TOWARDS BETTER PERFORMANCE WITH HETEROGENEOUS TRAINING DATA IN ACOUSTIC MODELING USING DEEP NEURAL NETWORKS

Yan Huang, Malcolm Slaney, Michael L. Seltzer, and Yifan Gong

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052

{yanhuang;mslaney;mseltzer;ygong}@microsoft.com

ABSTRACT

Modeling heterogeneous data sources remains a fundamental challenge of acoustic modeling in speech recognition. We call this the multi-condition problem because the speech data come from many different conditions. In this paper, we introduce the fundamental confusability problem in multi-condition learning, then discuss the problem formalization, the taxonomy, and the architectures for multi-condition learning. While the ideas presented are applicable to all classifiers, we focus our attention in this work on acoustic models based on deep neural networks (DNN). We propose four different strategies for multi-condition learning of a DNN that we refer to as a mixed-condition model, a condition-dependent model, a condition-normalizing model, and a condition-aware model. Based on the experimental results on the voice search and short message dictation task and the Aurora 4 task, we show that the confusability introduced when modeling heterogeneous data depends on the source of acoustic distortion itself, the front-end feature extractor, and the classifier. We also demonstrate the best approach for dealing with heterogeneous data may not be to let the model sort it out blindly, even with a classifier as sophisticated as a DNN.

Index Terms— Multi-task learning, deep learning, CD-DNN-HMM, noise robustness, channel compensation

1. INTRODUCTION

Modeling heterogeneous acoustic data sources coming from diverse acoustic environments—distinct channels, and different speakers with varying speaking styles or accents—is a fundamental challenge for acoustic modeling in large-vocabulary speech recognition. While recent advances in acoustic modeling using deep neural networks (DNN) have led to significant performance improvements on a variety of academic and industrial tasks [1, 2, 3, 4, 5], a recent study revealed that model robustness and effectively modeling heterogeneous acoustic sources remain important research problems for acoustic modeling [6].

We categorize the prior research on modeling heterogeneous acoustic data into two approaches. The *normalization* approach refers to methodologies which normalize or factor out the phonetically irrelevant heterogeneity. Cluster adaptive training [7, 8] and acoustic model factorization [9, 10] are examples of the normalization approach. The *explicit modeling* approach directly models the heterogeneous data source, such as the acoustic trajectory model, the generalized mixture HMM, and the synchronous HMM [14, 15, 16].

There is previous research that studies multi-condition learning in a deep neural network [17, 18, 19]. In this paper, we endeavor to further this research and provide a systematic view of modeling heterogeneous data. Specifically, we first describe the confusability problem when modeling multi-condition data. We then formalize the multi-condition acoustic modeling problem. In particular, we adopt the source-channel speech model, in which the observed speech is generated by passing a canonical speech signal through various components of an acoustic scene. Here, the term "acoustic scene" represents all phonetically irrelevant acoustic conditions, each causing a distortion of the original signal.

We thus elucidate four architectures for handling multi-condition data in the context of a neural network acoustic model: the *mixed-condition model*, the *condition-dependent model*, the *condition-normalizing model*, and the *condition-aware model*. We present experiments with these multi-condition learning approaches to illustrate the ideas and demonstrate their use in dealing with the different gender, the distinct acoustic channels, and the various environmental noise.

Two important outcomes of this paper are: the confusability issue in modeling heterogeneous data depends on the nature of the specific acoustic distortion and the invariant feature extraction capability of the model; the best approach for dealing with heterogeneous data is to not necessarily let the model sort it out blindly, even with a classifier as sophisticated as a DNN.

We further point out that the condition-dependent model can naturally solve the confusability issue; nevertheless practically it very often leads to the sub-optimal performance due to the data scarcity. In the partially condition-dependent model with condition-normalizing hidden layers, different layers of the neural network may be suitable for the modeling of different types of acoustic distortion factors.

The rest of this paper is organized as follows: Section 2 discusses the confusability problem in multi-condition modeling; Section 3 presents four multi-condition learning approaches within the context of a DNN acoustic model: the formulation, the implementation architecture, and the utility; These approaches are evaluated through a series of experiments in Section 4. Finally, we summarize our finding and present some conclusions in Section 5.

