Interspeech 2018
2-6 September 2018, Hyderabad

Densely Connected Networks for Conversational Speech Recognition

Kyu J. Han, Akshay Chandrashekaran, Jungsuk Kim, Ian Lane

Capio Inc., Belmont, CA, USA

{kyu,akshay,jungsuk,ian}@capio.ai

Abstract

In this paper we show how we have achieved the state-of-the-art performance on the industry-standard NIST 2000 Hub5 English evaluation set. We propose densely connected LSTMs (namely, dense LSTMs), inspired by the densely connected convolutional neural networks recently introduced for image classification tasks. It is shown that the proposed dense LSTMs would provide more reliable performance as compared to the conventional, residual LSTMs as more LSTM layers are stacked in neural networks. With RNN-LM rescoring and lattice combination on the 5 systems (including 2 dense LSTM based systems) trained across three different phone sets, Capio's conversational speech recognition system has obtained 5.0% and 9.1% on Switchboard and CallHome, respectively.

Index Terms: Densely connected LSTM, Switchboard, conversational speech recognition

1. Introduction

We have recently observed a series of leap-frog advancements in deep learning based acoustic and language modeling for conversational speech recognition. With the contributions mainly from convolutional neural networks (CNNs) and recurrent neural networks (RNNs), multiple research groups have continued to improve their system performance on the well-known, industry-standard NIST 2000 Hub5 English evaluation set[1] [1, 2, 3, 4, 5], approaching to the hypothesized human performance of the evaluation set. Achieving human parity has now become the topic of the speech recognition community, which nurtured interesting research works of contrasting transcriptions from human transcribers and conversational speech recognition systems [1, 2, 6]. It is reported in [6] that similar error patterns were found between human and machine transcriptions, hinting that the quality of machine transcriptions becomes closer to that of human transcribers. Conversational speech recognition, however, is still challenging, and we in [3] performed a comparative analysis on how vulnerable even the state-of-the-art conversational speech recognition system would be against real-world telephone conversations in the wild.

In this paper we propose a new neural network based acoustic model structure with dense connections between long short-term memory (LSTM) layers. Densely connected neural networks were originally introduced to avoid layer-wise vanishing gradient problems when CNNs are stacked in a very deep fashion, e.g., more than 100-layers, for image recognition tasks [7]. One can view dense connection as a variant from residual learning [8] or highway networks [9, 10]. In speech recognition, residual or highway connections have been applied to LSTMs, only between adjacent layers [11, 12, 13, 14]. Our dense LSTMs connect (almost) every layer to one another to further mitigate vanishing gradient effect between LSTM layers and help error signals propagated even back to the very first layer during back propagation in training. Benefiting from the proposed dense LSTMs, we were able to reach the marks of 5.0% and 9.1% in word error rate (WER) for the Switchboard and CallHome test sets, respectively, both of which are the best results reported thus far in the field.

This paper is organized as follows. Section 2 describes the proposed, densely connected LSTMs, accompanying empirical analysis on residual and dense LSTMs. Section 3, we detail the other components constituting Capio's conversational speech recognition system, such as language models and system combination. After presenting experimental results in a broader scale across individual systems in Section 4, we conclude this paper in Section 5 with the remarks on the contributions and future directions.

2. Densely Connected LSTM

Dense connection [7] was introduced for CNNs to yield the state-of-the-art performance on the CIFAR-10/100 data sets [15] for image classification, outperforming residual networks [8, 16] which had been the best performing neural network architecture in the domain. Like skip connections in residual learning, dense connections let error signals further back-propagated with less gradient vanishing effect between layers in a deep neural network. One notable difference between dense networks and residual networks is a connectivity pattern. Considering that $H_\ell(\cdot)$ is a general composite function of operations in the ℓ^{th} layer of a given neural network, a residual connectivity for the output of the ℓ^{th} layer, $\mathbf{x}_\ell$, can be written as

$$\mathbf{x}_\ell = H_\ell(\mathbf{x}_{\ell-1}) + \mathbf{x}_{\ell-1}, \tag{1}$$

while a dense connectivity can be represented as

$$\mathbf{x}_\ell = H_\ell([\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_{\ell-1}]), \tag{2}$$

where $[\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_{\ell-1}]$ is a concatenated vector of outputs from the first layer to the $(\ell-1)^{\text{th}}$ layer. The dense connectivity pattern accommodates more direct connections throughout layers while residual connections are only made between adjacent layers.

