TOWARDS UTTERANCE-BASED NEURAL NETWORK ADAPTATION IN ACOUSTIC MODELING

Ivan Himawan, Petr Motlicek, Marc Ferras Font, Srikanth Madikeri

Idiap Research Institute, Martigny, Switzerland
{ihimawan,motlicek,marc.ferras,srikanth.madikeri}@idiap.ch

ABSTRACT

Despite the superior classification ability of deep neural networks (DNN), the performance of DNN suffers when there is a mismatch between training and testing conditions. Many speaker adaptation techniques have been proposed for DNN acoustic modeling but in case of environmental robustness the progress is still limited. It is also possible to use techniques developed for adapting speakers to handle the impact of environments at the same time, or to combine both approaches. Directly adapting the large number of DNN parameters is challenging when the adaptation set is small. The learning hidden unit contributions (LHUC) technique for unsupervised speaker adaptation of DNN introduces speaker dependent parameters to the existing speaker independent network to increase the automatic speech recognition (ASR) performance of the target speaker using small amounts of adaptation data. This paper investigates the LHUC to adapt the speech recognizer to target speakers and environments where the impacts of speakers and noise differences are quantified separately. Our finding shows that the LHUC is capable of adapting to both speaker and noise conditions at the same time. Compared to the speaker independent model, about 9% to 13% relative word error rate (WER) improvement are observed for all test conditions using AMI meeting corpus.

Index Terms— Deep neural networks, acoustic model adaptation, environmental robustness, AMI corpus, LHUC

1. INTRODUCTION

Recognizing speech in wide range of conditions remains a challenging task. Most of today's ASR applications such as meetings, multi-party teleconferencing, or hands-free interfaces for controlling consumer-products will benefit from distant-talking operations [1]. As the distance between speaker and microphones increases, reverberation and noise dominate the direct sound. Also, the same applications may be operated by different speakers in variety of environmental conditions. Adapting existing model to unseen speakers and conditions will improve and guarantee consistent ASR performance.

Recently, the acoustic modeling based on DNN have gained remarkable success in speech recognition with substantial improvement of recognition accuracy of several ASR tasks. The DNN is a multi-layer perceptron with many hidden layers [2]. The deeper layers of DNN enable the acoustic model to learn complex boundaries between the hidden Markov model (HMM) states. Despite the progress using DNNs for acoustic modeling, the environmental robustness of DNN-based systems still gain a limited success. While large performance gain is obtained when the training and testing data have the same condition, degradation in performance can be severe in the mismatched conditions. Models trained on clean data may have difficulties recognizing the test data which is corrupted by noise [3].

The DNN has a large amount of parameters, thus large amounts of training data are usually required to train DNN. Also, the architecture of DNNs allows layers to be shared between tasks. For example, the hidden layers trained on multiple languages can lead to improved performance on a specific language in multilingual ASR tasks [4, 5]. In similar fashion, the noise robustness of DNN is usually achieved through multi-style training. In order to improve the recognition performance in the noisy environments, a DNN can be trained using various noise types and SNR levels [6, 3]. This strategy will enforce the DNN to be less sensitive to the change of the input by placing regularization on the cost function [3]. Adapting the DNN with small amounts of adaptation data is prone to over-fitting because of the large number of model parameters in DNN [7].

The popular adaptation techniques which estimate a set of linear transforms such as maximum likelihood linear regression (MLLR) [8, 9] and constrained MLLR [10] have been proposed for GMM/HMM frameworks. The transforms are used to adapt the means and covariances of Gaussian components of the acoustic models, such that they match the target speaker better [11]. These approaches however can not be directly applied to DNNs because of the different structure of modeling parameters. Nevertheless, there have been

This work was supported by the European Community under the project "DBox: A generic dialog box for multi-lingual conversational applications". This work was also partially supported by the EC FP7 funding, under "Speaker Identification Integrated Project (SIIP)".

some investigations of using feature-domain transform-based approaches such as feature-space MLLR (fMLLR) applied to DNNs [12, 13, 14]. Apart from speaker variabilities, variations in the audio recording process such as reverberations, speaker-to-microphone distance (e.g., close-talk or far-field), or recording devices can lead to significant differences in acoustic patterns. In most of cases, the techniques designed for reducing speaker variabilities could also be applied to minimize different channel or noise conditions [15, 11].

