ACOUSTIC MODEL ADAPTATION FOR PRESENTATION TRANSCRIPTION AND INTELLIGENT MEETING ASSISTANT SYSTEMS

Yan Huang and Yifan Gong

Microsoft Corporation

ABSTRACT

We present our solution for unsupervised rapid speaker adaptation in a state-of-art presentation and intelligent meeting transcription system. We adopt the Kullback-Leibler (KL) divergence regularized model adaptation paradigm. For the adaptation architecture, we found that the linear projection layer adaptation yields competitive performance with the additional benefit in its simplicity and robustness to small amount of adaptation data. To address the imperfect supervision, we use a supervision committee formed by multiple systems or single-system n-best to mask possibly mislabeled frames. To relieve the data sparsity issue, we apply noise and speaking rate perturbation data augmentation techniques to create a richer adaptation data set. In summary, the proposed solution consists of the KL-divergence regularized linear projection layer adaptation with frame masking and data augmentation. On a presentation transcription and a meeting transcription task, our proposed methodology yields 7.3 % and 7.9 % relative word error rate (WER) reduction against a strong baseline model trained from tens of thousand hour speech. To the best of our knowledge, this is a first reported work on rapid speaker adaptation on a state-of-art production system.

Index Terms— Acoustic model adaptation, speaker adaption, unsupervised adaptation, meeting transcription

1. INTRODUCTION

The increasingly more sophisticated neural network acoustic model trained from tens of thousand hour speech is believed to be relatively robust to speaker variability. The main challenges of the rapid unsupervised speaker adaptation are the imperfect supervision generated from the first-pass decoding and the limited amount of adaptation data. In this paper, we would like to answer the question whether the rapid unsupervised speaker adaptation is still beneficial for a state-of-art presentation and intelligent meeting transcription system.

There was abundant previous work on the neural network acoustic model adaptation. In a tandem system [1], the Maximum a-posterior (MAP) [2], and the Maximum likelihood linear regression (MLLR) [3, 4] adaptation techniques can be applied to the Gaussian mixture model (GMM) [5, 6]. In a hybrid or an end-to-end system, the neural network acoustic model are directly adapted. Given the large number of model parameters and limited amount of adaptation data, methodologies in this category mainly focus on different strategies to improve its robustness and avoid over-fitting [7–17]. For example, the transformation [18, 19], the learning hidden unit contribution (LHUC) [20, 21], the singular value decomposition (SVD) [22], and the factorized subspace adaptation [23], constrain the model adaptation in a reduced or highly compressed parameter space. Alternatively, the Kullback-Leibler (KL) divergence regularized adaptation [9, 24] and the Bayesian neural network adaptation [17, 25] make use of specially formulated objectives to prevent catastrophic forgetting and over-fitting. Lastly, the i-vector [26, 27] and the speaking code based adaptation [28] utilize the speaker-level representation as the auxiliary input for the conditioning model.

In this paper, we present our practical solution for the unsupervised rapid speaker adaptation for a bi-directional long short-term memory acoustic model (LSTM) in a presentation and intelligent meeting transcription system. We adopt the KL-divergence regularized model adaptation paradigm [9]. For the adaptation architecture, we compared adapting different component of the original network or additional speaker-specific network components. We found that the linear projection layer adaptation yields competitive performance comparing to the LHUC [20] or the factorized sub-space adaptation [23]. This is consistent with a previous study on a research system [29]. To address the imperfect supervision [30], we adopt a supervision committee formed by multiple systems or the single-system n-best list to mask the incorrectly labeled frames. To relieve the data sparsity, we applied different data augmentation techniques [31] to create a richer and relevant data set.

We compare different adaptation methodologies for both supervised and unsupervised adaptation. Our proposed solution, which consists of the KL-divergence regularized linear projection layer adaptation with frame masking and data augmentation, yields 7.3 % and 7.9 % relative WER reduction for the presentation transcription and the intelligent meeting transcription task respectively. The adaptation data amount ranges from 2 to 20 minutes per speaker.

