AN INVESTIGATION INTO USING PARALLEL DATA FOR FAR-FIELD SPEECH RECOGNITION

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

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

ABSTRACT

Far-field speech recognition is an important yet challenging task due to low signal to noise ratio. In this paper, three novel deep neural network architectures are explored to improve the far-field speech recognition accuracy by exploiting the parallel far-field and close-talk recordings. All three novel architectures use multi-task learning for the model optimization but focus on three different ideas: ***dereverberation and recognition joint-learning***, ***close-talk and far-field model knowledge sharing***, and ***environment-code aware training***. Experiments on the AMI single distant microphone (SDM) task show that each of the proposed method can boost accuracy individually, and additional improvement can be obtained with appropriate integration of these models. Overall we reduced the error rate by 10% relatively on the SDM set by exploiting the IHM data.

Index Terms— Far-field speech recognition, Deep neural network, Multi-task learning, Feature denoising, Parallel data

1. INTRODUCTION

Despite the significant advancement made in automatic speech recognition (ASR) after the introduction of deep neural network (DNN) based acoustic models [1, 2, 3], the far-field speech recognition remains a challenging problem [4]. In the distant talking scenarios, the speech signal is captured by one or more microphones located farther away from the speaker, which makes it susceptible to distortion from reverberation and additive noise.

Many technologies [5, 6, 7] have been proposed to handle the far-field speech recognition problem. Most of these existing methods can be grouped into two categories: front-end based and back-end based [8]. Front-end based approaches operate on the signal or the feature, and attempt to remove the corrupting reverberation or noise from the observations prior to recognition [7, 9]. Back-end based methods leave the observations unchanged and instead update the model parameters to match the corrupted speech in distant scenarios [6, 10].

The performance on far-field speech recognition degrades significantly even with DNN based acoustic models since robustness is still a concern in the DNN-HMM systems [11]. Several methods have been proposed in the DNN-HMM framework to improve the far-field speech recognition [6, 9, 10]. One type of methods learn a DNN based feature enhancement model using the time-synchronize parallel data, e.g., the artificially generated clean and noisy speech pairs, or the simultaneously collected close- and distant-talking speech pairs. An acoustic model is then built on the denoised feature [12, 13, 14, 15].

In this work, we investigate how to exploit parallel data to more effectively recognize far-field speech. We explore three novel deep neural network architectures which are optimized using the **multi-task learning** technique. In the **dereverberation and recognition joint-learning** architecture, we integrate the dereverberation and recognition into one structure. In the **close-talk and far-field model knowledge sharing** architecture, we use the close-talk model to influence the far-field model. In the **environment-code aware training** architecture we utilize the environmental representation extracted from a denoiser to improve the acoustic model. We evaluate and compare these novel architectures in detail. To our best knowledge, this is the first comprehensive work on exploiting parallel data for improving the far-field speech recognition.

The remainder of the paper is organized as follows. In Section 2 the novel architectures that exploit parallel data for the far-field speech recognition are proposed and described. In Section 3 experimental results on the AMI single distant microphone setup are reported and analyzed. We conclude the paper in Section 4.

2. STRATEGIES ON USING PARALLEL DATA

In this section, we propose and describe three novel DNN architectures that exploit close-talk and far-field parallel data for distant speech recognition.

2.1. Multi-task learning

Unlike the normal DNNs which optimize just one criterion (e.g., the cross-entropy criterion), DNNs in the multi-task learning framework jointly optimize more than one criteria in model training. For example, Chen et al. trained acoustic models that optimize both triphone and trigrapheme classification accuracy [16]. Heigold et al. trained multilingual systems that optimize for several languages simultaneously [17]. Chen et al. applied the multi-task learning technique to optimize phone and speaker classification accuracy at the same time [18]. Multi-task learning is typically used to regularize the model or to borrow knowledge from other information sources. In this work, multi-task learning is utilized in all proposed architectures that we will describe next.

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/1810/253405.

We would like to thank Dr. Shinji Watanabe, Dr. Michael L. Seltzer and Vijayaditya Peddinti for their useful discussions and suggestions.

