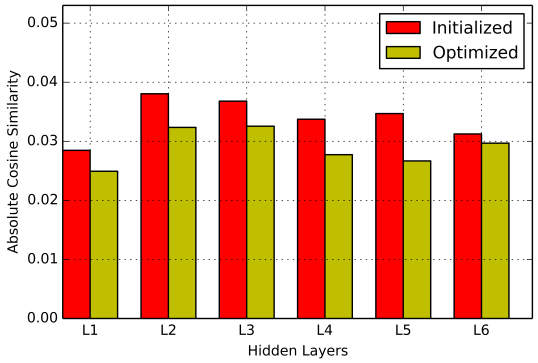

Figure 1: The absolute cosine similarity between the speaker and utterance bases for the model trained with speaker and utterance attributes.

Multi-attribute FHL. As it can be clearly seen, Multi-attribute FHL with speaker and utterance attributes outperforms the FHL adaptation at every stage. The best performance of 8.4% is reported when the optimized bases model is adapted for the test speaker which is a 29.0 % relative improvement over the baseline.

Figure 1 shows the absolute cosine similarity of the speaker and utterance bases for initialized and optimized bases stages. As it can be seen clearly, for all layers, absolute cosine similarity for optimized bases are smaller to that of the initialized bases. Therefore, during the optimization stage the speaker and utterance subspaces represented by the bases have become more orthogonal. This improvement of orthogonality allows to represent a wider range of speaker and utterance attribute combinations during the adaptation step. We believe this contributes to the performance gain we observe from "initialized + adapt" (8.8) to "OB + adapt" (8.4) stages.

In addition to the clean speech, the Aurora4 corpus has 6 noise types. All are recorded from two different channels. Therefore, for the next set of experiments we estimated i-vectors for 14 different conditions from the training data. The condition i-vectors that are estimated from the training data is used with both development and test data. Therefore, during testing we only required the condition label for each utterance, which is more practical than estimating the condition i-vector from the test data. The results for Multi-attribute FHL with speaker and these condition attributes are given in Table 3. As it can be clearly seen, Multi-attribute FHL with speaker and condition attributes slighly outperforms the FHL adaptation at every stage. The best performance of 8.7% is reported when the optimized bases model is adapted for the test speaker. However, the Multi-attribute FHL with speaker and utterance outperforms

Table 3: *WER % for Multi-attribute FHL with speaker and condition level information. The dimensionality of speaker and condition representations are 85 and 15 respectively.*

Stage	WER (%)				
	A	B	C	D	Avg.
Initialized	2.9	6.7	5.6	16.4	10.5
Initialized + Adapt	2.5	5.8	4.3	13.9	8.9
OB + Adapt	2.3	5.8	4.1	13.5	8.7

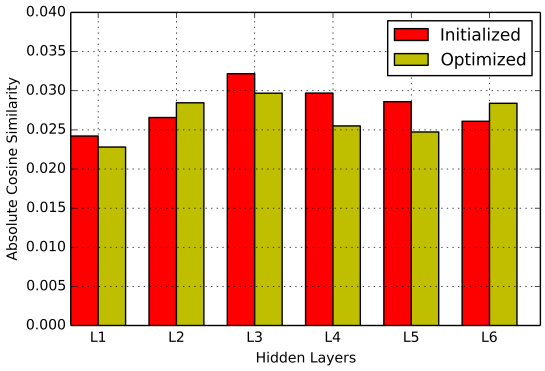

Figure 2: The absolute cosine similarity between the speaker and condition bases for the model trained with speaker and condition attributes.

the speaker and condition Multi-attribute FHL.

Figure 2 investigates the orthogonality of the speaker and condition subspaces. As it can be seen, the orthogonality reduces in layer 2 during the optimization of the bases. This may explain the smaller gains compared to the Multi-attribute experiments with speaker and utterance attributes. Moreover, in Aurora4 there is a mismatch of noise levels between training and testing data. The noise is added to the training data in the range of 10 - 20 dB with 1 dB steps while for test data the range is 5 -15 dB with 1 dB steps. Since, we are borrowing the condition i-vector from the training data, this mismatch can reduce the performance improvements. In addition, there is no mismatch between the clean condition i-vector for training and testing data. Therefore, this may explain why the speaker and condition Multi-attribute FHL model outperforms the speaker and utterance Multi-attribute FHL model for set A in the "OB + Adapt" stage.

