DATA AUGMENTATION FOR DEEP CONVOLUTIONAL NEURAL NETWORK ACOUSTIC MODELING

Xiaodong Cui, Vaibhava Goel, Brian Kingsbury

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA

ABSTRACT

This paper investigates data augmentation based on label-preserving transformations for deep convolutional neural network (CNN) acoustic modeling to deal with limited training data. We show how stochastic feature mapping (SFM) can be carried out when training CNN models with log-Mel features as input and compare it with vocal tract length perturbation (VTLP). Furthermore, a two-stage data augmentation scheme with a stacked architecture is proposed to combine VTLP and SFM as complementary approaches. Improved performance has been observed in experiments conducted on the limited language pack (LLP) of Haitian Creole in the IARPA Babel program.

Index Terms— convolutional neural networks, bottleneck features, data augmentation, vocal tract length perturbation, stochastic feature mapping.

1. INTRODUCTION

Data augmentation based on label-preserving transformations has been shown to be very effective at improving the robustness of deep neural networks [1][2][3][4], especially when the training data is limited. It is commonly used in image recognition where transformations such as translation, rotation, scaling and reflection [1][4] have led to significant improvements in recognition accuracy.

Data augmentation in speech-related applications is not a new practice. For instance, sometimes under the name of multi-style training [5], artificial noisy speech data is generated by adding noise to clean speech data for training noise robust acoustic models in automatic speech recognition (ASR). Another example is IMELDA [6] where multi-condition transforms are learned from tilted, noisy and undegraded speech data so that the sensitivity of the transforms to those conditions is reduced.

When it comes to deep neural network (DNN) or convolutional neural network (CNN) acoustic modeling, which has achieved the state-of-the-art performance in ASR nowadays, there is less reported work on data augmentation algorithms that are specifically designed to deal with speaker and acoustic variabilities for DNN or CNN training. Most recently, vocal tract length perturbation (VTLP) was proposed in [3] for augmenting data in CNN training. Experiments on the TIMIT database have shown decent improvements in phone error rate (PER). In [7], data augmentation using stochastic feature mapping (SFM) was proposed for DNN acoustic modeling. SFM augments training data by mapping speech features from a source speaker to a target speaker, which is equivalent to a special type of voice conversion in some designated feature space.

In this paper, we first show how SFM can be carried out in the log-Mel domain in CNN training and compare it with VTLP in this scenario. In addition, we propose a two-stage data augmentation scheme with a stacked architecture that combines VTLP and SFM as complementary approaches. In this scheme, a bottleneck CNN is first trained using data augmented by VTLP as a feature extractor in the first stage. The extracted bottleneck features are further normalized by speaker adaptive training. The speaker-adapted bottleneck features (with context) are again employed as input to build another DNN as the final classifier using data augmented by SFM in the second stage. Since VTLP and SFM generate data in different ways, this stacked architecture can make use of the merits of both approaches.

The remainder of the paper is organized as follows. Section 2 briefly reviews VTLP and SFM for DNN training in the speaker-adaptive feature space. Section 3 addresses VTLP and SFM for CNN training with log-Mel features as input. Section 4 gives the details about the proposed two-stage data augmentation approach that integrates both VTLP and SFM. Experimental results on limited language pack (LLP) of Haitian Creole are presented in Section 5 which is followed by a discussion in Sections 6 and a summary in Section 7.

2. DATA AUGMENTATION FOR DNNS

VTLP and SFM are investigated in [7] for DNN models in the speaker adaptive feature space as shown in Fig.1. In this feature processing pipeline 13-dimensional mean-normalized perceptual linear prediction (PLP) features with vocal tract length normalization (VTLN)[8] are used as the fundamental acoustic features. After taking into the context (CTX) by splicing adjacent 9 frames, linear discriminant analysis (LDA) is used to project the feature dimensionality down to 40. The components of LDA features are further decorrelated by a global semi-tied covariance (STC) matrix [9] and followed by speaker adaptive training (SAT) using feature space maximum likelihood linear regression (FMLLR). Finally, 9 frames (CTX2=± 4) of adjacent speaker-adapted features are used as input to the DNNs. In what follows, we will briefly review how VTLP and SFM are conducted in this scenario.