2. CONFUSABILITY WHEN MODELING HETEROGENEOUS DATA

The fundamental problem in modeling data from heterogeneous sources arises as the result of confusability at the classification boundary. Consider the two-class recognition problem shown at the top of Figure 1. For condition 1 and condition 2, there is a single decision boundary that can clearly divide the data into the two classes. In this case, one can build a classifier to label /a/ vs. /i/ for either condition, or the combination.

Now consider the alternative scenario shown at the bottom of Figure 1. Here the /a/ vs. /i/ boundary is clear for either condition in isolation, and yet taken together, the decision boundary is

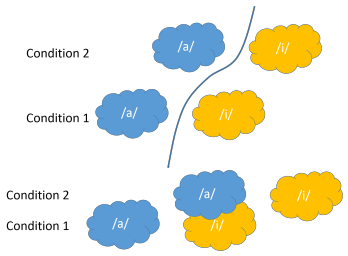

Fig. 1. *Two different types of confusability in an arbitrary feature space. In condition 1 the two vowels are distinguishable, even without knowing the condition (top). In condition 2 one must know or compensate for the condition to obtain perfect classification (bottom).*

unclear. Without knowing which condition generated the data, there is not a single good decision boundary. This illustrates the confusability problem in the multi-condition modeling. To solve the confusability problem in multi-condition modeling, one needs to either model the classification boundary for each condition separately, or a feature so that the "blurred" or "deformed" classification boundary in the multi-condition setup is well-separated again. Next, we will present the architecture and formulation of four different multi-condition learning methodologies and discuss how the confusability issue is handled in each of them.

3. FOUR MULTI-CONDITION LEARNING APPROACHES WITHIN THE CONTEXT OF A DEEP NEURAL NETWORK

We assume that the speech signal passes through various components of an acoustic scene (C) and we observe the signal (Y_C). The acoustic scene refers to a set of acoustic distortions that modify the canonical acoustic distribution and contain information irrelevant to phone discrimination, e.g. the environmental noise, channel, gender, speaker vocal tract, speaking style, etc. We treat each of them as one acoustic distortion channel and a subset of them jointly defines a particular acoustic scene in the multi-condition acoustic model framework. The remainder of this section will describe four multi-condition learning approaches within the context of the DNN-based acoustic model framework. These four approaches are illustrated in Figure 2.

3.1. Mixed-Condition Model

In the mixed-condition training we ignore the properties of the acoustic scene and pool the mixed-condition data to train a single model:

$$P(X|Y_{\boldsymbol{C}}) = P(X|Y). \tag{1}$$

Figure 2(a) is the widely-used multi-style training. Since this model mixes multi-condition data together, this type of model very often has "blurred" or "deformed" classification boundary. Whether the confusability issue described in Section 2 exhibits the top or the bottom scene as illustrated in Figure 1 dependents on the specific acoustic scene, the feature space, and the model. We further address this issue in our experiments.

3.2. Condition-Dependent Model

In the condition-dependent model, we partition the data into scene-dependent subsets and train scene-specific models:

$$P(X|Y_{\boldsymbol{C}}) = P_{\boldsymbol{C}}(X|Y). \tag{2}$$

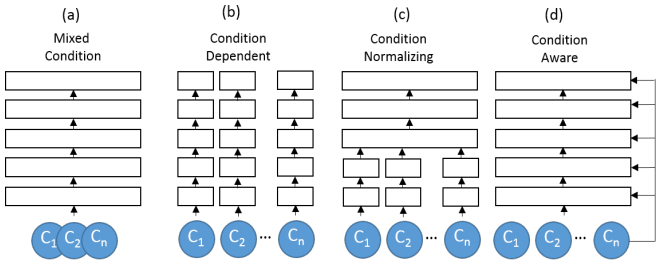

Fig. 2. *Four multi-condition learning architectures within the context of the DNN-based acoustic model framework. Each circle represents data from one acoustic condition.*

During the decoding stage, each utterance is processed by multiple models if details of the acoustic scene are unknown. The multiple models will compete to produce the final results:

$$P(X|Y_{\boldsymbol{C}}) = \underset{\boldsymbol{C}}{\text{argmax}}\ P_{\boldsymbol{C}}(X|Y). \tag{3}$$

The model separation mechanism in the condition-dependent model as in Figure 2(b) naturally resolves the confusability issue. However, for complicated acoustic scenes with many different acoustic factors, this approach is impractical. More importantly, this approach may have poor performance due to the data fragmentation, which will be further illustrated in our later experiments.

