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

Domain Adaptation of CNN based Acoustic Models under Limited Resource Settings

Masayuki Suzuki[1], *Ryuki Tachibana*[1], *Samuel Thomas*[2], *Bhuvana Ramabhadran*[2], *George Saon*[2]

[1]IBM Watson Multimodal, Tokyo, 103-8510, Japan
[2]IBM Watson Multimodal, Yorktown Heights, NY 10598
{szuk, ryuki}@jp.ibm.com, {sthomas, bhuvana, gsaon}@us.ibm.com

Abstract

Adaptation of Automatic Speech Recognition (ASR) systems to a new domain (channel, speaker, topic, etc.) remains a significant challenge, as often, only a limited amount of target domain data for adaptation of Acoustic Models (AMs) is available. However, unlike GMMs, to date, there has not been an established, efficient method for adapting current state-of-the-art Convolutional Neural Network (CNN)-based AMs. In this paper, we explore various training algorithms for domain adaptation of CNN based speech recognition systems with limited acoustic training data resources. Our investigations illustrate the following three main contributions. First, introducing a weight decay based regularizer along with the standard cross entropy criteria can significantly improve recognition performances with as little as one hour of adaptation data. Second, the observed gains can be improved further with the state-level Minimum Bayes Risk (sMBR) based sequence training technique. In addition to supervised training with limited amounts of data, we also study the effect of introducing unsupervised data at both the initial cross-entropy and subsequent sequence training stages. Our experiments show that unsupervised data helps with cross-entropy and sequence training criteria. Third, the effect of speaker diversity in the adaptation data is also investigated where our experiments show that although there can be large variance in final performance depending on the speakers selected, regularization is required to obtain significant gains. Overall, we demonstrate that with adaptation of neural network based acoustic models, we can obtain performance improvements of up to 24.8% relative.

Index Terms: Domain adaptation, Weight decay, sequence training, CNN

1. Introduction

Adaptation of acoustic or language models to the speaker, channel or topic (we refer to this collectively as the domain of interest) is a well known problem in speech recognition. Acoustic Model (AM) adaptation has a long history of research and there are well-established methods for adapting GMM-based AMs such as, Maximum Likelihood Linear Regression (MLLR) [1] or Maximum A Posteriori (MAP) adaptation [2] which can be applied both to the model and the feature space. With the advent of neural network based acoustic models [3, 4, 5, 6, 7], GMM-based AMs are no longer yield state-of-the-art performance in ASR systems. However, adapting the various configurations of neural network based acoustic models such as feed-forward Deep Neural Nets (DNNs), Convolutional Neural Networks (CNNs), or Long Short Term Memory (LSTM) networks, involves adapting a large number of parameters with a small amount of adaptation data, and remains a subject of research. Few approaches have been proposed in the literature. A simple approach, referred to as fine-tuning [8, 9, 10] that works quite well, includes an additional epoch of training some or all layers of the network with the adaptation data alone. This approach has been so powerful, that it has been used successfully to even adapt a multilingual neural network to the target language of interest [11, 12, 13]. A second approach involves the addition of another layer to perform a feature-space-like adaptation [14, 15, 16, 17, 18]. A third approach introduces a regularizer to control the extent by which the weights move from the baseline model [8, 9] when fine-tuned with the little adaptation data. Alternatively, the features (i-vectors or speaker-adapted features such as feature-space MLLR (fMLLR) have been used to serve as the medium of adaptation [19, 20] when the baseline acoustic model is speaker-dependent. An important factor in all of these training schemes is the underlying criterion used to train these networks, i.e,. Cross Entropy (CE) or the state-level Minimum Bayes Risk (sMBR) criterion [21]

In this paper, we investigate the impact of the regularizer when adapting different acoustic models, the impact of the training objective function, the impact of automatically produced labels on adaptation, and the role of speaker diversity in the adaptation data. We also illustrate a method to obtain gains when training with the sMBR criterion.

