Speaker Adaptation of Neural Networks with Learning Speaker Aware Offsets

Leda Sarı[1], *Samuel Thomas*[2], *Mark Hasegawa-Johnson*[1], *Michael Picheny*[2]

[1]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign
[2]IBM Research AI

{lsari2,jhasegaw}@illinois.edu, {sthomas,picheny}@us.ibm.com

Abstract

Although there have been several speaker adaptation studies for deep neural network based acoustic models, they are mostly focusing on adaptation of fully connected layers rather than long short-term memory (LSTM) layers. In this work, we present an unsupervised LSTM layer normalization technique that we call adaptation by speaker aware offsets (ASAO). These offsets are learned using an auxiliary network attached to the main senone classifier. The auxiliary network takes main network LSTM activations as input and tries to reconstruct speaker, (speaker,phone) and (speaker,senone)-level averages of the activations by minimizing the mean-squared error. Once the auxiliary network is jointly trained with the main network, during test time we do not need additional information for the test data as the network will generate the offset itself. We investigate the effect of ASAO of LSTM layers at different depths. We also show its performance when the inputs are already speaker adapted by feature space maximum likelihood linear regression (fMLLR). In addition, we compare ASAO with a speaker adversarial training framework. ASAO achieves higher senone classification accuracy and lower WER than both the unadapted models and the adversarial model on the HUB4 dataset.

Index Terms: speaker adaptation, speech recognition, neural networks

1. Introduction

Although deep neural networks (DNNs) are successfully used in automatic speech recognition, their performance is still affected by the variability inherent in speech. One of the main sources of variability is the mismatch between speakers. Techniques proposed to alleviate this problem include using speaker-informed input features to the DNNs [1, 2], adapting the model structure [3, 4] and using auxiliary adaptation models or features [5, 6, 7, 8, 9, 10, 11]. From another perspective, adaptation methods can also be classified as supervised and unsupervised based on whether they use additional text or labels for the test data in addition to audio.

In input feature adaptation systems, features are normalized using a transform such as feature-space maximum likelihood linear regression (fMLLR) [12, 1] or the features are augmented with speaker specific features such i-vectors [13, 2]. Other methods modify the speaker independent DNN model by introducing speaker adaptive layers [14]. For example, [15] investigated the use of learning an affine transform after LSTM activations at different layers of the network. Alternatively, the structure is kept the same but the weights are adapted based on speakers [3]. Recently, auxiliary feature or auxiliary network based adaptation methods become more popular as these methods usually require little or no adaptation data [9]. One such approach is to extract speaker invariant intermediate features by adversarial training [8]. In these systems, the auxiliary network performs speaker classification whereas the main network performs phone/senone classification. Auxiliary feature based systems are usually based on sequence summary vectors [16]. These methods are mainly applied to fully connected layers. However, recently some methods are extended for the adaptation of the LSTM layers. For example, [9] applied the sequence summary idea into encoder-decoder based end-to-end systems.

This work is an extension of our previous work [11] on auxiliary network based speaker adaptation. The proposed system is based on an auxiliary network and it is an unsupervised speaker adaptation method for hidden layers of the neural network based acoustic models, The auxiliary network takes the hidden layer activation from the unadapted main senone classifier and tries to reconstruct speaker level mean of the activations at the output. It has been shown that mean normalization can improve the classification performance in different applications [17, 18]. Therefore, here we use a similar idea at the speaker level for speaker normalization. Although during training time, we can access the speaker labels and hidden layer activations, during test time we do not have this information. Therefore, we propose to use the auxiliary network to predict these means which would not require additional information for the test data other than the acoustic input. Instead of predicting the means directly, we aim at predicting the shift of the average activation from the global average. Thus we would like to extract the speaker specific component within the activation explicitly using the auxiliary network.

In the joint training of the main and auxiliary networks, the bottleneck features from the auxiliary network is projected back into the hidden activation space and these vectors are used as an offset for the main network activations and are subtracted from them. Since the auxiliary network targets are speaker specific, our hypothesis is that the auxiliary network will explicitly learn speaker specific part of the activations. Therefore, we call our method as adaptation with speaker aware offsets (ASAO).

