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

Semi-Supervised Training in Deep Learning Acoustic Model

Yan Huang, Yongqiang Wang, and Yifan Gong

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052
{yanhuang; erw; ygong}@microsoft.com

Abstract

We studied semi-supervised training in a fully connected deep neural network (DNN), unfolded recurrent neural network (RNN), and long short-term memory recurrent neural network (LSTM-RNN) with respect to transcription quality, importance data sampling, and training data amount. We found that DNN, unfolded RNN, and LSTM-RNN exhibit increased sensitivity to labeling errors. One point relative WER increase in the training transcription translates to *a half point* WER increase in DNN and slightly more in unfolded RNN; while in LSTM-RNN it translates to *one full point* WER increase. LSTM-RNN is notably more sensitive to transcription errorss. We further found that the importance sampling has similar impact on all three models. In supervised training, importance sampling yields 2$\sim$3% relative WER reduction against random sampling. The gain is reduced in semi-supervised training. Lastly, we compared the model capacity with increased training data. Experimental results suggest that LSTM-RNN can benefit more from enlarged training data comparing to unfolded RNN and DNN.

We trained a semi-supervised LSTM-RNN using 2600 hours of transcribed and 10000 hours of untranscribed data on a mobile speech task. The semi-supervised LSTM-RNN yields 6.56% relative WER reduction against the supervised baseline trained from 2600 hours of transcribed speech.

Index Terms: semi-supervised learning, DNN, unfolded RNN, LSTM-RNN, importance data sampling

1. Introduction

Semi-supervised learning, as a classical machine learning problem, has been researched extensively in both theoretical [1, 2, 3, 4] and applied machine learning communities [5, 6, 7, 8, 9]. The motivation behind is simple: human labelled data is expensive and time consuming to obtain. In a speech service system, frequent model update with fresh data has been found to be important to achieve best production accuracy performance. Therefore, semi-supervised training is an ideal and economic acoustic model development strategy. This is especially true with the emerging new types of deep learning acoustic model with enlarged model capacity.

Self-training based semi-supervised training explicitly infers transcription for untranscribed data and use them for model training. It is widely adopted in large-scale semi-supervised acoustic model training [15, 16] due to its simplicity and scalability. We will primarily focus on this approach in this study.

In the past, there were good sources of semi-supervised training research in the Gaussian mixture hidden Markov model [10, 11, 12, 13] and in the fully connected deep neural network hidden Markov model [14, 15, 16]. In this paper, we answer the question with the emerging new types of deep learning acoustic model - what are the new challenges and what are the key strategies to address these problems.

Specifically, we studied three distinct factors of the semi-supervised training: the transcription quality, the importance data sampling, and the training data amount, in a fully connected deep neural network (DNN) [18], unfolded recurrent neural network (RNN) [19], and long short-term memory recurrent neural network (LSTM-RNN) [21].

We found that DNN, unfolded RNN, and LSTM-RNN exhibits increased sensitivity to labeling errors. One point WER increase in the training transcription translates to *a half point* WER increase in DNN; while in LSTM-RNN it translates to *one full point* WER increase. LSTM-RNN is significantly more sensitive to transcription errorss. For example, with the simulated erroneous training transcription at 5%, 10%, or 15% WER level, the semi-supervised DNN yields 2.37%, 4.84%, or 7.46% relative WER increase comparing to the baseline model trained with human transcription; in contrast, the corresponding WER increase is 2.53%, 4.89%, or 8.85% in an unfolded RNN and 4.47%, 9.38%, or 14.01% in an LSTM-RNN. Therefore, generating high quality inferred transcription or developing alternative LSTM neurons which is less sensitive to labeling errors are the keys to the success of semi-supervised LSTM training.

We further found that DNN, unfolded RNN, and LSTM-RNN can all benefit from importance data sampling similarly. In the supervised training setup, importance sampling yields 2%$\sim$3% relative WER reduction comparing to random sampling. The gain was reduced in semi-supervised training setup. Lastly, we compared the model capacity with increased amount of training data. LSTM-RNN can benefit most from enlarged training data among the three in the supervised setup. Nevertheless, in semi-supervised setting, the gain is significantly reduced due to its sensitivity to transcription errors in LSTM-RNN. We conducted a semi-supervised LSTM-RNN training using 2600 hours of transcribed and 10000 hours of untranscribed data on a mobile speech task. The semi-supervised LSTM-RNN yields 6.56% average relative WER reduction against the supervised baseline trained from 2600 hours of transcribed speech.

