INTEGRATED ADAPTATION WITH MULTI-FACTOR JOINT-LEARNING FOR FAR-FIELD SPEECH RECOGNITION

Yanmin Qian[1,2] *Tian Tan*[1] *Dong Yu*[3] *Yu Zhang*[4]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2] Cambridge University Engineering Department, Cambridge, UK
[3] Microsoft Research, Redmond, WA, USA
[4] MIT CSAIL, Cambridge, MA, USA

ABSTRACT

Although great progress has been made in automatic speech recognition (ASR), significant performance degradation still exists in distant talking scenarios due to significantly lower signal power. In this paper, a novel adaptation framework, named ***integrated adaptation with multi-factor joint-learning***, is proposed to improve the recognition accuracy for distant speech recognition. We explore and extract speaker, phone and environment factor representations using deep neural networks (DNNs), which are integrated into the main ASR DNN to improve classification accuracy. In addition, the hidden activations in the main ASR DNN are used to improve the factor extraction, which in turn helps the ASR DNN. All the model parameters, including those in the ASR DNN and factor extractor DNNs, are jointly optimized under the multi-task learning framework. Further more, unlike prior techniques, our novel approach requires no explicit separate stages for factor extraction and adaptation. Experiments on the AMI single distant microphone (SDM) task show that the proposed architecture can significantly reduce word error rate (WER) and additional improvement can be achieved by combining it with the i-vector adaptation. Our best configuration obtained more than 15% and 10% relative reduction on WER over the baselines using the SDM and close-talk data generated alignments, respectively.

Index Terms— Far-field speech recognition, Deep neural network, Factor representation, Multi-task learning, Integrated adaptation

1. INTRODUCTION

We have witnessed significant progress made in automatic speech recognition (ASR) in the last few years especially after the introduction of the deep neural network (DNN) based acoustic models [1, 2, 3]. These new advancements have reduced the word error rate (WER) to a level that passed the threshold for adoption in many close-talk scenarios (e.g., voice search on a smart phone). However, these systems still perform poorly under the distant (far-field) talking condition [4], where the speech signals are captured by one or more microphones located father away from the speaker. Low signal strength is the main cause of the problem in this scenario since it leads to low signal to noise ratio (SNR) and makes the system susceptible to reverberation and additive noise in normal environment.

Many techniques [5, 6, 7] have been proposed to deal with the far field speech recognition problem. Model adaptation [8], which automatically adjusts the model's behavior based on the testing condition, is one of the most important methods proposed. Popular model adaptation techniques include maximum likelihood linear regression (MLLR) [9, 10] in the GMM-HMM systems, and linear transformation based techniques, such as linear input network (LIN), linear output network (LON) and linear hidden network (LHN), in the DNN-HMM framework [11, 12].

More recently, factor-aware (e.g., noise-aware, speaker-aware) adaptation received great attention for the DNN-HMM systems. In this adaptation framework, a good factor representation, in addition to the speech feature vector, is fed into DNNs as auxiliary information. In most such systems, the auxiliary information is used to provide factor-dependent bias to the DNN so that the DNN's output depends on the factor value.

I-vector, originally proposed for speaker recognition [13], can be directly used as speaker and channel representation for factor-aware DNN adaptation [14, 15]. In [16] a speaker code is used and jointly optimized along with the DNN. In the noise-aware [17] and room-aware training [6], the average noise vector and $T60$ value are used as factor representations, respectively, to indicate the noise and room conditions. Multiple factors are extracted with joint factor analysis (JFA) [18] or vector Taylor's sequence (VTS) expansion [19] and used in [20].

In all the factor-aware methods, a factor representation, that is constant with regard to the speaker or utterance, has to be explicitly estimated before adaptation happens for both training and testing. The factor extraction process can be completely independent of the recognition task such as in [14, 15, 17], or highly coupled with the recognition task as in [16]. In either way, it introduces significant latency since the factor representation can only be reliably estimated after observing enough speech frames, in most cases, the whole utterance.

