UNSUPERVISED ADAPTATION OF STUDENT DNNS LEARNED FROM TEACHER RNNS FOR IMPROVED ASR PERFORMANCE

Lahiru Samarakoon, Brian Mak

Hong Kong University of Science and Technology

lahiruts@cse.ust.hk, mak@cse.ust.hk

ABSTRACT

In automatic speech recognition (ASR), adaptation techniques are used to minimize the training and testing mismatch. Many successful techniques are proposed for deep neural network (DNN) acoustic model (AM) adaptation. Recently, recurrent neural networks (RNNs) have shown to outperform DNNs in ASR tasks. However, the adaptation of RNN AMs are challenging and some cases when combined with the adaptation, DNN AMs outperform adapted RNN AMs. In this paper, we combine the student-teacher training and unsupervised adaptation to improve the ASR performance. First, RNNs are used as teachers to learn student DNNs. Then, these student DNNs are adapted in unsupervised fashion. Experimental results on AMI IHM and AMI SDM tasks show that student DNNs are adaptable with significant performance improvements for both frame-wise and sequentially trained systems. We also show that the combination of adapted DNNs with teacher RNNs can further improve the performance.

Index Terms— Acoustic model adaptation, Student-teacher training, Recurrent neural networks (RNNs), Deep neural networks (DNNs)

1. INTRODUCTION

The current state-of-the-art systems in automatic speech recognition (ASR) use recurrent neural network (RNN) architectures. RNNs are capable of modeling temporal dependencies of speech signals and therefore outperform simple feedforward deep neural networks (DNNs). However, all machine learning techniques including RNNs and DNNs are susceptible to performance degradation due to the training and testing mismatch. Adaptation techniques are developed to reduce this mismatch by transforming models to match testing conditions or by transforming the runtime features to match models.

Adaptation techniques are first developed for conventional Gaussian mixture model (GMM)–hidden Markov model (HMM) systems. The commonly used techniques include maximum a posteriori (MAP) [1] and maximum likelihood linear regression (MLLR) [2, 3]. In addition, speaker adaptive training (SAT) has been applied to GMM-HMM systems [4, 5]. Then, the adaptation techniques were developed for deep neural network (DNN)-HMM hybrid systems. Adaptation of DNNs has found to be effective as these methods improve the performance significantly [6, 7, 8, 9, 10, 11, 12]. However, unsupervised adaptation of RNN acoustic models (AMs) has been difficult with smaller gains [13, 14, 15]. This can be mainly because RNNs are more complex structures than DNNs. In [15], authors conjecture that the recurrent topology of LSTM-RNNs make it more effective to capture and normalize long-range speaker characteristics than DNNs. Consequently, this implicit normalization of the speaker variability reduces the adaptation gain of RNNs compared to the gains we observe in DNN adaptation. We also investigate the adaptability of RNN AMs in our experiments.

Recently, the student-teacher training is used to transfer knowledge between models [16, 17]. Student-teacher training is also known as knowledge distillation [18]. There are two steps to student-teacher training. In the first step, teacher models are trained and then the student models are learned to mimic the teacher. It is shown that student models perform better than a model of the same architecture when the latter is trained from scratch [19]. In [20], student-teacher training is used for domain adaptation. In that work, student-teacher training is used to avoid overfitting when the model is adapted with limited amount of data. The student-teacher paradigm is also used for speech enhancement [21]. In [21], teacher model is trained with the enhanced features while the student model learns to perform speech enhancement implicitly by mimicking the teacher's output distribution. Moreover, student-teacher training is successfully used to build multilingual systems in low-resource settings [19] and that work shows student models can achieve comparable recognition accuracy to teacher networks.