The main contributions of this paper are:

- First, we demonstrate the value of a regularizer (similar to the one presented in [8, 9] when adapting with very little, completely mismatched target domain data (See Section 3).
- Second, we show that using the Cross-Entropy (CE) adapted model as the basis of sMBR-based sequence training is better than using sequence training as the starting point for the CE adaptation step (See Section 5).
- Third, we demonstrate that while speaker diversity has an impact on the adaptation performance, there is a net improvement with the proposed approach regardless of the speaker distributions (See Section 4)
- Finally, we show, that proposed approach works well in an unsupervised training scenario (See Section 6).

The rest of this paper is organized as follows. The next section 2 provides a description of the data and the baseline systems used throughout this paper for adaptation. Section 3 presents the results from adaptation using the regularized approach using varied amounts of adaptation data using three different acoustic models. Section 4 studies the impact of speaker diversity in the data selected for adaptation. Section 5 presents

http://dx.doi.org/10.21437/Interspeech.2016-1161

the application of sequence training to the adaptation task. Comparison between supervised and unsupervised adaptation using the regularizer is presented in Section 6. The paper concludes with a summary of this empirical study in Section 7.

2. Data and System Descriptions

2.1. Baseline systems

In this section, we describe the baseline models and the experimental conditions as the basis for comparing various adaptation methods from the next section. We use three baseline models. One is a speaker-independent (SI) CNN AM which is trained with 1,975 hours of telephony speech data by using the sMBR criterion [22]. The CNNs operate on blocks of 11 consecutive 40-dimensional logmel frames augmented with first and second derivatives with 9×9 convolution windows. The convolution and pooling layer configuration is taken from [23]. The size of parameters for the network is $[243 \times 128, 1536 \times 256, 2048 \times 2048, 2048 \times 2048, 2048 \times 2048, 2048 \times 2048, 2048 \times 2048, 2048 \times 512, 512 \times 9300]$. A second baseline model is a speaker-adapted (SA) i-vector DNN AM which is also trained with the same 1,975 hours by using the sMBR criterion [19]. Each frame is represented by a feature vector of 13 perceptual linear prediction (PLP) cepstral coefficients which are mean and variance normalized per conversation side. Every 9 consecutive cepstral frames are spliced together and projected down to 40 dimensions using LDA. The LDA features are transformed with one feature-space MLLR (fMLLR) transform per conversation side at test time. 100-dimensional i-vectors extracted per conversation are concatenated and fed into DNN. A third baseline system is the speaker-independent (SI), i-vector based system which uses the afore-mentioned LDA features. All baseline systems are first trained with the cross-entropy criterion followed by sequence training. The size of parameters for the network is $[(11 \times 40 + 100) \times 2048, 2048 \times 2048, 2048 \times 2048, 2048 \times 512, 512 \times 32000]$. The vocabulary comprises of 250K words and the language model is a 4-gram LM with 200M n-grams.

2.2. Adaptation and Evaluation Data

In this paper, the adaptation data is drawn from telephone conversations from an internal data base of customer support calls recorded at various internal call centers. The data includes a diverse set of accents and channel characteristics that vary dramatically depending on the location of the call center. The adaptation data consists of three subsets of call center recordings; monaural recordings between agents and customers from Call Center A (CC-A), stereo recordings of the agent and customer from location B (CC-B) and their reference transcripts. These subsets contain 21 hours, and two hours of audio each from agents and customers, spanning a wide variety of accents. The baseline models do not contain any data from these call-centers and as such have not seen the channel, noise or speaker styles present in this data, thus serving as a good test set for exploring adaptation algorithms. The experiments reported in this paper, include a varying amount of data from the three adaptation data subsets, ranging from as little as 20 minutes to 25 hours, while maintaining distribution of the three subsets. The evaluation data comprises of a total of 6 hours of audio (63K words), and also matches the distribution of the adaptation data sets. There is no overlap of speakers between the adaptation and evaluation corpora.