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

2.2. Dereverberation and recognition joint-learning

The first architecture is motivated by the recent work, which used the parallel data to train a DNN based denoising front-end [12, 13, 19]. In this work, we integrate the denoising and classification components into one unified structure with multi-task learning. More specifically we optimize two tasks: the dereverberation task and the recognition task. In the dereverberation task we design a regression model to estimate the close-talk speech given the far-field speech. This regression model can be optimized using the time-synchronized close-talk and far-field speech pairs by minimizing the mean squared error (MMSE)

$$E_{mse} = \sum_{t=1}^{T} ||\overline{\mathbf{x}}_t - \mathbf{x_t}||_2^2 \tag{1}$$

between the DNN outputs $\overline{\mathbf{x}}_t$ and the referenced close-talk features $\mathbf{x_t}$ at each frame t. In the recognition task we learn a discriminative model to classify senones by optimizing the cross-entropy (CE) criterion

$$E_{ce} = \sum_{t=1}^{T} \mathbf{D_t} \log(\mathbf{P_t}), \tag{2}$$

where $\mathbf{D_t}$ and $\mathbf{P_t}$ represent the target state probabilities and the estimated state posteriors at frame t, respectively.

Two architectures, denoted as the *parallel* structure and the *front-back* structure, are developed and compared for the dereverberation and recognition joint-learning as illustrated in Figure 1. The parallel structure at left is widely used in other multi-task learning tasks [16, 17, 18], including the REVERB Challenge [6]. This parallel structure is consisted of fully shared hidden layers at the bottom and two task-dependent softmax layers on top.

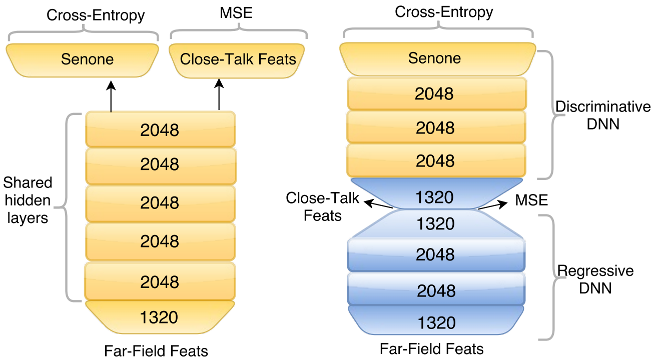

Fig. 1. The dereverberation and recognition joint-learning framework: parallel (left) and front-back (right) structures.

The front-back structure at right is inspired by the work in [15] and extended for the far-field speech recognition task. In this structure, the front-end DNN is optimized as a dereverberation model and the back-end DNN is learned to classify senones. Because the FBANK features are used as the target outputs of the front-end DNN, the back-end DNN can be concatenated on top seamlessly to form a complete DNN structure. In contrast to the work in [15], which optimizes the classification DNN after the denoising DNN is trained, our approach optimizes two components jointly with the multi-task learning framework.

Although these two structures are different, they are both jointly trained in the same way to optimize the interpolated CE and MMSE criteria

$$E(\theta) = E_{ce}(\theta) + \lambda E_{mse}(\theta) \tag{3}$$

where θ represents the whole DNN model parameter set, $E_{ce}(\theta)$ and $E_{mse}(\theta)$ are the cross-entropy and mean square error objective functions defined in equations (1) and (2), respectively, and λ is a *mixing factor* to balance these two criteria.

2.3. Close-talk and far-field model knowledge sharing

As we know, the performance of the close-talk ASR system is much better than that of the far-field system. For example, we can achieve 27% word error rate (WER) on the close-talk setup but only 56% WER on the single distant microphone (SDM) setup in the AMI meeting transcription task. An interesting question is whether the far-field model can learn from the close-talk model to achieve better recognition accuracy, similar to the work in [20] where the small-size DNN is taught by the large-size DNN to get a better performance.

Our proposed structure that encourages knowledge sharing between close-talk and far-field models is illustrated in Figure 2. In this novel structure a close-talk DNN and a far-field DNN are linked together to enable knowledge sharing. The close-talk and far-field models are trained with cross-entropy (CE) criterion. The knowledge sharing is achieved by minimizing the mean square error (MSE) between the outputs of the corresponding two hidden layers in the two parallel models.

