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

On the Use of Gaussian Mixture Model Framework to Improve Speaker Adaptation of Deep Neural Network Acoustic Models

Natalia Tomashenko[1,2,3]*, Yuri Khokhlov*[3] *and Yannick Estève*[1]

[1]LIUM, University of Le Mans, France
[2]ITMO University, Saint-Petersburg, Russia
[3]STC-innovations Ltd, Saint-Petersburg, Russia
natalia.tomashenko@univ-lemans.fr, khokhlov@speechpro.com, yannick.esteve@univ-lemans.fr

Abstract

In this paper we investigate the Gaussian Mixture Model (GMM) framework for adaptation of context-dependent deep neural network HMM (CD-DNN-HMM) acoustic models. In the previous work an initial attempt was introduced for efficient transfer of adaptation algorithms from the GMM framework to DNN models. In this work we present an extension, further detailed exploration and analysis of the method with respect to state-of-the-art speech recognition DNN setup and propose various novel ways for adaptation performance improvement, such as, using bottleneck features for GMM-derived feature extraction, combination of GMM-derived with conventional features at different levels of DNN architecture, moving from monophones to triphones in the auxiliary GMM model in order to extend the number of adapted classes, and finally, using lattice-based information and confidence scores in maximum a posteriori adaptation of the auxiliary GMM model. Experimental results on the TED-LIUM corpus show that the proposed adaptation technique can be effectively integrated into DNN setup at different levels and provide additional gain in recognition performance.

Index Terms: speaker adaptation, deep neural networks (DNN), MAP, fMLLR, CD-DNN-HMM, GMM-derived (GMMD) features, speaker adaptive training (SAT), confidence scores

1. Introduction

Nowadays, deep neural networks (DNNs) have replaced conventional Gaussian mixture models (GMM) HMMs in most state-of-the-art automatic speech recognition (ASR) systems, because it has been shown [1] that DNN-HMM models outperform GMM-HMMs in different ASR tasks. However, various adaptation algorithms that have been developed for GMM-HMM systems [2, 3] cannot be easily applied to DNNs because of the different nature of these models. Various adaptation methods have been developed for DNNs. Most of them can be classified into several types: linear transformation, regularization techniques, auxiliary features, multi-task learning, combining GMM and DNN models and other model-space adaptation techniques.

Linear transformation can be applied at different levels of the DNN-HMM system: to the input features, as in linear input network transformation (LIN) [4, 5, 6] or feature-space discriminative linear regression (fDLR) [7, 8]; to the activations of hidden layers, as in linear hidden network transformation (LHN) [4, 5]; or to the softmax layer, as in LON [6] or in output-feature discriminative linear regression [8]. The authors of [9] describe a method based on linear transformation in the feature space and principal components analysis (PCA).

The second type of adaptation consists in re-training the entire network or only a part of it using special *regularization techniques* for improving generalization, such as L2-prior regularization [10], Kullback-Leibler divergence regularization [11], conservative training [12]. In [13] only a subset of the hidden units is retrained. The number of speaker-specific parameters is reduced in [14] through factorization based on singular value decomposition. Regularized adaptive training of subsets of DNN parameters is explored in [15].

The concept of *multi-task learning* (MTL) has recently been applied to the task of speaker adaptation in several works [16, 17, 18, 19] and has been shown to improve the performance of different model-based DNN adaptation techniques, such as LHN [17] and learning speaker-specific hidden unit contributions (LHUC) [18].

Using auxiliary features is another approach in which the acoustic feature vectors are augmented with additional speaker-specific or channel-specific features computed for each speaker or utterance at both training and test stages. An example of effective auxiliary features is i-vectors [20, 21, 22, 23]. Alternative methods are adaptation with speaker codes [24] and factorized adaptation [25].