2.3. Performance of the baseline systems

Table 1: WERs of baseline systems on CC-A, Agent: CC-B , Customer: CC-B)

	CC-A	CC-B$_{agent}$	CC-B$_{cust.}$	Ave.
SI CNN	28.6	35.0	46.0	36.5
SI i-vec. DNN	30.9	29.5	42.0	34.1
SA i-vec. DNN	29.2	26.9	41.0	32.4

The Word Error Rates (WERs) of the two baseline systems on the evaluation corpora are presented in Table 1. The out-of-the-box i-vector based, speaker-adapted DNN system has better performance on the average than a speaker-independent CNN system. While no adaptation data is used here, the speaker-adapted, i-vector based DNN system, inherently has an adaptation component to it, as the statistics for the i-vector computation are derived from the evaluation corpora.

3. Supervised adaptation with varying amounts of data

The adaptation algorithm used in this section, is similar to the one proposed in [8]. This scheme resembles MAP adaptation, with the adapted weight updates arrived at from using a weighted combination of the updates from adaptation data and the baseline model. Unlike the work in [8] where adaptation was performed at a speaker level, in this work, the entire adaptation data is pooled and the algorithm is used as an overall domain adaptation scheme, as given by Equation 1.

$$\Delta \mathbf{w}_t = -\alpha \nabla_{\mathbf{w}} E(\mathbf{w}_t) - \beta(\mathbf{w}_{t-1} - \mathbf{w}_0) \quad (1)$$

where α is a learning rate, β is a regularization parameter, and $\mathbf{w}_0$ is model parameters of the initial model. The network was adapted using the cross-entropy training criterion and trained to convergence after 20 epochs.

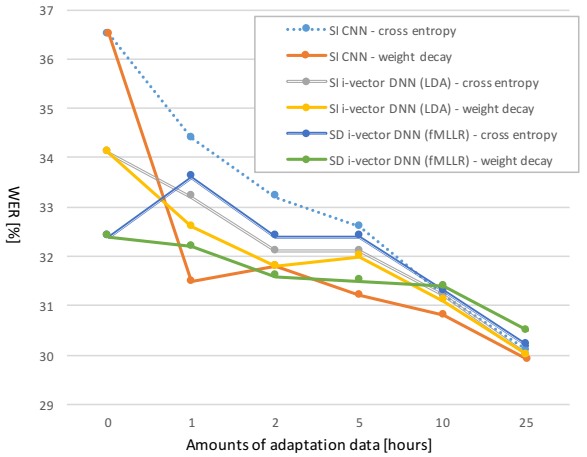

Figure 1: WERs with various amounts of adaptation data

Figure 1 illustrates the behavior of the three systems when adapted with varying amounts of data using the cross-entropy training criterion. As expected, as the amount of adaptation data is increased, all baseline systems provide improved performance. With 25 hours of adaptation data, the speaker-independent

CNN system outperforms all other systems albeit by a narrow margin. However, with only one hour of adaptation data, there is a lot more to be gained from adapting a speaker independent CNN system (dotted blue line), 6.7% relative improvement, than the speaker-independent i-vector DNN system (solid grey line), which yields only 2.6% relative improvement. On the other hand, the speaker-dependent (SD), i-vector DNN system which has the best performance with no adaptation, gets worse (solid blue line) when adapted with one hour of adaptation data. This could be attributed to the fact that the feature space transform estimation does not have sufficient statistics to render the adaptation stable.

Next, we study the impact of using a regularizer. The regularization algorithm implemented in this paper is in the form of a weight-decay based adaptation scheme. It outperforms simple adaptation (fine-tuning) on all three very different baseline systems. A different weight decay factor, that was empirically determined to maximize the classification performance on held out data, was used depending on the amount of adaptation data available. For example, a decay factor of 0.01 was used for up to 2 hours of adaptation data, 0.001 for 5 to 10 hours and 0.0001 for up to 25 hours. With increasing amounts of adaptation data, the network parameters can be refined reliably, and this is illustrated by the weight assigned to the updates derived from the adaptation data. The SI CNN system benefits the most with one hour of adaptation, obtaining a 16% relative improvement when using the regularizer (solid red line versus dotted blue line). In contrast, the i-vector based, SI DNN system improves by 1.6% relative (grey versus yellow line). The i-vector based SD DNN system, now shows a very marginal gain of less than 1% relative over the baseline performance. This clearly illustrates the importance of controlling the weight updates when very little adaptation data is available. As the training data is increased up to 25 hours, we can see that all models converge to within 2% relative performance of each other. However, in most real-world applications, it is desirable to see the best performance with as little as one hour of adaptation data. In such cases, it appears the use of a speaker-independent system provides maximum flexibility, with the CNN models outperforming the DNN models. Given that the best performance was obtained with an SI CNN system, we chose to explore the sequence training analysis using the SI-CNN system[1]