Recently, [16] proposed a speaker adaptation method for DNNs named LHUC where speaker dependent parameters are introduced to transform the speaker independent (SI) feature space to the speaker dependent (SD) feature space using only a small amount of adaptation data. The LHUC technique is unsupervised and an improved ASR performance is reported. In this paper, apart from adapting to target speakers, LHUC is used to adapt the DNN to different noise conditions to increase environmental robustness. The objective is to improve the ASR performance for unseen noise conditions with small amounts of training data. In contrast, the method of adapting DNN such as adding condition-specific layers requires large amounts of data for training the mixed-condition model [17].

This paper investigates the LHUC technique for speaker adaptation and environmental robustness using AMI meeting corpus with an objective to quantify the ASR improvement in noisy conditions. Following Section 2 discusses related work. Section 3 describes DNN acoustic modeling along with review of techniques for adapting neural network (NN) acoustic models and the LHUC adaptation technique. Experimental setup is described in Section 4. Results are presented and discussed in Section 5. Finally, the study is concluded in Section 6.

2. RELATED WORK

One of first methods for adapting hybrid NN/HMM system was to augment the existing NN with an extra input layer with a linear activation function [18]. The adaptation layer could also be added before the final output activation functions (i.e., softmax) [19]. These transforms are typically trained while keeping the rest of the network parameters fixed, thus the NN is trained with relatively few parameters. This is preferable when adapting the NN of large number of parameters while the adaptation data is limited [19, 20].

Different multi-condition learning architectures within the context of the DNN-based acoustic model framework to explicitly model different conditions were explored in [17]. For example, the channel specific layer can be trained for model adaptation while keeping the top layers, which have been trained with mixed-channels data, fixed. This has been shown to reduce WER over the baseline multi-condition model. Another alternative is to model the posterior given the speech observation and the acoustic scene, where both are added to the neural network as an input such as noise aware training in [21]. While the noise robustness of the model is improved, this approach requires a priori noise information as input features for training the DNN. The new DNN may need to be trained for unseen conditions.

The DNN weights can be adapted directly [22, 23]. However, modifying the entire DNN weights with small adaptation data leads to over-fitting and also results in extremely large speaker dependent parameter sets. As an alternative, small subsets of the DNN weights may be modified such as adapting the top hidden layer [24]. In [25], singular value decomposition is performed to the DNN model and an additional layer with smaller dimension is added to store speaker information. Most of the aforementioned techniques however are designed for speaker adaptation of DNNs, and there have been limited evaluations on environmental robustness.

3. DNN ACOUSTIC MODELING

In a DNN/HMM hybrid system, the emission probabilities of the HMM states are estimated with a DNN [26]. Using notation from [4], the output of the l-th layer, $\mathbf{u}_l$ is obtained as:

$$\mathbf{u}_l = \sigma(\mathbf{W}_l \mathbf{u}_{l-1} + \mathbf{b}_l), \qquad \text{for } 1 \leq l < L \tag{1}$$

where $\mathbf{W}_l$ denotes the matrix of connection weights between $l-1$-th and l-th layers, $\mathbf{b}_l$ is the additive bias vector at the l-th layer, and $\sigma(x) = 1/(1 + exp(-x))$ is a sigmoid activation function.

The DNN is trained using the standard error back-propagation procedure and the optimization is done through stochastic gradient descent by minimizing a negative log posterior probability cost function over the set of training examples $O = \{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$, with $\mathbf{u}_0 = \mathbf{o}_t$ [27, 28]:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} - \sum_{t=1}^{T} logP(s_t|\mathbf{o}_t), \tag{2}$$

where $\theta = \{\mathbf{W}_1, \ldots, \mathbf{W}_L, \mathbf{b}_1, \ldots, \mathbf{b}_L\}$ is the set of parameters of the network. In the training stage, the forced alignment of the acoustics with the transcript is performed to obtain s_t that is the most likely state s at time t.