In this paper, we investigate the adaptability of student DNN AMs learned from RNN teachers. First, we propose to employ the student-teacher paradigm to train a student DNN from RNN AMs as teachers. Then a well-developed DNN adaptation technique is used to adapt student DNNs. Since the adaptability of RNNs is low compared to that of DNNs, it is not clear whether the student DNNs trained to mimic RNNs are adaptable. Therefore, it is interesting to investi-

gate the adaptability of student DNNs. Moreover, this knowledge transfer from RNNs to DNNs has other benefits. DNNs are more efficient in terms of latency and computational resources than RNNs that is desirable in real-time decoding applications. In addition, it is shown that DNN acoustic models can be pruned [22, 23, 24] so as to reduce the deployment costs and improve the latency. Moreover, it is more efficient to perform sequence-discriminative training for DNNs than RNNs. Finally, this study also aim to discover some insights on whether there is potential to develop effective techniques for RNN adaptation. We have evaluated our approach in two benchmark ASR tasks: the Augmented Multi-party Interaction (AMI) [25] individual headset microphone (IHM) and the AMI single distant microphone (SDM) tasks, respectively.

The rest of the paper is organized as follows. Section 2 briefly describes the student-teacher training and details of its usage in this paper. In Section 3 we give the details of our experimental setup. The results are reported in Section 4 and we conclude our work in Section 5.

2. STUDENT-TEACHER TRAINING

The first work of student-teacher training was proposed to investigate the effectiveness of depth in deep neural networks [16]. In [17], this method was used to compress a large DNN to a smaller DNN which can be deployed in devices with limited computational and storage resources. Hinton et al. [18] coined the term "knowledge distillation" and provided further evidence to the effectiveness of the student-teacher training algorithm.

In general, frame-level cross-entropy (CE) is used as the training criterion:

$$\mathcal{F}_{CE} = -\sum_{t}\sum_{i=1}^{C} P^{ref}(i|\mathbf{x}_t)\,\log(P^{model}(i|\mathbf{x}_t)) \quad (1)$$

where C is the total number of context dependent (CD) HMM states and $P^{ref}(i|\mathbf{x}_t)$ is the probability of feature frame $\mathbf{x}_t$ belonging to class i in the reference distribution while $P^{model}(i|\mathbf{x}_t)$ is the probability of feature frame $\mathbf{x}_t$ belonging to class i according to the model being trained.

In standard training, the reference distribution is obtained from the forced alignment of the training data. In that case, $P^{ref}(i|\mathbf{x}_t)$ becomes a one-hot vector which is also known as training with hard labels. The simplified formulation is given below:

$$\mathcal{F}_{CE-Hard} = -\sum_{t} \log(P^{model}(i|\mathbf{x}_t)). \quad (2)$$

In student-teacher training, instead of using the hard labels, a student model is trained to mimic the distribution of the teacher network as given below:

$$\mathcal{F}_{CE-Soft} = -\sum_{t}\sum_{i=1}^{C} P^{teacher}(i|\mathbf{x}_t)\,\log(P^{model}(i|\mathbf{x}_t)). \quad (3)$$

In general [20, 21], student network is trained to minimize the following loss function which an interpolation between the soft and hard CE losses:

$$\mathcal{F} = (1-\alpha)\mathcal{F}_{CE-Hard} + \alpha\mathcal{F}_{CE-Soft} \quad (4)$$

where α is the interpolation weight.

In this work, instead of combining an ensemble of teachers to form one teacher distribution, we use multiple streams of teacher distributions to train the student model. This approach essentially increases the training data by making multiple copies and each copy uses the labels from the corresponding teacher distribution. We believe this may help the student model to learn from multiple views. Furthermore, we do not interpolate teacher labels with original hard targets. The training criterion used in this paper to train student models is given below:

$$\mathcal{F} = -\sum_{j}\sum_{t}\sum_{i=1}^{C} P^{j}(i|\mathbf{x}_t)\,\log(P^{model}(i|\mathbf{x}_t)) \quad (5)$$

where $P^{j}(i|\mathbf{x}_t)$ is the probability of feature frame $\mathbf{x}_t$ belonging to class i in the teacher j's distribution.

