INVESTIGATION OF UNSUPERVISED ADAPTATION OF DNN ACOUSTIC MODELS WITH FILTER BANK INPUT

Takuya Yoshioka[†,‡], *Anton Ragni*[†], *Mark J. F. Gales*[†]

†Cambridge University Engineering Department, Cambridge, UK
‡NTT Communication Science Laboratories, Kyoto, Japan

ABSTRACT

Adaptation to speaker variations is an essential component of speech recognition systems. One common approach to adapting deep neural network (DNN) acoustic models is to perform global constrained maximum likelihood linear regression (CMLLR) at some point of the systems. Using CMLLR (or more generally, generative approaches) is advantageous especially in unsupervised adaptation scenarios with high baseline error rates. On the other hand, as the DNNs are less sensitive to the increase in the input dimensionality than GMMs, it is becoming more popular to use rich speech representations, such as log mel-filter bank channel outputs, instead of conventional low-dimensional feature vectors, such as MFCCs and PLP coefficients. This work discusses and compares three different configurations of DNN acoustic models that allow CMLLR-based speaker adaptive training (SAT) to be performed in systems with filter bank inputs. Results of unsupervised adaptation experiments conducted on three different data sets are presented, demonstrating that, by choosing an appropriate configuration, SAT with CMLLR can improve the performance of a well-trained filter bank-based speaker independent DNN system by 10.6% relative in a challenging task with a baseline error rate above 40%. It is also shown that the filter bank features are advantageous than the conventional features even when they are used with SAT models. Some other insights are also presented, including the effects of block diagonal transforms and system combination.

***Index Terms*—** Deep neural network, acoustic model adaptation, hybrid, tandem, stacked hybrid.

1. INTRODUCTION

Recently acoustic models based on deep neural networks (DNNs) have been successful in many speech recognition tasks [1–3]. Since the DNNs are less sensitive to the increase in the input dimensionality than GMMs, they allow us to exploit a richer set of features than conventional low-dimensional feature vectors, such as MFCCs and PLP coefficients. In particular, log mel-filter bank channel outputs, or FBANK features, have been shown to provide improved recognition accuracy [2, 4].

There are two popular configurations of acoustic models based on DNNs. In the first configuration, which is called a *hybrid*, the DNNs are utilized to predict context-dependent HMM states [1, 5]. The second model configuration, which is called a *tandem*, exploits the DNNs to perform nonlinear discriminative feature transformation [6, 7]. The transformed features are then input to an acoustic model based on GMM-HMMs or another DNN-HMM hybrid model. The latter allows multiple DNNs to be stacked [8] and is especially referred to as a *stacked hybrid*.

It has been shown that DNN-based acoustic models can benefit from speaker adaptation and speaker adaptive training (SAT), although they are more robust against speaker variability than GMM-HMMs. Regularized re-training approaches modify DNN parameters to better recognize adaptation utterances while keeping the adapted model not to deviate too much from the unadapted model [9, 10]. Instead of modifying network parameters, several methods, such as linear input network (LIN) [11], feature discriminative linear regression (fDLR) [12], and the speaker code approach [13], discriminatively convert input features to improve DNN's classification accuracy for the adaptation data. Alternatively, such speaker-adapted features may be obtained by using conventional techniques, such as global constrained maximum likelihood linear regression (CMLLR) [12, 14]. Because these input transformation approaches handle the speaker variability in the input space, they enable the use of SAT to be simple. One advantage of the CMLLR-based approaches as opposed to discriminative methods, such as LIN, is that the resultant speaker transforms are less sensitive to errors in supervision labels [15]. Stacked hybrid and tandem DNN acoustic models can also utilize CMLLR in the nonlinearly transformed feature space.

There has been limited investigation as to which configuration (hybrid, stacked hybrid, or tandem) is suitale for building a DNN SAT acoustic model using global CMLLR when FBANK features are used as an input. It is also debatable whether the FBANK features provide better recognition performance than the MFCCs and PLP coefficients when they are used in SAT systems. This paper examines these issues on three different data sets with different error rate ranges, assuming SAT and run-time unsupervised adaptation scenarios.