3.1. Condition-specific Layers Adaptation

Adaptation of DNN to a specific condition can be achieved by training new or existing layers with the adaptation data [17]. This approach will adapt the network to the target condition while not completely eliminating the classification ability of previously trained layers. In this paper, only the bottom layer ($l = 1$) of the clean (SI) model is adapted with the noise-corrupted development data [22]. In our experiments, adapting the top layer ($l = L$) instead of the bottom layer results in worse WER performance. Specifically, using the existing model with L number of layers, the learning rate of l-th layer,

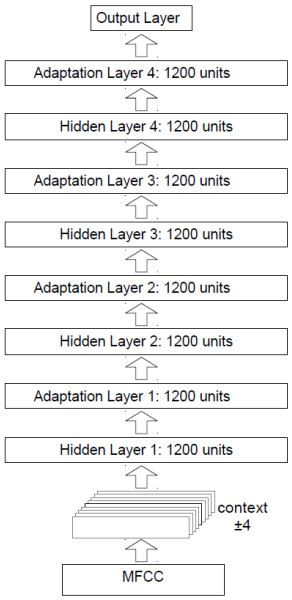

Fig. 1. The adaptation layer is added on top of each SI DNN's hidden layer after sigmoid non-linearity. No sigmoid non-linearity is applied at the output of adaptation layer. Note that the condition-specific layers adaptation uses DNN architecture with four hidden layers without adaptation layers.

for $1 < l$, was set to zero to simply reuse the learned weights. Only a bottom layer is adapted using the development data.

3.2. Learning Hidden Unit Contributions (LHUC)

The aim of speaker-dependent adaptation is to adjust the SI model parameters so that they generalize better to unseen speakers. Using some amount of adaptation data $(\mathbf{o}_t^k, s_t^k)_{t=1}^{T^k}, T^k << T$ for speaker k, the posterior distribution $P(s_t|\mathbf{o}_t)$ is adjusted by introducing speaker dependent parameters $\theta^k = \{\mathbf{r}_l^k, \ldots, \mathbf{r}_L^k\}$ where $\mathbf{r}_l^k$ denotes the vector of the speaker dependent parameters for the l^{th} hidden layer [16].

The posterior distribution $P(s_t|\mathbf{o}_t^k; \theta, \theta^k)$ for speaker model k is comprised of the network with the hidden layer output:

$$\mathbf{u}_l^k = f(\mathbf{r}_l^k) \circ \sigma(\mathbf{W}_l \mathbf{u}_{l-l}^k + \mathbf{b}_l), \quad (3)$$

where $\circ$ is an element-wise multiplication.

During training, the learned feature detectors from SI network are frozen, so that it is more robust against overfitting [16]. In our LHUC implementation, an adaptation layer is inserted for each hidden layer of the SI DNN, and the weights are set to an identity matrix. Thus, the SD and SI models are equivalent before adaptation. Figure 1 shows the architecture of SD DNN with added adaptation layers. The function $f(.)$ has been chosen to constrain the weights of the adaptation layers to be greater than or equal to zero before performing forward propagation and error back-propagation. Note that $\mathbf{r}_l^k$ is a full matrix in our case. Most of the non-diagonal weights after adaptation have extremely small values and a threshold may be set to force them to zero for stringent storage requirements.

Table 1. WERs[%]: Performance of DNN/HMM systems on IHM AMI evaluation set for different SNR (dB) levels. Cond.spec refers to condition-specific layers adaptation.

	Test SNR			
Trained on	CLEAN	20dB	15dB	10dB
IHM (CLEAN) (SI)	32.4	45.3	55.0	67.4
+Cond.spec (20dB)	36.8	42.2	49.2	59.5
+Cond.spec (15dB)	39.2	41.6	47.4	56.3
+Cond.spec (10dB)	43.2	42.7	47.1	54.8
IHM (20dB)	39.1	36.9	40.1	46.7
IHM (15dB)	44.0	38.2	39.8	44.2
IHM (10dB)	51.1	41.4	41.5	43.9

4. DATA AND SYSTEM SETUPS

Our experiments are carried out on the AMI corpus which contain meetings recorded in equipped instrumented meeting rooms at three sites in Europe (Edinburgh, IDIAP, TNO) [29]. The ASR experiments employ headset recordings (IHM). The speech makes about 67 hours for each audio stream (after performing voice activity detection) available for training, and holds around 7 hours for development and evaluation sets. The experiments use the suggested AMI corpus partitions for training, development, and evaluation (test) sets [30].

The Kaldi toolkit is used to develop DNN/HMM systems [31]. The IHM systems are trained on 39-dimensional MFCC features including their delta and acceleration versions using 9-frame temporal context. The DNNs are trained to estimate posterior probabilities of roughly 4K tied-state (senone) targets by employing four 1200-neuron hidden layers. The AMI pronunciation dictionary of approximately 23K words is used in the experiments, and the Viterbi decoding is performed using a 2-gram language model, previously built for NIST RT'07 corpora [29].