3. EXPERIMENT SETUP

In this paper, we use the AMI corpus which contains about 100 hours of meetings conducted in English. The speech is recorded by multiple microphones, including one IHM and a uniform microphone circular array. In the experiments, we use the IHM data and the speech from the first microphone in the array which is known as the SDM. We use the ASR split [26] of the corpus where 78 hours of the data are used for training while about 9 hours each are used for evaluation and development. We use 90% of the training set for training, and the rest is used as the validation set. The results are reported on the evaluation set.

For both IHM and SDM datasets, we extract the Mel-frequency cepstral coefficients (MFCCs) from the speech using a 25 ms window and a 10 ms frame shift. Then the linear discriminant analysis (LDA) features are obtained by first splicing 7 frames of 13-dimensional MFCCs and then projecting downwards to 40 dimensions using LDA. A single semi-tied covariance (STC) transformation [27] is applied on top of the LDA features. Also, we extract speaker-normalized CMLLR (also known as fMLLR) features after applying speaker

specific CMLLR transforms on top of these LDA+STC features. The GMM-HMM system for generating the alignments for DNNs and RNNs is trained on these 40 dimensional CMLLR features. We train the DNN-HMM baselines on the CMLLR features that span a context of 11 neighboring frames. Before being presented to the DNN, cepstral mean variance normalization (CMVN) is performed on the features globally. DNNs have 6 sigmoid hidden layers with 2048 units per layer, and around 4000 senones as the outputs.

We train bidirectional long short-term memory with projection (BLSTMP) with 3 layers of 1024 memory cells (512 forward and 512 backword) with a 300 dimensional projection. We also trained bidirectional residual memory networks (BRMN) as described in [28]. We use latency controlled bidirectional training as proposed in [29]. The input feature is a single frame for both BLSTMP and BRMN. These BLSTMP and BRMN models are used as teachers in student-teacher training.

We conduct experiments on models trained to optimize the cross-entropy criterion as well as the state-level minimum Bayes risk (sMBR) criterion. All the DNNs and RNNs are trained using CNTK [30]. Kaldi [31] is used to build GMM-HMM systems and for i-vector extraction. The UBM consists of 128 full Gaussians. For decodings, we use the trigram language model as used in Kaldi, which is an interpolation of trigram language models trained on AMI and Fisher English transcripts.

4. RESULTS

DNN Adaptation vs RNN Adaptation

First, we highlight the difficulty in adapting RNN acoustic models by comparing the effectiveness of unsupervised adaptation on DNN vs LSTMP-RNN acoustic models (Table 1). These results are reported on LDA+STC features for the IHM task. RNN adaptations are performed on unidirectional LSTMP-RNN AMs. The baseline results are given in the row where the method is "None". The LSTMP-RNN model outperforms the corresponding DNN baseline. As can be clearly seen, the speaker-aware training (SaT) with speaker-dependent (SD) bias after the second pass of the adaptation [7], improves the performance consistently for DNNs as well as LSTMP-RNNs. In addition the gains we observe from the CMLLR features is considerably reduced for LSTMP-RNNs in comparison when CMLLR is used with DNNs. For ease of comparison, the best performances of the adaptation are listed for both DNNs and LSTMP-RNNs. We get the best performance of LSTMP-RNN adaptation (25.7%) which is worse than the performance of the best DNN result (25.1%). More details of these comparisons and adaptation techniques are given in [13].

As can be clearly seen from Table 1, the relative gains of the LSTMP-RNN adaptation is considerably smaller to that of DNN adaptation. Therefore, we can claim that the adaptation of LSTMP-RNN is more difficult than adaptation of DNN. One of the reasons for this is that LSTMP-RNNs are more complex models than the feedforward DNNs. This increased complexity of LSTMP-RNNs makes adaptation more difficult. In addition, LSTMP-RNNs may already capture and normalize the speaker characteristics. Therefore, in rest of this paper, adaptation experiments are performed only on the DNNs for the IHM and SDM tasks.