The remainder of this paper is organized as follows: Section 2 discusses the transcription quality factor in semi-supervised DNN, unfolded RNN, and LSTM-RNN; Section 3 discusses the data sampling factor; Section 4 discusses the training data amount factor; Section 5 concludes this study.

2. Transcription Quality

In this section, we study how transcription quality affects deep learning acoustic model in DNN, unfolded RNN, and LSTM.

2.1. Model Formulation

The three deep learning based acoustic models studied in this paper share the similar stacked layer-wise deep structure with the only differences in whether a recurrent network path exists and the specific type of neuron used.

http://dx.doi.org/10.21437/Interspeech.2016-1596

A DNN [13] is a fully connected feed-forward neural network. The input signal x_t is forward-propagated through the hidden layers (W_l, b_l) until it reaches the last layer (L), where the sigmoid non-linearity (σ) is replaced by the softmax (ϕ):

$$\begin{cases} h_0 = \mathbf{x}_t \\ h_l = \sigma(W_l h_{l-1} + b_l) & 1 \leq l \leq L \\ y_t = \phi(W_L h_{L-1} + b_L) & l = L \end{cases} \quad (1)$$

An RNN [20] uses both the current frame (x_t) and the previous frames encoded as a history vector (h_{t-1}) to predict the output (y_t):

$$\begin{cases} h_t = \sigma(W_{xh} x_t + W_{hh} h_{t-1}) \\ y_t = \phi(W_{hy} h_{t-1}) \end{cases} \quad (2)$$

The unfolded RNN [19] is a feed-forward network via unrolling an RNN with certain time steps. It can be thought of either as a feed-forward neural network with special temporal network parameter tying or as a truncated simplified RNN.

LSTM-RNN is a special type of recurrent neural network with specially designed memory cell. A set differentiable gates, namely the input gate (i_t), forgetting gate (f_t), output gate (o_t), and control gate (c_t), are used to determine what to store and when to read and write. The first three gates are parameterized by a set of weight matrix ($W_{\cdot x}, W_{\cdot m}, W_{\cdot c}$) connecting with the input (x), recurrent cell activation (m), control gate (c) respectively together with the bias. The control gate (c_t) is determined by the previous state of itself, the forgetting gate, and the input gate. We adopted a similar LSTM-RNN structure as in [21]:

$$\begin{cases} i_t = \sigma(W_{ix} x_t + W_{im} m_{t-1} + W_{ic} c_{t-1} + b_i) \\ f_t = \sigma(W_{fx} x_t + W_{fm} m_{t-1} + W_{fc} c_{t-1} + b_f) \\ c_t = f_t \odot c_{t-1} + i_t \odot \sigma(W_{cx} x_{t-1} + W_{cm} m_{t-1} + b_c) \\ o_t = \sigma(W_{ox} x_t + W_{om} m_{t-1} + W_{oc} c_t + b_o) \\ m_t = o_t \odot h(c_t) \\ y_t = \phi(Wymm_t + b_y) \end{cases} \quad (3)$$

The error back-propagation is used for optimization. In the cross-entropy objective, a frame-level error signal is calculated and the gradient is back-propagated through the network for optimization. When a transcription error happens, an incorrect gradient will be generated and back-propagated in optimization.

2.2. Simulation Experiment

To empirically study the impact of the transcription quality, we conducted a simulation experiment in semi-supervised DNN, unfolded RNN, and LSTM-RNN on a mobile speech task.

We first use a recognizer to decode 400 hours of mobile speech training data and generate erroneous transcription. The recognizer was intentionally configured at an accuracy mode which allows us to simulate erroneous transcription at a range of quality levels. We randomly select a subset of the training data which uses machine transcription. We then mixed it with the rest of data with human transcription. By adjusting the mixing rate of machine transcription and human transcription, we can effectively simulate erroneous training transcription at a desired quality level with realistic error patterns in typical machine inferred transcription. We thus obtain four versions of simulated transcription at 2%, 5%, 10%, and 15% WER level for the 400 hours of training data. The subset to be used with machine transcription were randomly selected to ensure no sampling bias between different versions of simulated data sets.

Table 1: Specification of the DNN, unfolded RNN, and LSTM-RNN models and the supervised baseline accuracy. WERR refers to relative WER reduction.