We propose *densely connected LSTMs* (namely, dense LSTMs) in this paper, inspired by the success of dense connection for CNNs. In speech recognition, there has been a limited number of effort to exploit residual connection or its variants, e.g., highway connection, to LSTMs with minor differences in implementation [11, 12, 13, 14], but none using dense connection yet. To understand how dense LSTMs would work as layers get deeper, let us take a look at Figure 1. For the experiments, we trained (uni-directional) LSTMs with the cell dimension of 128 using a small portion of our entire training data, i.e., 300hr Switchboard-1 Release 2 corpus from LDC (LDC97S62), and tested them against the NIST 2000 Hub5

[1] As known as Switchboard, but it actually consists of the two testsets of Switchboard and CallHome.

10.21437/Interspeech.2018-1486

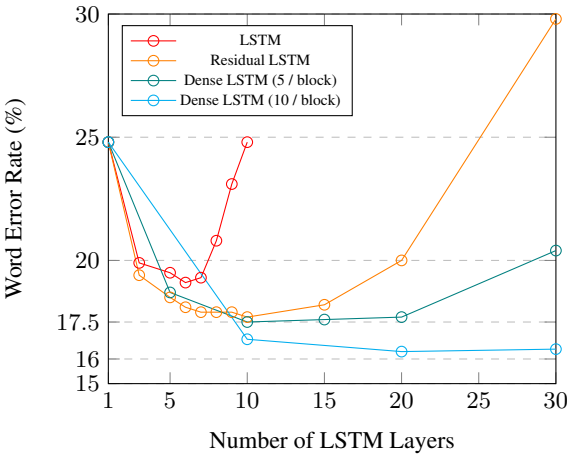

Figure 1: *WER comparison between residual and dense connection for LSTMs with the cell dimension of 128.*

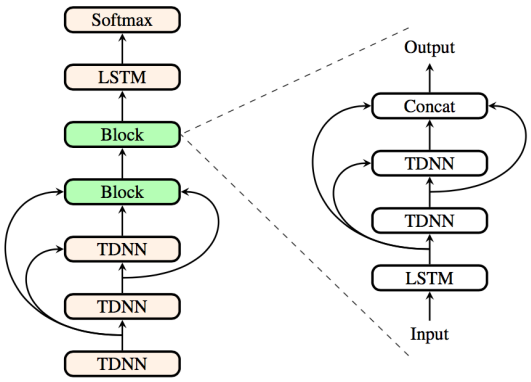

Figure 2: *Structure of a dense TDNN-LSTM acoustic model. Each dense block outputs 1,024 dimensional non-linear activation vectors.*

English evaluation set (Switchboard and CallHome combined). The red curve indicates that normal LSTMs would not obtain any benefit after the 6th layer where the lowest WER of 19.1% is reached. The performance of residual LSTMs, depicted as the orange curve, seems further improved until the 10th layer where 17.7% is marked and then continues to degrade thereafter as more layers are added. This validates that residual learning makes LSTMs perform better with more layers as has been reported in [11, 12, 13, 14], but we also see that there is a clear limitation. In contrast, dense LSTMs are shown continuously benefited as more layers are stacked even after the 10th layer, further pulling the lowest possible WER down to around 16% at the number of LSTM layers of 20 (light blue curve). There are a couple of notes on the dense LSTMs experimented. Due to the connectivity pattern in Eq. (2) of concatenating vectors coming out of the previous layers, the dimension of an input vector for the ℓ^{th} LSTM layer with the cell dimension of d is $(\ell-1)\times d$, which would keep increasing as ℓ goes larger. Thus we grouped LSTM layers into blocks where dense connections are applied only within the same block, and linked blocks by a transitional layer. This dense block concept was also exploited in the original paper for densely connected CNNs [7], but with other purposes. The green curve in Figure 1 is based on the dense LSTMs where every group of 5 LSTM layers belong to one block while the light blue curve comes from the dense LSTMs with 10 LSTM layers per block.

Dense connection can be easily applied to existing LSTM-based neural network architectures for speech recognition, thanks to a simple connectivity pattern. It can improve performance as more LSTM layers are added, since it helps alleviate layer-wise vanishing gradient effect. Based on our experimental validation from Figure 1, we propose two dense LSTM architectures for conversational speech recognition, which are detailed in the next subsections.