This paper demonstrates the success of rapid unsupervised speaker adaptation in a state-of-art system. To the best of our knowledge, this is a first reported work on rapid speaker adaptation on an up-to-date production scale system; previous study was primarily conducted on research benchmarks. By carefully selecting and combining several existing technologies with extensions, we establish a practical solution which advances the rapid unsupervised speaker adaptation in a practical speech service system.

The rest of this paper is organized as follows: Section 2 introduces our proposed methodology; Section 3 presents the experiments and results; Section 4 concludes this paper.

2. METHODOLOGIES

In this section, we present our proposed methodology for rapid unsupervised adaptation.

2.1. Adaptation Architecture

The unstructured neural network model information distribution makes it difficult to identify specific network component for speaker variability. Therefore, most speaker adaptation solutions empirically associate certain existing sub neural network or introduce additional network component as the speaker signature for model adaptation.

In the sub-network adaptation, certain component of the original network is chosen as the speaker signature to be adapted. It does not introduce additional model parameters or modify the original network structure. For training, we identify the selected nodes to receive gradient update and keep others unchanged during back-propagation. When using the adapted model in a practical system, switching to the speaker adapted model only involves swapping some neural network layers to the speaker adapted layers. This approach does not introduce additional run-time latency and is simple to implement from the engineering perspective.

For adaptation with additional network component, we studied the activation function based adaptation [20, 21] and the factorized subspace-based adaptation [23] for comparison. The activation function based adaptation learns a speaker-specific hidden-unit contribution. As illustrated in Eq.(1), h_m^l is the re-weighted hidden unit activation for the m-th speaker at the l-th layer; W^l is the speaker independent parameters of the l-th layer; $\phi^l(.)$ is the non-linearity function of the l-th layer; r_m^l is the speaker-dependent parameters used to re-weight the hidden-unit contribution of the l-th layer for m-th speaker and $\alpha(.)$ is used to constrain the range of r_m^l:

$$h_m^l = \alpha(r_m^l) \cdot \phi^l((W^l)^T h_m^{l-1}). \quad (1)$$

The factorized subspace adaptation [23] formulates the speaker-specific parameter residue using the factorized low-rank representation. As illustrated in Eq.(2), W_m^l is the adapted model, W_0^l is the baseline model, $\Gamma_m^l D_m^l \Psi_m^l$ is the low-rank representation of the speaker-specific residue matrix. The rank of Γ_m^l and Ψ_m^l defines the dimension of the adaptation parameter space:

$$W_m^l = W_0^l + \Gamma_m^l D_m^l {\Psi_m^l}^T. \quad (2)$$

Both of these two approaches represent the speaker-specific network component in a low-dimensional space and thus significantly reduce the number of additional model parameters. Empirically, the adaptation performance is often determined by the representation capacity and the appropriate trade-off with the data amount.

2.2. Supervision Committee

One key issue in the unsupervised model adaptation is the imperfect supervision. Incorrect gradient generated from the incorrect supervision can lead to catastrophic parameter update during adaptation.

We propose to use a hypothesis committee to mask likely mislabeled frames. We first use the baseline and the alternative systems (or single system n-best) to generate multiple word-level hypotheses; then obtain the senone-level hypothesis via aligning the word-level hypotheses with the baseline model. The degree of agreement is used to determine whether a specific frame should be used or partially used. Formally, the committee-based KL-regularized objective is defined as

$$CE_{KL,Committee} = \sum_i (\hat{p}_i \log p_i) f(l_i^{(0)}, l_i^{(1)}, \cdots, l_i^{(M)}), \quad (3)$$

where p_i is the posterior of the adapted model, $\hat{p}_i$ is the KL-regularized target, M is the total number of alternative systems, i is the frame index. The KL-regularized target ($\hat{p}_i$) is the linear combination of the posterior from the baseline model ($\bar{p}_i$) and the 0-1 hard label $\tilde{p}_i$, α is the combination weight:

$$\hat{p}_i = (1-\alpha)\tilde{p}_i + \alpha \bar{p}_i. \quad (4)$$

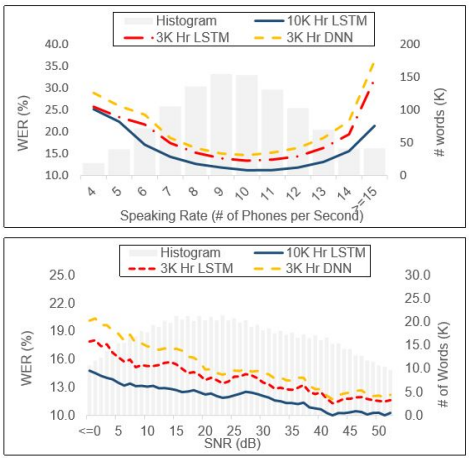

Fig. 1. System performance with respect to different speaking rate (top) and SNR (bottom) for 10K hr LSTM, 3K hr LSTM, and 3K hr DNN. The top figure depicts the WER performance of the three systems with respect to different speaking rate. The bottom figure depicts the performance with respect to SNR.

$f(.)$ is a function which measures the degree of agreement between the primary hypothesis ($l_i^{(0)}$) generated by the baseline system and the alternative hypothesis ($l_i^{(j)}$) from other systems in the committee:

$$f(l_i^{(0)}, l_i^{(1)}, \cdots, l_i^{(M)}) = \left(\frac{\sum_{j=1}^{M} \delta(l_i^{(0)} = l_i^{(j)})}{M} \right)^{\beta}, \quad (5)$$

where δ is a delta function, β (≥ 1) is a warping parameter used to further dampen the weight for frames with only partial agreement. Larger β penalizes frames with partial agreement more severely. We simply set $\beta = 1$ in all experiments throughout this paper. Eq. (5) provides a soft weighting for frames with partial hypothesis committee agreement. When only one alternative system is used, Eq. (5) becomes a simple 0-1 valued function. A frame is discarded if the alternative system disagrees with the primary system.

2.3. Data augmentation

Data sparsity is another barrier in rapid speaker adaptation for a large-scale model with massive number of parameters. Inspired by the study on the model robustness to varied speaking rate and SNR, we proposed to apply the duration and noise perturbation techniques to generate multiple samples of faster, slower, or noisier speech to address the data sparsity. These techniques are not new, which have been used in previous work [31]. As an initial study, we trained a pair of DNN and LSTM using 3K hour speech and another LSTM with 10K hour speech. All models share the same 80-dimension log-filter bank (LFB) front-end and 9404 tied senone states. We measure the speaking rate and SNR dependent WER using 100 hour test data similarly to [32]. As depicted in Figure 1, in a model with more advanced structure and enlarged training data, the robustness pattern respect to different speaking rate and SNR remains similar.

3. EXPERIMENTS AND RESULTS

In this section, we present experimental results. All experiments were conducted on anonymized data with personally identifiable information removed.

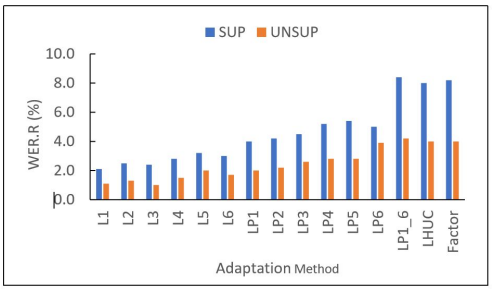

Fig. 2. Performance comparison of different model adaptation structure in supervised adaptation.