Among the adaptation methods developed for DNNs, a few take advantage of robust adaptability of GMMs [7, 26, 27, 28, 29, 30]. The most common way of *combining GMM and DNN models* for adaptation is using GMM-adapted features, for example fMLLR, as input for DNN training [7, 26, 27]. Other methods include temporally varying weight regression [29] and GMM-derived features [31, 32].

There are also several *model-space adaptation* methods, that do not fall into the above categories, such as LHUC [33], where an amplitude parameter is introduced for each hidden unit, tied on a per-speaker basis, and then estimated using adaptation data; the adaptation parameters estimation via maximum a posteriori (MAP) linear regression [34]; and hierarchical MAP approach [35]. In [36] the shape of the activation function is changed to better fit the speaker-specific characteristics.

In this paper we investigate the GMM framework for adaptation of DNN-HMM acoustic models. Our approach is based on using features derived from a GMM model for training DNN models [31, 32, 37] and GMM-based adaptation techniques. In the previous works it was shown that GMM log-likelihoods can be effectively used as features for training a DNN HMM model, as well as for the speaker adaptation task. Experiments were reported on the small (15 hours) WSJ corpora with medium-size vocabularies (5K and 20K words).

http://dx.doi.org/10.21437/Interspeech.2016-1230

The purpose of this work is a detailed exploration and analysis of the method with respect to a state-of-the-art speech recognition DNN setup on a large vocabulary speech recognition task on a large corpus, to highlight the strengths and weaknesses of the proposed approach. We propose various novel ways for adaptation performance improvement. Firstly, we improve GMM-derived (GMMD) feature extraction by using BN features for training of the auxiliary GMM model, which is used for GMMD feature extraction. Secondly, we experiment with combination of GMMD features and conventional features at different levels of DNN architecture in order to discover the best possible configuration. Finally, we explore a novel approach for the combination of MAP and fMLLR techniques for SAT with GMMD features. In addition, we apply lattice-based information and confidence scores in MAP adaptation of the auxiliary GMM model to improve the adaptation performance.

The rest of the paper is organized as follows. In Section 2, SAT for DNN-HMM based on GMMD features is introduced. Section 3 describes MAP adaptation algorithm using lattices scores. The experimental results are given in Section 4. Finally, conclusions are presented in Section 5.

2. GMM framework for adaptation of DNN acoustic models

Construction of GMM-derived features for adapting DNNs was proposed in [31, 32], where it was demonstrated, using MAP and fMLLR adaptation as an example, that this type of features makes it possible to effectively use GMM-HMM adaptation algorithms in the DNN framework. In this work we improve the previously proposed scheme for GMM-derived feature extraction and apply the concept of GMMD features with adaptation to state-of-the art DNN architecture.

One of the main differences in GMMD feature extraction procedure in comparison with the previous works consists in changing the type of basic acoustic features for the auxiliary GMM model. In the past an auxiliary GMM model was trained on MFCC features, and then this GMM model was used to extract GMMD features for further training DNN model. In this work we investigate the effectiveness of the proposed approach on another level of DNN architecture. We use BN features from DNN to train GMM model for GMMD feature extraction. The motivation for using BN features in this approach is that for the better source features we can obtain better adaptation results. BN features allow us to capture long term spectro-temporal dynamics of the signal with GMDD features and are proven to be effective both for GMM and DNN acoustic model training [38].

The scheme for training DNN model with GMM adaptation framework is shown in Figure 1. First, 40-dimensional log-scale filterbank features concatenated with 3-dimensional pitch-features, are spliced across 11 neighbouring frames (5 frames on each side of the current frame), resulting in 473-dimensional (43×11) feature vectors. After that a DCT transform is applied and the dimension is reduced to 258. Then a DNN model for 40-dimensional BN features is trained on these features.

