Residual LSTM: Design of a Deep Recurrent Architecture for Distant Speech Recognition

Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee

Abstract

In this paper, a novel architecture for a deep recurrent neural network, residual LSTM is introduced. A plain LSTM has an internal memory cell that can learn long term dependencies of sequential data. It also provides a temporal shortcut path to avoid vanishing or exploding gradients in the temporal domain. The proposed residual LSTM architecture provides an additional spatial shortcut path from lower layers for efficient training of deep networks with multiple LSTM layers. Compared with the previous work, highway LSTM, residual LSTM reuses the output projection matrix and the output gate of LSTM to control the spatial information flow instead of additional gate networks, which effectively reduces more than 10% of network parameters. An experiment for distant speech recognition on the AMI SDM corpus indicates that the performance of plain and highway LSTM networks degrades with increasing network depth. For example, 10-layer plain and highway LSTM networks showed 13.7% and 6.2% increase in WER over 3-layer baselines, respectively. On the contrary, 10-layer residual LSTM networks provided the lowest WER 41.0%, which corresponds to 3.3% and 2.8% WER reduction over 3-layer plain and highway LSTM networks, respectively. Training with both the IHM and SDM corpora, the residual LSTM architecture provided larger gain from increasing depth: a 10-layer residual LSTM showed 3.0% WER reduction over the corresponding 5-layer one.

Index Terms

ASR, LSTM, GMM, RNN, CNN

I. Introduction

Over the past year, the emergence of deep neural networks has fundamentally changed the design of automatic speech recognition (ASR). Neural network-based acoustic models presented significant performance improvement over the prior state-of-the-art Gaussian mixture model (GMM) [1], [2], [3]. Advanced neural network-based architectures further improved ASR performance. For example, convolutional neural networks (CNN) which has been huge success in

image classification and detection were effective to reduce environmental and speaker variability in acoustic features [4], [5], [6], [7]. Recurrent neural networks (RNN) were successfully applied to learn long term dependencies of sequential data [8], [9], [10].

The recent success of a neural network based architecture mainly comes from its deep architecture. However, training a deep neural network is a difficult problem due to vanishing or exploding gradients. Furthermore, increasing depth in recurrent architectures such as gated recurrent unit (GRU) and long short-term memory (LSTM) is significantly more difficult because they already have a deep architecture in the temporal domain.

There have been two successful architectures for a deep feed-forward neural network: residual network and highway network. Residual network [11] was successfully applied to train more than 100 convolutional layers for image classification and detection. The key insight in the residual network is to provide a shortcut path between layers that can be used for an additional gradient path. Highway network [12] is an another way of implementing a shortcut path in a feed-forward neural network. [12] presented successful MNIST training results with 100 layers.

Highway LSTM [13], [14] is a recurrent version of highway network. LSTM [15] has internal memory cells that provide shortcut gradient paths in the temporal direction. Highway LSTM reused them for a highway shortcut in the spatial domain. It also introduced new gate networks to control highway paths from the prior layer memory cells. [13] presented a highway LSTM for far-field speech recognition and showed improvement over plain LSTM. However, [13] also showed that highway LSTM degraded with increasing depth.

In this paper, a novel highway architecture, residual LSTM is introduced. The key insights of a residual LSTM are summarized as below.

- Highway connection between layer outputs instead of internal memory cells: LSTM internal memory cells are used to deal with gradient issues in the temporal domain. Reusing it again for the spatial domain could make it more difficult to train a network in both temporal and spatial domains. The proposed residual LSTM uses a layer output for the spatial shortcut connection instead of an internal memory cell, which can less interfere with a temporal grandient flow.
- Each layer output at the residual LSTM learns residual mapping not learnable from highway path. Therefore, each new layer does not need to waste time or resource to generate similar outputs from prior layers.
- Residual LSTM reuses LSTM projection matrix as a gate network. For a usual LSTM

network size, more than 10% learnable parameters can be saved from residual LSTM over highway LSTM.