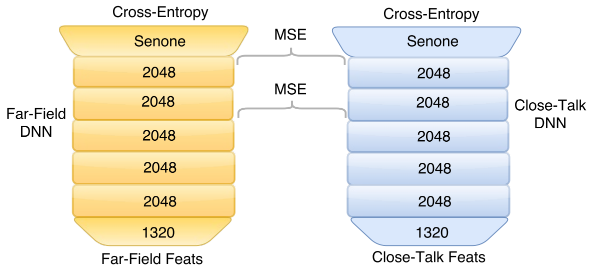

Fig. 2. Close-talk and far-field model knowledge sharing through the links bridging corresponding hidden layers

The model parameters of the entire architecture are jointly learned to optimize the interpolated objective function

$$E(\theta) = E_{ce\_far}(\theta) + E_{ce\_close}(\theta) + \lambda E_{mse}(\theta) \tag{4}$$

where $E_{ce\_far}(\theta)$ and $E_{ce\_close}(\theta)$ are the cross-entropy criteria for the far-field and close-talk DNNs respectively, $E_{mse}(\theta)$ is the mean square error between the hidden layer outputs of the close-talk and far-field DNNs, and λ is the *mixing factor*.

After model training, the close-talk DNN and the links between two DNNs can be discarded. Only the far-field DNN is used in the decoding.

2.4. Environment-code aware training

In this section, we propose the environment-code aware training. Several methods have been proposed to extract the environment or speaker features and to use them as auxiliary information to improve the speech recognition accuracy. Notable works include the IVector based adaptation [21], noise-aware training [22] and room-aware training [6].

Different from these previous works, where the auxiliary information is coded as a constant value across the whole utterance, in our architecture the environment code is dynamically estimated using a neural network. Specifically the synchronized parallel data are utilized to learn an environment-related representation. In Figure 3, the neural network at left is used to extract the room-dependent environment-code. 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 some room-dependent information that is related to reverberation and device. We used the outputs of a bottleneck layer as the representation of the environment-code.

This environment-code can be fed to the input layer (similar to the augmented feature discussed in the works [6, 21, 22]), to the hidden layer, or to the output layer as shown in Figure 3. Different from previous works such as [23], in our approach the environment-code extractor and the recognition DNN are jointly learned by optimizing the interpolated criteria

$$E(\theta) = E_{ce}(\theta) + \lambda E_{mse}(\theta) \tag{5}$$

instead of training the environment (or speaker)-code extractor first and then the recognition model.

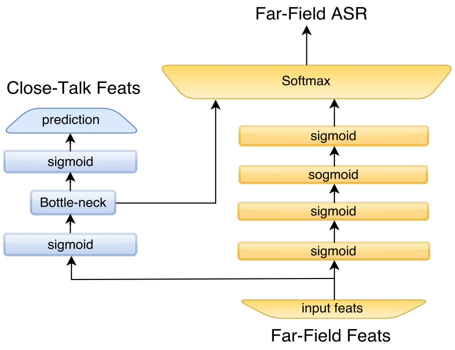

Fig. 3. The environment-code aware training framework.

Note that the previous work using static auxiliary feature ([6, 21]) requires a separate process to estimate the auxiliary information before the acoustic score calculation can be started. In our proposed approach the environment-code is embedded in the acoustic model and no extra pre-computation step is needed for the realtime decoding.

3. EXPERIMENTS AND RESULTS

3.1. Experimental setup and baseline systems

To evaluate the proposed approaches, a series of experiments were performed on AMI corpus, which contains around 100 hours of meetings recorded in specifically equipped instrumented meeting rooms at three sites in Europe (Edinburgh, IDIAP, TNO) [4]. Acoustic signal is captured and synchronized by multiple microphones including individual head microphones (IHM, close-talk), lapel microphones, and one or more microphone arrays. For the far-field speech recognition in this work, the condition using the single distant microphone (SDM, the first microphone in the primary array) is evaluated, and the simultaneously recorded IHM (close-talk) data are used to form the parallel data pairs. Our experiments adopted the suggested AMI corpus partition that contains about 80 hours and 8 hours in training and evaluation sets respectively [5].

In this work, we exploited Kaldi [24] for building speech recognition systems and CNTK [25] 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 the other 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, a simple way to exploit the close-talk data for improving the far-field speech recognition is to use the IHM model to generate the senone alignment and use it to train the SDM model. The performance of these two baselines, which are comparable with other works [5, 10], are presented in Table 1. From the table, we can clearly observe that a substantial improvement on the SDM set can be achieved by using alignments from the synchronized IHM data.

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

In this subsection we report our evaluation on the novel strategies we described in Section 2. In all the experiments reported below we used the IHM alignment since it is better than the SDM alignment as shown in Table 1. The same 1320-dim contextually expanded FBANK features are used as the inputs in all the novel architectures. The performance comparisons on the baselines and the new strategies for the SDM subset setup are illustrated in Table 2.