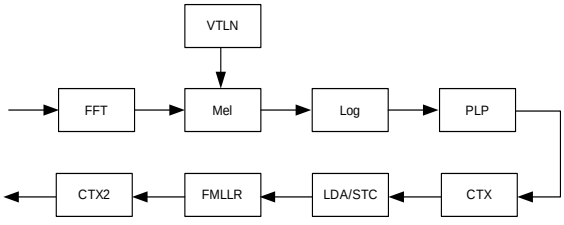

Fig. 1. Speaker adaptive feature space for DNN models.

2.1. Vocal Tract Length Perturbation (VTLP)

VTLP was first proposed in [3], where for each utterance in the training set a warping factor α is randomly chosen from $[0.9, 1.1]$ to warp the Mel-frequency axis to generate a new replica of the original data. In [7], a modified version of VTLP is used where the warping factor is perturbed deterministically according to Eq.1

$$\alpha \mapsto \{\alpha-4,\ \alpha-2, \alpha+2,\ \alpha+4\} \tag{1}$$

Eq.1 follows the notation in the IBM Attila toolkit [10] in which the vocal tract length warping factor is quantized between $[0.8, 1.25]$. As a result, the estimated warping factor α is an integer between $[0, 20]$ with 10 equivalent to the neutral warping factor 1.0.

According to Eq.1, VTLN warping factor α for a speaker is first estimated and then perturbed in both positive and negative directions by small shifts (± 2 and ± 4) to give 4 more warping factors. The perturbed warping factors, if beyond $[0.8, 1.25]$, are clipped to 0.8 or 1.25, which corresponds to integer 0 or 20, respectively,

2.2. Stochastic Feature Mapping (SFM)

To conduct SFM, one needs to choose a source speaker, a target speaker and a desired feature space. A speaker dependent model is first built for the target speaker. A mapping between the two speakers is then estimated in the chosen feature space based on the feature sequences from the source speaker and the speaker-dependent model of the target speaker under a selected statistical criterion.

In [7], SFM is designed with respect to the speaker adaptive feature space in Fig.1 for DNN models. The LDA space is the feature space in which the mapping is created. The speaker dependent model for the target speaker B in the LDA space $\boldsymbol{\lambda}_{\text{LDA}}^{(B)}$ is estimated by model space maximum likelihood linear regression (MLLR) [11]. A linear mapping (FMLLR) between the source and the target speakers is estimated under the maximum likelihood (ML) criterion

$$\{\tilde{\mathbf{A}}, \tilde{\mathbf{b}}\} = \underset{\{\mathbf{A},\mathbf{b}\}}{\operatorname{argmax}} \log P\big(\mathbf{A}\mathbf{O}_{\text{LDA}}^{(\text{S})} + \mathbf{b}|\boldsymbol{\lambda}_{\text{LDA}}^{(B)}\big). \tag{2}$$

After the linear transformation $\{\tilde{\mathbf{A}}, \tilde{\mathbf{b}}\}$ is estimated, the LDA feature sequence for the target speaker B can be obtained by

$$\mathbf{O}_{\text{LDA}}^{(\text{B})} = \tilde{\mathbf{A}}\mathbf{O}_{\text{LDA}}^{(\text{S})} + \tilde{\mathbf{b}} \tag{3}$$

and the final speaker-adapted feature sequence of the target speaker can be obtained as

$$\mathbf{O}_{\text{FMLLR}}^{(\text{B})} = \mathbf{A}^{(\text{B})}(\tilde{\mathbf{A}}\mathbf{O}_{\text{LDA}}^{(\text{S})} + \tilde{\mathbf{b}}) + \mathbf{b}^{(\text{B})} \tag{4}$$

where $\{\mathbf{A}^{(\text{B})}, \mathbf{b}^{(\text{B})}\}$ are speaker adaptive linear transformation (FMLLR) in Fig.1 for the target speaker.