The experimental result on the AMI SDM corpus [16] showed 10-layer plain and highway LSTMs had 13.7% and 6.2% increase in WER over 3-layer baselines, respectively. On the contrary, 10-layer residual LSTM presented the lowest WER 41.0%, which corresponds to 3.3% and 2.8% WER reduction over 3-layer plain and highway LSTMs, respectively. For an experiment with IHM and SDM corpora, 10-layer residual LSTM showed 3.0% WER reduction over 5-layer one, whereas the experiment with only SDM corpus presented 1% reduction.

The rest of this paper is organized as follows. Section II will review existing highway architectures. Section III will introduce residual LSTM. Section IV will explain experimental setup and provide results on AMI distant speech recognition. Finally, this paper ends with conclusion at section V.

II. Revisiting Highway Networks

In this section, we give a brief review of LSTM and three existing highway architectures.

A. Residual Network

Residual network [11] provides an identity mapping by shortcut paths. Since the identity mapping is always on, function output only needs to learn residual mapping. Formulation of this relation can be expressed as:

$$y = F(x; W) + x \tag{1}$$

y is a layer output, x is a layer input and $F(x; W)$ is a function with an internal parameter W. Without a shortcut path, $F(x; W)$ should represent y from input x, but with an identity mapping x, $F(x; W)$ only needs to learn residual mapping, $y - x$. As layers are stacked up, if no new residual mapping is needed, a network can bypass identity mappings without training, which could greatly simplify training of a deep network.

B. Highway Network

Highway network [12] provides another way of implementing a shortcut path for a deep neural-network. Layer output $H(x; W_h)$ is multiplied by a transform gate $T(x; W_T)$ and before

going into the next layer, a highway path $x \cdot (1 - T(x; W_T))$ is added. Formulation of a highway network can be summarized as:

$$y = H(x; W_h) \cdot T(x; W_T) + x \cdot (1 - T(x; W_T)) \tag{2}$$

Transform gate is defined as:

$$T(x; W_T) = \sigma(W_T x + b_T) \tag{3}$$

Unlike a residual network, a highway path is not always turned on. For example, a highway network can ignore a highway path if $T(x; W_T) = 1$, or bypass a layer output when $T(x; W_T) = 0$.

C. Long Short-Term Memory (LSTM)

Long short-term memory (LSTM) [15] was proposed to resolve vanishing or exploding gradients for a recurrent neural network. LSTM has an internal memory cell that is controlled by forget and input gate networks. A forget gate in an LSTM determines how much of prior memory value should be passed into the next time step. Similarly, an input gate scales a new input to a memory cell. Depending on the states of both gates, LSTM can represent long-term or short-term dependency of sequential data. The formulation of an LSTM is as follows:

$$i_t^l = \sigma(W_{xi}^l x_t^l + W_{hi}^l h_{t-1}^l + w_{ci}^l c_{t-1}^l + b_i^l) \tag{4}$$

$$f_t^l = \sigma(W_{xf}^l x_t^l + W_{hf}^l h_{t-1}^l + w_{cf}^l c_{t-1}^l + b_f^l) \tag{5}$$

$$c_t^l = f_t^l \cdot c_{t-1}^l + i_t^l \cdot \tanh(W_{xc}^l x_t^l + W_{hc}^l h_{t-1}^l + b_c^l) \tag{6}$$

$$o_t^l = \sigma(W_{xo}^l x_t^l + W_{ho}^l h_{t-1}^l + W_{co}^l c_t^l + b_o^l) \tag{7}$$

$$r_t^l = o_t^l \cdot \tanh(c_t^l) \tag{8}$$

$$h_t^l = W_p^l \cdot r_t^l \tag{9}$$

l represents layer index and i_t^l, f_t^l and o_t^l are input, forget and output gates respectively. They are component-wise multiplied by input, memory cell and hidden output to gradually open or close their connections. x_t^l is an input from $(l-1)^{th}$ layer (or an input to a network when l is 1), h_{t-1}^l is a l^{th} layer output at time $t-1$ and c_{t-1}^l is an internal cell state at $t-1$. W_p^l is a projection matrix to reduce dimension of h_t^l.