Model	DNN	unfolded RNN	LSTM-RNN
Front-end	LFB	LFB	LFB
# of Senones	5980	5980	5980
# of Hidden Layers	5	4	4
# of Parameters	30M	5M	20M
WER	19.4%	18.2%	17.1%
WERR	NA	6.3%	12.2%

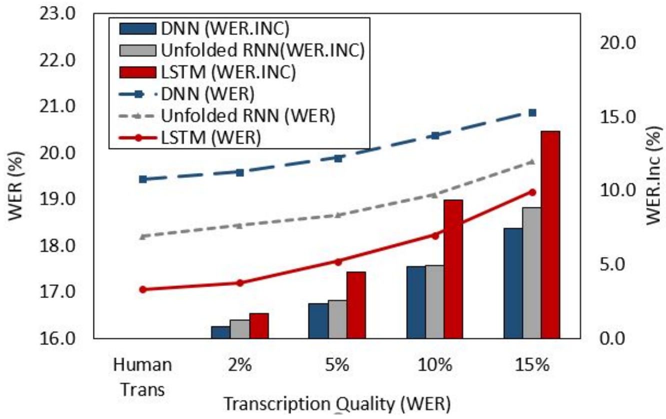

Figure 1: Performance comparison of the semi-supervised DNN, unfolded RNN, and LSTM-RNN. The models were trained using the same 400 hours of training data with different transcription quality measured by the transcription WER. WERR and WER.INC refer to relative WER reduction and relative WER increase respectively.

We train the semi-supervised DNN, unfolded RNN, and LSTM-RNN using the same 400 hours of mobile speech data with simulated transcription at different quality level separately. We also trained a supervised baseline model with the same data as a reference. All models were trained using the cross entropy (CE) criteria. The front-end is the 87-dimension log filter bank (LFB) feature with a context window of 11 frames. All models share the same senone states, state alignment model, and unsupervised RBM pre-training. The model specification and the supervised baseline model accuracy are summarized in Table 1. The models were evaluated on a mobile speech test set with 5 hours of speech. All models throughout this paper were trained using the CNTK toolkit [22].

Figure 1 presents the semi-supervised DNN, unfolded RNN, and LSTM-RNN accuracy performance at different transcription quality in comparison with the supervised baseline:

- In DNN, one point WER increase in the training transcription translates to *a half point* WER increase in the resulting model accuracy performance. The semi-supervised training generates 0.77%, 2.37%, 4.84%, or 7.46% relative WER increase (WER.INC) for the simulated transcription at 2%, 5%, 10%, or 15% WER level.
- In LSTM-RNN, one point WER increase in the training transcription roughly translates into *one full point* WER increase in the resulting LSTM-RNN. We observed 1.69%, 4.47%, 9.38%, or 14.01%, nearly doubled relative WER increase, comparing to the semi-supervised DNN, at the same simulated transcription

Table 2: Model accuracy performance comparison of the importance sampling versus the random sampling in semi-supervised DNN and LSTM. WERR is relative WER reduction.

SEMI-SUP	Baseline	Random$_{+400hrs}$	Import.$_{+400hrs}$
DNN(WER)	19.43	17.63	17.31
DNN(WERR)	NA	9.26	10.91
LSTM(WER)	17.06	16.06	15.86
LSTM(WERR)	NA	5.86	7.03
SUP	Baseline	Random$_{+400hrs}$	Import.$_{+400hrs}$
DNN(WER)	19.43	17.56	17.01
DNN(WERR)	NA	9.62	12.45
LSTM(WER)	17.06	15.17	14.66
LSTM(WERR)	NA	11.08	14.07

quality level. LSTM-RNN is significantly more sensitive to transcription errors.

- The unfolded RNN is slightly more sensitive to transcription errors comparing to the DNN. We observed 1.26%, 2.53%, 4.89%, or 8.85% relative WER increase accordingly.

It is to be noted that transcription simulation via mixing human transcription with machine transcription used in this paper may result in slightly different model training error pattern as comparing to real machine generated transcription.

2.3. Discussion

The simulation experiments reveal a distinct fact that LSTM-RNN is significantly more sensitive to transcription errorss comparing to DNN and unfolded RNN.

One simple strategy is to weight the contribution of the error signal based on the frame-level confidence score. This allows the poorly transcribed frames to be "blackened out" and excluded from the gradient estimation.

Given the fact that unfolded RNN is only moderately more sensitive to transcription errors comparing to DNN, we believe that the memory cell in the LSTM-RNN is the root cause. We can parametrize the control gate and the forgetting gate as a function of the frame-level confidence to relieve the adverse impact of poorly transcribed frames.