3.3. Condition-Normalizing Model

Figure 2(c) demonstrates a partial condition-dependent model structure that utilizes the condition-normalizing layers to explicitly model different conditions. This structure allows maximal data sharing across various conditions and yet keeps the modeling capability for heterogeneous data. The layers can be either linear or non-linear.

The condition-normalizing layer placed on the bottom hidden layer can be viewed as a nonlinear feature-normalization layer. The input is forward propagated through the appropriate condition-dependent lower branch of the model to normalize the input into a representation which removes the condition-specific irrelevant variability. By training the shared upper layers and condition-dependent lower layers jointly, we ensure the condition-dependent lower layers learn to perform the required normalization while the shared upper layers can focus on learning distinctions between phonetic classes.

The condition-normalizing layer can be also placed on the top or on the middle hidden layers. The choice of the placement for the condition-normalizing layer depends on the specific condition to be models and the neural network invariant feature extraction. The condition-normalizing hidden layer is typically trained via conducting model adaptation from a condition-independent model [11, 12, 13]. Usually re-training of the shared condition independent portion of the network generates further performance gain.

3.4. Condition-Aware Model

The condition-aware model explicitly models the posterior given the speech observation and the acoustic scene. Information about the acoustic scene is added to the neural network input as illustrated in Figure 2(d).

Below is the generalized formulation for the condition-aware

model:

$$P(X|Y_C) = \sum_c P(X, C|Y_C) = \sum_c P(C|Y_C)P(X|Y_C, C). \quad (4)$$

When the condition is known *a priori*, the model is simplified to $P(X|Y_C, C)$ which is illustrated in Figure 2(d).

The selection of the acoustic conditions to be modeled and their representation are the two key factors in the success of the condition-aware model. The acoustic scene in general should consist of the most distinct acoustic distortion factors in the task. Regarding the condition-aware representation, it remains as unresolved due to lack of a principled guide. The rule of thumb is that the representation should provide sufficient activation capacity to guide the neural network learning towards better class separation.

We further note that the acoustic scene information does not have to be descriptive information about the distortion, it can also be information estimated from the speech observation [17].

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, we present experimental results with the four models. We will demonstrate their usage in modeling gender, channel, and background noise.

Specifically, we first illustrate that the confusability behavior can change due to different modeling techniques by comparing the gender dependent and independent models in the GMM and the DNN; then demonstrate the confusability behavior also depends on the nature of the condition itself. We show that even with deep learning techniques the confusability issue still exists across distinct channels. We further demonstrate using the condition-normalizing layer can remedy the data fragmentation issue in the channel-specific model. Lastly, we present the condition-aware DNN modeling noise.

All of our experiments were conducted on a mobile voice search and short message dictation task (VS/SMD) except that the condition-aware model experiments were conducted on the Aurora 4, a medium-vocabulary corpus for evaluating noise robustness. All experiments used a Viterbie-based decoder and a 3-gram language model for decoding.

4.1. Mix-condition versus Condition-dependent Model for Gender in the GMM and DNN

In this experiment, we compare mixed-condition and the condition-dependent models with respect to gender using the GMM and DNN on the short message dictation task (SMD). The GMM is a discriminative model using the feature-space minimum phone error rate (fMPE) [24] and the boosted MMI (bMMI) [23] criterion with a 39-dimension MFCC front-end. The DNN model is a cross entropy (CE) model with 87-dimension log filter bank (LFB) front-end feature with 11-frame context windows. The GMM and the DNN share the same 400 hour training data, the decision tree, and the MLE seed model.

Table 1 compares the gender-dependent versus the gender-independent GMM and DNN models evaluated on a 50 hour SMD test set. With a GMM, the GD model outperforms the GI model with 8.7% word error rate reduction (WERR). This clearly demonstrates the condition-dependent model performs better than mixed-condition training for gender in the GMM. Nevertheless, with a DNN, the GI model performs almost as well as the GD model with a negligible performance difference.