D. Highway LSTM

Highway LSTM [13], [15] reused LSTM internal memory cells for spatial domain highway connections between stacked LSTM layers. Equations (4), (5), (7), (8), and (9) do not change for a highway LSTM. Equation (6) is updated to add a highway connection:

$$c_t^l = d_t^l \cdot c_t^{l-1} + f_t^l \cdot c_{t-1}^l + i_t^l \cdot \tanh(W_{xc}^l x_t^l + W_{hc}^l h_{t-1}^l + b_c^l) \quad (10)$$

$$d_t^l = \sigma(W_{xd}^l x_t^l + W_{cd}^l c_{t-1}^l + w_{cd}^l c_t^{l-1} + b_d^l) \quad (11)$$

Where d_t^l is a depth gate that connects c_t^{l-1} in the $(l-1)^{th}$ layer to c_t^l in the l^{th} layer. [13] showed that an acoustic model based on the highway LSTM improved far-field speech recognition compared with a plain LSTM. However, [13] also showed that word error rate (WER) degraded when the number of layers in the highway LSTM increases from 3 to 8.

III. Residual LSTM

In this section, a novel architecture for a deep recurrent neural network, residual LSTM is introduced. Residual LSTM starts with an intuition that the separation of a spatial-domain shortcut path with a temporal-domain cell update may give better flexibility to deal with vanishing or exploding gradients. Unlike a highway LSTM, residual LSTM does not accumulate a highway path on an internal memory c_t^l. Instead, a shortcut path is added to an LSTM output layer. Figure III describes a cell diagram of a residual LSTM. h_t^{l-1} is a shortcut path from $(l-1)^{th}$ layer output that is added to a projection output m_t^l. Although a shortcut path can be any lower layer output, in this paper, we used a previous layer output. Equations (4), (5), (6) and (7) do not change for a residual LSTM. The updated equations are as follows:

$$r_t^l = \tanh(c_t^l) \quad (12)$$

$$m_t^l = W_p^l \cdot r_t^l \quad (13)$$

$$h_t^l = o_t^l \cdot (m_t^l + W_h^l x_t^l) \quad (14)$$

Where W_h^l can be replaced by an identity matrix if the dimension of x_t^l matches that of h_t^l. For a matched dimension, Equation (14) can be changed into:

$$h_t^l = o_t^l \cdot (m_t^l + x_t^l) \quad (15)$$

Since a highway path is always turned on for a residual LSTM, there should be a scaling parameter on the main path output. For example, linear filters in the last CNN layer of a residual network are reused to scale the main path output. For a residual LSTM, a projection matrix W_p^l is reused in order to scale the LSTM output. Consequently, the number of paramemters for a residual LSTM does not increase compared with a plain LSTM. Simple complexity comparison between residual LSTM and highway LSTM is as follows. If the size of the internal memory cells is N and the output layer dimension after projection is $N/2$, the total number of reduced parameters for a residual LSTM becomes $N^2/2 + 4N$. For example, if N is 1024 and the number of layers is more than 5, residual LSTM has approximately 10% less network parameters compared with a highway LSTM.

One thing to note is that a highway path should be scaled by an output gate as in Equation (14). Empirical experiments without an output gate presented significant performance loss. This is because highway paths are accumulated as the number of layers increases. Without proper scaling, the variance of an LSTM output keeps increasing as the number of layers grows.