To augment the training data, for each speaker in the training set a number of speakers are randomly chosen from the same training set as target speakers. All feature sequences of this speaker are mapped to those target speakers.

3. DATA AUGMENTATION FOR CNNS

CNNs are known to be more invariant to pattern variabilities due to the normalization effect of convolutions in local receptive fields and subsequent pooling. For instance, speaker variability caused by vocal tract differences can be effectively reduced by CNNs. This property makes CNNs especially attractive when the training data is sparse.

In this section, we extend our previous data augmentation effort to CNN acoustic modeling. Since the input to CNNs needs to be topographical, normalized log-Mel features with context are usually used where the outputs of the Mel-frequency filter bank after VTLN are taken the logarithm and their speaker-dependent mean is computed and subtracted. The normalized log-Mel features are spliced with their left and right 5 adjacent frames to form a feature map. Two other feature maps are created by computing deltas and double deltas.

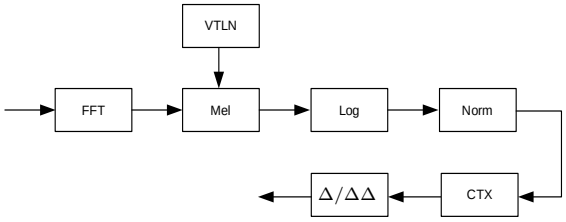

Fig. 2. Normalized log-Mel feature space for CNN models.

Given the normalized log-Mel input features, the extension of VTLP to CNN is straightforward. Eq.1 can be directly applied.

To apply SFM in the log-Mel feature space, one can still follow the procedure in Section 2.2 to first build a speaker-dependent model in the log-Mel feature space and then estimate a linear transformation in that space to transform the data. However, different from the speaker-adaptive feature space in the DNN scenario, dimensions of the log-Mel features, which are the outputs of Mel-frequency filter bank, are strongly correlated. Since the standard FMLLR estimation assumes diagonal covariances in GMMs [12], it can not be directly applied to the log-Mel feature space. One way to cope with this problem is to diagonalize the log-Mel feature space before estimating the FMLLR transformation. After the features are transformed in the diagonalized space they are transformed back to the original log-Mel space. The diagonalization is accomplished by a global semi-tied covariance (STC) transformation [9]. This mapping from the source speaker S to the target speaker B is indicated in Eq.5:

$$\mathbf{O}_{\text{LogMEL}}^{(\text{B})} = \mathbf{C}^{-1} \cdot \mathbf{F} \cdot \mathbf{C} \cdot \mathbf{O}_{\text{LogMEL}}^{(\text{S})} \tag{5}$$

where $\mathbf{C}$ is the STC transformation and $\mathbf{C}^{-1}$ is its inverse. $\mathbf{F}$ is the (augmented) FMLLR transformation in the diagonalized log-Mel feature space. Note that in order to estimate the FMLLR transformation in the diagonalized space, the speaker dependent model of the target speaker $\boldsymbol{\lambda}^{(B)}$ has to be trained with STC. This diagonalization approach in Eq.5 has been previously used in speaker normalization for CNN inputs [13].

4. A TWO-STAGE DATA AUGMENTATION SCHEME

While both VTLP and SFM augment training data based on label-preserving transformations, they augment in different ways. VTLP attempts to create "new" speakers by perturbing the vocal tract length of a speaker, which appears to be especially effective for systems that use Mel-frequency as their final feature space. SFM does not create new speakers but by statistically mapping feature sequences between speakers it can improve acoustic richness in the training data. Therefore, there is a reason to believe that the two approaches can be complementary.

In this section, we propose a two-stage data augmentation scheme to take advantage of both VTLP and SFM in a stacked architecture which is illustrated in Fig.3.