4. Speaker diversity

In this section, we compare the performance of the weight decay regularization scheme using a very limited amount of adaptation data, namely, 20 minutes selected from different speakers. The adaptation data comprises of 10 minutes from CC-A and 10 minutes from CC-B, each of which contains two speakers. In order to study the impact of the diversity of speakers in the adaptation data, we experimented with five different trials, selecting 10 minutes from each location but a different set of speakers each time. The results are tabulated in Table 2.

It can be observed that regardless of the distribution of speakers in the adaptation corpora, regularization is needed to obtain significant gains. Without regularization, adaptation can actually hurt performance, as we are trying to when very little adaptation data is available. This is not a surprising result as we are using very little adaptation data to refine a large number of parameters in the network. Also, a rather large variance in the final performance can be seen with WERs ranging between 31.9% and 34.0%, depending on the speakers selected for adaptation.

[1]Although, not presented in this paper, we have seen similar trends, albeit much smaller improvements with the speaker dependent and independent i-vector based DNN systems.

Table 2: WERs with only 20 minutes and 4 speakers adaptation data

	CC-A	CC-B_{agent}	CC-$B_{cust.}$	Ave.
Baseline CNN	28.6	35.0	46.0	36.5
Trial 1	31.5	29.0	45.9	35.5
Trial 1 (weight decay)	26.9	26.5	42.3	31.9
Trial 2	29.6	28.4	43.9	34.0
Trial 2 (weight decay)	27.2	26.9	42.1	32.1
Trial 3	29.8	26.3	43.5	33.2
Trial 3 (weight decay)	27.7	25.5	42.0	31.7
Trial 4	27.9	27.9	43.0	32.9
Trial 4 (weight decay)	26.8	26.5	42.1	31.8
Trial 5	28.3	31.4	46.2	35.3
Trial 5 (weight decay)	27.0	30.2	44.8	34.0

5. Adaptation using sMBR sequence training

Sequence training using the sMBR training criterion has shown consistent improvements over training the network's parameters with the cross entropy criterion. In this section, we explore the value of sequence training for adaptation. It is important to note here that the baseline systems used in this paper are all sequence trained. With MAP-like adaptation, some of the benefits of sequence training using a discriminative criterion are lost (smoothed out). In order to investigate if further improvements can be obtained with sequence training, we used the baseline system as the starting point, similar to Section 3. We first adapt using the cross-entropy criterion, followed by sequence training with the sMBR criterion. Figure 2 illustrates this process and the results are presented in Table 3. The SI CNN model is adapted with different amounts of adaptation data, training to covergence after 20 epochs. Subsequently, lattices are produced using this system and the network is trained with the sMBR training criterion, converging after 10 epochs through the adaptation data. Table 3 shows that sequence training provides consistent gains for all amounts of adaptation data, with the final adaptation performance with 25 hours of adaptation, exceeding the result in Figure 1 by 5.6% relative.

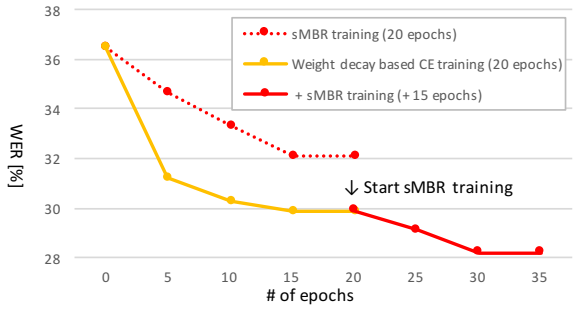

Figure 2: WERs of sequence training with 25 hours adaptation data

Table 5: WERs when 200 hours of unsupervised data and 2 hours of supervised data are available