The main difference from [11] is that here we show that the ASAO approach is also applicable to the LSTM based systems in addition to fully connected networks. Thus, we show that the proposed method is a flexible approach. We also perform our experiments on a different dataset and provide comparison with an adversarial training approach.

Miao et al. [6] also use an auxiliary network but they apply the offset only to the input features rather than the hidden layers of the main network. And they use i-vectors as input to the auxiliary network rather than hidden layer activations. In [8], an adversarial multitask objective is used to extract speaker-invariant deep features. Here, we also use a multitask objective but our aim is to learn speaker dependent information explicitly through our adaptation network which is then used as an offset to get speaker invariant features. Veselỳ et al. [16] use summary vectors computed over the utterance as additional input features to their main network. In our study, we use speaker or pho-

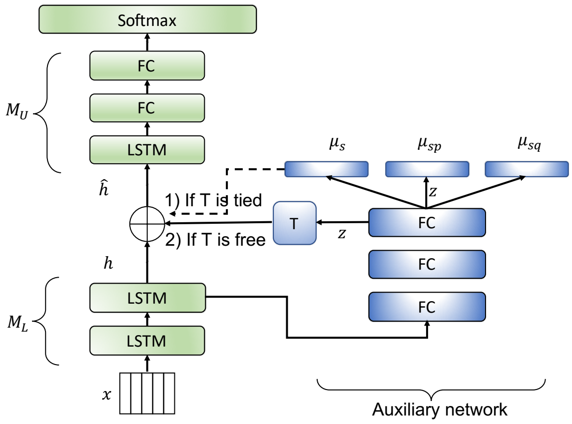


Figure 1: *Flowchart of the ASAO system with its auxiliary network*

netic level averages to summarize speaker and phonetic content but we use them as training targets for our auxiliary network instead of using them as input.

In Section 2 of the paper, we describe the method to extract speaker-aware offsets. We present the experimental setup and the results in Section 3. In Section 3.3, we compare our method to previous work and we conclude the paper in Section 4.

2. Adaptation with Speaker-Aware Offset

The proposed ASSO system consists of two major components. The main network performs senone classification and the auxiliary network tries to reconstruct speaker and phonetic level averages of the hidden layer activations of the main network which is to be adapted. As shown in Fig. 1, there is also another component T that is the bridge between the auxiliary and the main networks and it transforms auxiliary network output and generates the offset for the main network. In the following subsections, details of the main and auxiliary networks will be given.

2.1. Main Network

Let the input speech features of the network be denoted by $X = \{x_1, x_2, \dots, x_T\}$ where T denotes the duration and the corresponding senone labels as $Y = \{y_1, y_2, \dots, y_T\}$. First, we train a main senone classifier with inputs X and outputs Y. Then we choose an adaptation layer l and divide the network into lower (M_L) and upper (M_U) parts. Thus, the main network outputs are written as $M_U(M_L(x_t))$. The hidden layer activations h^l at layer l are then used as input to the auxiliary network which tries to learn the speaker dependent component of h^l. In this study, the main network consists of several LSTM layers followed by fully-connected layers. However, this does not restrict the choice of l because given an utterance of length T, we can compute the LSTM output for each time step t which is a vector that can be adapted as in adapting the fully connected layers.

2.2. Auxiliary Network

The aim of the auxiliary network is to extract speaker dependent information of the hidden layer activations h^l. Let the hidden layer activation at layer l at for input at time t is denoted by h_t. Also let s_t, p_t and q_t denote the speaker, phone and senone label of input x_t, respectively. Then the average activations are computed as in Eqs. 1-4. In these equations, $\mathbf{1}[.]$ is the indicator function that evaluates to 1 when its argument holds. The averages are computed over time instances which has the particular label or label pair. Similar to μ_s, we also define μ_p and μ_q.