2.1. Dense TDNN-LSTM

The first proposed network with dense connection is dense TDNN-LSTM. It has the common network skeleton with the model configuration used in [17], consisting of 7-layer time delay neural networks (TDNNs) combined with 3-layer LSTMs. The model architecture is depicted in Figure 2, where 3 TDNNs are followed by a couple of dense blocks and 1 LSTM in the final layer before the softmax layer. Each green-highlighted dense block contains 1 LSTM and 2 TDNNs with the dense connectivity pattern. The final layer in each block is to concatenate all the outputs of the neural layers inside the block.

Table 1 shows a WER comparison between the original TDNN-LSTM [17] and the proposed dense TDNN-LSTM. For this experiment, we utilized the training data of Fisher English Training Part 1 and 2 (LDC2004S13, LDC2005S13) and the aforementioned Switchboard corpus. The total amount of the data used for training is approximately 2,000hrs. We report the performance of the trained models on Switchboard and CallHome separately. It is noticeable in the table that there is a statistically meaningful improvement (by around 5%, relative) on the CallHome testset by the proposed model.

2.2. Dense CNN-bLSTM (bi-directional LSTM)

We propose another dense LSTM architecture in dense CNN-bLSTM, as shown in Figure 3. This architecture has 3 CNN layers followed by 2 dense blocks (blue-highlighted), each of which contains $N = 7$ LSTM layers being connected densely to one another. The final layer in each block concatenates the output vectors from all the layers inside.

Table 1 also presents the performance of the proposed dense CNN-LSTM model, which exceeded the performance of the dense TDNN-LSTM model introduced in Section 2.1 for both of the Switchboard and CallHome test set.

2.3. Acoustic Model Training

The Kaldi toolkit [18] was used to train these dense networks. Lattice-free maximum mutual information (LF-MMI) was chosen as an objective function for network training. The cross entropy objective function was also applied as an extra regularization as well as leaky HMM to avoid overfitting [19]. The learning rate was gradually adjusted from 10^{-3} to 10^{-5} over the course of 4 epochs.

The total 140K word tokens to cover the entire words contained in the training data were mapped to the PronLex pronunciation lexicon (LDC97L20). The phone dictionary consists of 42 non-silence phones with lexical stress markers on vowels as well as two hesitation phones, making the total phones to 44.

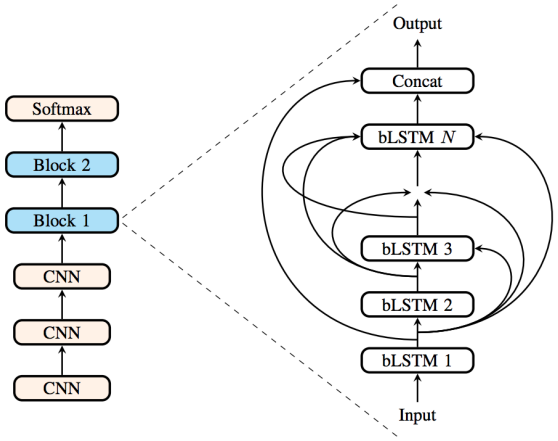

Figure 3: *Structure of a dense CNN-bLSTM acoustic model. Each dense block has 256 dimensional non-linear activation vectors.*

Table 1: *WERs for the proposed dense LSTM acoustic models.*

Acoustic Model	SWBD	CH
TDNN-LSTM	7.3%	13.8%
Dense TDNN-LSTM	7.3%	13.0%
Dense CNN-bLSTM	7.1%	12.5%

2.4. Acoustic Model Adaptation

Neural networks are well known to be hard for adaptation due to a huge number of parameters to be tuned, unlike statistical frameworks such as GMMs. As a result there have been alternative approaches to update only a small part of a neural network model [20, 21, 22, 23] to obtain adaptation benefits. In this paper, we utilizes a simple model adaptation scheme exploiting *parameter averaging*.

The idea is similar to how Kaldi's NNET3 acoustic model training handles the models updated across multiple GPUs throughout iterations [24]. Kaldi's NNET3 training strategy lets each GPU do stochastic gradient descent (SGD) separately with different randomized subsets of a training data and, after processing a fixed number of samples, averages the parameters of the models across all the jobs and re-distribute the result to each GPU. We borrow this concept of parameter averaging to average the parameters of a seed neural network model and its adapted version.

In order to update a seed model with adaptation data before parameter averaging, we applied the same training technique in Section 2.3 with the LF-MMI objective function, but with no cross entropy objective. The learning rate was set to be gradually decreased from 10^{-5} to 10^{-7} over the course of 4 epochs.