Generating high quality derived transcription and developing alternative LSTM neurons less sensitive to labeling errors are the key to high quality semi-supervised LSTM-RNN.

3. Importance Data Sampling

Data are not equally valuable, which has been an important observation in our practice in semi-supervised acoustic model training [16]. Besides the transcription quality, data sampling difference is another fundamental difference between the machine supervised/selected data and human transcribed data.

In this section, we study how importance data sampling affects DNN and LSTM, both in the supervised and semi-supervised training setting. We adopted this simple importance data sampling based on confidence as suggested in [16] in this study. Starting with the same baseline models as in Section 2.2, we trained the DNN and LSTM-RNN with additional 400 hours of mobile data via random sampling or the importance sampling. The transcription quality of the machine supervised/selected data is at 5% WER level.

Table 2 summarizes the accuracy performance comparison of the importance sampling versus the random sampling in semi-supervised DNN and LSTM:

- In supervised setup, adding 400 hours of machine supervised data via random or importance sampling yield 12.45% or 9.62% relative WER reduction against the baseline trained from 400 hours of transcribe data. In LSTM-RNN, the corresponding relative WER reduction is 11.08% or 14.07%. We observe around 3% additional WER reduction with importance sampling both in DNN and LSTM-RNN.
- In semi-supervised setup, adding 400 hours of human transcribed data via random or importance sampling yield 9.26% or 10.91% relative WER reduction. In LSTM-RNN, the corresponding relative WER reduction is 7.03% or 5.86%. The benefit of importance sampling drops to 1$\sim$2%.

Both DNN and LSTM-RNN can benefit similarly from importance data sampling in the supervised and semi-supervised setup with small but consistent gain. We didn't observe distinct systematic differences between these two models in this regard.

The gain from the importance sampling is smaller in the semi-supervised setup. Here the value of the data itself and the quality of the inferred transcription jointly determine how much it can benefit from the importance sampling. The gain is reduced due to the fact that the more valuable data are usually "harder" to recognize and typically with lower accuracy.

Overall, data sampling can yield additional moderate but consistent gain in semi-supervised LSTM. We think that the improved transcription quality and an effective strategy to reduce the model sensitivity to transcription error can help maximize the benefit from the importance sampling.

4. Training Data Amount

In this section, we study how increased training data affects the supervised and semi-supervised neural network acoustic model.

4.1. Simulation Experiments

We adopted the similar set of DNN and LSTM-RNN models as described in Section 2 in this study. The baseline models are the supervised baseline DNN and LSTM-RNN models trained from 400 hours of mobile speech training data. We added 400, 800, or 1200 hours of mobile speech with human transcription in the supervised training or with machine transcription in the semi-supervised training. No importance data sampling was applied.

The transcription quality of the machine supervised/selected data is at around 5% WER level. Note that the average quality of the semi-supervised training data degrades as more machine supervised data was mixed with the fixed amount of baseline transcribed training data.

Figure 2 presents supervised and semi-supervised DNN and LSTM-RNN model results.

- In supervised training setup, LSTM-RNN can benefit more from enlarged training data comparing to DNN. For example, the LSTM-RNN yields 11.08%, 13.89%, and 16.60% relative WER reduction against the baseline with additional 400, 800, and 1200 hours of training data. In comparison, the corresponding WER reduction in DNN is 9.62%, 12.40%, or 14.50%.
- In semi-supervised training setup, DNN continues to benefit from enlarged training data with around 1$\sim$3% gap in WER reduction comparing to the supervised counterpart. For example, the corresponding semi-supervised DNN yields 9.26%, 10.91%, and 12.97% at

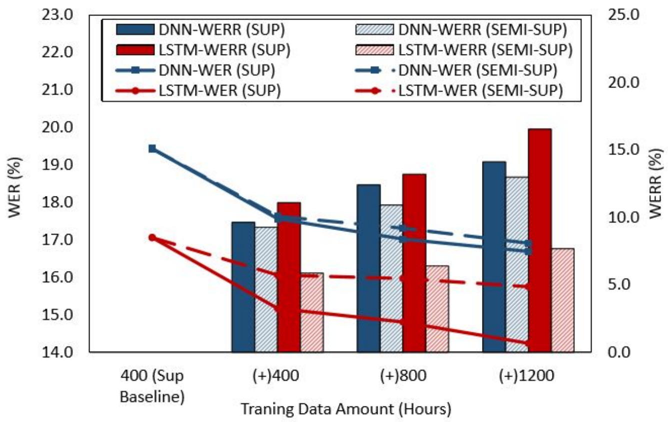

Figure 2: Performance of the supervised and semi-supervised DNN and LSTM-RNN with increased training data. WERR is the relative WER reduction.

the same training data points. The gap between the supervised and semi-supervised training exhibits slightly increased trend due to the lower average transcription quality with more machine transcription mixed.