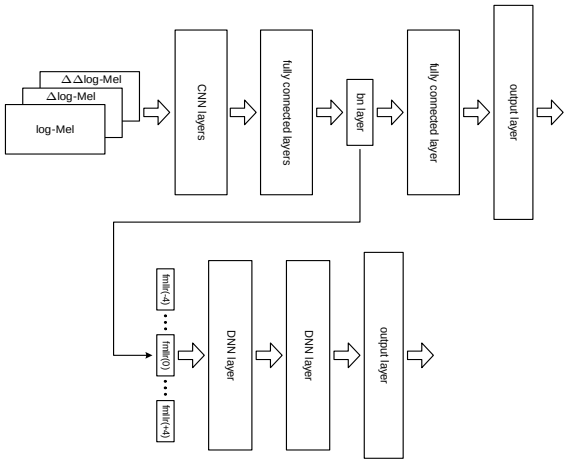

Fig. 3. The stacked architecture that combines VTLP and SFM for two-stage data augmentation.

In the first stage, a bottleneck CNN is built with mean normalized log-Mel features as input. This CNN is trained with training data augmented using VTLP. The bottleneck layer is one layer below the last fully connected layer in the network, whose details will be described in the experimental section. After the bottleneck CNN is trained, it is used as a feature extractor where the input to the sigmoid nonlinear activation function of the bottleneck layer is used as the features. The reason to use the input rather than the output of the sigmoid nonlinearity is to ensure a good dynamic range of the features and furthermore the resulting linear features are roughly normally distributed, which will benefit the speaker adaptive GMM training later. Since the CNN is trained with VTLP, the features extracted this way is expected to be more speaker invariant than the original features.

Upon the extracted bottleneck features, an ML speaker-adaptive model based on FMLLR is estimated. It has been observed that speaker adaptation on the bottleneck features helps the stacked bottleneck architecture [14].

In the second stage, a DNN is built whose input is the speaker-adaptive features coming from the feature space of the speaker-adaptive model. This DNN is learned with training data augmented using SFM, through which the acoustic richness of the training data is further improved.

5. EXPERIMENTAL RESULTS

Experiments are conducted on the IARPA Babel Haitian Creole LLP. It comprises 23.8 hours of telephony data for the training data set and 20.1 hours of telephony data for the development set. The training data set consists of scripted and conversational speech while the development set consists of conversational speech only. Specifically, the training set is composed of 19.9 hours of conversational data and 3.9 hours of scripted data. Most of the data is sampled at 8 KHz. A small portion of the data is originally sampled at 48KHz but downsampled to 8KHz for training. Approximately 40%-50% of the audio is speech.

We will compare the performance of VLTP and SFM under the DNN and CNN architectures and also the performance of the proposed two-stage data augmentation scheme under the stacked CNN architecture.

5.1. DNN Experiments

The baseline DNN acoustic model has 5 hidden layers of 1024 hidden units with sigmoid activation functions and a softmax output layer. The input to the network is 9 adjacent frames of 40 dimensional speaker adaptive features. Therefore the total dimensionality of the input is 360. The network is initialized with layer-wise discriminative pre-training. After the pre-training, it is first optimized by 15 iterations of cross-entropy (CE) training followed by 30 iterations of Hessian-free (HF) sequence training based on the state-level minimum Bayes risk (sMBR) criterion [15]. In the case of data augmentation, both VTLP and SFM generate 4 replicas of the original data, which makes the augmented training data 5 times larger than the original training data.

model	data augmentation	CE	sMBR
DNN	none	66.9	62.8
	VTLPx4	62.2	59.6
	SFMx4	62.3	59.1

Table 1. Word error rates (WERs) of DNN baseline and data augmented DNN acoustic models.

Table 1 shows the WERs of the baseline DNN model without data augmentation and the WERs of the DNN models trained using VTLP and SFM. After HF sequence training, the WER of the baseline DNN is 62.8% while the WERs of data augmented DNN models under VTLP and SFM are 59.6% and 59.1%, respectively. Both data augmentation techniques improve the ASR performance. SFM in this case is 0.5% absolute better than VTLP.