We took an advantage of the CallHome American English Speech corpus (LDC97S42) for our experiments on acoustic model adaptation. According to the 2000 Hub5 Evaluation result reported by NIST [25], this corpus was listed as one of publicly available training materials. We only used a training portion of the corpus which contains 80 telephone conversations between native English speakers of around 13 speech hours. There is no overlap in data itself as well as speaker between this adaptation data and the CallHome portion of the NIST 2000 Hub5 English evaluation set, but it is expected for adapted models to perform better than before adaptation, at least against the CallHome test set.

Table 2: *Acoustic model adaptation results in WER. Before parameter averaging.*

Dense Model	SWBD	CH
TDNN-LSTM	7.3% → 7.7%	13.0% → 12.2%
CNN-bLSTM	7.1% → 7.5%	12.5% → 12.1%

Table 3: *Acoustic model adaptation results in WER. After parameter averaging.*

Dense Model	SWBD	CH
TDNN-LSTM	7.7% → 7.2%	12.2% → 12.1%
CNN-bLSTM	7.5% → 7.1%	12.1% → 11.9%

Tables 2 and 3 show the experimental results with and without parameter averaging in model adaptation. Although there exists a consistent improvement for the CallHome test set across the updated dense LSTM models in Table 2, the performance against the Switchboard test set is all degraded. This indicates that the models updated with the adaptation data from the CallHome corpus have the parameters shifted toward CallHome-specific regions in a parameter space, but farther away from a Switchboard-specific domain. The proposed parameter averaging method is shown in Table 3 to balance out the biases in the updated models. It seems to pull the Switchboard WERs back to the range before the model adaptation while preserving the benefit for the CallHome side. The overall improvement for the CallHome test set across the updated models is approximately 5% (relative).

3. System Descriptions

3.1. Other Acoustic Models

To achieve the state-of-the-art performance on the NIST 2000 Hub5 English evaluation set, we trained 3 CNN-bLSTM models across three different phone sets (CMU[2], MSU[3], PronLex) in addition to the aforementioned dense LSTM models. The CMU set consists of 39 phones with three lexical stress markers. The MSU set has 36 phones with no stress distinctions. Different trees were formed for the CNN-bLSTMs during the training stages, which could provide diversity to a combined system later.

For all the CNN-bLSTMs, log-mel features were fed into the three convolutional layers and a 3×3 kernel was applied with the filter size of 32 throughout the layers. The filtered signals were then passed to the 7-layer bLSTMs with the cell dimension of 1,024 after being appended with 100-dimensional i-vectors. Each neural network layer is followed by non-linear ReLU activation.

3.2. Language Models

The 4-gram language model (LM) was trained with the open-source library of SRILM [26] on a combination of publicly available data, including Fisher English Training Part 1 and 2 (LDC2004T19, LDC2005T19), Switchboard-1 Release 2 (LDC97S62), CallHome American English Speech (LDC97T14), Switchboard Cellular Part 1 (LDC2001T14), TED-LIUM [27], British Academic Spoken English (BASE) [28], Michigan Corpus of American Spoken English (MICASE)

[2]http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict

[3]http://www.isip.piconepress.com/projects/switchboard/releases/sw-ms98-dict.text

Table 4: *Experimental evaluation in WER for the 5 individual systems and their combinations. d is the dimension of LSTM cells.*

Acoustic Model	d	SWBD		CH		RT-02		RT-03	
		N-gram	RNN	N-gram	RNN	N-gram	RNN	N-gram	RNN
CNN-bLSTM (CMU)	1,024	6.8%	5.9%	11.5%	10.7%	10.4%	9.2%	9.9%	9.0%
CNN-bLSTM (MSU)	1,024	6.8%	5.9%	11.3%	10.5%	10.0%	9.1%	9.7%	8.9%
CNN-bLSTM (PronLex)	1,024	6.4%	5.6%	11.4%	10.7%	9.9%	9.0%	9.8%	9.1%
Dense CNN-bLSTM	256	7.1%	6.1%	11.9%	11.1%	10.5%	9.6%	10.3%	9.5%
Dense TDNN-LSTM	1,024	7.2%	6.1%	12.1%	11.0%	10.7%	9.5%	10.5%	9.5%
3 CNN-bLSTMs Combined	-	-	5.1%	-	9.7%	-	8.2%	-	8.5%
5 System Combination	-	-	**5.0%**	-	**9.1%**	-	**8.1%**	-	**8.0%**

[29] and English Gigaword (LDC2003T05). We used this LM for the 2nd-pass LM rescoring. For the 1st pass decoding, we pruned the trained 4-gram LM with the pruning thresholds of 1.0e-8, 1.0e-7, and 1.0e-6 for bigrams, trigrams, and 4-grams, respectively.