In order to simulate the noise-corrupted speech, noise that is recorded from a robot[1] (noise is generated by robot's cpu fan and motor) is extracted. To obtain the noisy speech files, the noise and speech root mean square (RMS) powers over the whole files were first computed. The gain factor applied to the noise file was then computed to scale the noise file prior to being added to a clean speech file.

[1] https://www.aldebaran.com/en/humanoid-robot/nao-robot

Table 2. WERs[%]: Performance of DNN/HMM systems on IHM AMI evaluation set - LHUC applied to adapt IHM clean model to different SNR (dB) levels. AD: Adaptation and Decoding. D: Decoding only.

	Test SNR				
IHM (CLEAN) model (SI)	Mode	CLEAN	20dB	15dB	10dB
+ LHUC-S	AD	31.1	x	x	x
	D	x	41.9	52.9	67.8
+LHUC-NS	AD	x	41.8	50.2	61.8
+LHUC-S+LHUC-N	AD	x	39.4	48.1	61.5
+fMLLR	AD	28.4	38.8	46.9	58.9
+ LHUC-S	AD	27.6	x	x	x
	D	x	35.5	42.3	54.5
+LHUC-NS	AD	x	37.5	44.2	54.6
+LHUC-S+LHUC-N	AD	x	35.0	41.8	51.6

5. EXPERIMENTAL RESULTS

5.1. Experiments using Condition-specific Layers

Table 1 shows the WER results of DNN/HMM systems trained with clean and noise-corrupted speech of different SNR levels. Note that the important numbers are shown in darker color. In the same table, the results of condition-specific layers adaptation using clean (IHM) model are presented. We can observe that the adaptation to specific SNR condition improves ASR performance on noisy speech but the performance degradation is observed when recognizing the clean test set. Compared to the performance of IHM clean model, the condition-specific layers adaptation achieves improvement by 3.1% absolute WER (from 45.3% to 42.2%) at SNR of 20dB, 7.6% absolute WER (from 55.0% to 47.4%) at SNR of 15dB and 12.6% absolute WER (from 67.4% to 54.8%) at SNR of 10dB. Note that the best performance for test at SNR of 20dB and 15dB is not equal to the SNR level used for model adaptation. The best ASR performance is obtained when the training condition matches the testing condition.

5.2. Experiments using LHUC

In order to demonstrate the modeling capacity of LHUC, an oracle experiment was performed in which the reference transcripts were used to align the adaptation data. This supervised adaptation yields 48% relative WER improvement. Besides this experiment, all the following LHUC experiments are unsupervised.

The speech recognizer can be adapted to target speakers and environment factors separately, or in the same iteration. Three types of LHUC adaptation procedures are devised to quantify the impact of adaptation to speakers and noise conditions either separately or jointly. The model trained on clean (IHM) data is used for performing LHUC with multiple adaptation and decoding iterations, and different target conditions, as described below:

- LHUC-S: (1) adaptation parameters estimated per speaker from clean recordings; (2) adapted DNN used for decoding the target data.
- LHUC-NS: (1) adaptation parameters estimated per speaker and noise (from target data); (2) adapted DNN used for decoding the target data.
- LHUC-S+LHUC-N: (1) adaptation parameters estimated per speaker from clean recordings; (2) adapted DNN used for decoding the target data; (3) adaptation parameters estimated per speaker and noise (from target data) using transcript inferred from previous decoding phase on target data; (4) final adapted DNN used for decoding the target data;

Further, conventional speaker adaptive fMLLR training [32] is also investigated when deployed in combination with LHUC for noise adaptation. Note that for LHUC-S, the DNN prior for computing the likelihoods used in subsequent decoding is estimated using the clean (IHM) training data. For other LHUC experiments, the prior is estimated from the target data for given noise conditions. DNN adaptation performance using LHUC is given in Table 2, including results when LHUC is combined with feature-based adaptation (fMLLR).

The LHUC adaptation (LHUC-S) improves over the SI model by 1.3% (from 32.4% to 31.1%) absolute WER (about 4% relative). Recognizing test set with SNR of 20dB improves over the SI model but slightly degrades when recognizing test set with SNR of 10dB. This shows that the LHUC adapts the speakers well but may have difficulties to recognize utterances with a severe noise condition.