Table 1. *Word error rates (WER %) comparison of DNN vs LSTMP-RNN adaptation results for models trained of LDA+STC features for the IHM task. Relative improvement are given in the brackets [13].*

Method	DNN	LSTMP-RNN
None	29.0 (-)	28.1 (-)
SaT	27.0 (6.9)	26.2 (6.8)
CMLLR	26.3 (9.3)	26.3 (6.4)
Best	25.1 (13.5)	25.7 (8.5)

Table 2. *WER % for various adaptation techniques applied to the DNN baseline models trained on CMLLR features.*

Model	IHM	SDM
DNN Baseline	26.3	53.2
+ LHUC	24.9	52.6
+ SaT	26.0	52.8
+ SVD-Bottleneck	25.2	52.1
+ FHL	24.3	50.6

Table 2 presents results when different adaptation techniques are applied to the DNN baselines for the IHM and SDM tasks. We compare four state-of-the-art DNN adaptation techniques: namely, learning hidden unit contributions (LHUC) [32], SaT [10, 9, 33], singular value decomposition (SVD) based bottleneck adaptation [34, 35, 36] and factorized hidden layer (FHL) [37, 38, 39]. As can be clearly seen, all adaptation techniques improves the performance significantly. FHL reports the best performance with 2.0% and 2.6% absolute improvements over the corresponding baseline for the IHM and SDM tasks, respectively. Therefore, for the rest of experiments, we select FHL as the adaptation technique.

Student-Teacher Training and Adaptation

Table 3 shows the results for the baseline models trained on the IHM and SDM tasks on CMLLR features. Both BLSTMP and BRMN models outperform DNN baselines significantly. For the IHM task, BRMN outperforms the BLSTMP whereas for the SDM task BLSTMP outperforms the BRMN model. This observation suggests that BLSTMP is more effective when used in more challenging reverberant conditions. This is expected as superior modeling of temporal dependencies

Table 3. *WER % for various baseline models trained on CM-LLR features.*

Model	IHM	SDM
DNN Baseline	26.3	53.2
BLSTMP	24.6	48.2
BRMN	24.4	49.0
Student DNN	25.5	51.5

Table 4. *WER % for FHL adaptation of Student DNN models.*

Model	IHM	SDM
DNN Baseline	26.3	53.2
+ FHL	24.3	50.6
Student DNN	25.5	51.5
+ FHL	23.6	50.0
BLSTMP	24.6	48.2
BRMN	24.4	49.0

by RNN models report more gains in reverberant conditions. The last row of the Table 3 shows the result for student DNNs trained when both BLSTMP and BRMN models are used as teachers. As can be seen, for both IHM and SDM tasks, the student DNN improves the performance significantly over the baseline. However, for both tasks, teacher models outperform the respective student DNN.

It is worth highlighting that for the IHM task, the performance of FHL adapted DNN is similar to the performances of BLSTMP and BRMN models. This observation may suggest that recurrent models are implicitly adapted and able in generalizing to different speakers which may explains the difficulties in adapting RNN acoustic models as conjectured in [15]. However, for the SDM task, recurrent models outperform the FHL adapted DNN. We believe this is because SDM task provides more room for improvement due to reverberation. This is because BLSTMP and BRMN models are efficient in modeling temporal dependencies. Later in this paper, we further analyse this via system combinations.

Table 4 presents the results of student DNN adaptation. As can be clearly seen, student DNN reports significant improvements after FHL adaptation. For the IHM task, FHL adapted student DNN significantly outperforms BLSTMP (1.0% absolute) and BRMN (0.8% absolute) models. However, for the SDM task, BLSTMP and BRMN models perform better than the FHL adapted student DNN. As mentioned before, we believe this observation is because recurrent models are more effective in handling reverbarant conditions. We can conclude that student DNNs learned from RNN teachers are adaptable with significant performance improvements.