The RNN LM built with the CUED-RNNLM tookit [30] was trained on a subset of the aforementioned text data, consisting of only Fisher, Switchboard and CallHome with 2M sentences and 24M word tokens. We used variance regularization [31] as the optimization criterion of the objective function for the RNN LMs with 1,000 nodes in each of two hidden layers.

3.3. System Combination

In order to combine the systems, we applied a lattice combination that conducts a union of lattices from individual systems and searches the best path from the extended lattices using minimum Bayes risk decoding [32]. The combination weights were found through a hyper-parameter optimization algorithm, called a tree-structured Parzen estimator [33], using a held-out development set.

4. Experimental Results

We evaluated the performance of the total 5 individual systems across the three different phone sets against the Switchboard and CallHome test set of the NIST 2000 Hub5 English evaluation set. The performances in terms of WER, before and after RNN LM rescoring, are shown in Table 4.

Among the acoustic models, the PronLex based CNN-bLSTM outperformed the other models, marking 5.6% for Switchboard, while the MSU based CNN-bLSTM reached the lowest WER of 10.5% for CallHome. These two numbers are the best reported WERs achieved by any single system so far. The CNN-bLSTM models obtained WERs 0.2%-0.5% (absolute) for Switchboard and 0.3%-0.6% (absolute) for CallHome lower than the dense acoustic models. The proposed dense LSTMs significantly contributed to system combination. By comparing the WERs of the 3 CNN-bLSTMs combined and that of the final 5 systems including the two dense networks, the improvements resulted from the dense networks are shown to be approximately 5% (relative) across the test sets. We observed that RNN LM rescoring provided consistent improvement in all the testing cases with absolute improvements between 0.7% and 1.2%, and a maximum reduction in WER of up to 10% (relative) in the case of the dense TDNN-LSTM model on RT-02 (from 11.3% to 10.1%).

In the table, we also report the two additional test sets in RT-02 and RT-03 other than Switchboard and CallHome, expecting them to offer a diverse view of our systems. They are another evaluation sets with telephone conversations, publicly available, which are exclusive with the NIST 2000 Hub5 English evaluation set, but some portions of RT-02 and RT-03 come from the Switchboard data collection projects. For more details, one might refer to LDC2004S11 (RT-02) and LDC2007S10 (RT-03).

5. Conclusions

In this paper we have proposed a couple of densely connected LSTM architectures, bringing the dense connectivity that was successful for CNNs in image classification tasks to the LSTM framework for conversational speech recognition. This allowed LSTMs to have more direct connections between layers such that layer-wise vanishing gradient effect can be further alleviated even as more layers are stacked in deep neural network models. Also we have introduced parameter averaging for acoustic model adaptation that averages the parameters of a seed neural network acoustic model and its adapted one, in order to balance out between domain adaptation and generalization.

We note that in a comparison of the reported numbers of 5.1% and 9.9% from [5] our combined system has made a significant improvement on the CallHome portion of the NIST 2000 Hub5 English evaluation set, mainly due to the acoustic model adaptation using the CallHome training data of approximately 13 hours of speech. This shows that domain specific data which has similar acoustics and lexical information would have more direct impact on performance improvement. As discussed in [3], even the best conversational speech recognition system could suffer from higher error rates when it is tested against real-world data with a number of unseen dynamics in data characteristics. To have systems more robust to unseen testing conditions, given limited resources of audio data and the corresponding reference transcripts, unsupervised learning that can continuously improve the recognition coverage of a given speech recognition system would be required. In addition, various testing materials beyond the NIST 2000 Hub5 English evaluation set or RTs by LDC would be able to provide deeper insights on how systems could be generalized against real-world data.

6. References

[1] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, and A. Stolcke, "Achieving human parity in conversational speech recognition," in *MSR-TR-2016-71*, 2016.

[2] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L. Lim, B. Roomi, and P. Hall, "English conversational telephone speech recognition by humans and machines," in *Proc of INTERSPEECH 2017*, 2017, pp. 132–136.