$$\mu_g = \frac{1}{T}\sum_t h_t \tag{1}$$

$$\mu_s = \frac{1}{\sum_t \mathbf{1}[s_t = s]}\sum_t \mathbf{1}[s_t = s] h_t \tag{2}$$

$$\mu_{sp} = \frac{1}{\sum_t \mathbf{1}[s_t = s, p_t = p]}\sum_t \mathbf{1}[s_t = s, p_t = p] h_t \tag{3}$$

$$\mu_{sq} = \frac{1}{\sum_t \mathbf{1}[s_t = s, q_t = q]}\sum_t \mathbf{1}[s_t = s, q_t = q] h_t \tag{4}$$

In order to extract speaker specific component of h, we use the deviation of speaker dependent mean activations from the global averages. Given the above definitions, the three linear output layers of the auxiliary network can be written as $(\mu_s - \mu_g)$ that captures the deviation in the speaker mean from the global mean, $(\mu_{sp} - \mu_p)$ that captures speaker and phone level variation, and $(\mu_{sq} - \mu_q)$ that captures speaker and senone level variation. These three output layers share the auxiliary network parameters at the lower layers and differ only in their final layers. By parameter sharing and joint training of these outputs, we extract speaker dependent information in the network.

The hidden speaker dependent information for the t-th frame, z_t is computed from the last common hidden layer in the auxiliary network. Once z_t is computed, an affine transformation T is applied to the speaker dependent features z_t. We call these transformed features as speaker aware offsets and subtract them from the hidden layer activations of the main network as in Eq. (5). Thus, we aim at obtaining speaker independent component of h^l such that we achieve better senone classification accuracy in this speaker invariant space.

$$\widehat{h}_t^l = h_t^l - \mathrm{T}(z_t), \qquad n \in \{1, 2, \dots, T\} \tag{5}$$

If we augment our main network with this auxiliary network, the outputs of the main network are written as $M_U(\widehat{h}_t^l)$.

2.3. Training Procedure

Initially, the main senone classifier is trained using the cross entropy objective. After choosing the layer l, we compute the average activations for different units as shown in Eqs. 1-4, these give us the training targets for the auxiliary network. Finally, the auxiliary network and the transform T are attached to the main network and joint training is performed. The overall training objective $\mathcal{L}$ is a linear combination of the cross entropy loss from the main network and the total mean squared error (MSE) from the three output layers of the auxiliary network.

$$\mathcal{L} = \mathcal{L}_{\text{xent}}(\widehat{y}, y) + \mathcal{L}_{\text{MSE}}(s) + \mathcal{L}_{\text{MSE}}(sp) + \mathcal{L}_{\text{MSE}}(sq) \tag{6}$$

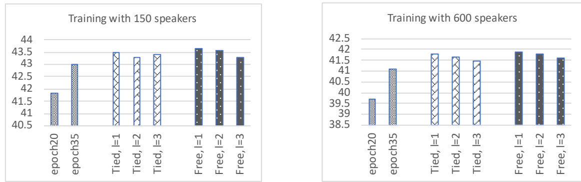
In the above equation, $\widehat{y}$ and y denote the estimated and true senone sequences and the MSE terms can be described using the auxiliary network outputs o_t as

$$\mathcal{L}_{\text{MSE}}(sp) = \frac{1}{T}\sum_t ||o_{sp,t} - (\mu_{sp,t} - \mu_{p,t})||_2^2 \tag{7}$$

where the subscripts denote the speaker and phone (or senone) label for frame at t.

Table 1: *Training and heldout data settings*

# Train Spk	Utt/Spk	# Heldout Spk
150	40	1241
600	10	1336



(a) *Training with 150 speakers* (b) *Training with 600 speakers*

Figure 2: *Senone classification accuracy of the main network on the heldout data before and after adaptation for logmel inputs*

3. Experiments

We performed our experiments on the Hub4 Broadcast News dataset [19, 20]. As the number of speakers in this dataset is large, using all the speakers in training lead to speaker independent model that conceals the effect of adaptation. Therefore, we performed training under two conditions shown in Table 1. We kept the total number of utterances the same but vary the number of speakers and the number of utterances per speaker is adjusted accordingly. In all settings, training and heldout sets are disjoint in terms of speakers. We did not perform larger experiments with higher number of speakers for two reasons: 1) Having about 2000 speakers will inherently lead to a speaker independent model that does not necessarily require adaptation. 2) In our framework, since we compute all (speaker,senone)-level means to get the training targets for the auxiliary network, it becomes computationally expensive to compute the averages for a large number of speakers.