Table 1. *Performance comparison of the gender-dependent (GD) and gender-independent (GI) GMM and DNN on the short message dictation task (SMD).*

	GI	GD	WERR (GD vs.GI)
GMM	21.8	19.9	8.7
DNN	14.9	14.7	1.3

This experiment indicates given a specific acoustic distortion factor, whether the confusability in the multi-condition learning exhibits the top or the bottom case in Figure 1 depends on the specific classifier (model) and the feature space. The deep neural network with the layer-wise nonlinear feature extraction may learn better invariant and selective features and therefore can turn a confusable problem in the GMM into a well-separable one in the DNN. The gender distortion is such a good example.

4.2. Channel Confusability and Multi-model for Distinct Channels

Many distortion factors do not belong to the above category and the confusability problem persists in the deep learning acoustic model. Channel variability is such an example. In a recent study [6], we found that the DNN improves the ASR performance of speech coming from all different devices comparing to the GMM. Nevertheless, the performance gap across different devices remains large. In this section, we illustrate that the confusability problem exists in deep learning based mixed-channel acoustic models.

We trained three channel-specific DNNs with 40 hours of speech from each mobile device channel. We also trained two mixed-channel DNNs, one using 40 hours randomly selected mixed channel data and the other one using all 120 hours mixed-channel data. All five DNNs share the same decision tree and the MLE seed model. Furthermore, for reference, we distilled "A+B+C.Multi" from the three channel-specific models. The test combines three 4 hour test set for each channel.

As shown in Table 2, the channel-specific DNN outperforms mixed-channel DNN trained using the same 40 hours of data with 4$\sim$5 % WERR for all three devices. This indicates confusability exists in the mixed-channel DNN. The notably different observation on channel as compared to gender suggests that whether the confusability existing as a distinct problem not only depends on the modeling technology, but also depends on the nature of the acoustic condition itself. In the DNN, certain acoustic conditions, such as the channel variation, does not get normalized as successfully as others, such as the gender.

If we increase the training data by mixing all 40 hour training data from the three devices, the resulting model (A+B+C.120hrs) has significant performance boost for each device. This shows the data fragmentation issue in the multi-model approach as described in Section 4.2.

Table 2. *Performance comparison of the mixed-condition versus the condition-dependent model for channel.*

	A	B	C	AVG
A.40hrs	**27.6**	32.7	26.7	
B.40hrs	32.0	**28.9**	28.8	
C.40hrs	29.7	31.5	**25.9**	
A+B+C.Multi	**27.6**	**28.9**	**25.9**	27.4
A+B+C.40hrs	28.3	29.7	26.9	28.4
A+B+C.120hrs	22.9	23.5	21.8	22.7

4.3. Condition Specific Layers for Channel Normalization

In this section, we present an experiment using the condition-normalizing layers to model specific channels and yet allow maximum data sharing across channels. This directly addresses the data fragmentation issues in a channel-dependent model.

Specifically, we train a mixed-channel DNN (A+B+C.120hrs) with 120 hours mixed channel data with the same setup as before, then add a channel-specific bottom hidden layer (linear layer) for each channel. The channel specific layer was trained via model adaptation using the 40 hour channel-specific data with the top layers fixed. We then evaluate the resulting model (A+B+C.120hrs.ChanLayer) with channel specific layers on the corresponding matched-channel test set. As shown in Table 3, "A+B+C.120hrs.ChanLayer" yields 0.7%, 0.5%, and 0.6% absolute word error rate reduction for device A, B, and C respectively as compared to the baseline mixed-channel model (A+B+C.120hrs). On average, a 2.6% reduction in WER was achieved. Furthermore, we conduct model re-training for the shared top layers using the channel normalized features derived from the channel dependent bottom layer. The resulting model (A+B+C.120hrs.ChanLayer.RT) yields an additional 0.2% absolute word error rate reduction. In total, 3.5% WER reduction is obtained over to the baseline multi-condition model (A+B+C.120hrs). We further note that the placement of the condition specific layer depends on the specific conditions to be modeled, the layer-wise feature learning, and the information distribution in the deep neural network. The multi-lingual DNN work in [20, 21] suggests an alternative choice for the placement of the condition specific layer.