- In the semi-supervised LSTM-RNN, we observe 5.86%, 6.91%, and 7.69% relative WER reduction, roughly only half of the gain comparing to the supervised LSTM-RNN counterpart. The gap between the supervised and semi-supervised training also exhibits a more dramatic increased trend as the average training transcription quality drops with more machine derived transcription mixed. The root cause here is the sensitivity to transcription errors in LSTM-RNN, which is consistent with our previous study on the transcription quality sensitivity study.

LSTM-RNN has larger modeling capacity comparing to DNN and can potentially benefit from large amount of training data. Nevertheless, in the semi-supervised training setup, the performance gain can be largely reduced due to its high sensitivity to transcription error. It is to be noted that we did experiment with increased model size and observed similar results.

4.2. Large Scale Semi-Supervised LSTM-RNN

We conducted an initial experiment on a large scale semi-supervised LSTM-RNN training on the mobile speech task. We automatically supervised and selected 10000 hours of untranscribed data from our production traffic using a multi-view learning approach similar to [16] and the simple confidence-based importance sampling. The selected machine supervised transcription is at around 4% WER level.

The baseline LSTM was trained from 2600 hours of transcribed data. The semi-supervised LSTM-RNN was trained on 12600 hours of data in total. The LSTM-RNN has similar model structure as in Table 1, except with a larger senone set (9404 senone states).

In order to complete this large scale semi-supervised LSTM training in a reasonable turnaround time, we used the BMUF algorithm [23] implemented in CNTK [22]. Our training job is running on a high-performance GPU cluster, in which each node is equipped with 4 Nivdia Telsa K40 GPUs. All the computing nodes in the GPU cluster are inter-connected via InfiniBand (IB); all our training data are stored on a shared Hadoop distributed file system (HDFS). 16 K40 GPUs were used to train

Table 3: Accuracy performance of the 12600 hours of semi-supervised LSTM-RNN and the supervised baseline trained from 2600 hours of transcribed data. WERR is the relative WER reduction.

Test Sets	Sup LSTM	Semi-sup LSTM	WERR
Test A	14.48	13.62	5.94
Test B	14.17	13.16	7.13
Average	14.33	13.39	6.56

the LSTM-RNN models. 9 full sweepings were performed on all the data followed by additional 3 sweepings of 2600 hours of transcribed data. The described configuration allowed us to finish the training in 5 days.

Two test sets collected during different period of time from production traffic were used to evaluate the models. Test A consists of 25 hours of speech, which was collected around two years earlier than the time period when the untranscribed data were harvested; Test B consists of 17 hours of speech, which was collected about half year later than the time period when the untranscribed data were harvested. The untranscribed training data are strictly separated from the testing data.

Table 3 presents the accuracy performance of the large scale semi-supervised LSTM training. On Test A, the WER drops from 14.48% to 13.62% or 5.94% relative WER reduction comparing to the supervised baseline. On test B, the WER drops from 14.17% to 13.16% or 7.13% relative WER comparing to the supervised baseline.

5. Conclusion

In conclusion, we studied the transcription quality, the importance data sampling, and the training data amount, in a fully connected deep neural network (DNN), unfolded recurrent neural network (RNN), and long short-term memory recurrent neural network (LSTM-RNN).

We found that LSTM-RNN exhibits high sensitivity to transcription errors. One point WER increase in the training transcription translates to *one full point* WER increase in LSTM-RNN, comparing to *a half point* WER increase in DNN. All three models benefit from importance data sampling with similar 2∼3% relative WER reduction comparing to the random sampling. Regarding the training data amount, LSTM-RNN can benefit more from enlarged data comparing to unfolded RNN and DNN in the supervised setup. In the semi-supervised setup, the gain from enlarged training data in the LSTM-RNN shrinks significantly due to its sensitivity to transcription errors. Therefore, we conclude generating high quality transcription and effectively suppressing effect of erroneous transcription is the key to the success of high quality large scale semi-supervised LSTM-RNN acoustic model training. The importance data sampling can yield consistent moderate accuracy gain.