[3] K. J. Han, S. Hahm, B. Kim, J. Kim, and I. Lane, "Deep learning-

based telephony speech recognition in the wild," in *Proc of INTERSPEECH 2017*, 2017, pp. 1323–1327.

[4] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The Microsoft 2017 conversational speech recognition system," in *MSR-TR-2017-39*, 2017.

[5] G. Kurata, B. Ramabhadran, G. Saon, and A. Sethy, "Language modeling with highway LSTM," in *Proc of ASRU 2017*, 2017.

[6] A. Stolcke and J. Droppo, "Comparing human and machine errors in conversational speech transcription," in *Proc of INTERSPEECH 2017*, 2017, pp. 137–141.

[7] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *arXiv:1608.06993*, 2016.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc of CVPR 2016*, 2016, pp. 770–778.

[9] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc of NIPS 2015*, 2015, pp. 2377–2385.

[10] ——, "Highway networks," in *Proc of ICML 2015*, 2015.

[11] Y. Zhang, G. Chen, D. Yu, K. Yao, S. Khudanpur, and J. Glass, "Highway long short-term memory RNNs for distant speech recognition," in *Proc of ICASSP 2016*, 2016, pp. 5755–5759.

[12] G. Pundak and T. Sainath, "Highway-LSTM and recurrent highway networks for speech recognition," in *Proc of INTERSPEECH 2017*, 2017.

[13] J. Kim, M. El-Khamy, and J. Lee, "Residual LSTM: Design of a deep recurrent architecture for distant speech recognition," in *Proc of INTERSPEECH 2017*, 2017.

[14] L. Huang, J. Sun, J. Xu, and Y. Yang, "An improved residual LSTM architecture for acoustic modeling," in *Proc of ICCCS 2017*, 2017.

[15] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," in *Tech Report*, 2009.

[16] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *arXiv:1605.07146*, 2009.

[17] G. Cheng, V. Peddinti, D. Povey, V. Manohar, S. Khudanpur, and Y. Yan, "An exploration of dropout with LSTMs," in *Proc of INTERSPEECH 2017*, 2017.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc of ASRU 2011*, 2011.

[19] D. Povey, V. Peddinti, G. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free MMI," in *Proc of INTERSPEECH 2016*, 2016.

[20] B. Li and K. C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems," in *Proc of INTERSPEECH 2010*, 2010.

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

[22] I. Himawan, P. Motlicek, M. F. Font, and S. Madikeri, "Towards utterance-based neural network adaptation in acoustic modeling," in *Proc of ASRU 2015*. IEEE, 2015, pp. 289–295.

[23] L. Lu, "Sequence training and adaptation of highway deep neural networks," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 461–466.

[24] D. Povey, X. Zhang, and K. Sanjeev, "Parallel training of DNNs with natural gradient and paramter averaging," in *Proc of ICLR 2015*, 2015.

[25] J. F. William, W. M. Fisher, A. F. Martin, M. A. Przybocki, and D. S. Pallett, "NIST evaluation of conversational speech recognition over the telephone: English and mandarin performance results," in *NIST*, 2000.

[26] A. Stolcke, "SRILM – An extensible language modeling toolkit," in *Proc of ICSLP 2002*, 2002, pp. 901–904.

[27] A. Rousseau, P. Deléglise, and Y. Esteve, "TED-LIUM: an automatic speech recognition dedicated corpus." in *LREC*, 2012, pp. 125–129.

[28] P. Thompson and H. Nesi, "Research in progress, the British Academic Spoken English (base) corpus project." *Language Teaching Research*, vol. 5, no. 3, 2001.

[29] R. C. Simpson, S. L. Briggs, J. Ovens, and J. M. Swales, "The Michigan Corpus of Academic Spoken English." *Ann Arbor, MI: The Regents of the University of Michigan*, 2002.

[30] X. Chen, X. Liu, Y. Qian, M. J. F. Gales, and P. C. Woodland, "CUED-RNNLM - An open-source toolkit for efficient training and evaluation of recurrent neural network language models," in *Proc of ICASSP 2016*, 2016, pp. 6000–6004.

[31] Y. Shi, W. Zhang, M. Cai, and J. Liu, "Variance regularization of RNNLM for speech recognition," in *Proc of ICASSP 2014*, 2014.

[32] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Comp. Speech and Lang.*, vol. 4, pp. 802–828, 2011.

[33] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Proc of NIPS 2011*, 2011.