1) ***Multi-Condition Training***: The most straightforward method to use the IHM and SDM parallel data is the multi-condition training (also named multi-style training [22]), which just pools all the data from different conditions to train the model. Shown as the first two lines in the table, multi-condition training can only provide a small gain over the baseline. This is consistent to the conclusion in [10].

2) ***Dereverberation and Recognition Joint-Learning (DRJL)***: As described in Section 2.2, we built a parallel and a front-back structure. In the parallel structure the bottom 6 hidden layers are shared for both dereverberation and recognition. In the front-back structure 3 hidden layers are used for dereverberation whose output is fed into another 3 hidden layers used for recognition. Both structures provide meaningful gains for the far-field scenario while the front-back structure (with a 3.7% WER reduction over the baseline) outperformed the parallel structure (with a 2.4% WER reduction over the baseline). We believe this is because the front-back structure can exploit the dereverberation results more directly.

Table 2. WER (%) comparisons on the proposed strategies for using parallel data to improve the SDM condition. The 10k-utterance subset and IHM alignment are used in all setups. (**DRJL** denotes approach #1: Dereverberation and Recognition Joint-Learning; **CFMKS** denotes approach #2: Close-talk and Far-field Model Knowledge Sharing; **Env-code** denotes approach #3: Environment-code aware training)

System			WER (%)
DNN-HMM			65.2
IHM+SDM Multi-Cond DNN-HMM			64.2
DRJL	DRJL-Structure	Parallel	62.8
		Front-Back	**61.5**
CFMKS	MSE-Position	Low-Hidden	66.2
		Mid-Hidden	64.5
		High-Hidden	**61.7**
Env-code	Code-Integration	Input-Layer	63.2
		Hidden-Layer	62.0
		Output-Layer	**61.2**

*3) **Close-talk and Far-field Model Knowledge Sharing (CFMKS)***: This architecture is constructed following Section 2.3. Both the close-talk and far-field DNNs have 6 hidden layers. The related results are shown in the middle block of Table 2. In this study we compared the performance of systems where the MSE constraint (i.e., knowledge sharing) is added at different hidden layers: from the lower layers to the higher ones. Results indicate that enforcing knowledge sharing closer to the output layer performs better since higher layers have stronger influence to the output.

*4) **Environment-code aware training (Env-code)***: In this setup, the DNN for the environment representation has 4 hidden layers, with a 100-dimension bottleneck layer in the third layer. The ASR DNN also has 6 hidden layers. The environment-code is explored to be integrated into the ASR DNN at different layers. The results using this strategy are illustrated at the bottom block in Table 2. The results indicate that the proposed environment-dependent representation is useful, and integrating this code into the acoustic model is effective for the far-field speech recognition. Among the different integration strategies, connecting environment-codes to the output layer achieves the best performance. This is likely because at the output layer the environment code can have more direct effect to the estimated posteriors.

Figure 4 summarizes the comparison of the proposed strategies using the parallel data on the subset, and the new architectures all get a large improvement for the distant speech recognition.

At the first glance it seems that the proposed novel architectures are quite complicated. In fact, however, there is no additional computational cost in the **DRJL** and **CFMKS** architectures during decoding, and there is only slight additional cost in the **Env-code** architecture caused by the code-representation DNN. In other words all these new models are suitable for real-time applications.

3.3. Evaluation on the full set

In this subsection we evaluate and compare the best configuration in each proposed strategy on the full AMI SDM corpus. The results listed in Table 3 show that the gains we observed on the subset can be carried over to the full set although the improvement becomes smaller.

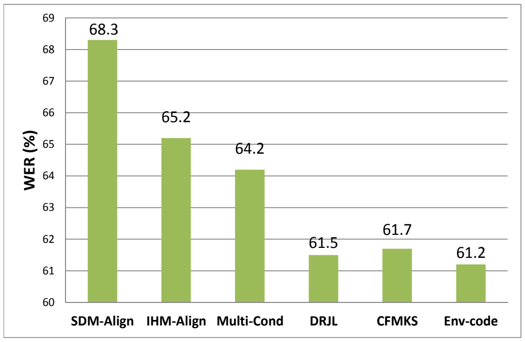

Fig. 4. Comparison of the proposed strategies using the parallel data.