We conducted a semi-supervised LSTM-RNN training with 2600 hours of transcribed and 10000 hours of untranscribed data on a mobile task. The semi-supervised LSTM-RNN yields 6.56% relative WER reduction against the supervised baseline.

Ongoing work includes the transcription error robust semi-supervised LSTM-RNN and its sequence training.

6. Acknowledgements

The authors would like to thank Dr. Jinyu Li for the help in the unfolded RNN model setup and Project Philly team for infrastructure support.

7. References

[1] Chapelle, O. , Scholkopf, B. , and Zien, A. , “Semi-Supervised Learning,” MIT Press, 2006.

[2] Zhu, X. , “Semi-supervised Learning Literature Survey,” Technical Report, Computer Science, University of Wisconsin-Madison, 2005.

[3] Basu, S. , “Semi-supervised Clustering: Probabilistic Models, Algorithms and Experiments,” Doctoral Thesis, University of Texas, Austin, 2005.

[4] Kingma, D. P. , Rezende, D. J. , Mohamed, S. , and Welling, M. , “Semi-Supervised Learning with Deep Generative Models,” In Advances in Neural Information Processing Systems (NIPS), 2014.

[5] Joachims, T. , “Transductive Inference for Text Classification Using Support Vector Machines,” In Proceeding of the International Conference on Machine Learning (ICML), 1999.

[6] Nigam, K. , McCallum, A. , Thrun, S. , and Mitchell, T. , “Text Classification From Labeled and Unlabeled Documents Using EM,” Machine Learning, 39(2/3):103134, 2000.

[7] Liang, P. , “Semi-supervised Learning for Natural Language Processing,” PhD Thesis, Massachusetts Institute of Technology, 2005.

[8] Fergus, R. , Weiss, Y. , and Torralba, A. , “Semi-Supervised Learning in Gigantic Image Collections,” In Advances in Neural Information Processing Systems (NIPS), 2009.

[9] Guillaumin, M. , Verbeek, J. , Schmid, C. , “Multimodal Semi-Supervised Learning for Image Classification,” In Proceeding of CVPR 2010.

[10] Lamel, L. , Gauvain, J. L. , and Adda, G. , “Lightly Supervised and Unsupervised Acoustic Model Training,” Computer Speech and Language 2002.

[11] Wessel, F. and Ney, H. , “Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition,” IEEE Transactions on Speech and Audio Processing, 2005.

[12] Ma, J. and Schwartz, R. M. ,“Unsupervised versus Supervised Training of Acoustic Models,” in Proceeding of Interspeech 2008.

[13] Yu, K. , Gales, M. , Wang, L. , and Woodland, P. C. , “Unsupervised Training and Directed Manual Transcription for LVCSR, Speech Communication, 2010.

[14] Manohar, V. M. , Povey, D. , and Khudanpur, S. , “Semi-Supervised Maximum Mutual Information Training of Deep Neural Network Acoustic Models,” In Proceeding of Interspeech 2015.

[15] Liao, H. , McDermott, E. , and Senior A. , “Large Scale Deep Neural Network Acoustic Modeling with Semi-Supervised Training Data for YouTube Video Transcription,” in Proceeding of IEEE ASRU 2013.

[16] Huang, Y. , Yu, D. , Gong, Y. , and Liu, C. , “Semi-Supervised GMM and DNN Acoustic Model Training with Multi-System Combination and Condence Re-calibration”, in Proceeding of Interspeech 2013.

[18] 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 - Special Issue on Deep Learning for Speech and Language Processing, 2012.

[19] Saon, G. , Soltau, H. , Emami, A. , and Picheny, M. , “Unfolded Recurrent Neural Networks for Speech Recognition,” In the Proceedings of Interspeech 2014.

[20] Robinson, A., J., “An Application of Recurrent Nets to Phone Probability Estimation, Neural Networks, IEEE Transactions on, vol. 5, no. 2, pp. 298305, 1994.

[21] Sak, H. , Senior, A. , and Beaufays, F. , “Long Short-term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition,” in Proceeding of Interspeech 2014.

[22] Agarwal, A. , et al. , “An Introduction to Computational Networks and the Computational Network Toolkit”, Microsoft Technical Report MSR-TR-2014-112, 2014.

[23] Chen, K. and Huo, Q. , “Scalable Training of Deep Learning Machines by Incremental Block Training with Training Block Parallel Optimization and Blockwise Model-Update Filtering,” In Proceeding of ICASSP 2016.