In this work, we develop a DNN based approach to extract multiple factor representations. The extracted factors are integrated into and used to adapt the main DNN that conducts speech recognition. At the same time, the hidden layers in the main DNN is fed into and used to adapt the factor extractors. The model parameters in all the factor extractors and the main DNN are jointly trained under the multi-task learning framework. Unlike aforementioned works, in our proposed adaptation framework factor representations are dynamically estimated after observing each speech frame in the same pace speech recognition happens. There is no separate factor ex-

The work reported here was started at JSALT 2015 in UW, Seattle, and was supported by JHU via grants from NSF (IIS), DARPA (LORELEI), Google, Microsoft, Amazon, Mitsubishi Electric, and MERL. Yanmin Qian was partly supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology). Supplementary data for this paper is available at https://www.repository.cam.ac.uk/handle/1810/253406.

traction process. This eliminates the high latency problem often observed in other approaches. We evaluate our proposed approach on the far-field speech recognition task.

The remainder of the paper is organized as follows. In Section 2 we introduce the novel integrated adaptation framework with multi-factor joint-learning and describe factor extraction, factor integration and joint training in detail. We report experimental results in Section 3 and conclude the paper in Section 4.

2. MULTI-FACTOR INTEGRATION AND JOINT-TRAINING

2.1. Factor Extraction

Different from models in which the auxiliary information is coded as a constant vector across the whole utterance or speaker session, the factor representations are dynamically estimated using a DNN in our proposed framework. Our work is inspired by the previous works which used DNNs to extract speaker representation and achieved a good performance on both speaker [21, 22] and speech recognition [23, 24, 25]. Here, we extend this basic idea to extract not only speaker representation but also phone and environment representations, which are believed to be helpful for acoustic modelling.

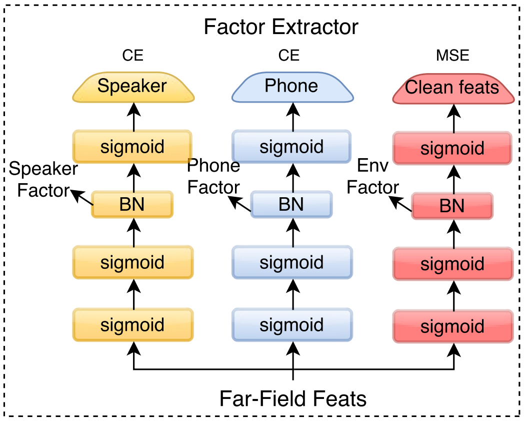

Fig. 1. The structure of the factor extractors, each of which is a four-layer DNN with a bottleneck layer in the middle. The speaker and phone extractors are trained using the cross-entropy criterion and the environment-factor remover is trained using the mean square error (MSE) criterion to construct close-talk features from the far-field features.

The structures of the factor extractors are illustrated in Figure 1. Each factor extractor is a four-layer DNN with a bottleneck layer in the middle. The output of the bottleneck layer is used as the representation of the specific factor the DNN is trained for. For different factors different targets and objective functions are used for model optimization. More specifically, the speaker and phone labels are used in speaker and phone factor extractors, respectively. These two DNNs are trained to differentiate among speakers and phones and are optimized using the cross-entropy criterion. The synchronized parallel far-field and close-talk data are utilized to learn an environment-remover representation. This DNN takes the far-field feature as the input and the close-talk feature as the reference target. In other words it is trained to learn the transformation from the far-field feature to the close-talk feature. We believe that this learned transformation encodes knowledge to remove room-dependent information related to reverberation and device from the input signal.

2.2. Factor Integration

The extracted factor representations can be fed into the input layer (similar to the way the augmented features are used in [6, 14, 17]), the hidden layer, or the output layer to aid the main ASR DNN to conduct speech recognition. In addition, information from the main ASR DNN can help extracting better factor representations. Figure 2 shows the architecture in which all the factors are fed into the output layer of the main ASR DNN while the hidden layer output of the main ASR DNN is fed as auxiliary information into the factor extractors. This later information flow, named cross-connections in this paper, is shown as the red line in Figure 2 and it is novel. With this design of factor integration and cross-connection, the main ASR DNN and factor extractors can benefit from each other and improve the performance.