The remainder of this paper is organized as follows. Section 2 describes the three DNN SAT acoustic model configurations (hybrid, stacked hybrid, and tandem) based on global CMLLR and discusses their relative merits. Section 3 reports experimental results, followed by our conclusion in Section 4.

2. SPEAKER ADAPTIVE ACOUSTIC MODELS BASED ON DNNS

We consider three DNN SAT model configurations, i.e., the hybrid, stacked hybrid, and tandem, which can be used with an FBANK input. This section describes the ways in which we built our SAT systems for each model configuration and discusses the characteristics

This work was partly supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012.The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

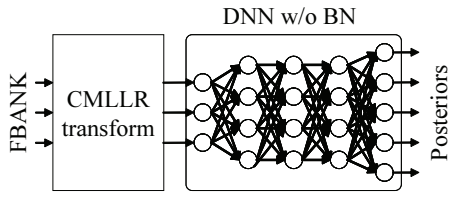

Fig. 1. Hybrid SAT model.

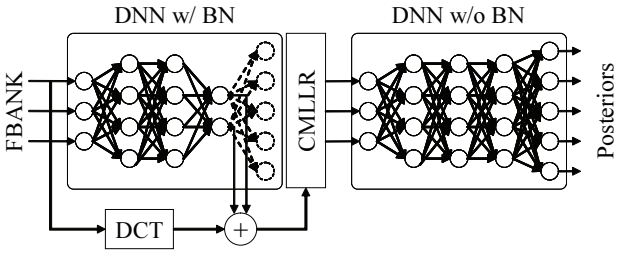

Fig. 2. Stacked hybrid SAT model.

of the three configurations.

2.1. Hybrid SAT Model

A hybrid DNN acoustic model consists of a DNN and a context-dependent HMM. In this configuration, the DNN predicts the context-dependent HMM states based on a set of input features, such as FBANK features, MFCCs, and PLP coefficients. Specifically, the DNN accepts a sequence of extended feature vectors, each consisting of several frames of input features within a context window. With the extended feature vector denoted by $\boldsymbol{o}_n$, the DNN estimates the posterior probability, $p(s|\boldsymbol{o}_n)$, of HMM state s with multiple layers of nonlinear processing. The posterior probability is converted to the state likelihood by $p(\boldsymbol{o}_n|s) \propto p(s|\boldsymbol{o}_n)p(s)$ to perform Viterbi decoding, where $p(s)$ is the prior probability. The DNN parameters are initialized with discriminative pre-training [12], followed by fine-tuning based on a negative cross entropy criterion. There are alternative methods for initialization (e.g., generative pre-training [5]) and fine-tuning (e.g., sequence training [16]). Training the DNN on raw, i.e., speaker un-normalized, features yields a hybrid speaker independent (SI) acoustic model.

A SAT model can be realized by adding a mechanism of adapting a system to each speaker present in the training and test sets. With this configuration, adaptation is performed in the input space by applying a speaker-specific global CMLLR transform to the input features of each speaker as shown in Fig.1. Both full and block-diagonal transforms can be used. The latter may be more suitable when the amount of adaptation utterances is small. When reliable CMLLR transforms are obtained, the hybrid configuration is advantageous than the other two in that no information is lost during speaker normalization processing and therefore the DNN can fully exploit the information that the input features have.

When FBANK features are used as an input, an accurate GMM-HMM acoustic model is necessary for the FBANK features to estimate global CMLLR transforms. Such a model can be obtained by performing single pass retraining (SPR) from a GMM-HMM model of MFCCs or PLP coefficients. Since FBANK features are highly correlated with each other, it is essential to use full covariance matrices or semi-tied covariance (STC) transforms [17]. In our experiments, we used a global STC transform[1]. Thus, SPR is configured to produce both FBANK-space GMM parameters and a global STC transform. When using block diagonal CMLLR transforms, the STC transform also needs to have a block diagonal structure.