An auxiliary triphone or monophone GMM model is used to transform BN feature vectors into log-likelihoods vectors. At this step, speaker adaptation of the auxiliary SI GMM-HMM model is performed for each speaker in the training corpus and a new speaker-adapted (SA) GMM-HMM model is created in order to obtain SA GMM-derived features. For a given BN feature vector, a new GMM-derived feature vector is obtained by calculating log-likelihoods across all the states of the auxiliary

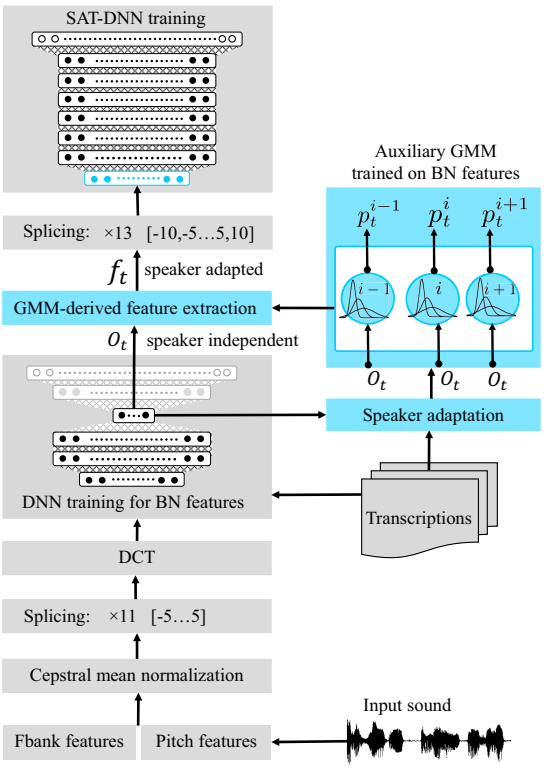

Figure 1: *Using speaker adapted GMM-derived features for SAT DNN-HMM training.*

GMM model on the given vector. Suppose o_t is the BN feature vector at time t, then the new GMM-derived feature vector f_t is calculated as follows:

$$f_t = [p_t^1, \ldots, p_t^n], \tag{1}$$

where n is the number of states in the auxiliary GMM-HMM model,

$$p_t^i = \log\left(P(o_t \mid s_t = i)\right) \tag{2}$$

is the log-likelihood estimated using the GMM-HMM. Here s_t denotes the state index at time t. Then the features are spliced in time taking a context size of 13 frames: [-10,-5...5,10]. We will refer to these resulting features as GMMD features. These features are used as the input for training the DNN. The proposed approach can be considered a feature space transformation technique with respect to DNN-HMMs trained on GMMD features.

3. MAP adaptation using lattices scores

The use of lattice-based information and confidence scores [39, 40] is a well-known method for improving the performance of unsupervised adaptation. In this work we use the MAP adaptation algorithm for adapting the SI GMM-HMM model [3]. Speaker adaptation of a DNN-HMM model built on GMMD features is performed through the MAP adaptation of the auxiliary GMM-HMM model, which is used for calculating GMMD features.

We modify the traditional MAP adaptation algorithm by using lattices instead of alignment to the 1-best hypothesis of the first decoding pass as follows. Let m denote an index of a Gaus-

sian in the SI acoustic model (AM), and μ_m the mean of this Gaussian. Then the MAP estimation of the mean vector is

$$\widehat{\mu}_m = \frac{\tau\mu_m + \sum_t \gamma_m(t) p_s(t) o_t}{\tau + \sum_t \gamma_m(t) p_s(t)}, \quad (3)$$

where τ is the parameter that controls the balance between the maximum likelihood estimate of the mean and its prior value; $\gamma_m(t)$ is the posterior probability of Gaussian component m at time t; and $p_s(t)$ is the confidence score of state s at time t in the lattice obtained from the first decoding pass by calculating arc posteriors probabilities. The forward-backward algorithm is used to calculate these arc posterior probabilities from the lattice as follows:

$$P(l|O) = \frac{\sum_{q \in Q_l} p_{acc}(O|q)^{\frac{1}{\alpha}} P_{lm}(w)}{P(O)}, \quad (4)$$

where α is the language model scale factor (the optimal value for α is found empirically); q is a path through the lattice corresponding to the word sequence w; Q_l is the set of paths passing through arc l; $p_{acc}(O|q)$ is the acoustic likelihood; $P_{lm}(w)$ is the language model probability; and $p(O)$ is the overall likelihood of all paths through the lattice.