Table 5. *WER % for various system combinations.*

Combination	IHM	SDM
FHL + BLSTMP	22.6	47.4
FHL + BRMN	22.6	47.8

Table 6. *WER % for the adaptation of sequnce trained DNN models.*

Model	IHM	SDM
DNN Baseline (CE)	26.3	53.2
+ sMBR	24.5	50.3
Student DNN (CE)	25.5	51.5
+ sMBR	23.8	48.0
+ FHL	21.7	45.9

System Combinations

Next in Table 5 we presents the results for system combination. For the purpose of this paper, we combine systems by interpolating decoding lattices. For both IHM and SDM tasks, we interpolate FHL adapted student DNN lattices with the BLSTMP or BRMN decoding lattices. As can be clearly seen, lattice interpolations report significant performance improvements. For the IHM task, the combination of FHL adapted DNN with the BLSTMP model reports a 2.0% absolute improvement over the BLSTMP model. Similarly, FHL adapted student DNN combination with the BRMN reports 1.8% absolute improvement over the BRMN. The gains of system combinations for the SDM task is smaller compared to IHM mainly because there is a considerable performance gap between the FHL adapted student DNN and teacher models for SDM. The gains of these system combinations suggests that there is potential for the adaptation of BRMN and BLSTMP models.

Results on Sequence-discriminative models

Finally, we report the results on sMBR sequence-trained models for both IHM and SDM tasks in Table 6. Even though the models are trained sequentially using sMBR, we used the cross-entropy criterion for the second pass adaptation. As can be clearly seen, sMBR training results in absolute improvements of 1.7% and 3.5% on IHM and SDM tasks respectively. For both IHM and SDM tasks, the gains of sMBR training is consistent among the DNN baseline and the student DNN. The FHL adaptation reports further absolute improvements with 1.9% and 2.1% over strong sMBR student DNNs for IHM and SDM tasks, respectively. We have not conducted experiments on sMBR trained BLSTMP and BRMN models due to resource limitations. Therefore, the approach presented in this paper to improve the performance by adapting DNNs learned from recurrent teachers can be used to avoid

the sequential discriminative training of the RNNs.

5. CONCLUSIONS

In this paper, we investigated the adaptability of student deep neural networks (DNNs) trained where recurrent neural networks (RNNs) are used as teachers. Since the adaptability of RNN acoustic models (AMs) are lower compared to that of the DNNs, this approach first allows to learn a better performing DNN using student-teacher training and then later improve the performance further by unsupervised adaptation. We used bidirectional long short-term memory with projection (BLSTMP) and bidirectional residual memory networks (BRMN) as teacher networks. Factorized hidden layer (FHL) is used as the adaptation method. Experimental results on AMI IHM and AMI SDM tasks show that student DNNs are adaptable with significant performance improvements for both frame-wise and sequentially trained systems. We also show that the combination of adapted DNNs with teacher RNNs can further improve the performance. These improvements due to system combinations also suggest that there is potential to develop adaptation techniques for RNN AMs with significant performance gains.

6. REFERENCES

[1] J. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

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

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

[4] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *ICSLP*. ISCA, 1996, vol. 2, pp. 1137–1140.

[5] M.J.F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 417–428, 2000.

[6] L. Samarakoon and K.C. Sim, "Learning factorized transforms for speaker normalization," in *ASRU*. IEEE, 2015.

[7] L. Samarakoon and K.C. Sim, "On combining i-vectors and discriminative adaptation methods for unsupervised speaker normalization in DNN acoustic models," in *ICASSP*. IEEE, 2016.

[8] 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 *ICASSP*. IEEE, 2013, pp. 7893–7897.

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

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

[11] 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 *ICASSP*. IEEE, 2013, pp. 7942–7946.

[12] Dong Yu, , and Li Deng, *Automatic Speech Recognition - A Deep Learning Approach*, Springer London, New York, 2015.