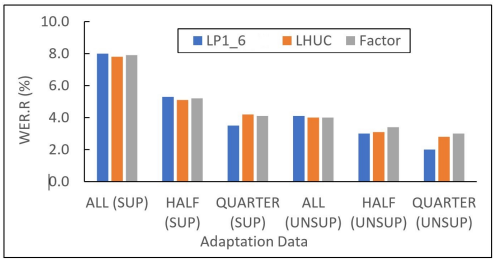

Fig. 3. Adaptation performance with reduced data amount. HALF and QUARTER refer to using 50 % or 25 % of ALL for adaptation.

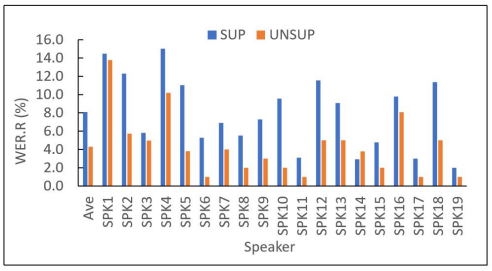

Fig. 4. Comparison of supervised and unsupervised adaptation measured by relative WER reduction (WER.R). The only difference between the sup/unsup adaptation compared here is whether the human transcription or the first-pass decoding is used.

3.1. Experimental Setup

The presentation transcription consists of 19 speakers, each with 2 to 20 minutes of speech. For each speaker, the first half of speech is used for training and the rest for testing. We conducted supervised and unsupervised adaptation for comparison on this task.

The intelligent meeting transcription task consists of 14 meetings and 100 meeting speakers. Each meeting consists of 3 to 15 speakers and each speaker has 3 to 20 minutes of speech. The multi-channel far-field speech first passes through our far-field multi-channel audio processing to generate the single-channel enhanced speech. The speech is segmented into speaker homogeneous regions using the audio-visual signal and the speaker profile [33]. We only conduct the unsupervised adaptation on this task.

The baseline is a bidirectional LSTM model trained from tens of thousand hour speech. It has 6 bidirectional LSTM layers followed by a fully connected top layer. Each layer has 2048 hidden units. The input consists of a 80-dim log-filter bank feature (LFB). The output layer has 9404 senone states.

3.2. Adaptation Structure

We first compare adapting different component of the original network illustrated in Figure 2. L_i refers to adapting the i-th recurrent layer; LP_i refers to adapting the projection layer of the i-th recurrent layer, e.g. LP_1 refers to adapting the projection layer of the bottom recurrent layer, LP_{1_6} refers to adapting the projection layer of all six recurrent layers. LHUC refers to the learning hidden unit contribution based adaptation. Factor refers to the factorization-based adaptation. We use the relative WER reduction (WER.R) to measure the adaptation performance throughout this paper.

We found that the linear projection layer adaptation (LP_{1_6}) yields robust adaptation performance for both supervised and unsupervised adaptation. In particular, adapting the recurrent layer is not as effective as adapting the projection linear layer. It is likely due to the fact that the recurrent layer is harder to optimize, especially with limited amount of data. On the other hand, the linear projection layer adaptation provides a good trade-off between the representation capacity and ease of training.

We further implemented the LHUC and the factorized sub-space adaptation. We found that the LHUC and the factorized sub-space adaptation yield similar adaptation performance. Comparing to the linear projection layer adaptation, they are not notably better. We hypothesize that the LHUC and the factorized sub-space adaptation may perform better when the adaptation data is even smaller. We therefore reduce the data to half or quarter of the current setup and repeat the same experiments. As shown in Fig 3, with the reduced adaptation data, the LHUC and the factorized sub-space adaptation indeed performs slightly better than the linear projection layer adaptation. Due to the adaptation training run-time cost, we choose to only adapt speakers with reasonable amount of data and discard others with too little data (e.g. $\leq$ 3 min). In that operating point, the projection layer adaptation performs the best.