	CC-A	CC-B_{agent}	CC-$B_{cust.}$	Ave.
Weight decay based CE adaptation (unsupervised 200h) + sMBR (supervised 2h)	25.8	22.8	40.8	29.8

Table 3: WERs of weight decay based adaptation (20 epochs) + sMBR sequence adaptation (10 epochs) with various amounts of training data

	CC-A	CC-B_{agent}	CC-$B_{cust.}$	Ave.
Baseline SI CNN	28.6	35.0	46.0	36.5
1h (weight decay)	24.6	27.8	42.1	31.5
1h (+sMBR)	24.3	27.3	41.0	30.9
2h (weight decay)	25.1	27.1	43.2	31.8
2h (+sMBR)	24.1	26.4	41.5	30.7
5h (weight decay)	24.4	26.5	42.8	31.2
5h (+sMBR)	24.0	25.6	40.6	30.1
10h (weight decay)	24.2	25.8	42.3	30.8
10h (+sMBR)	23.6	24.8	41.2	29.9
25h (weight decay)	23.8	23.6	42.3	29.9
25h (+sMBR)	22.4	22.8	39.3	28.2

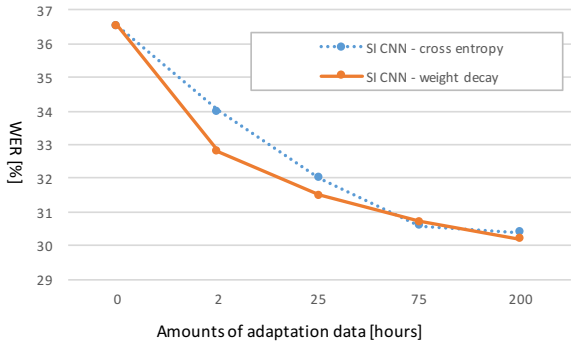

Figure 3: WERs of unsupervised adaptation with various amounts of training data

6. Unsupervised adaptation

In many real-world applications, acquiring data with reference (manually annotated) transcripts for the target domain is impractical and expensive. Therefore, it is essential that unsupervised adaptation algorithms achieve performance levels by unsupervised adaptation. In this section, we compare supervised and unsupervised adaptation using the approach presented in Sections 3 and 5. For unsupervised domain adaptation, we explore the use of up to an additional 200 hours of unlabeled data. The transcripts for adaptation were automatically generated using the SI CNN baseline system. Table 4 presents the WERs obtained by adapting the SI CNN system with varying amounts of unsupervised data. The first three rows of this table are a direct comparison with Table 3, i.e., the 2 hour and 25 hour adaptation data runs use the same data in both tables: one uses the reference transcripts and the other uses the automated transcripts. While significant improvements can still be obtained with the regularized unsupervised adaptation, the relative improvements are less than those obtained with supervised adaptation. This result is in-line with what is normally reported in the literature when using unlabeled data. For example, increasing the adaptation data from 2 hours to 25 hours, results in an improvement of 8% relative (WER reduction from 30.7% to 28.2%) for the supervised case, while the unsupervised case gains 4.3% relative (WER reduction from 32.8% to 31.4%). When the unsupervised data is increased to 75 hours and 200 hours, further improvements can be seen. However, with 200 hours of unsupervised data, the best WER obtained is 30.3%. In contrast, the best WER obtained with 25 hours of supervised data is 28.2%. Using just the unsupervised data allows us to approach the performance of a system adapted with only 5 hours of supervised data. Next, we explored the impact of replacing just 2 hours from the 200 hours of unsupervised data with reference transcripts. These results are shown in Table 5. The WER from this adapted system is now 29.8% which is closer to the performance of a system adapted with 10 hours of supervised data. This clearly indicates that even a small amount of supervised data, can be very impactful despite having 20 times more unsupervised data to adapt on.