3.1. Architectures

Input features are 40-dimensional log-mel features along with their deltas and delta-deltas, resulting in 120-dimensional input. The main network consists of three LSTM layers followed by two fully-connected layers with 256 and 512 nodes. LSTM layers are unidirectional and contain 128 cells. Number of output units is 2000 corresponding to each senone. The auxiliary network has three fully connected layers with 512, 256 and 128 units. Therefore, z has dimension of 128. When we adapt the LSTM layers, each of the three output layers have dimension 128 as the LSTMs have 128 cells. All fully-connected layers except the one generating z have rectified linear unit nonlinearity.

The networks are trained using PyTorch [21] with Adam optimization with learning rate 0.001. In the first stage, the main senone classifier is randomly initialized by Xavier method [22] and trained for 20 epochs. Then the hidden features h and their speaker, (speaker, phone) and (speaker,senone) level averages are computed. In joint network training, the main network is initialized from the senone classifier and the auxiliary network targets are taken to be the average activations. Joint training is performed for 15 epochs. Feature extraction, HMM training and decoding are performed using the Attila toolkit [23].

In joint training, the auxiliary network is utilized in two different ways. Either the T matrix shared its weights with the speaker-level output layer of the auxiliary network, therefore the output is directly used as the offset or T is kept to be a general affine transformation of the last hidden layer of the auxiliary network z. In both cases, the offset will carry speaker dependent information.

3.2. Results

Fig. 2 shows the frame-level senone classification accuracy on the heldout data for the networks trained with logmel features. We compare the main network trained for 20 epochs (epoch20), with the two adaptation methods (T is Tied or Free) for the adaptation of different LSTM layers individually (l=1,2,3). Since joint training is performed for 15 epochs, we also compare the performance of the 35th epoch (epoch35) of the main network training without any adaptation. We present the results under two training conditions (150 vs 600 training speakers).

In all adaptation experiments, we performed better than the initial unadapted model at epoch20 on the heldout set which resulted in 41.84% and 39.71% accuracy for 150 and 600-speaker conditions. For both 150 and 600 speaker conditions, the adapted models performed slightly better than the unadapted epoch35 model which achieved 43% and 41.07% on the heldout set in 150 and 600 speaker conditions, respectively. If we compare the performance of the adaptation of different LSTM layers, we saw that the improvements over the unadapted model decrease as we go deeper in the network (l gets larger). This might be due to the fact that as we get deeper and therefore closer to the output level, the hidden features become less speaker dependent and more phone dependent. Therefore, the difference activations h and $\hat{h}$ becomes smaller as we go deeper in the network and it reduces the effectiveness of adaptation. For the best performing layer (l=0), we achieved 43.64% and 41.9% in the 150 and 600-speaker conditions, respectively.

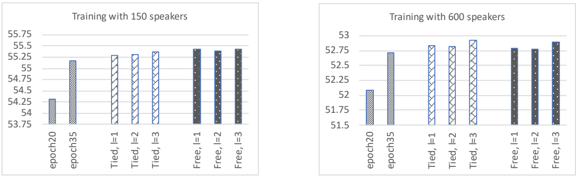
When we compare the performance of two architectures (T is Tied or Free), i.e. second and third group of bars in Figs. 2a and 2b, we see that with the Free setup, we got slight improvements over using Tied meaning that adaptation with $\mathrm{T}(z)$ is better. This also supports the observation made in [11].

If we compare the results of two training conditions (Fig. 2a vs. Fig. 2b), we see that in the 600-speaker case, our unadapted model has lower accuracy than 150-speaker condition possibly because of higher number of speakers in the validation set. In the 600-speaker case, we see that the relative improvement in senone accuracy with adaptation is larger than that of the 150-speaker condition.