In a particular case, when $p_s(t) = 1$ for all states and t, formula (3) represents the traditional MAP adaptation. In addition to this frame-level weighting scheme, we apply a confidence base selection scheme, when we use in (3) only those observations, for which confidence scores exceed a given threshold. For adaptation of a DNN AM, first, MAP adaptation of an auxiliary GMM model is performed and a new speaker-adapted (SA) GMM model is obtained, as described above. Second, at the recognition stage, SA GMMD features for DNN are calculated using this SA GMM.

4. Experimental results

4.1. Data sets

The experiments were conducted on the TED-LIUM corpus [41]. We used the last (second) release of this corpus [41]. This publicly available data set contains 1495 TED talks that amount to 207 hours (141 hours of male, 66 hours of female) speech data from 1242 speakers, 16kHz. For experiments with SAT and adaptation we removed from the original corpus data for those speakers, who had less than 5 minutes of data, and from the rest of the corpus we made four data sets: training set, development set and two test sets. Characteristics of the obtained data sets are given in Table 1. The motivation for creating the new test and development data sets was to obtain data sets, that are more representative and balanced in characteristics (gender, duration) than the original ones and more suitable for adaptation experiments.

For evaluation we use 150K word vocabulary and publicly available trigram language model *cantab-TEDLIUM-pruned.lm3*[1].

4.2. Baseline system

We used the open-source Kaldi toolkit [42] and followed mostly TED-LIUM Kaldi recipe to train the baseline system.

For training DNN models, first the initial GMM model was trained using 39-dimensional MFCC features with delta and acceleration coefficients. Linear discriminant analysis (LDA) followed by maximum likelihood linear transform (MLLT) and fMLLR transformation was then applied over these MFCC features to build a GMM-HMM system. Discriminative training with the boosted maximum mutual information (BMMI) objective was finally performed on top of this model.

[1] http://cantabresearch.com/cantab-TEDLIUM.tar.bz2

Table 1: *Data sets statistics*

Characteristic	Data set			
	Training	Development	Test$_1$	Test$_2$
Duration, hours				
Total	171.66	3.49	3.49	4.90
Male	120.50	1.76	1.76	3.51
Female	51.15	1.73	1.73	1.39
Duration per speaker, minutes				
Mean	10.0	15.0	15.0	21.0
Minimum	5.0	14.4	14.4	18.3
Maximum	18.3	15.4	15.4	24.9
Number of speakers				
Total	1029	14	14	14
Male	710	7	7	10
Female	319	7	7	4
Number of words	-	36672	35555	51452

4.2.1. SAT DNN on fMLLR features

Then a DNN was trained for BN feature extraction. The DNN system was trained using the frame-level cross entropy criterion and the senone alignment generated from the GMM system. For training this DNN, 40-dimensional log-scale filterbank features concatenated with 3-dimensional pitch-features, were spliced across 11 neighbouring frames, resulting in 473-dimensional (43×11) feature vectors. After that a DCT transform was applied and the dimension was reduced to 258. A DNN model for extraction 40-dimensional BN features was trained with the following topology: a 258-dimensional input layer; four hidden layers (HL), where the third HL was a BN layer with 40 neurons and other three HLs were 1500-dimensional; the output layer was 2390-dimensional. On the obtained BN features we trained the GMM model, which is used to produce forced alignment, and then SAT-GMM model was trained on fMLLR-adapted BN features. Then for training the final DNN model, fMLLR-adapted BN features were spliced in time with the context of 13 frames: [-10,-5...5,10]. The final DNN had a 520-dimensional input layer; six 2048-dimensional HLs with logistic sigmoid activation function, and a 4184-dimensional softmax output layer, with units corresponding to the CD states. The DNN parameters were initialized with stacked restricted Boltzmann machines (RBMs) by using layer by layer generative pre-training. The DNN was trained with an initial learning rate of 0.008 using the cross-entropy objective function.