When LHUC is performed on the noisy test sets (LHUC-NS), that is to adapt on speakers as well as on noise conditions simultaneously, ASR improvements are observed across all

test noise conditions. Compared to LHUC-S, small improvement is observed for test with SNR of 20dB (0.1%), modest improvement for test with SNR of 15dB (2.1%), but large gain is observed for test with SNR of 10dB (6.0%). When the speaker adapted model (LHUC-S) is applied to adapt the noise (LHUC-S+LHUC-N), further improvements are observed. It also yields better performance than LHUC-NS across all test conditions. Compared to SI model, LHUC-S+LHUC-N models improve ASR performance by about 6% to 7% absolute WER (about 9% to 13% relative) for all test conditions.

The model that is adapted to target speakers and subsequently to the noise conditions (LHUC-S+LHUC-N) performs better than the condition-specific layers adaptation for test with SNR of 20dB. The ASR performance improves by 2.2% (from 41.6% to 39.4%) absolute WER (5.3% relative). The performance for test SNR of 15dB is close to the gain acquired using condition-specific layers adaptation, with the difference of 1% absolute WER (48.1% for LHUC-S+LHUC-N versus 47.1% for condition-specific layers adaptation). Note that condition-specific model is adapted using about 7 hours of development data while LHUC adaptation added more layers. Although LHUC shows improvements across all noise conditions, there is still a large gap between condition-specific layers adaptation and LHUC adaptation, particularly for low SNR condition (i.e., 61.5% versus 54.8% for test SNR of 10dB).

As shown in Table 2, the feature-based per-speaker adaptation fMLLR yields better results compared to LHUC-only adaptation using clean (IHM) model across all test noise conditions. Combining fMLLR and LHUC for speaker adaptation improves ASR performance by 0.8% (from 28.4% to 27.6%) absolute WER (about 2.8% relative). This shows that LHUC is complementary to fMLLR [16]. The LHUC speaker adaptation in combination with fMLLR on the noise-corrupted test set yields further improvements across all test SNR levels. Note that results show LHUC-S are better than LHUC-NS when decoding test sets of different SNR levels. This may be due to a better pseudo-transcript used for LHUC-S adaptation. Compared to the conventional speaker adaptive fMLLR results, combining fMLLR and LHUC-S+LHUC-N improves ASR performance by about 4% to 7% absolute WER (about 10% to 12% relative) for all test conditions.

To investigate how the amount of adaptation data affects the WER, the per-speaker ASR performance is evaluated for three categories according to amounts of adaptation data. Figure 2 shows the averaged WER improvement for LHUC-S+LHUC-N with respect to SI model for test set with SNR of 20dB. The WER reduces with increased availability of adaptation data. With less than 5 minutes of data, the averaged WER improvement is about 3% compared to about 7% when more than 10 minutes of data is available.

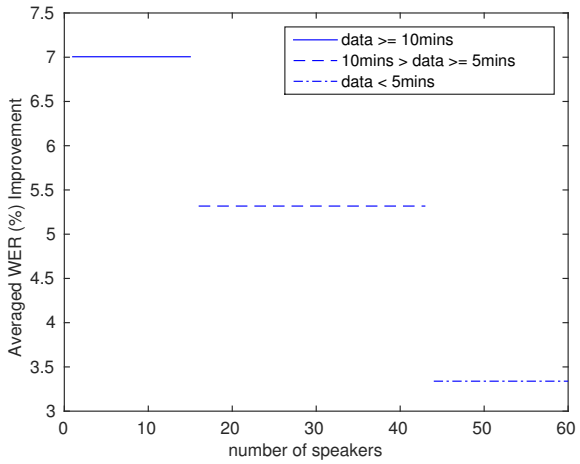

Fig. 2. Averaged absolute WER (%) improvement for LHUC-S+LHUC-N with respect to SI model and the amount of unsupervised adaptation data available for test set with SNR of 20dB.

6. CONCLUSION

This paper proposes LHUC adaptation applied in DNN acoustic modeling which can efficiently reduce variability caused by both speaker and environment. Experimental results reveal that LHUC is able to adapt to target speakers as well as to noise conditions at the same time. Large performance gain is observed when the clean model is adapted to speaker first and then to noise. Our findings show that for high SNR condition (20dB), LHUC yields better performance compared to condition-specific layers adaptation. Combining fMLLR with LHUC further improves the ASR performance across all test SNR levels. Our future work aims at improving transcriptions from the first pass decoding outputs such as using word confidence scores to select the most informative examples. The suitability of LHUC technique for practical ASR applications will also be investigated.