5.2. CNN Experiments

The baseline CNN model has two convolutional layers followed by five fully connected feedforward layers. All hidden layers use sigmoid activation functions and the output layer is softmax. The input features to the first convolutional layer are 40-dimensional log-Mel features with VTLN and their deltas and double deltas. The temporal context is 11 frames. There are 128 hidden units (feature maps) in the first convolutional layer, the local receptive field has an overlapping window of 9x9 with a shift of 1 in both temporal and spectral domains, which results in 32x3 windows for each feature map. On top of that, max pooling is applied in a 3x1 non-overlapping window which results in 11x3 windows for each feature map. There are 256 hidden units (feature maps) in the second convolutional layer, the local receptive field has an overlapping window of 4x3 with a shift of 1 in both temporal and spectral domains which results in 8x1 windows for each feature map. Following the second convolutional layer are five fully connected feedforward layers, each containing 1,024 units. The training of the CNN is similar to that of the DNN described in Section 5.1 which is composed of 15 iterations of CE training followed by 30 iterations of HF sMBR sequence training.

Table 2 shows the WERs of the baseline CNN model without data augmentation and the WERs of the CNN models trained using VTLP and SFM. After HF sequence training, the WER of the baseline CNN is 61.2% while the WERs of data augmented DNN models under VTLP and SFM are 58.4% and 58.7%, respectively.

model	data augmentation	CE	sMBR
CNN	none	64.6	61.2
	VTLPx4	61.9	58.4
	SFMx4	61.2	58.7

Table 2. Word error rates (WERs) of CNN baseline and data augmented CNN acoustic models.

The CNN baseline model (61.2%) is 1.6% absolute better than the DNN baseline (62.8%), which indicates that the CNN model is better than the DNN model given the sparse training data. Similar to the DNN scenario, both data augmentation techniques improve the ASR performance for CNN models. VTLP in this case is 0.3% absolute better than SFM.

5.3. Two-stage Data Augmentation Experiments

For the proposed two-stage data augmentation scheme, the bottleneck CNN has two convolutional layers followed by six fully connected feedforward layers among which the second topmost layer is a bottleneck layer. Other than the bottleneck layer, all other layers including both convolutional and fully connected layers have the same setup as that in Section 5.2. The bottleneck layer consists of 40 hidden units. The training of this bottleneck CNN is composed of 15 iterations of CE training followed by 30 iterations of HF sMBR sequence training. VTLP is used in the training of this stage where 4 replicas of the original data are generated.

The input to the sigmoid nonlinearity in the bottleneck layer is chosen as the features for the next stage DNN training. The reason behind it is that after comparing the performance using the input and output of the sigmoid nonlinearity of the bottleneck layer we find the input to the sigmoid has a better dynamic range as features which benefits the speaker adaptive training in the later stage.

The DNN training in the second stage employs the speaker adapted bottleneck features as input. There are 2 hidden layers in the DNN and each layer has 1,024 hidden units with sigmoid nonlinear activation functions. The DNN model is first trained using 15 iterations of CE training then followed by 30 iterations of HF sMBR sequence training. SFM is applied in this stage together with VTLP. So a total of 8 replicas of the original data are generated.

model	data augmentation	CE	sMBR
CNN baseline	none	64.6	**61.2**
CNN	VTLPx4	61.9	**58.4**
CNN-bn40	VTLPx4	62.9	59.1
CNN-bn40-fmllr DNN	VTLPx4	60.1	58.0
CNN-bn40-fmllr DNN	VTLPx4+SFMx4	59.2	**57.1**

Table 3. Word error rates (WERs) of baseline CNN and CNN/DNN using two-stage data augmentation.

Table 3 shows the WERs of the baseline CNN model without data augmentation and the WERs of the stacked architecture that uses the proposed two-stage data augmentation. The WER of the CNN baseline is 61.2% and with VTLP it is reduced to 58.4%, which has already been reported in Table 2. When adding a bottleneck layer, the WER of the bottleneck CNN model with VTLP is 59.1% which is 0.7% worse than without using the bottleneck layer. However, when using the speaker adapted bottleneck features to train the DNN using SFM on top of VTLP, the final WER is 57.1%. Therefore, after combining VTLP and SFM using the stacked architecture, this two-stage data augmentation scheme is 1.3% absolute better than CNN using VTLP and 1.6% absolute better than CNN using SFM.