Output gate is a trainable network which can learn a proper range of an LSTM output. For example, if an output gate is set as $\frac{1}{\sqrt{2}}$, an l^{th} layer output becomes

$$h_t^l = \sum_{k=1}^{l} (\frac{1}{\sqrt{2}})^{(l-k+1)} m_t^k + (\frac{1}{\sqrt{2}})^l x_t \tag{16}$$

Where, x_t is an input to LSTM at time t. If m_t^l and x_t are independent each other for all l and have fixed variance of 1, regardless of layer index l, the variance of layer l^{th} output becomes 1. Since variance of a layer output is variable in the real scenario, a trainable output gate will better deal with exploding variance than a fixed scaling factor.

IV. Experiments

A. Experimental Setup

AMI meeting corpus [16] is used to train and evaluate a residual LSTM. AMI corpus consists of 100 hours of meeting recordings. For each meeting, three to four people have free conversation in English. Frequently, overlapped speaking from multiple speakers happens and for that case, the training transcript always follows a main speaker. Multiple microphones are used to synchronously record conversations in different environments. Individual headset microphone (IHM) recorded clean close-talking coversation and single distant microphone (SDM) recorded

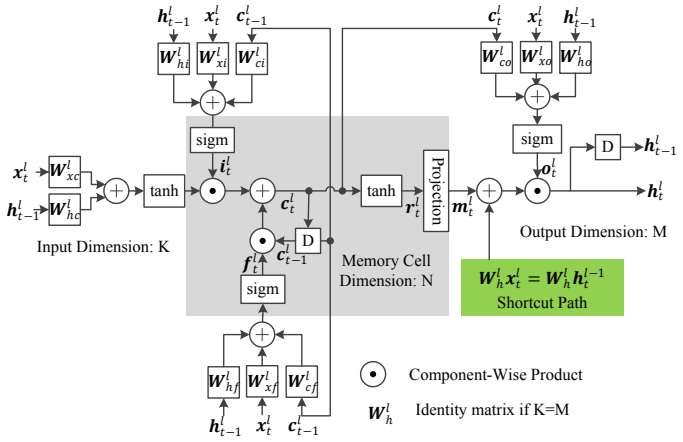

Fig. 1. Residual LSTM: A shortcut from a prior layer output h_t^{l-1} is added to a projection output m_t^l. W_h^l is a dimension matching matrix between input and output. If K is equal to M, it is replaced with an identity matrix.

far-field noisy conversation. In this paper, SDM is used to train a residual LSTM at Section IV-B and IV-C and combined SDM and IHM corpora are used at Section IV-D.

Kaldi [17] is a toolkit for speech recognition that is used to train a context-dependent LDA-MLLT-GMM-HMM system. The trained GMM-HMM generates forced aligned labels which are later used to train a neural network-based acoustic model. Three neural network-based acoustic models are trained: plain LSTM without any shortcut path, highway LSTM and residual LSTM. All three LSTM networks have 1024 memory cells and 512 output nodes for experiments at Section IV-B, IV-C and IV-D.

The computational network toolkit (CNTK) [18] is used to train and decode three acoustic models. Truncated back-propagation through time (BPTT) is used to train LSTM networks with 20 frames for each truncation. For parallel processing, 40 utterances are simultaneously processed to update parameters. Cross-entropy loss function is used with L2 regularization.

For decoding, reduced 50k-word fisher dictionary is used for lexicon and based on this lexicon, tri-gram language model is interpolated from AMI training transcript. As a decoding option, word error rate (WER) can be calculated based on non-overlapped speaking or overlapped speaking. Recognizing overlapped speaking is to decode up to 4 concurrent speeches. Decoding overlapped speaking is a big challenge considering a network is trained to only recognize a main speaker. Following sections will provide WERs for both options.

B. Training Performance with increasing Depth

Figure 2 compares phone error rates (PER) with increasing depth. Cross-validation (CV) PER is measured with a separate data that is not used for regression. Both training and CV PERs are shown in the plots. Figure 2 does not include a plain LSTM due to a limited space. Its PERs were always worse than those for highway or residual LSTM.