7. REFERENCES

[1] T. Yoshioka et al., "Making Machines Understand Us in Reverberant Rooms: Robustness Against Reverberation for Automatic Speech Recognition," *IEEE Signal Processing Magazine*, Nov. 2012.

[2] Dong Yu and Li Deng, *Automatic Speech Recognition - A Deep Learning Approach*, Springer Publishing Company, Incorporated, 2014.

[3] Shi Yin et al., "Noisy training for deep neural networks in speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, 2015.

[4] Arnab Ghoshal, Pawel Swietojanski, and Steve Renals, "Multilingual training of deep neural networks," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[5] Jui-Ting Huang et al., "Cross-Language Knowledge Transfer Using Multilingual Deep Neural Network With Shared Hidden Layers," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013.

[6] Chao Weng, Dong Yu, Michael L. Seltzer, and Jasha Droppo, "Single-Channel Mixed Speech Recognition using Deep Neural Network," in *Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing*, 2014.

[7] 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 *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.

[8] C. Legetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[9] P. C. Woodland, "Speaker Adaptation for Continuous Density HMMs: A Review," in *Proceedings ISCA ITR-Workshop on Adaptation Methods for Speech Recognition*, 2001.

[10] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, 1998.

[11] Yongqiang Wang and M. J. F. Gales, "Speaker and Noise Factorization for Robust Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, 2012.

[12] Frank Seide et al., "Feature Engineering in Context-Dependent Deep Neural Networks for Conversational Speech Transcription," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[13] Dong Yu et al., "Feature learning in deep neural networks - studies on speech recognition tasks," *in Proceedings of the International Conference on Learning Representations*, 2013.

[14] Takuya Yoshioka, Xie Chen, and Mark J. F. Gales, "Impact of single-microphone dereverberation on DNN-based meeting transcription systems," in *Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing*, 2014.

[15] D. Kim and M. J. F. Gales, "Adaptive training with noisy constrained maximum likelihood linear regression for noise robust speech recognition," *in Proceedings of Interspeech*, 2009.

[16] Pawel Swietojanski and Steve Renals, "Learning Hidden Unit Contributions for Unsupervised Speaker Adaptation of Neural Network Acoustic Models," in *Proceedings of IEEE Spoken Language Technology Workshop*, 2014.

[17] Yan Huang, M. Slaney, M. L. Seltzer, and Y. Gong, "Towards Better Performance with Heterogeneous Training Data in Acoustic Modeling using Deep Neural Networks," in *Proceedings of Interspeech*, 2014.

[18] V. Abrash, H. Franco, A. Sankar, and M. Cohen, "Connectionist Speaker Normalization and Adaptation," in *Proceedings Eurospeech*, 1995, pp. 2183–2186.

[19] Bo Li and Khe Chai Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems," in *Proceedings of Interspeech*, 2010.

[20] R. Price, K.-I. Iso, and K. Shinoda, "Speaker adaptation of deep neural networks using a hierarchy of output layers," in *Proceedings of IEEE Spoken Language Technology Workshop*, 2014.

[21] Michael L. Seltzer et al., "An investigation of deep neural networks for noise robust speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[22] Hank Liao, "Speaker adaptation of context dependent deep neural networks," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013.

[23] Dong Yu et al., "KL-Divergence Regularized Deep Neural Network Adaptation for Improved Large Vocabulary Speech Recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[24] Kaisheng Yao et al., "Adaptation of Context-Dependent Deep Neural Networks for Automatic Speech Recognition," in *Proceedings of IEEE Spoken Language Technology Workshop*, 2012.

[25] Jian Xue et al., "Singular Value Decomposition Based Low-Footprint Speaker Adaptation and Personalization for Deep Neural Network," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.

[26] H. Bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic, 1994.

[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct. 1986.

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

[29] Thomas Hain et al., "Transcribing Meetings With the AMIDA Systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 2, pp. 486–498, 2012.

[30] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals, "Hybrid Acoustic Models for Distant and Multichannel Large Vocabulary Speech Recognition," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013.

[31] Daniel Povey et al., "The Kaldi speech recognition toolkit," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[32] Tasos Anastasakos, John McDonough, and John Makhoul, "Speaker adaptive training: A maximum likelihood approach to speaker normalization," in *Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing*, 1997.