4.2.2. Baseline SI-DNN

This model was trained in a similar way as the SAT DNN described above, but without fMLLR adaptation.

4.3. Impact of the auxiliary GMM and parameter τ on GMMD features

An auxiliary GMM is used for GMMD feature extraction. In order to speed up these preliminary experiments, we performed them on a smaller (85 hours) subset of the training dataset.

We aim to explore two factors related to this GMM: (1) the topoplogy of the model and (2) the type of input features for training this model, and choose the configuration, which is more effective for GMMD feature extraction. We experimented with the following parameters of GMM model: the total number of Gaussians and their distributions between states. Also GMM models were trained on two different types of input features: 39-dimensional MFCC and BN features, extracted as described in Section 4.2.1. In addition we extracted features with different values of adaptation parameter τ (in formula (3)). The performance results in terms of Word Error Rate (WER) for DNN models, used for BN feature extraction, are presented in Table 2. Parameter $Power$ in the table is the exponent for number of Gaussians according to occurrence counts. We can see that for GMMD feature extraction it is better to train an auxiliary GMM model on BN features than on MFCC, and that equal distribution of number of Gaussians between states ($Power = 0$) performs worse than distribution which is dependent on occurrence counts. We set parameter $\tau = 5$ for all the following experiments.

Table 2: *Impact of the parameters of the auxiliary model topology and τ on GMMD feature extraction (on the development set)*

Features	Gaussians	τ	Power	WER,%
MFCC	2500	5	0.5	13.89
	2500	5	0.0	14.05
	3800	5	0.5	13.75
	3800	5	0.0	13.65
BN	2500	1	0.5	13.69
	2500	3	0.5	13.51
	2500	5	0.5	13.34
	2500	7	0.5	13.33
	2500	10	0.5	13.40
	2500	5	0.0	13.34
	3800	5	0.5	13.33
	3800	5	0.0	13.48
	10200	5	0.5	13.92

4.4. Adaptation results

The adaptation experiments were conducted in an unsupervised mode on the test data using transcripts obtained from the first decoding pass. For this set of experiments all the training corpus is used. We empirically studied the approach described in Section 2 and applied it at different levels to the conventional recipe. The performance results in terms of WER for SI and SAT DNN-HMM models are presented in Table 3. The first two lines of the table correspond to the baseline SI and SAT DNNs, which were trained as described in Section 4.2. For the adaptation experiments we first (line 3) trained a SAT-DNN using only MAP adaptation, as shown in Figure 2 (adapted features AF_1), where the adapted GMMD features were concatenated with the conventional unadapted BN features. Second (line 4), we trained a SAT-DNN using MAP and fMLLR adapted features, as shown in Figure 3 (AF_2). For this model we used a triphone auxiliary GMM model instead of a monophone GMM, but kept for GMMD feature extraction only the most frequent states, as we noticed that this type of auxiliary GMM model gave slightly better results. Finally, we performed state-level system combination experiments for two SAT-DNNs adapted with MAP and with fMLLR (line 7). For this purpose we trained a DNN on fea-

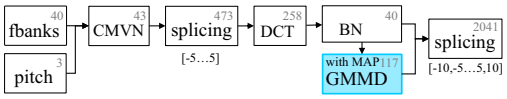

Figure 2: *Adapted features AF_1*

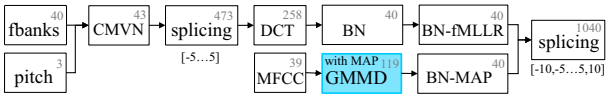

Figure 3: *Adapted features AF_2*

tures AF_1, but using the same units in the softmax output layer as in model 2. In this work combination results are shown for the mean of outputs of two DNNs. In addition we performed an adaptation experiment with using lattices scores instead of alignment (line 5), as described in Section 3. We can see that MAP adaptation on GMMD features can be complementary to fMLLR adaptation on conventional BN features.