Figure 2a shows PERs for a highway LSTM. Increasing depth from 3 to 5 does not show much difference in both training and CV PERs. However, 10-layer highway LSTM showed significant degradation. The converged training PER is much higher than 3 or 5-layer models in spite of increased complexity. Since CV PER also degraded, loss from training PER did not come from better generalization. Therefore, training loss is purely due to increased depth. For a 10-layer highway LSTM, it has 3.6% CV PER loss and 15% training PER loss compared with 3-layer highway LSTM.

Figure 2b shows PERs for a residual LSTM. CV PERs consistently get better with increasing depth. For a training PER, it is slightly degraded for a 10-layer network. However, it is not a training loss from increased depth because CV PER for a 10-layer network improved. Specifically, the regularization weight for a residual LSTM was set to be 2-4 times higher than that of a highway LSTM because a residual LSTM is more susceptible to overfitting. Smaller regularization weights can easily decrease training PER but generalization performance will be much worse.

C. WER Evaluation with SDM corpus

Table I compares WER for LSTM, highway LSTM and residual LSTM with increasing depth. All three networks were trained by SDM AMI corpus. Both overlapped and non-overlapped WERs are shown. For each layer, internal memory cell size is set to be 1024 and output node size is fixed as 512. A plain LSTM performed worse with increasing layers. Especially, the 10-layer LSTM degraded up to 13.7% over the 3-layer LSTM for non-overlapped WER. A highway LSTM showed better performance over a plain LSTM but still could not avoid degradation with increasing depth. The 10-layer highway LSTM presented 6.2% increase in WER over the 3-layer network.

On the contrary, a residual LSTM improved with increasing layers. 5-layer and 10-layer residual LSTMs have 1.2% and 2.2% WER reduction over the 3-layer network, respectively.

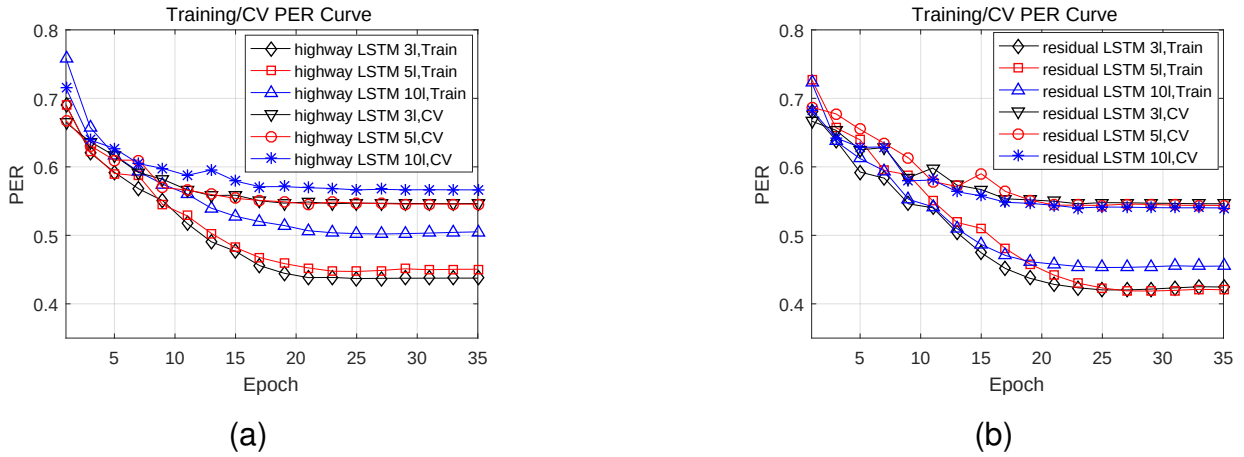

Fig. 2. Training and cross validation (CV) phone error rates (PER) for each epoch on AMI SDM corpus. (a) shows phone error rates for a highway LSTM. When depth increases from 3 to 10, both training and cross validation errors severely degraded. (b) shows phone error rates for a residual LSTM. Training/CV PER does not show significant degradation. On the contrary, CV PER got improved with 10-layer model.