Table 3. *Performance of the condition-normalization DNN with channel-specific layers.*

	A	B	C	AVG
A+B+C.120hrs	22.9	23.5	21.8	22.7
A+B+C.120hrs.ChanLayer	22.2	23.0	21.2	22.1(2.6)
A+B+C.120hrs.ChanLayer.RT	22.1	22.7	21.0	21.9(3.5)

4.4. Condition-Aware Model Experiments

In order to study the potential of a DNN-based acoustic model to exploit information about the acoustic scene, we performed an experiment using Aurora 4, a medium-vocabulary corpus based on WSJ0. In this corpus, the multi-condition training set consists of 7137 utterances from 83 speakers. One half of the utterances was recorded by a high-quality close-talking microphone and the other half was recorded using a variety of secondary microphones. Both halves include a combination of clean speech and speech corrupted by one of six different types of noise (street traffic, train station, car, babble, restaurant, airport) at a range of signal-to-noise ratios (SNR) between 10–20 dB. The evaluation set consists of 330 utterances from 8 speakers. This test set was recorded by the primary microphone and a number of secondary microphones. These two sets are then each corrupted by the same six noises used in the training set at SNRs between 5–15 dB, creating a total of 14 test sets. These 14 test sets can then be grouped into 4 subsets, based on the type of distortion: none (A), additive noise only (B), channel distortion only (C), and noise + channel (D).

In this experiment, we defined the acoustic scene as the noise environment of the utterance. There were six noise environments plus the clean condition for a total of seven acoustic scenes. To make the network aware of the scene, a 7-dimensional vector that performs a one-hot encoding of the acoustic scene (noise condition) is added to every hidden layer of the network. This one-hot vector can be considered a dynamic bias at each layer that changes based on the acoustic scene. In this experiment, we assumed that knowledge of which environment an utterance was in was known a priori, in both training and test. Knowledge about the particular SNR was not used.

Table 4 compares the performance of two DNN-based speech recognition systems. In both cases, the networks were trained with a context window of 11 frames of 24-dimensional log filterbank features, augmented with the log energy. Static, delta, and delta-delta features were used and mean normalization was applied. The network was trained with three hidden layers with 2048 hidden units each, and an output layer representing 3202 senones. In the context-aware network, all three hidden layers were augmented with the encoding of the acoustic context. We note that the baseline system performance is slightly different from [17] due to a different model training setup.

As the table shows, only a marginal improvement in overall accuracy was obtained by this method of augmenting the input features with information about the acoustic context. Improvements in performance in additive noise conditions (B) were negated by a degradation in clean conditions (A). We note that an ASR system that includes speaker information shows promise using a similar condition-aware model approach [19].

Table 4. *Performance comparison of a standard DNN and a DNN augmented with an encoding of the acoustic scene.*

	A	B	C	D	AVG
DNN (3×2048)	5.9	10.4	9.8	22.4	15.2
+ scene dynamic bias	6.1	10.2	9.8	22.4	15.1

5. CONCLUSION

We studied the multi-condition learning problem in a deep learning acoustic modeling framework. We first described the fundamental confusability problem in the multi-condition learning; then discussed the mixed-condition, condition-dependent, condition-normalizing, and condition-aware four types of methodologies on the formulation, implementation architecture, and practical utility.

We introduced the "acoustic scene" concept to represent different acoustic distortion channels for canonical speech. We showed that the degree of confusability introduced in modeling heterogeneous data depends on the specific acoustic distortion factor itself, the acoustic front-end feature extractor, and the classifier. The deep neural network with the layer-wise nonlinear feature extraction may turn some confusable problems (e.g. gender) in the GMM into a well-separable one. Nevertheless, many other heterogeneities (e.g. channel variation) cannot be normalized blindly as successfully even in the DNN. We illustrated that the condition specific layers can be used to normalize the channel variation in the DNN. The best placement of the condition specific layer in the network may depend on the nature of the acoustic condition itself.

Lastly, we studied the condition-aware DNN model which explicitly models the posterior given the speech observation and the acoustic scene. The success of the condition-aware DNN largely depends on the effective representation of the acoustic scene which can be the simple identifier or descriptiive information about the specific acoustic condition estimated from the speech observation. The best approach for dealing with heterogeneous data is not to necessarily always let the model sort it out blindly, even with a classifier as sophisticated as a DNN.

6. REFERENCES

[1] Dahl, G.E., Yu, D., Deng, L., and Acero, A., "Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition," IEEE Transactions on Audio, Speech, and Language Processing (TASLP) — Special Issue on Deep Learning for Speech and Language Processing, Volume: 1, No. 1, Page(s): 33–42, Jan 2012.