[13] L. Samarakoon, B. Mak, and K.C. Sim, "Learning factorized transforms for unsupervised adaptation of LSTM-RNN acoustic models," in *Interpeech*. ISCA, 2017.

[14] Chaojun Liu, Yongqiang Wang, Kshitiz Kumar, and Yifan Gong, "Investigations on speaker adaptation of lstm rnn models for speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5020–5024.

[15] Y. Zhao, J. Li, K. Kumar, and Y. Gong, "Extended low-rank plus diagonal adaptation for deep and recurrent neural networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[16] Jimmy Ba and Rich Caruana, "Do deep nets really need to be deep?," in *Advances in neural information processing systems*, 2014, pp. 2654–2662.

[17] Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong, "Learning small-size dnn with output-distribution-based criteria," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[19] J. Cui, B. Kingsbury, B. Ramabhadran, G. Saon, T. Sercu, K. Audhkhasi, A. Sethy, M. Nussbaum-Thom, and A. Rosenberg, "Knowledge distillation across ensembles of multilingual models for low-resource languages," in *2017 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP), 2017.

[20] T. Asami, R. Masumura, Y. Yamaguchi, H. Masataki, and Y. Aono, "Domain adaptation of dnn acoustic models using knowledge distillation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[21] S. Watanabe, T. Hori, J. Le Roux, and J. R. Hershey, "Student-teacher network learning with enhanced features," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[22] Tara N Sainath, Brian Kingsbury, Vikas Sindhwani, Ebru Arisoy, and Bhuvana Ramabhadran, "Low-rank matrix factorization for deep neural network training with high-dimensional output targets," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013.

[23] J. Xue, J. Li, and Y. Gong, "Restructuring of deep neural network acoustic models with singular value decomposition.," in *INTERSPEECH*. ISCA, 2013, pp. 2365–2369.

[24] G. Mantena and K. C. Sim, "Entropy-based pruning of hidden units to reduce dnn parameters," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, 2016.

[25] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, et al., "The AMI meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, 2005, vol. 88.

[26] P. Swietojanski, A. Ghoshal, and S. Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in *ASRU*. IEEE, 2013, pp. 285–290.

[27] M.J.F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[28] M. Baskar, M. Karafit, L. Burget, K. Vesel, F. Grzl, and J. H. ernock, "Residual memory networks: Feed-forward approach to learn long-term temporal dependencies," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[29] Yu Zhang, Guoguo Chen, Dong Yu, Kaisheng Yaco, Sanjeev Khudanpur, and James Glass, "Highway long short-term memory RNNs for distant speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016.

[30] D. Yu, A. Eversole, M. Seltzer, K. Yao, Z. Huang, B. Guenter, O. Kuchaiev, Y. Zhang, F. Seide, H. 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.

[31] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The kaldi speech recognition toolkit," in *ASRU*. IEEE, 2011.

[32] Pawel Swietojanski, Jinyu Li, and Steve Renals, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1450–1463, 2016.

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

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

[35] K. Kumar, C. Liu, K. Yao, and Y. Gong, "Intermediate-layer DNN adaptation for offline and session-based iterative speaker adaptation," in *INTERSPEECH*. ISCA, 2015.

[36] Y. Zhao, J. Li, and Y. Gong, "Low-rank plus diagonal adaptation for deep neural networks," in *ICASSP*. IEEE, 2016, pp. 5005–5009.

[37] L. Samarakoon and K.C. Sim, "Factorized hidden layer adaptation for deep neural network based acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[38] L. Samarakoon and K.C. Sim, "Multi-attribute factorized hidden layer adaptation for DNN acoustic models," in *Interpeech*. ISCA, 2016.

[39] L. Samarakoon and K.C. Sim, "Low-rank bases for factorized hidden layer adaptation of DNN acoustic models," in *SLT*. IEEE, 2016.