6. DISCUSSION AND RELATION TO PRIOR WORK

VTLP was first reported in [3] being applied to CNN models using log-Mel input features. In [7], SFM was proposed for DNN models using speaker adaptive input features. In this work, we first compare the performance of VTLP and SFM for both DNN and CNN models. In principle, the label-preserving transformations under VTLP and SFM for data augmentation are transparent to the models and only have to do with the input feature spaces. For VTLP, the VTL perturbation in either the log-Mel feature space or speaker adaptive feature space is exactly the same. For SFM, care needs to be taken when estimating the mapping between the source and the target speakers using FMLLR in the log-Mel feature space as input for CNN models. The FMLLR transformation has to be performed in a covariance-diagonalized feature space given the strong correlation between the Mel filter bank outputs. The covariance diagonalization is accomplished by STC in this work.

Furthermore, a novel two-stage data augmentation scheme based on a stacked CNN architecture is proposed in this work, motivated by the speculation that VTLP and SFM could be complementary given their nature on generating label-preserving transformations. The bottleneck CNN trained with VTLP as a feature extractor is expected to generate features that are more speaker invariant. On top of that, the DNN model trained with SFM in the speaker adapted bottleneck feature space can further improve the acoustic richness in the training data. We find that stage-wise augmentation of the training data using VTLP and SFM can obtain better performance than VTLP or SFM alone for augmenting the training data for either DNNs or CNNs. This two-stage data augmentation scheme so far has yielded the best performance on Haitian Creole LLP for our systems.

7. SUMMARY

In this paper, we extended our previous work on data augmentation using SFM from DNN models to CNN models and compared the performance of VTLP and SFM in both DNN and CNN models. We proposed a novel two-stage data augmentation scheme based on a stacked CNN architecture that takes advantage of the complementary nature of VTLP and SFM. By improving the transformation invariance under VTLP and SFM stage-wise, superior performance is obtained for Haitian Creole LLP.

8. ACKNOWLEDGEMENTS

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government. This effort uses the IARPA Babel Program language collection release IARPA-babel201b-v0.2b limited language pack.

9. REFERENCES

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *International Conference on Document Analysis and Recognition (ICDAR)*, 2003, pp. 958–963.

[3] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *International Conference on Machine Learning (ICML) Workshop on Deep Learning for Audio, Speech, and Language Processing*, 2013.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Neural Information Processing Systems (NIPS)*, 2012, pp. 1106–1114.

[5] F.-H. Liu, Y. Gao, L. Gu, and M. Picheny, "Noise robustness in speech to speech translation," in *Eurospeech*, 2003.

[6] M. J. Hunt and C. Lefebvre, "Distance measures for speech recognition," *Aeronautical Note, NAE-AN-57*, 1989.

[7] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5582–5586.

[8] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 49–60, 1998.

[9] 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.

[10] H. Soltau, G. Saon, and B. Kingsbury, "The IBM Attila speech recognition toolkit," in *Spoken Language Technology Workshop (SLT)*, 2010, pp. 97–101.

[11] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[12] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[13] T. N. Sainath, B. Kingsbury, A. r. Mohamed, G. E. Dahl, G. Saon, H. Soltau, T. Beran, A. Y. Aravkin, and B. Ramabhadran, "Improvements to deep convolutional neural networks for LVCSR," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2013, pp. 315–320.

[14] M. Karafiat, F. Grezl, K. Vesely, M. Hannemann, I. Szoke, and J. Cernocky, "BUT 2014 Babel system: Analysis of adaptation in NN based systems," in *Interspeech*, 2014.

[15] B. Kingsbury, T. N. Sainath, and H. Soltau, "Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization," in *Interspeech*, 2012.