Note that, our architecture does not rule out integration of factors extracted using existing techniques. For example, auxiliary features such as i-vector, $T60$ and speaking-rate can all be concatenated with the raw acoustic feature to form the model inputs.

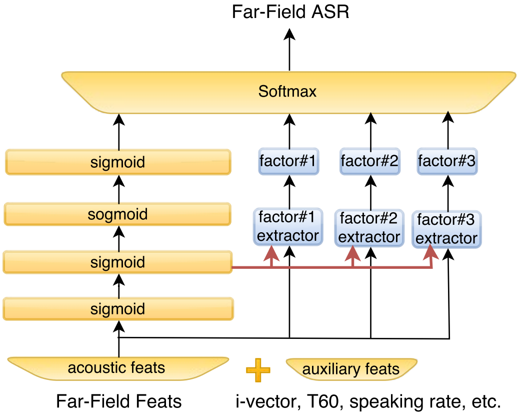

Fig. 2. The multi-factor integration and joint-learning architecture in which the factor representations are fed into the main ASR DNN's output layer while the hidden layer of the main ASR DNN is fed into the factor extractors.

2.3. Multi-task Learning

The model parameters in both factor extractors and the main ASR DNN are jointly learned under the multi-task learning framework [26, 27] from the randomly initialized model. More specifically, the objective function

$$E(\theta) = E_{asr}(\theta) + \lambda_1 E_{phn}(\theta) + \lambda_2 E_{spk}(\theta) + \lambda_3 E_{env}(\theta) \quad (1)$$

for the proposed model is a weighted sum of four criteria: the cross-entropy (CE) criterion $E_{asr}(\theta)$ used for the senone classification in the main ASR DNN, the CE criterion $E_{phn}(\theta)$ for the phone factor extraction, the CE criterion $E_{spk}(\theta)$ for the speaker factor extraction, and the MSE criterion $E_{env}(\theta)$ for the environment-factor remover. θ represents all the DNN parameters. λ_1, λ_2, and λ_3, which are set to 0.1, 0.1, and 0.01 in our study, are the *mixing weights* for the three factor extractors.

The multi-task joint learning proposed here is another key difference between our approach and prior arts that use DNNs to extract information representation. For example, in [23, 24] the authors firstly trained a DNN to extract a speaker code and then optimized the main ASR DNN with the speaker representation extractor fixed. In contrast, in our approach the factor extractors and the main ASR DNN are tightly coupled into one integrated framework and jointly optimized. There is no explicitly separated factor extraction and adaptation stage in both training and decoding. During decoding, only the senone softmax layer is needed which can be easily computed using the conventional feed-forward algorithm.

3. EXPERIMENTS AND RESULTS

3.1. Experimental setup and baseline systems

To evaluate the proposed approaches, a series of experiments were performed on the AMI single distant microphone (SDM) task. There are about 80 hours and 8 hours in training and evaluation sets respectively [5].

In this work, we exploited Kaldi [28] for building speech recognition systems and CNTK [29] for training our novel DNN architectures. We first followed the officially released kaldi recipe to build an LDA-MLLT-SAT GMM-HMM model. This model uses 39-dim MFCC feature and has roughly 4K tied-states and 80K Gaussians. We then use this acoustic model to generate the senone alignment for neural network training. In the DNN-HMM systems, 40-dimensional log mel-filter bank features with delta and delta-delta are used. The DNN input layer is formed from a contextual window of 11 frames or 1320 units. The DNN baseline has 6 hidden layers with 2048 Sigmoidal units in each layer. The networks are trained using the stochastic gradient descent (SGD) based backpropagation (BP) algorithm, with minibatch size of 256.

For decoding, we used the 50K-word AMI dictionary and a trigram language model interpolated from the one created using the AMI training transcripts and that using the Fisher English corpus. During the decoding we followed the standard AMI recipe and did not rule out overlapping segments. About 10% absolute WER reduction can be achieved if we don't consider these segments.