Table 3: *Summary of the adaptation results for DNN models.*

	Model	Features	WER,%		
			Dev	Test_1	Test_2
1	SI	BN	13.16	11.94	15.43
2	SAT	fMLLR-BN	11.72	10.88	14.21
3	SAT	AF_1	11.56	10.51	13.94
4	SAT	AF_2	11.44	10.81	14.14
5	SAT	AF_1 +lattice	11.43	10.45	13.80
6	as 3, state tying from 2		11.40	10.65	14.03
7	posterior fusion: 2 and 6		11.01	10.25	13.52

5. Conclusions

In this paper we have investigated GMM framework for adaptation of DNN-HMM acoustic models and proposed various novel ways for adaptation performance improvement, such as, using bottleneck features for GMM-derived feature extraction, combination of GMM-derived with conventional features at different levels of DNN architecture, and finally, using lattice-based information and confidence scores in MAP adaptation of the auxiliary GMM model for DNN acoustic model adaptation. Experimental results on the TED-LIUM corpus demonstrate that, in an unsupervised adaptation mode, the proposed adaptation technique can provide approximately, a 14–16% relative WER reduction on different adaptation sets, compared to the SI DNN system built on conventional features, and a 3–6% relative WER reduction compared to the SAT-DNN trained on fMLLR adapted features. Experiments with different types of fusion show that MAP adaptation on GMMD features can be complementary to fMLLR adaptation on conventional BN features, and the most efficient type of fusion is the combination of DNN models on the state level.

6. Acknowledgements

This work was partially funded by the European Commission through the EUMSSI project, under the contract number 611057, in the framework of the FP7-ICT-2013-10 call.

7. References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, Jaitly *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.

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

[3] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech and Audio Proc.*, vol. 2, pp. 291–298, 1994.

[4] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. De Mori, "Adaptation of hybrid ANN/HMM models using linear hidden transformations and conservative training," in *Proc. ICASSP*, 2006.

[5] 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," 1995.

[6] B. Li and K. C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems," pp. 526–529, 2010.

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

[8] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Proc. SLT*. IEEE, 2012, pp. 366–369.

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

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

[11] 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. ICASSP*, 2013, pp. 7893–7897.

[12] D. Albesano, R. Gemello, P. Laface, F. Mana, and S. Scanzio, "Adaptation of artificial neural networks avoiding catastrophic forgetting," in *Proc. IJCNN'06*. IEEE, 2006, pp. 1554–1561.

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

[14] 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 *Proc. ICASSP*. IEEE, 2014, pp. 6359–6363.

[15] T. Ochiai, S. Matsuda, X. Lu, C. Hori, and S. Katagiri, "Speaker adaptive training using deep neural networks," in *Proc. ICASSP*. IEEE, 2014, pp. 6349–6353.

[16] S. Li, X. Lu, Y. Akita, and T. Kawahara, "Ensemble speaker modeling using speaker adaptive training deep neural network for speaker adaptation," in *Proc. INTERSPEECH*, 2015.

[17] Z. Huang, J. Li, S. M. Siniscalchi, I.-F. Chen, J. Wu, and C.-H. Lee, "Rapid adaptation for deep neural networks through multi-task learning," in *Proc. INTERSPEECH*, 2015.

[18] P. Swietojanski, P. Bell, and S. Renals, "Structured output layer with auxiliary targets for context-dependent acoustic modelling," in *Proc. INTERSPEECH*, 2015.