3.3. Comparison

We repeated similar experiments described above for the case where the inputs are already speaker adapted by fMLLR [12]. To get the inputs, we computed 40-d fMLLR features and then concatenated the vectors within ± 5 context. The main and auxiliary network structures kept the same except the input layer which has 440-dimensional inputs rather than 120. The goal is to show if the proposed method is complementary to fMLLR based input adaptation by comparing the performances of the adapted and unadapted fMLLR models.

Most of the conclusions from logmel features also hold for this case. As shown in Figs. 3a-3b, adapted models are always better than epoch20 model. As compared to epoch35, we still observe increase in the adapted models although the gains are



(a) *Training with 150 speakers* (b) *Training with 600 speakers*

Figure 3: *Senone classification accuracy of the main network on the heldout data before and after adaptation for fMLLR input*

Table 2: *Senone classification accuracy comparison of unadapted model with adversarial adaptation of [8] and ASAO*

	150	600
Epoch20	41.84	39.71
Epoch35	43.00	41.07
Adversarial	42.24	40.50
ASAO	**43.39**	**41.48**

much smaller. For the fMLLR case, there is in general not a pattern among adapting different LSTM layers, in these experiments adapting $l = 2$ or 3 can perform better than adapting $l = 1$. This might be due to the fact that now input features are already speaker normalized so $l = 1$ is more speaker independent as compared to the logmel experiments. Since the adaptation in these systems still lead to improvement, it can be concluded that the fMLLR based adaptation and the proposed ASAO methods are complementary.

As mentioned in Section 1, there are studies that make use of an auxiliary network in different ways. Among those, the one that is closest to ours is the adversarial training method of [8] which tries to obtain speaker independent hidden layer activations using multitasking. In this system, the auxiliary network takes the hidden layer activation (similar to our h) as input but the outputs of the auxiliary network are speaker labels rather than speaker dependent means. They also do not utilize an additional projection matrix T. Their goal is to maximize the senone classification accuracy of the main network while minimizing the speaker classification accuracy by gradient reversal.

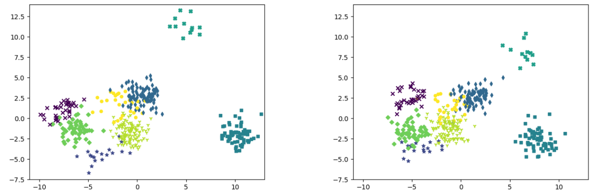
We use the pretrained senone classifier with logmel inputs and then attach the speaker classifying auxiliary network and perform joint training. Different from [8] which adapt only a fully connected network, we tried the same approach on LSTM layers. In the logmel based system, we adapted $l = 2$ for which our system has disadvantage. Table 2 compares the senone classification performances of the unadapted model, adversarial adaptation and ASAO methods. It is shown that ASAO leads to the highest senone classification accuracy.

According to Table 2, the adversarial training method improves the accuracy over the unadapted method at epoch20. Although the joint training is performed for 15 iterations, it cannot reach the performance of the epoch35 model of the unadapted main network. Therefore, we can conclude that our ASAO method is more effective for speaker normalization of the hidden layer activations of the main network.

In Table 3, we present the word error rates (WER) on the heldout datasets for the unadapted model, adversarial training and best senone classifiers from the ASAO method. For the logmel features, ASAO method achieves 9.2% and 6.7% relatively lower WER than the unadapted model in 150 and 600-speaker conditions, respectively. It also performs better than the adversarial training method.