Next, we present the results for Multi-attribute FHL with speaker, utterance and condition level information in Table 4. Interestingly, the final performance for this model (8.7) is worse than the that of Multi-attribute FHL model with speaker and utterance attributes (8.4). This can also be due to the noise level mismatch in training and testing data as discussed above. Furthermore, the utterance and condition i-vectors may not be complementary.

To alleviate the issue of noise mismatch for condition i-vectors, we replace the condition attribute with the channel attribute in the next set of experiments. To obtain the channel i-vector, we only used the clean speech recorded from that channel. The results are presented in Table 5. As it can be seen clearly, this combination of attributes outperformed the Multi-attribute FHL for speaker, utterance and condition attributes

Table 4: *WER % for Multi-attribute FHL with speaker, utterance and condition level information. The dimensionality of speaker, utterance and condition representations are 85, 15 and 15 respectively.*

Stage	WER (%)				
	A	B	C	D	Avg.
Initialized	2.7	6.6	5.6	16.2	10.4
Initialized + Adapt	2.4	5.8	4.2	13.9	8.9
OB + Adapt	2.3	5.6	4.1	13.5	8.7

Table 5: *WER % for Multi-attribute FHL with speaker, utterance and channel level information. The dimensionality of speaker, utterance and channel representations are 85, 15 and 15 respectively.*

Stage	WER (%)				
	A	B	C	D	Avg.
Initialized	2.6	6.6	5.4	16.1	10.3
Initialized + Adapt	2.5	5.6	4.2	13.5	8.7
OB + Adapt	2.5	5.5	4.1	13.0	8.4

at every stage. However, the performance gain is similar to the Multi-attribute FHL model with speaker and utterance attributes. This can be due to the fact that in Aurora4 we have only two different representations for channel attribute. Therefore, compared with the utterance and speaker attributes the channel attribute provides less information for normailzation.

6. Conclusion and Future Work

In this paper, we have extended the previously proposed Factorized Hidden Layer (FHL) method for speaker adaptation to normalize multiple variabilities of the speech signal. In an FHL, low-dimensional speaker representations are used as interpolation weight vectors to linearly combine a set of rank-1 matrices and a set of vectors to construct an SD transformation matrix and an SD bias, respectively. These matrices and vectors are learned adaptively with the speaker representations, which are initialized using the speaker i-vectors and later updated using unsupervised adaptation. In Multi-attribute FHL, in addition to the speaker variability, linear combinations of matrices and vectors are estimated to other attributes of speech like condition, channel and utterance using corresponding i-vector for each attribute as the interpolation weight vector. Experimental results on Aurora4 task showed 26.0% relative improvement over the baseline when standard FHL adaptation is used for speaker adaptation. The Multi-attribute FHL adaptation improved over the standard FHL adaptation where improvements reached upto 29.0% relative to the baseline.

As we have shown in our orthogonality analysis, the orthogonality between different attribute subspaces should be increased to further improve the performance of the Multi-attribute FHL method. Therefore, as future work, we are planning to include a constrain to the training objective to incorporate it. Moreover, it is also possible to estimate factorized i-vectors as proposed in [39] to obtain orthogonal repesentations for different attributes. We believe factorized i-vector representations may improve the Multi-attribute FHL method over the standard i-vector representations.

7. References

[1] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[2] C. Leggetter and P. 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.

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

[4] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *ICSLP*, vol. 2. ISCA, 1996, pp. 1137–1140.

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

[6] L. Samarakoon and K. Sim, "Learning factorized transforms for speaker normalization," in *ASRU*. IEEE, 2015.

[7] ——, "On combining i-vectors and discriminative adaptation methods for unsupervised speaker normalization in DNN acoustic models," in *ICASSP*. IEEE, 2016.

[8] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *ICASSP*. IEEE, 2013, pp. 7893–7897.

[9] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, "I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription," in *ICASSP*. IEEE, 2014, pp. 6334–6338.

[10] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *ASRU*. IEEE, 2013, pp. 55–59.

[11] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in *ICASSP*. IEEE, 2013, pp. 7942–7946.

[12] D. Yu and L. Deng, *Automatic Speech Recognition - A Deep Learning Approach*. New York: Springer London, 2015.

[13] L. Samarakoon and K. Sim, "Factorized hidden layer adaptation for deep neural network based acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[14] N. Parihar, J. Picone, D. Pearce, and H. Hirsch, "Performance analysis of the Aurora large vocabulary baseline system," in *EUSIPCO*. IEEE, 2004, pp. 553–556.