2.2. Stacked Hybrid SAT Model

A stacked hybrid model includes two DNNs as shown in Fig. 2. The fundamental idea is to perform global CMLLR in a nonlinearly compressed feature space obtained with the first DNN and then input the speaker-normalized features into the second DNN. To perform nonlinear feature compression, the first DNN has a bottleneck (BN) layer [7], which consists of a limited number of units. Each input feature vector is forwarded through this DNN. The linear outputs from the BN layer are computed and further converted by a global STC transform. The generated features are combined with HLDA features derived from MFCCs or PLP coefficients to form a set of new features, which are called TANDEM features[2]. Then, global CMLLR is applied to the TANDEM features and the resultant speaker-normalized TANDEM features are then fed into the second DNN. Both DNNs can be trained in the same way as the hybrid DNN, i.e., by pre-training followed by fine-tuning. Note that the block diagram shown in Fig. 2 omit some elements, such as HLDA, STC, and liftering.

Unlike the hybrid SAT model, the stacked hybrid SAT model allows very high dimensional feature vectors (e.g., those produced by a bank of Gabor filters [19]) to be exploited since the CMLLR matrix size is independent of the input feature vector dimensionality. One potential drawback of this approach is that the three components, i.e., the two DNNs and CMLLR transforms, are estimated separately with different criteria, which may limit the adaptation effect.

2.3. Tandem SAT Model

In a tandem SAT model, the second DNN of a stacked hybrid SAT model is replaced by a conventional GMM-HMM acoustic model. That is, the GMM-HMM acoustic model is trained on speaker-normalized TANDEM features generated by a DNN with a BN layer followed by global CMLLR transforms. One desirable property of this approach is that CMLLR transforms are estimated in such a way that the likelihood of the GMM-HMM model is increased.

Since decoding is performed with the conventional GMM-HMM model, various adaptation techniques can be applied, including MLLR and cluster adaptive training [20]. In our experiments, we adapted the GMM means to each test speaker with global MLLR at decoding time in addition to global CMLLR. Furthermore, existing discriminative feature mapping techniques, such as fMPE [21], may be applied on top, although this is not investigated in this work.

3. EXPERIMENTS

We conducted a series of experiments in three different unsupervised adaptation tasks with different error rate ranges. First, we compare the three DNN SAT acoustic model configurations in one task. We also contrast FBANK-based systems with MFCC-based systems to examine whether the FBANK features provide improved recognition performance when they are used in CMLLR-based SAT systems. Then, we extend our evaluation to the other two tasks.

[1]The configuration of jointly using STC and CMLLR was first presented at ASRU 2013 [18] for convolutional neural networks when this paper was under review.

[2]The term "tandem" is capitalized when referring to a feature vector and written in lower case when referring to a model configuration.

Table 1. WERs (%) of SI systems.

System		WER (%)		
Model	Feature	Dev	Eval	Avg
GMM-HMM	MFCC	52.4	52.7	52.6
DNN-HMM hybrid	MFCC	42.5	42.8	42.7
	FBANK	42.6	40.2	41.4

Table 2. Comparison of different configurations of DNN SAT models with FBANK input.

System		WER (%)		
Model	Transform	Dev	Eval	Avg
Hybrid	Full	37.4	37.4	37.4
	Block	37.3	36.6	37.0
Stacked hybrid	Full	40.3	40.2	40.3
Tandem	Full	38.8	38.9	38.9

3.1. Meeting Transcription Using Single Distant Microphone

Our first set of experiments was conducted in a meeting transcription task based on single distant microphone (SDM). We used the AMI corpus [22], which consists of recordings of meetings, each conducted by four participants. Many meeting participants were non-native speakers and thus had distinct acoustic characteristics. Although the meetings were recorded with table-top eight-channel microphone arrays, only the first microphone data were used. To mitigate the impact of reverberation caused by sound reflection in rooms, the data were pre-processed by single-channel dereverberation [23], which provided a 4-5 % relative improvement to a well-trained DNN acoustic model [24]. We split the corpus as described in [25], which resulted in 59, 2.7, and 2.6 hours of data for training, a development test, and an evaluation test, respectively. The training set consisted of 175 speakers while each test set contained 8 speakers. Speech segments and speaker identities were obtained from the provided labels and overlapping speech segments were ignored.