Besides the standard full training set, a randomly selected 10K-utterance subset (about 10 hours) is used for fast model training and evaluation. The training procedures and test sets are identical in the sub- and full-set experiments. Since the IHM and SDM data are synchronized and the quality of the IHM data is much higher than that of the SDM data, we trained another SDM baseline using the IHM model generated senone alignment. The performance of these two baselines, which are comparable with other works [5, 30], are presented in Table 1.

Table 1. WER (%) of the Baseline Systems on the SDM Data

System	Alignment	Sub Set	Full Set
DNN-HMM	SDM	68.3	58.8
DNN-HMM	IHM	65.2	55.9

3.2. Evaluation of the proposed strategies

The proposed integrated adaptation architecture with multi-factor joint-learning is evaluated in this subsection. In all the experiments reported below we used the IHM alignment since it is better than the SDM alignment as shown in Table 1 and since the IHM data are also used to train the environment-factor remover. The same 1320-dim contextually expanded FBANK features are used as the inputs for both the main ASR DNN and factor extractor DNNs. The main ASR DNN is configured to have 6 hidden layers with 2048 units per layer. All the factor extractors have 4 hidden layers with the 100 dimension bottleneck in the third layer. The output dimensions of all the modules are shown below:

- **Main ASR DNN**: with 4K units, which is the number of senones in the HMM model.
- **Speaker factor**: with 547 units, which corresponds to the number of speakers in the AMI training set.
- **Phone factor**: with 176 units, which is the number of position-dependent phones in AMI dictionary.
- **Environment-factor remover**: with 1320 units, the same size as the input feature since it tries to estimate the close-talk context-expansion FBANK features.

With this model size configuration, the entire architecture is trained following the multi-task learning procedure described in Section 2.

We first compared the performance of systems in which only one factor is used and the factor representation is integrated at input, hidden, and output layers of the main ASR DNN. The results achieved using the 10k-utterance SDM subset are illustrated in Table 2. From the table we can observe that all factors are helpful when they are used alone no matter which layer they are integrated to. This demonstrates that the neural network based factor extraction is effective for the far-field speech recognition task we evaluated. On the other hand, at which layer the integration happens does matter. In fact, integrating the factors at the output layer consistently and significantly outperforms the system where the integration happens at the input and hidden layers. This is likely because at the output layer the factors can have more direct effect to the estimated posteriors. Among the three factors, the environment-remover representation, which is believed to be especially important for the far-field scenarios, seems to perform best although the difference is not significant when integrated at the output layer.

Table 2. WER (%) comparisons of the proposed multi-factor joint-learning DNN (denoted as MF-DNN) using different factors on the SDM 10k-utterance subset. The IHM generated alignment is used in all setups. Different factor-integration layers are investigated.

System	Factor	Integration	WER(%)
DNN	—	—	65.2
MF-DNN	Speaker	Input	63.8
		Hidden	63.7
		Output	**61.6**
	Phone	Input	64.1
		Hidden	63.7
		Output	**61.4**
	Env	Input	63.2
		Hidden	62.0
		Output	**61.2**

We then investigated the efficacy of adding the cross-connection in the architecture. Since the best integration position is the output layer, we added the cross-connection only to that configuration. The results on the 10k-utterance SDM subset are presented in Table 3. These results clearly show that adding cross-connections can provide consistent improvements for every factor. This confirmed our conjecture that with this cross-connection the main ASR DNN and the factor DNNs can benefit from each other.

We further integrated the multiple factors into one framework to build the final multi-factor joint-learning architecture. As shown in the bottom rows of Table 3, significant improvement can be obtained with multi-factor integration compared to the single-factor integration. Compared to the baseline trained using the IHM alignment, our proposed multi-factor assisted joint-learning method gets 8% relative reduction on WER.

Table 3. WER (%) comparisons of the proposed multi-factor joint-learning DNN with and without the cross-connection (denoted as X-connection) on the 10k-utterance SDM subset. The IHM alignment is used in all setups.