3.3. Hypothesis Committee

To illustrate the impact of supervision quality, we conduct a pair of supervised and unsupervised model adaptation, which only differs in whether human transcription or the first-pass decoding result is used as supervision. As depicted in Figure 4, imperfect supervision in unsupervised adaptation results in significant performance gap between sup/unsup adaptation.

Table 1 presents the results of adaptation with different hypothesis committees. The hypothesis committee consists of complimentary acoustic models with varied model structures or different front-end. For example, S0 is the baseline system, S1 is a uni-direction LSTM with 40-dim LFB feature, S2 is a structurally different bi-direction LSTM with fewer number of layers and memory cells. 2 -system (S0/trans) refers to joining the human transcription with the transcription generated from the baseline first pass decoding; 2-system (S0/S1) and 3-system (S0/S1/S2) refer to joining the transcription generated from the alternative systems (S1 or S1/S2) with the transcription from the baseline system (S0).

As discussed in Section 2.2, the unsupervised adaptation reduces the gain of the supervised adaptation by more than half, i.e. 4.1 % relative WER reduction for the unsupervised adaptation and 8.0 % for the supervised counterpart. 2-system (S0/trans) joins the unsupervised transcription with the human transcription and thus perfectly masks the incorrectly labeled frames. It yields 7.6 % relative WER reduction, which is the oracle performance upper bound given the unsupervised transcription quality.

Instead of using the human transcription, we introduce a second system (S1) to generate alternative transcription and form the supervision committee. 2-system (S0/S1) yields 6.3 % relative WER reduction. Further adding a third system (S2), 3-system (S0/S1/S2) yields 6.4 % relative WER reduction, only slightly better. It is to be noted that the gain here is due to the effective masking of the incorrectly labeled frames. It is different from rover as our supervision committee formulation in Eq.(3) does not use transcription from the alternative systems for adaption. As running an alternative system introduces additional cost, we apply the same idea to the single-system n-best hypothesis. We use the n-best of the baseline system to form the hypothesis committee for frame masking. As in Table 1, 1-best, 3-best, and 5-best yield 6.1 %, 6.5 %, and 6.1 % relative WER reduction respectively.

Table 1. Adaptation performance of applying the supervision committee for the unsupervised adaptation.

Model	WER.R
SUP (KL=0.5)	8.1
UNSUP (1-system, 1-best) (KL=0.8)	4.0
UNSUP (2-System, S0/trans)(KL=0.8)[1]	7.6
UNSUP (2-System, S0/S1) (KL=0.8)	6.3
UNSUP (3-System, S0/S1/S2) (KL=0.8)	6.4
UNSUP (1-System, 2-best) (KL=0.8)	6.1
UNSUP (1-System, 3-best) (KL=0.8)	6.5
UNSUP (1-System, 5-best) (KL=0.8)	6.1
UNSUP (1-system, 1-best, conf) (KL=0.8)	4.6
UNSUP (3-System, S0/S1/S2, rover)	3.5

We further experimented with the single-system confidence based data selection similar to [34], but only achieve moderate additional gain. This is likely due to the sub-optimal confidence classifier. We also conducted rover without adaptation. The standard rover using S0/S1/S3 results in 3.5 % relative WER reduction.

3.4. Data Augmentation

For the speaking rate perturbation, two sets of speaking rate augmentation range (e.g. 0.9-1.1 and 0.8-1.2) were used with a post-processing to filter out the utterances with speaking rate out of normal speaking rate range. For the noise perturbation, we generate one or two copies of the data with additive noise and combine with the original data for adaptation. As the noise perturbation is primarily for regularization, we only add mild noise without significantly reducing the SNR. Table 2 presents supervised and unsupervised adaptation result. We observed consistent gains using the duration augmentation for both the supervised and the unsupervised adaptation. The noise perturbation yields smaller but still consistent performance gain. Adding more noisy simulations might make the original data underrepresented in the data and therefore does not help further.

Table 2. Accuracy performance of data augmentation for supervised and unsupervised adaptation.