Our recognition systems were built as follows. First, we applied speaker and meeting-level mean and variance normalization to 52-dimensional features, consisting of MFCCs and their first to third-order delta parameters. Then, they were projected onto a 39-dimensional feature space by HLDA. These HLDA feature vectors were used to train a baseline MPE SI acoustic model, which consisted of cross-word triphone HMMs with approximately 4,000 context-dependent states and 16 Gaussians per state. The HLDA features were further processed by speaker-level global and full CMLLR to generate speaker-normalized HLDA feature vectors. Based on these features, we trained a baseline MPE SAT model. Using this SAT model, we performed forced alignment to produce frame-level state labels. With these supervision labels, we trained a hybrid SI model, a hybrid SAT model, and a tandem SAT model with a context window of nine frames as described in Section 2. We used FBANK features as an input, where each FBANK feature vector consisted of 24-channel log mel-filter bank outputs plus their delta parameters up to the third order. The FBANK-space GMM-HMM acoustic model needed for performing SAT was obtained with SPR from the baseline SI system. We also trained a stacked hybrid model based on the speaker-normalized TANDEM features generated during the tandem SAT model construction. The DNN used in the hybrid systems consisted of five hidden layers each with 1,500 hidden units and a softmax output layer with approximately 4,000 targets. The DNN used for TANDEM feature extraction had a BN layer with 26 units just before the output layer. The GMM-HMM model for the tandem SAT system used the same number of states and Gaussians as the baseline model and was trained with an MPE criterion.

Table 3. Impact of FBANK and TANDEM-space transforms.

Transform space		WER (%)		
FBANK	TANDEM	Dev	Eval	Avg
-	✓	38.8	38.9	38.9
✓	✓	37.0	37.4	37.2

Table 4. FBANK vs. MFCC comparison results. Tandem systems used speaker-normalized inputs.

System		WER (%)		
Model	Feature	Dev	Eval	Avg
Hybrid	MFCC	40.0	39.7	39.9
	FBANK	37.4	37.4	37.4
Tandem	MFCC	38.3	38.8	38.6
	FBANK	37.0	37.4	37.2

Decoding was performed in the same way as described in [24], i.e., by bigram lattice generation with a 40K-word language model, followed by trigram lattice rescoring and confusion network rescoring. For the SAT systems, the hybrid SI system was used to generate adaptation supervision hypotheses.

Table 1 shows the word error rates (WERs) of the SI systems. As expected, the DNN acoustic model substantially improved the recognition performance. We can also see that using FBANK features yielded a lower average WER than using MFCCs. The objective of our SAT systems is to further improve the performance of this FBANK-based SI system.

Table 2 compares the WERs of the three FBANK-based DNN SAT models described in Section 2. We can see that all of the considered systems outperformed the DNN SI system, demonstrating the benefit of SAT using CMLLR. Of the three model configurations, the hybrid SAT model achieved the lowest WER. This would be because the hybrid SAT model retains all information inherent in the input features during speaker normalization processing whereas the other two configurations do not due to the nonlinear feature compression performed by the DNN with a BN layer. Using a block diagonal CMLLR transform provided further improved performance, where we used two 48×48 transforms, the first one of which was applied to the set of static and first-order delta FBANK parameters while the other one was applied to the second and third-order delta parameter set. When we compare the stacked hybrid and tandem models, the latter achieved a lower error rate owing to MLLR adaptation performed at decoding time. (Without MLLR, the tandem SAT system yielded WERs of 39.8% and 40.2% for the development and test sets, respectively, which were comparable to the WERs of the stacked hybrid system.) Considering the relatively limited performance gain obtained with the stacked hybrid configuration, our further investigation was conducted with the hybrid and tandem models.

When reliable CMLLR transforms of input features can be obtained as in this task, the tandem SAT system can benefit from using the speaker-normalized input features. An experiment was performed to evaluate the impact of global CMLLR transforms of input (FBANK) features on the tandem SAT system. As shown in Table 3, performing global CMLLR in the input space also provided a meaningful performance gain to the tandem SAT acoustic model, thereby achieving almost the same WER as the hybrid SAT system. Therefore, we decided to use speaker-nomalized inputs with both the hybrid and tandem models in the following experiments.

Our final experiment on the SDM meeting data set was to compare the FBANK and MFCC features with the SAT setups. The

Table 5. Results on AMI IHM data set. Tandem systems used speaker-normalized FBANK features as input.