Table 4: WERs of unsupervised adaptation with various amounts of training data

	CC-A	CC-B_{agent}	CC-$B_{cust.}$	Ave.
Baseline SI CNN	28.6	35.0	46.0	36.5
2h (weight decay)	26.7	28.7	43.0	32.8
2h (+sMBR)	27.3	28.7	42.3	32.8
25h (weight decay)	25.9	26.2	42.4	31.5
25h (+sMBR)	26.6	26.2	41.5	31.4
75h (weight decay)	26.4	24.0	41.8	30.7
75h (+sMBR)	26.7	23.8	40.7	30.4
200h (weight decay)	25.8	23.3	41.6	30.2
200h (+sMBR)	26.4	23.0	41.4	30.3

7. Conclusion

In this paper, we have demonstrated how neural network systems can be efficiently adapted with limited transcribed acoustic data starting from a well trained network using a regularized form of cross entropy based neural network training. The performance of these adapted models are significantly improved further with sMBR based sequence training, resulting in a net gain of 24.8% relative. With adaptation data likely to biased with only a few speakers we present empirical evidence that although there can be large variance because of the distribution of speakers, the proposed recipe can still be effective. As unsupervised data is more likely to be available for adaptation, the paper also explores the effect of unsupervised data on both the cross-entropy and sequence training parts of the adaptation recipe. With its empirical results the paper gives insights on how to best collect and use domain adaptation data given the cost and time effort behind data collection in new domains.

8. References

[1] Christopher J Leggetter and Philip 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.

[2] Jean-Luc Gauvain and Chin-Hui Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Speech and audio processing, ieee transactions on*, vol. 2, no. 2, pp. 291–298, 1994.

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

[4] Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Mike Seltzer, Geoffrey Zweig, Xiaodong He, Julia Williams, et al., "Recent advances in deep learning for speech research at microsoft," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8604–8608.

[5] Tara N Sainath, Brian Kingsbury, George Saon, Hagen Soltau, Abdel-rahman Mohamed, George Dahl, and Bhuvana Ramabhadran, "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks*, vol. 64, pp. 39–48, 2015.

[6] Tara N Sainath, Oriol Vinyals, Andrew Senior, and Hasim Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4580–4584.

[7] Andrew Senior, Hasim Sak, and Izhak Shafran, "Context dependent phone models for lstm rnn acoustic modelling," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4585–4589.

[8] Hank Liao, "Speaker adaptation of context dependent deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7947–7951.

[9] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and Frank Seide, "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7893–7897.

[10] Akio Kobayashi, Kazuo Onoe, Manon Ichiki, and Shohei Sato, "Comparison of unsupervised sequence adaptations for deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016.

[11] Jia Cui, Xiaodong Cui, Bhuvana Ramabhadran, Jung-Ho Kim, Brian Kingsbury, Jonathan Mamou, Lidia Mangu, Michael Picheny, Tara N Sainath, and Abhinav Sethy, "Developing speech recognition systems for corpus indexing under the iarpa babel program," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6753–6757.

[12] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7304–7308.

[13] Ngoc Thang Vu, David Imseng, Daniel Povey, Petr Motlicek, Tanja Schultz, and Hervé Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 7639–7643.

[14] Frank Seide, Gang Li, Xie Chen, and Dong Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 24–29.

[15] Kaisheng Yao, Dong Yu, Frank Seide, Hang Su, Li Deng, and Yifan Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 366–369.

[16] Pawel Swietojanski and Steve Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 171–176.

[17] Tsubasa Ochiai, Shigeki Matsuda, Xugang Lu, Chiori Hori, and Souichi Katagiri, "Speaker adaptive training using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6349–6353.

[18] Jian Xue, Jinyu Li, Dong Yu, Mike Seltzer, and Yifan Gong, "Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 6359–6363.

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

[20] Takuya Yoshioka, Anton Ragni, and Mark JF Gales, "Investigation of unsupervised adaptation of dnn acoustic models with filter bank input," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6344–6348.

[21] Brian Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 3761–3764.

[22] George Saon, Hong-Kwang J Kuo, Steven Rennie, and Michael Picheny, "The ibm 2015 english conversational telephone speech recognition system," *arXiv preprint arXiv:1505.05899*, 2015.

[23] Tara N Sainath, Abdel-rahman Mohamed, Brian Kingsbury, and Bhuvana Ramabhadran, "Deep convolutional neural networks for lvcsr," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8614–8618.