Model	WER.R	Model	WER.R
SUP (KL=0.5)	8.0	UNSUP (KL=0.8)	4.1
SUP (Dur0.9-1.1)	9.0	UNSUP (Dur0.9-1.1)	5.1
SUP (Dur0.8-1.2)	8.6	UNSUP (Dur0.8-1.2)	5.0
SUP (Noise1copy)	8.5	UNSUP (Noise1copy)	4.6
SUP (Noise2copy)	7.8	UNSUP (Noise2copy)	4.4

We combine the supervision committee with data augmentation. As reported in Table 3, 2-system combination with noise and speaking-rate perturbation yields 7.3 % relative WER reduction. To avoid running two systems, using 3-best from 1 system only combined with data augmentation can also yield 7.1 % relative WER reduction. We didn't further experiment with combining 3-system with data augmentation as it is lack of interest to us due to the additional computation cost in a practical solution.

[1] As the human transcription is used in the hypothesis committee, it provides a performance upper bound for the committee-based approach.

Table 3. Performance of unsupervised model adaptation with supervision committee and data augmentation.

Model	WER.R
SUP (KL=0.5)	8.0
UNSUP (KL=0.8)	4.1
2 System (S0/S1)	6.3
2 System (S0/S1) + Dur0.9-1.1 + Noise1copy	7.3
1 System (S0/3-best)	6.5
1 System (S0/3-best) + Dur0.9-1.1 + Noise1copy	7.1

3.5. Speaker Adaptation for Meetings

For intelligent meeting transcription, the speaker adaption is performed on speaker segmented speech after speaker diarization. Table 4 presents the overall unsupervised speaker adaptation result. On 14 meetings with around 100 meeting speakers, the adaptation yields 7.9 % average relative WER reduction for speakers with at least 10 minutes speech. For speakers with too little data, we choose not to adapt for cost-effective concern. It is to be noted that the diarization error can also affect the adaptation performance. The reported result is based on an end-to-end run of the meeting transcription system [33] with real diarization. We also evaluate the adaption based on the ground truth diarization and obtain slightly better results.

Table 4. Performance of unsupervised speaker adaptation on the intelligent meeting transcription system.

Meeting	WER.R	Meeting	WER.R
Meeting001	4.7	Meeting008	10.9
Meeting002	6.0	Meeting009	2.6
Meeting003	15.0	Meeting010	3.0
Meeting004	6.1	Meeting011	7.7
Meeting005	8.1	Meeting012	10.0
Meeting006	6.9	Meeting013	2.0
Meeting007	4.7	Meeting014	4.6
Average = 7.9			

4. CONCLUSION

In summary, we presented our acoustic model adaptation solution for a practical presentation and meeting transcription system. We found the simple linear projection layer adaptation with supervision committee and data augmentation can yield competitive adaptation performance. Our results suggest that the state-of-art neural network acoustic model can still benefit from rapid unsupervised speaker adaptation. It is practical to deploy this technology for an offline speech transcription system. For online streaming adaptation, we think it is a very interesting but still extremely challenging topic.

5. REFERENCES

[1] N. Morgan and H.A. Bourlard, "Neural networks for statistical recognition of continuous speech," *IEEE*, pp. 742 – 772, 1995.

[2] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE/ACM Transaction on Audio Speech Language Processing*, pp. 291 – 298, 1994.

[3] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, pp. 171 – 185, 1995.

[4] M. J. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, pp. 75 – 98, 1998.

[5] Z. Tuske, P. Golik, R. Schluter, and H. Ney, "Speaker adaptive joint training of gaussian mixture models and bottleneck features," in *Proceedings of ASRU*, 2015.

[6] Y. Q. Wang, C. Zhang, M.J.F. Gales, and P.C. Woodland, "Speaker adaptation and adaptive training for jointly optimized tandem systems," in *Proceedings of Interspeech*, 2018.