System	Factor	Integration	WER(%)
DNN	—	—	65.2
MF-DNN	Speaker	Output	61.6
		+X-connection	**61.0**
	Phone	Output	61.4
		+X-connection	**60.8**
	Env	Output	61.2
		+X-connection	**60.7**
	Spk+Phn+Env	Output	60.4
		+X-connection	**60.1**

As we mentioned in Section 2, our proposed approach demands no separate factor estimation stage. It is interesting to see whether the proposed approach can be combined with other adaptation technologies. In this work, we extract a 128 dimensional i-vector for each speaker using a 2048-component GMM and concatenate the i-vector with the raw acoustic feature as the inputs to both the ASR DNN and the factor extractor DNNs (shown in Figure 2). The results using the combined feature is summarized in Table 4. These results show that both the proposed approach and the i-vector based approach can achieve substantial gains. These two approaches are also complementary as additional improvement can be obtained by combining two. The best result is achieved when the environment-remover and phone factor are combined with i-vector. No further gain is observed when DNN based speaker factor is also used. This is because the speaker information has been well represented in i-vector.

Table 4. WER (%) of different combinations of the proposed multi-factor joint-learning DNN with the i-vector based adaptation on the SDM 10k-utterance subset. The IHM alignment is used in all setups.

System	Factor	WER(%)
DNN	—	65.2
	i-vector	62.5
MF-DNN	i-vector+Env	57.9
	i-vector+Env+Phn	**57.1**
	i-vector+Env+Phn+Spk	57.9

Finally, the proposed multi-factor joint-learning using the best structure and configuration is evaluated on the full AMI SDM corpus, and the results are listed in Table 5. The conclusion on the full corpus is consistent with that on the subset: significant gain can be observed when using multi-factor joint-learning and additional improvement can be obtained when further combining i-vector. Overall, on the SDM full set we reduced the WER from 58.8% to 55.9% by using the IHM generated alignment, and further reduced it to 50.0% with the proposed approach. This translates to 15% and 10% relative error reduction over the baselines using the SDM and IHM alignments, respectively. In addition, even larger improvement is observed on the subset, which demonstrates that the proposed approach can be especially useful when only small training set is available, e.g., when building a new system for a new language or a new task.

Table 5. WER (%) comparisons of the proposed architecture on the full set, all with IHM alignment

System	Sub Set	Full Set
DNN	65.2	55.9
DNN+i-vector	62.5	52.0
MF-DNN	60.1	53.5
MF-DNN+i-vector	**57.1**	**50.0**

4. CONCLUSION

In this paper we proposed a novel integrated adaptation framework with multi-factor joint-learning for the far-field speech recognition. Several useful factors, including speaker, phone and environment, are explored for the distant scenarios. DNNs are used to extract factor representations, which can be integrated with the ASR DNN. In this unified framework, the factors are fed into the main ASR DNN and the hidden layer of the main ASR DNN is fed into the factor extractors so that they benefit from each other. Different from previous works in which all the modules are trained separately, in our proposed approach all the DNN parameters are jointly optimized under the multi-task learning framework. In addition, our approach requires no separation of the factor estimation and adaptation stages in both training and decoding. Both factor estimation and adaptation are embedded inside the framework. We observe that the best result is achieved when the factors are integrated to the main ASR DNN at the output layer, and adding the cross-connection from the main ASR DNN to the factor extractors helps.

This novel integrated adaptation architecture can also be easily combined with other factor-based adaptation techniques. The final best multi-factor joint-learning architecture combined with i-vector adaptation obtains more than 15% and 10% relative reduction on WER for the AMI SDM task over the baselines using the SDM and IHM alignments, respectively.

5. REFERENCES

[1] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdelrahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, and Brian Kingsbury, “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, 2012.

[2] Frank Seide, Gang Li, and Dong Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proceedings of Interspeech*, 2011, pp. 437–440.

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

[4] Thomas Hain, Luká Burget, John Dines, Philip N Garner, František Grézl, Asmaa El Hannani, Marijn Huijbregts, Martin Karafiat, Mike Lincoln, and Vincent Wan, "Transcribing meetings with the amida systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 486–498, 2012.

[5] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in *Proceedings of ASRU*, 2013, pp. 285–290.