Table 3: *WER for 150 and 600 speaker training conditions under different models*

	150		600
Unadapted	21.8	Unadapted	26.9
Adversarial [8]	21.1	Adversarial	26.0
ASAO, Free, l=1	19.8	ASAO, Free, l=1	25.1



(a) h (b) $\hat{h}$

Figure 4: *2D LDA projection of h and $\hat{h}$ for the logmel condition with ASAO, Free, l=1 model for a phone; colors represent different speakers*

We also visualize the unnormalized (h) and normalized $\hat{h}$ hidden activations. We randomly selected 10 heldout speakers and for each phone, we plotted 2-dimensional linear discriminant analysis (LDA) projection of the activations. For the model with the largest improvement, i.e. 150-speaker logmel condition with Free, l=1 adaptation, we get Fig. 4 for a phone. In the figure, each color represent a different speaker. In the speaker normalized space, we expect that the activations related to a certain phone to be closer. As shown in the figure, after normalization, $\hat{h}$ from different speakers get closer and harder to separate. For example, the one on the top right or the bottom right get closer to the larger cluster on the lower left.

4. Conclusions

In this work, we presented a neural network layer speaker adaptation scheme using an auxiliary network. This auxiliary network which is trained to reconstruct speaker, (speaker-phone), and (speaker-senone)-level averages, generated a speaker aware offset that is subtracted from the main network activations. The main advantage of the auxiliary network is that once it is trained, we do not need additional data for test speakers as the auxiliary network will automatically generate the speaker-aware offsets. We showed that the proposed model can be used to adapt LSTM layers in addition to fully-connected layers which was shown earlier in [11]. Experimental results showed that if the input features are speaker independent logmel features, adapting lower layers of the network is more helpful and also using a free projection (T matrix), we perform better than tying it to the speaker dependent output layer of the auxiliary network. We also showed that ASAO can slightly improve the senone classification accuracy when the inputs to the main network are speaker adapted fMLLR features showing that the two methods can be complementary. We also compared our multitask learning with the adversarial training method of [8] and showed that ASAO is more effective in speaker normalization.

5. References

[1] S. P. Rath, D. Povey, K. Veselỳ, and J. Cernockỳ, "Improved feature processing for deep neural networks." in *Interspeech*, 2013, pp. 109–113.

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

[3] P. Swietojanski and S. 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.

[4] 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 *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 6359–6363.

[5] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7942–7946.

[6] Y. Miao, H. Zhang, and F. Metze, "Speaker adaptive training of deep neural network acoustic models using i-vectors," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 11, pp. 1938–1949, 2015.

[7] S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q. Liu, "Fast adaptation of deep neural network based on discriminant codes for speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1713–1725, 2014.

[8] Z. Meng, J. Li, Z. Chen, Y. Zhao, V. Mazalov, Y. Gong *et al.*, "Speaker-invariant training via adversarial learning," *arXiv preprint arXiv:1804.00732*, 2018.

[9] M. Delcroix, S. Watanabe, A. Ogawa, S. Karita, and T. Nakatani, "Auxiliary feature based adaptation of end-to-end asr systems," *Proc. Interspeech 2018*, pp. 2444–2448, 2018.

[10] X. Cui, V. Goel, and G. Saon, "Embedding-based speaker adaptive training of deep neural networks," *Proc. Interspeech 2017*, pp. 122–126, 2017.

[11] L. Sarı and M. Hasegawa-Johnson, "Speaker adaptation with an auxiliary network," in *Machine Learning in Speech and Language Processing Workshop (MLSLP)*, 2018.

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

[13] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[14] F. Seide, G. Li, X. Chen, and D. 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] M. Kitza, R. Schlüter, and H. Ney, "Comparison of blstm-layer-specific affine transformations for speaker adaptation," *Proc. Interspeech 2018*, pp. 877–881, 2018.

[16] K. Veselỳ, S. Watanabe, K. Žmolíková, M. Karafiát, L. Burget, and J. H. Černockỳ, "Sequence summarizing neural network for speaker adaptation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5315–5319.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[18] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[19] "1996 English Broadcast News Speech (HUB4)," https://catalog.ldc.upenn.edu/LDC97S44.

[20] "1997 English Broadcast News Speech (HUB4)," https://catalog.ldc.upenn.edu/LDC98S71.

[21] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[22] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[23] H. Soltau, G. Saon, and B. Kingsbury, "The ibm attila speech recognition toolkit," in *2010 IEEE Spoken Language Technology Workshop*. IEEE, 2010, pp. 97–102.