We further investigate whether additional WER reduction can be obtained by combining different strategies. From the last two rows in Table 3 we can see that these architectures may be complementary, and we do get additional gains when combining DRJL with CFMKS. However, no additional improvement is observed when further integrating the Env-code strategy.

Overall, by exploiting the IHM data, we reduced the WER from 58.8% to 53.2%, a 10% relative reduction. Half of the gain is from using the IHM alignment and half of the gain is from using our novel architectures that exploit IHM data.

Table 3. WER (%) Comparisons of the Proposed Strategies on the Full Set, all with IHM alignment

System	Sub Set	Full Set
DNN-HMM	65.2	55.9
DRJL	61.5	53.8
CFMKS	61.7	54.0
Env-code	61.2	54.0
DRJL+CFMKS	60.1	**53.2**
DRJL+CFMKS+Env-code	**59.7**	53.3

4. SUMMARY

In this paper we proposed several novel architectures for exploiting the parallel data for improving far-field speech recognition. The key ideas we explored include the dereverberation and recognition joint-learning that integrates the dereverberation and classification components into the same architecture, the close-talk and far-field model knowledge sharing that enables IHM model to transfer knowledge to the SDM model by adding constraints between hidden layers of two DNNs, and the environment-code aware training that utilizes the parallel data to extract the environment representation and use it in the recognition DNN. All these novel architectures are trained with the multi-task learning strategy by jointly optimizing multiple criteria. All the proposed architectures are effective and can improve the far-field speech recognition accuracy. Overall, we reduced the WER on the SDM set by 10% relatively over the baseline by exploiting the IHM data.

In this study, we explored the novel architectures upon the DNN model. We will apply the same idea to the long short-term memory (LSTM) recurrent neural networks as the next step.

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] Marc Delcroix, Takuya Yoshioka, Atsunori Ogawa, Yotaro Kubo, Masakiyo Fujimoto, Nobutaka Ito, Keisuke Kinoshita, Miquel Espi, Takaaki Hori, Tomohiro Nakatani, et al., "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge," in *Proceedings of REVERB challenge workshop*, 2014.

[8] 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.

[9] Takashi Yoshioka, Xie Chen, and Mark JF Gales, "Impact of single-microphone dereverberation on dnn-based meeting transcription systems," in *Proceedings of ICASSP*. IEEE, 2014, pp. 5527–5531.

[10] 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.

[11] Yan Huang, Dong Yu, Chaojun Liu, and Yifan Gong, "Multi-accent deep neural network acoustic model with accent-specific top layer using the kld-regularized model adaptation," in *Proceedings of Interspeech*, 2014, pp. 2977–2981.

[12] Jun Du, Qing Wang, Tian Gao, Yong Xu, Lirong Dai, and Chin-Hui Lee, "Robust speech recognition with speech enhanced deep neural networks," in *Proceedings of Interspeech*, 2014, pp. 616–620.

[13] Kun Han, Yanzhang He, Deblin Bagchi, Eric Fosler-Lussier, and DeLiang Wang, "Deep neural network based spectral feature mapping for robust speech recognition," in *Proceedings of Interspeech*, 2015, pp. 2484–2488.

[14] Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara, "Speech dereverberation using long short-term memory," in *Proceedings of Interspeech*, 2015, pp. 2435–2439.

[15] Tian Gao, Jun Du, Lirong Dai, and Chin-Hui Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Proceedings of ICASSP*, 2015, pp. 4375–4379.

[16] 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.

[17] Georg Heigold, Vincent Vanhoucke, Alan Senior, Patrick Nguyen, Marc'Aurelio Ranzato, Matthieu Devin, and Jeffrey Dean, "Multilingual acoustic models using distributed deep neural networks," in *Proceedings of ICASSP*, 2013, pp. 8619–8623.

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

[19] Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara, "Deep autoencoders augmented with phone-class feature for reverberant speech recognition," in *Proceedings of ICASSP*, 2015, pp. 4365–4369.

[20] Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong, "Learning small-size dnn with output-distribution-based criteria," in *Proceedings of Interspeech*, 2014, pp. 1910–1914.

[21] 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.

[22] 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.

[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] 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.

[25] Dong Yu, Adam Eversole, Mike Seltzer, Kaisheng Yao, Zhiheng Huang, Brian Guenter, Oleksii Kuchaiev, Yu Zhang, Frank Seide, Guoguo Chen, 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.