System		Feature	WER (%) Dev	Eval	Avg	
GMM-HMM	SI	MFCC	35.2	30.2	32.7	
Hybrid	SI	MFCC	28.2	25.1	26.7	
		FBANK	27.8	24.2	26.0	
	SAT	MFCC	25.0	22.2	23.6	S1
		FBANK	23.8	21.6	22.7	S2
Tandem	SAT	MFCC	24.1	22.0	23.1	S3
		FBANK	23.1	21.4	22.3	S4

Table 6. System combination results.

Systems combined	WER (%) Dev	Eval	Avg
S2 $\oplus$ S4	22.0	20.3	21.2
S3 $\oplus$ S4	22.3	20.8	21.6
S1 $\oplus$ S2 $\oplus$ S3 $\oplus$ S4	21.7	20.0	20.9

results are shown in Table 4. We can see that using the FBANK features resulted in lower WERs regardless of the employed model configuration. This means that the additional information provided by the FBANK features is useful for discriminating different HMM states even with the SAT setups. In summary, the lowest WER was achieved by the hybrid SAT system using block-diagonal CMLLR transforms, which improved the FBANK-based hybrid SI system by 10.6 % relative in this challenging unsupervised adaptation task.

3.2. Meeting Transcription Using Headset Microphones

The second set of experiments was also based on the AMI corpus but used individual headset microphones (IHM). Since the microphones were close to the speakers' mouths, these data are little affected by reverberation and background noise. We split the corpus into training and test sets in the same way as in the SDM task. The system setups were almost identical to those used in the SDM task, which enables us to assess the robustness of DNN acoustic models against environmental distortion. The only difference was that the third-order delta FBANK parameters were not used in the IHM task.

Table 5 summarizes the WERs obtained on this data set. The same trend was observed as in the SDM task. The DNN acoustic model made a substantial improvement to the baseline GMM-HMM system. Also, the FBANK-based systems consistently achieved better performance than the MFCC-based systems with both SI and SAT setups. It is noteworthy that, comparing the SDM and IHM results, there is a large performance gap between these two data sets even with the best performing DNN SAT configuration. This suggests the necessity of further research to improve the environmental robustness of DNN acoustic models.

On this data set, the DNN tandem SAT system (using speaker-normalized FBANK features) outperformed the hybrid SAT system. When we inspected the per-speaker WERs, we found that these two model configurations exhibited different performance patterns. For a few speakers the hybrid SAT system achieved lower WERs while the tandem SAT system performed better for other speakers, which indicates that these systems are complementary. To confirm this, we performed a set of system combination experiments using the confusion network combination method [26]. The results are shown in Table 6. Combining the hybrid and tandem SAT systems resulted in a WER of 21.2%. (A similar improvement was observed in [27].) On the other hand, combining the FBANK and MFCC-based systems also improved the performance by 3.1% relative, achieving a WER of 21.6%. Combining the hybrid and tandem SAT systems with MFCC and FBANK inputs achieved the lowest WER of 20.9%, which was better than the performance of the hybrid SI system by 19.6% relative.

Table 7. FBANK vs. PLP comparison with tandem SAT configurations on Vietnamese FLP.

System		Transform space Input	TANDEM	TER (%)
GMM-HMM	SI	-	N/A	65.9
PLP tandem	SAT	-	✓	54.9
FBANK tandem		-	✓	53.6
		✓	✓	52.8

3.3. Vietnamese Speech Recognition

We also investigated the effect of CMLLR-based adaptation of DNN acoustic models on a more challenging Vietnamese speech recognition task. We used the Vietnamese full language pack (FLP) of the IAPRA BABEL Program[3]. The Vietnamese FLP consists of scripted and conversational speech data and contained 108.8 hours of speech for training and 9.58 hours of speech for evaluation. We employed a phone set based on X-SAMPA and a language model trained with the Vietnamese FLP transcripts. We used the same tools and configurations as those used in our meeting transcription experiments to produce our Vietnamese systems except for a baseline GMM-HMM system having been built using features consisting of pitch features (pitch+Δ+$\Delta\Delta$) and PLP coefficients. The pitch features were also appended to the TANDEM features.

In this task, we considered only the tandem SAT configuration. With this configuration, we examined the degree to which the recognition performance was increased by using FBANK features as the input into DNNs instead of PLP coefficients.