[15] V. Abrash, H. Franco, A. Sankar, and M. Cohen, "Connectionist speaker normalization and adaptation," in *Eurospeech*. ISCA, 1995, pp. 2183–2186.

[16] B. Li and K. Sim, "Comparison of discriminative input and output transformation for speaker adaptation in the hybrid NN/HMM systems," in *INTERSPEECH*. ISCA, 2010, pp. 526–529.

[17] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, and T. Robinson, "Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system," in *INTERSPEECH*. ISCA, 1995.

[18] J. Trmal, J. Zelinka, and L. Müller, "Adaptation of a feedforward artificial neural network using a linear transform," in *Text, Speech and Dialogue*. Springer, 2010, pp. 423–430.

[19] 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 *INTERSPEECH*. ISCA, 2012.

[20] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. D. Mori, "Adaptation of hybrid ANN/HMM models using linear hidden transformations and conservative training," in *ICASSP*. IEEE, 2006, pp. 1189–1192.

[21] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *ASRU*. IEEE, 2011, pp. 24–29.

[22] S. Xue, H. Jiang, and L. Dai, "Speaker adaptation of hybrid NN/HMM model for speech recognition based on singular value decomposition," in *ISCSLP*. IEEE, 2014, pp. 1–5.

[23] J. Xue, J. Li, D. Yu, M. Seltzer, and Y. Gong, "Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network," in *ICASSP*. IEEE, 2014, pp. 6359–6363.

[24] K. Kumar, C. Liu, K. Yao, and Y. Gong, "Intermediate-layer DNN adaptation for offline and session-based iterative speaker adaptation," in *INTERSPEECH*. ISCA, 2015.

[25] S. Dupont and L. Cheboub, "Fast speaker adaptation of artificial neural networks for automatic speech recognition," in *ICASSP*. IEEE, 2000, pp. 1795–1798.

[26] J. Stadermann and G. Rigoll, "Two-stage speaker adaptation of hybrid tied-posterior acoustic models," in *ICASSP*. IEEE, 2005, pp. 977–980.

[27] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *ICASSP*. IEEE, 2013, pp. 7947–7951.

[28] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *SLT*. IEEE, 2014, pp. 171–176.

[29] Y. Huang and Y. Gong, "Regularized sequence-level deep neural network model adaptation," in *INTERSPEECH*. ISCA, 2015.

[30] T. Tian, Q. Yanmin, Y. Maofan, Z. Yimeng, and K. Yu, "Cluster adaptive training for deep neural network," in *ICASSP*. IEEE, 2015, pp. 4325–4329.

[31] C. Wu and M. J. F. Gales, "Multi-basis adaptive neural network for rapid adaptation in speech recognition," in *ICASSP*. IEEE, 2015, pp. 4315–4319.

[32] L. Lee and R. Rose, "Speaker normalization using efficient frequency warping procedures," in *ICASSP*, vol. 1. IEEE, 1996, pp. 353–356.

[33] A. Senior and I. Lopez-Moreno, "Improving DNN speaker independence with i-vector inputs," in *ICASSP*. IEEE, 2014, pp. 225–229.

[34] H. Huang and K. Sim, "An investigation of augmenting speaker representations to improve speaker normalization for DNN-based speech recognition," in *ICASSP*. IEEE, 2015, pp. 4610–4613.

[35] S. Kundu, G. Mantena, Y. Qian, T. Tan, M. Delcroix, and K. Sim, "Joint acoustic factor learning for robust deep neural network based automatic speech recognition," in *ICASSP*. IEEE, 2016.

[36] M. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[37] D. Yu, A. Eversole, M. Seltzer, K. Yao, Z. Huang, B. Guenter, O. Kuchaiev, Y. Zhang, F. Seide, H. Wang *et al.*, "An introduction to computational networks and the computational network toolkit," Tech. Rep. MSR, Microsoft Research, 2014, http://codebox/cntk, Tech. Rep., 2014.

[38] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The Kaldi speech recognition toolkit," in *ASRU*. IEEE, 2011.

[39] P. Karanasou, Y. Wang, M. Gales, and P. Woodland, "Adaptation of deep neural network acoustic models using factorised i-vectors," in *INTERSPEECH*. ISCA, 2014.