[19] R. Price, K.-i. Iso, and K. Shinoda, "Speaker adaptation of deep neural networks using a hierarchy of output layers," in *Proc. SLT*. IEEE, 2014, pp. 153–158.

[20] P. Karanasou, Y. Wang, M. J. Gales, and P. C. Woodland, "Adaptation of deep neural network acoustic models using factorised i-vectors," in *Proc. INTERSPEECH*, 2014, pp. 2180–2184.

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

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

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

[24] S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q. Liu, "Fast adaptation of deep neural network based on discriminant codes for speech recognition," *Audio, Speech, and Language Processing, IEEE/ACM Trans. on*, vol. 22, no. 12, pp. 1713–1725, 2014.

[25] J. Li, J.-T. Huang, and Y. Gong, "Factorized adaptation for deep neural network," in *Proc. ICASSP*. IEEE, 2014, pp. 5537–5541.

[26] S. P. Rath, D. Povey, K. Veselỳ, and J. Cernockỳ, "Improved feature processing for deep neural networks." in *Proc. INTERSPEECH*, 2013, pp. 109–113.

[27] H. Kanagawa, Y. Tachioka, S. Watanabe, and J. Ishii, "Feature-space structural MAPLR with regression tree-based multiple transformation matrices for DNN," 2015.

[28] X. Lei, H. Lin, and G. Heigold, "Deep neural networks with auxiliary Gaussian mixture models for real-time speech recognition," in *Proc. ICASSP*. IEEE, 2013, pp. 7634–7638.

[29] S. Liu and K. C. Sim, "On combining DNN and GMM with unsupervised speaker adaptation for robust automatic speech recognition," in *Proc. ICASSP*. IEEE, 2014, pp. 195–199.

[30] B. Murali Karthick, P. Kolhar, and S. Umesh, "Speaker adaptation of convolutional neural network using speaker specific subspace vectors of SGMM," 2015.

[31] N. Tomashenko and Y. Khokhlov, "Speaker adaptation of context dependent deep neural networks based on MAP-adaptation and GMM-derived feature processing," in *Proc. INTERSPEECH*, 2014, pp. 2997–3001.

[32] N. Tomashenko and Y. Khokhlov, "GMM-derived features for effective unsupervised adaptation of deep neural network acoustic models," in *Proc. INTERSPEECH*, 2015, pp. 2882–2886.

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

[34] Z. Huang, J. Li, S. M. Siniscalchi, I.-F. Chen, C. Weng, and C.-H. Lee, "Feature space maximum a posteriori linear regression for adaptation of deep neural networks," in *Proc. INTERSPEECH*, 2014, pp. 2992–2996.

[35] Z. Huang, S. M. Siniscalchi, I.-F. Chen, J. Li, J. Wu, and C.-H. Lee, "Maximum a posteriori adaptation of network parameters in deep models," in *Proc. INTERSPEECH*, 2015.

[36] S. M. Siniscalchi, J. Li, and C.-H. Lee, "Hermitian polynomial for speaker adaptation of connectionist speech recognition systems," *Audio, Speech, and Language Processing, IEEE Trans. on*, vol. 21, no. 10, pp. 2152–2161, 2013.

[37] J. P. Pinto and H. Hermansky, "Combining evidence from a generative and a discriminative model in phoneme recognition," IDIAP, Tech. Rep., 2008.

[38] F. Grézl, M. Karafiát, and K. Vesely, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *Proc. ICASSP*. IEEE, 2014, pp. 7654–7658.

[39] L. Uebel and P. C. Woodland, "Improvements in linear transform based speaker adaptation," in *Proc. ICASSP*, 2001, pp. 49–52.

[40] C. Gollan and M. Bacchiani, "Confidence scores for acoustic model adaptation," in *Proc. ICASSP*, 2008, pp. 4289–4292.

[41] A. Rousseau, P. Deléglise, and Y. Estève, "Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks." in *Proc. LREC*, 2014, pp. 3935–3939.

[42] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.