Table 7 shows the token error rate (TER) results, where a token is defined as a Vietnamese syllable. We can see that the use of FBANK features resulted in a significant performance improvement and that using speaker-normalized FBANK features provided a further small gain. The best performance was achieved by the FBANK-based tandem system performing speaker normalization at both the FBANK and TANDEM levels. This is consistent with the results of the two meeting transcription experiments.

4. CONCLUSION

In this paper, we explored global CMLLR-based aproaches to building a DNN SAT acoustic model with an FBANK input. Our observations can be summarized as follows. (1) The hybrid SAT model and the tandem SAT model with a speaker-normalized input yielded large performance gains, although other model configurations was also effective. We also discussed the characteristics of different model configurations and the situations in which each configuration is suitable for use. (2) The FBANK-based DNN SAT models consistently outperformed the MFCC and PLP-based DNN SAT models. (3) Combining systems with different model configurations and different input feature types further improved the recognition performance.

[3]This effort uses the IARPA Babel Program Vietnamese language collection release babel107b-v0.7. The Full Languaue Pack was used. (http://www.iarpa.gov/Programs/ia/Babel/babel.html)

5. 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. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, 2012.

[2] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero, "Recent advances in deep learning for speech research at Microsoft," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 8604–8608.

[3] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kinsgbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[4] A.-r. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 4273–4276.

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

[6] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2000, pp. 1635–1638.

[7] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2007, pp. IV–757–IV–760.

[8] K. M. Knill, M. J. F. Gales, S. P. Rath, P. C. Woodland, C. Zhang, and S.-X Zhang, "Investigation of multilingual deep neural networks for spoken term detection," in *Proc. Workshop. Automat. Speech Recognition, Understanding*, 2013, pp. 138–143.

[9] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7947–7951.

[10] 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 *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7893–7897.

[11] J. P. Neto, C. Martins, and L. B. Almeida, "Speaker-adaptation in a hybrid HMM-MLP recognizer," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1996, pp. 3382–3385.

[12] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-depencent deep neural networks for conversational speech transcription," in *Proc. Workshop. Automat. Speech Recognition, Understanding*, 2011, pp. 24–29.

[13] 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 *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7942–7946.

[14] S. P. Rath, D. Povey, K. Veselý, and J. H. Černocký, "Improved feature processing for deep neural networks," in *Proc. Interspeech*, 2013, pp. 109–113.

[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*, 2010, pp. 526–529.

[16] 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*, 2012.

[17] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. Speech, Audio Process.*, vol. 7, no. 3, pp. 272–281, 1999.

[18] T. N. Sainath, B. Kingsbury, A.-r. Mohamed, G. E. Dahl, G. Saon, H. Soltau, B. Beran, A. Y. Aravkin, and B. Ramabhadran, "Improvements to deep convolutional neural networks for LVCSR," in *Proc. Workshop. Automat. Speech Recognition, Understanding*, 2013, pp. 315–320.

[19] S.-Y. Chang, B. T. Meyer, and N. Morgan, "Spectro-temporal features for noise-robust speech recognition using power-law nonlinearity and power-bias subtraction," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7063–7067.

[20] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," 2000.

[21] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "FMPE: Discriminatively trained features for speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2005, pp. 961–964.

[22] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post andD. Reidsma, and P. Wellner, "The AMI meeting corpus: a pre-announcement," in *Proceedings of Int. Worksh. Machine Learning for Multimodal Interaction*, 2006, pp. 28–39.

[23] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, 2012.

[24] T. Yoshioka, X. Chen, and M. J. F. Gales, "Impact of single-microphone dereverberation on DNN-based meeting transcription systems," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, accepted.

[25] C. Breslin, KK Chen, M. J. F. Gales, and K. Knill, "Integrated online speaker clustering and adaptation," in *Proc. Interspeech*, 2011, pp. 1085–1088.

[26] G. Evermann and P. C. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proc. NIST Speech Transcription Worksh.*, 2000.

[27] P. Swietojanski, A. Ghoshal, and S. Renals, "Revisiting hybrid and GMM-HMM system combination techniques," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 6744–6748.