[7] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proceedings of ICASSP*, 2013.

[8] K. Yao, D. Yu, F. Seide, H. Su, Li. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Proceedings of SLT*, 2012.

[9] D. Yu, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proceedings of ICASSP*, 2013.

[10] L. Samarakoon and K. C. Kim, "Factorized hidden layer adaptation for deep neural network based acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 2241–2250, 2016.

[11] Y. Miao, H. Zhang, and F. Metze, "Towards speaker adaptive training of deep neural network acoustic models," in *Proceedings of Interspeech*, 2014.

[12] C. Liu, Y. Wang, K. Kumar, and Y. Gong, "Investigations on speaker adaptation of lstm rnn models for speech recognition," in *Proceedings of ICASSP*, 2016.

[13] P. Swietojanski and S. Renals, "Differentiable pooling for unsupervised speaker adaptation," in *Proceedings of ICASSP*, 2015.

[14] M. Delcroix, K. Kinoshita, T. Hori, and T. Nakatani, "Context adaptive deep neural networks for fast acoustic model adaptation," in *Proceedings of ICASSP*, 2016.

[15] T. Tan, Y. Qian, and K. Yu, "Cluster adaptive training for deep neural network based acoustic model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, p. 459–468, 2016.

[16] M. Delcroix, S. Watanabe, A. Ogawa, S. Karita, and T. Nakatani, "Auxiliary feature based adaptation of end-to-end asr systems," in *Proceedings of Interspeech*, 2018.

[17] X. Xie, X. Liu, T. Lee, S. Hu, and L. Wang, "BLHUC: Bayesian learning of hidden unit contributions for deep neural network speaker adaptation," in *Proceedings of ICASSP*, 2019.

[18] 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," in *Proceedings of Europspeech*, 1995.

[19] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. De Mori, "Linear hidden transformations for adaptation of hybrid ann/hmm models," *Speech Communication*, pp. 827–835, 2007.

[20] P. Swietojanski, J. Li, and S. Renal, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE/ACM Transaction on Audio Speech Language Processing*, pp. 1450 – 1463, 2016.

[21] S. M. Siniscalchi, J. Li, and C. H. Lee, "Hermitian based hidden activation functions for adaptation of hybrid HMM/ANN models," in *Proceedings of Interspeech*, 2012.

[22] 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 *Proceedings of ICASSP*, 2014.

[23] L. Samarakoon and K. C. Sim, "Factorized hidden layer adaptation for deep neural network based acoustic modeling," in *Proceedings of ICASSP*, 2014.

[24] Y. Huang and Y. Gong, "Regularized sequence-level deep neural network model adaptation," in *Proceedings of Interspeech*, 2015.

[25] 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 *Proceedings of Interspeech*, 2015.

[26] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proceedings of ASRU*, 2013.

[27] A. Senior and I. Lopez-Moreno, "Improving dnn speaker independence with i-vector inputs," in *Proceedings of ICASSP*, 2014.

[28] 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 *Proceedings of ICASSP*, 2013.

[29] M. Kitza, R. Schluter, and H. Ney, "Comparison of blstm-layer-specic afne transformations for speaker adaptation," in *Proceedings of ICASSP*, 2018.

[30] V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Semi-supervised training of acoustic model using a lattice-free MMI," in *Proceedings of ICASSP*, 2018.

[31] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transaction on Audio Speech Language Processing*, pp. 1469 – 1477, 2015.

[32] Y. Huang, D. Yu, C. Liu, and Y. Gong, "A comparative analytic study on the gaussian mixture and context dependent deep neural network hidden markov models," in *Proceedings of Interspeech*, 2014.

[33] T. Yoshioka and et.al., "Online audio-visual meeting transcription," in *Proceedings of ASRU*, 2019.

[34] G. Gollan and M. Bacchiani, "Confidence scores for acoustic model adaptation," in *Proceedings of ICASSP*, 2008.