[6] Ritwik Giri, Michael L Seltzer, Jasha Droppo, and Dong Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," in *Proceedings of ICASSP*, 2015, pp. 5014–5018.

[7] Takuya Yoshioka, Armin Sehr, Marc Delcroix, Keisuke Kinoshita, Roland Maas, Tomohiro Nakatani, and Walter Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[8] Seyedmahdad Mirsamadi and John HL Hansen, "A study on deep neural network acoustic model adaptation for robust far-field speech recognition," in *Proceedings of Interspeech*, 2015, pp. 2430–2434.

[9] Christopher J Leggetter and Philip C 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.

[10] Mark JF Gales and Philip C Woodland, "Mean and variance adaptation within the mllr framework," *Computer Speech and Language*, vol. 10, no. 4, pp. 249–264, 1996.

[11] 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, pp. 526–529.

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

[13] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[14] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proceedings of ASRU*, 2013, pp. 55–59.

[15] Penny Karanasou, Yongqiang Wang, Mark JF Gales, and Philip C Woodland, "Adaptation of deep neural network acoustic models using factorised i-vectors," in *Proceedings of Interspeech*, 2014, pp. 2180–2184.

[16] Shaofei Xue, Ossama Abdel-Hamid, Hui Jiang, Lirong Dai, and Qingfeng Liu, "Fast adaptation of deep neural network based on discriminant codes for speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1713–1725, 2014.

[17] Michael L Seltzer, Dong Yu, and Yongqiang Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proceedings of ICASSP*, 2013, pp. 7398–7402.

[18] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[19] Pedro J Moreno, *Speech recognition in noisy environments*, Ph.D. thesis, Carnegie Mellon University Pittsburgh, 1996.

[20] Jinyu Li, Jui-Ting Huang, and Yifan Gong, "Factorized adaptation for deep neural network," in *Proceedings of ICASSP*. IEEE, 2014, pp. 5537–5541.

[21] Yuan Liu, Yanmin Qian, Nanxin Chen, Tianfan Fu, Ya Zhang, and Kai Yu, "Deep feature for text-dependent speaker verification," *Speech Communication*, vol. 73, pp. 1–13, 2015.

[22] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Jorge Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proceedings of ICASSP*, 2014, pp. 4052–4056.

[23] Hengguan Huang and Khe Chai Sim, "An investigation of augmenting speaker representations to improve speaker normalisation for dnn-based speech recognition," in *Proceedings of ICASSP*, 2015, pp. 4610–4613.

[24] Yulan Liu, Penny Karanasou, and Thomas Hain, "An investigation into speaker informed dnn front-end for lvcsr," in *Proceedings of ICASSP*, 2015, pp. 4300–4304.

[25] Marc Ferras and Hervé Bourlard, "Mlp-based factor analysis for tandem speech recognition," in *Proceedings of ICASSP*, 2013, pp. 6719–6723.

[26] Dongpeng Chen, Brian Mak, Cheung-Chi Leung, and Sunil Sivadas, "Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition," in *Proceedings of ICASSP*, 2014, pp. 5592–5596.

[27] Nanxin Chen, Yanmin Qian, and Kai Yu, "Multi-task learning for text-dependent speaker verification," in *Proceedings of Interspeech*, 2015, pp. 185–189.

[28] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *Proceedings of ASRU*. Dec. 2011, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.

[29] Dong Yu, Adam Eversole, Mike Seltzer, Kaisheng Yao, Zhiheng Huang, Brian Guenter, Oleksii Kuchaiev, Yu Zhang, Frank Seide, Huaming Wang, et al., "An introduction to computational networks and the computational network toolkit," Tech. Rep., Tech. Rep. MSR, Microsoft Research, 2014, http://codebox/cntk, 2014.

[30] Ivan Himawan, Petr Motlicek, David Imseng, Blaise Potard, Namhoon Kim, and Jaewon Lee, "Learning feature mapping using deep neural network bottleneck features for distant large vocabulary speech recognition," in *Proceedings of ICASSP*, 2015, pp. 4540–4544.