[2] Seide, F., Li, G., and Yu, D., "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks," in the Proceedings of Interspeech 2012.

[3] Kingsbury, B., Sainath, N. T., and Soltau, H., "Scalable Minimum Bayes Risk Training of Deep Neural Network Acoustic Models Using Distributed Hessian-free Optimization," in the Proceedings of Interspeech 2012.

[4] Jaitly, N., Nguyen, P., Senior, A., and Vanhoucke, V., "Application of Pretrained Deep Neural Networks to Large Vocabulary Speech Recognition," in the Proceedings of Interspeech 2012.

[5] Deng, L., Li, J., Huang, J., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J., Gong, Y., and Acero, A., "Recent Advances in Deep Learning for Speech Research at Microsoft," in the Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[6] Huang, Y., Yu, D., Liu, C., and Gong, Y., "A Comparative Analytic Study on the Gaussian Mixture and Context Dependent Deep Neural Network Hidden Markov Models," submitted to the Interspeech 2014.

[7] Gales, M. J. F., "Cluster adaptive training of hidden Markov models," IEEE Trans. Speech Audio Processing, vol. 8, no. 4, pp. 417428, July 2000.

[8] Yu, K., "Adaptive Training for Large Vocabulary Continuous Speech Recognition," Ph.D. Thesis, Cambridge University, 2006.

[9] Gales, M. J. F., "Acoustic Factorisation," in Proceedings of ASRU 2001.

[10] Wang, Y., Gales, M. J. F., "Speaker and Noise Factorization for Robust Speech Recognition," in IEEE Transactions on Audio, Speech, and Language Processing (TASLP), VOL. 20, NO. 7, September 2012.

[11] Seide, F., Li, G., Chen, X., and Yu, D., "Feature Engineering in Context-Dependent Deep Neural Networks for Conversational Speech Transcription," in the Proceedings of the ASRU, 2011.

[12] Li, B. and Sim, K. C., "Comparison of Discriminative Input and Output Transformations for SpeakerAdaptation in the Hybrid NN/HMM Systems," in the Proceedings of Interspeech 2010.

[13] Hank, L., " Speaker Adaptation of Context Dependent Deep Neural Networks," in the Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[14] Gong, Y., "Stochastic trajectory modeling and sentence searching for continuous speech recognition," IEEE Transactions on Speech Audio Processing, vol. 5, no. 1, pp. 3344, 1997.

[15] Korkmazskiy, F., Juang, B. H., and Soong, F. K. Generalized mixture of HMMs for continuous speech recognition, in the Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1997.

[16] Zhao, Y. and Juang, B. H., "Modeling Heterogeneous Data Sources for Speech Recognition Using Synchronous Hidden Markov Models," in the Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[17] Seltzer, M., Yu, D., Wang Y., "An Investigation of Deep Neural Networks for Noise Robust Speech Recognition," in the Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[18] Abdel-Hamid, O. and Jiang, H., "Fast Speaker Adaptation of Hybrid NN/HMM Model for Speech Recognition Based on Discriminative Learning of Speaker Code," in the Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[19] Saon, G., Soltau, H., Nahamoo, D., and Picheny M., "Speaker Adaptation of Neural Network Acoustic Models Using I-Vectors," in the Proceedings of the ASRU 2013.

[20] Huang, J., Li, J., Yu, D., Deng, L., and Gong, Y., "Cross-Language Knowledge Transfer Using Multilingual Deep Neural Network With Shared Hidden Layers," in the Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[21] Arnab G., Pawel, S. and Steve, R., "Multilingual training of deep neural networks," in the Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[22] Yu, D., Seltzer, M., Li, J. ,Huang, J., and Seide, F., "Feature Learning in Deep Neural Networks – Studies on Speech Recognition Tasks," in the Proceedings of 2013 International Confernece on Learning Representation, 2013.

[23] Povey, D, Kingsbury, B., Ramabhadran, B., Saon, G., Soltau H., and Visweswariah, K., "Boosted MMI for model and feature-space discriminative training," in the Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2008.

[24] Povey, D, Kingsbury, B., Mangu, L., Saon, G., Soltau, H., and Zweig, G., "fMPE: Discriminatively Trained Features for Speech Recognition," in the Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2005.