The 10-layer residual LSTM showed the lowest 41.0% WER, which corresponds to 3.3% and 2.8% WER reduction over 3-layer plain and highway LSTMs.

D. WER Evaluation with SDM and IHM corpora

Table II compares WER of highway and residual LSTMs trained with combined IHM and SDM corpora. With increased corpus size, the best performing configuration for a highway LSTM is changed into 5-layer with 40.7% WER. However, 10-layer highway LSTM still suffered from training loss from increased depth: 6.6% increase in WER (non-overlapped). On the contrary, 10-layer residual LSTM showed the best WER of 39.3%, which corresponds to 3.1% WER reduction (non-overlapped) over the 5-layer one, whereas the prior experiment trained only by SDM corpus presented 1% improvement. Increasing training data provides larger gain from a deeper network. Residual LSTM enabled to train a deeper LSTM network without any training loss.

V. CONCLUSION

In this paper, we proposed a novel architecture for a deep recurrent neural network: residual LSTM. A residual LSTM provides a shortcut path between adjacent layer outputs. Unlike a highway network, a residual LSTM does not assign dedicated gate networks for a shortcut

Acoustic Model	Layer	WER (overlapped)	WER (non-overlapped)
Plain LSTM	3	51.1%	42.4%
	5	51.4%	42.5%
	10	56.3%	48.2%
Highway LSTM	3	50.8%	42.2%
	5	51.0%	42.2%
	10	53.5%	44.8%
Residual LSTM	3	50.8%	41.9%
	5	50.0%	41.4%
	10	50.0%	41.0%

TABLE I

ALL THREE LSTM NETWORKS HAVE THE SAME SIZE OF LAYER PARAMETERS:1024 MEMORY CELLS AND 512 OUTPUT NODES. FIXED-SIZE LAYERS ARE STACKED UP WHEN THE NUMBER OF LAYERS INCREASES.

Acoustic Model	Layer	WER (overlapped)	WER (non-overlapped)
Highway LSTM	3	51.3%	42.3%
	5	49.5%	40.7%
	10	52.1%	43.4%
Residual LSTM	3	50.8%	41.9%
	5	49.4%	40.5%
	10	48.7%	39.3%

TABLE II

TWO NETWORKS ARE TRAINED WITH COMBINED SDM AND IHM CORPORA. THE NETWORK CONFIGURATION IS THE SAME AS BEFORE: 1024 MEMORY CELLS AND 512 OUTPUT NODES FOR EACH LAYER.

connection. Instead, projection matrix and output gate are reused for a shortcut connection, which provides roughly 10% reduction of network parameters compared with a highway LSTM. Experiments on AMI corpus showed that a residual LSTM improved significantly with increasing depth, meanwhile 10-layer plain and highway LSTMs severely suffered from training loss.

REFERENCES

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Large vocabulary continuous speech recognition with context-dependent dbn-hmms," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 4688–4691.

[2] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[3] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[4] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4580–4584.

[5] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1120–1124, 2014.

[6] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP)*. IEEE, 2012, pp. 4277–4280.

[7] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for lvcsr," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8614–8618.

[8] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 273–278.

[9] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014.

[10] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." in *INTERSPEECH*, 2014, pp. 338–342.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[12] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[13] Y. Zhang, G. Chen, D. Yu, K. Yaco, S. Khudanpur, and J. Glass, "Highway long short-term memory rnns for distant speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5755–5759.

[14] K. Yao, T. Cohn, K. Vylomova, K. Duh, and C. Dyer, "Depth-gated lstm," in *Presented at Jelinek Summer Workshop on August*, vol. 14, 2015, p. 1.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 28–39.

[17] 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 *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[18] D. Yu, A. Eversole, M. Seltzer, K. Yao, Z. Huang, B. Guenter, O. Kuchaiev, Y. Zhang, F. Seide, H. Wang *et al.*, "An introduction to computational networks and the computational network toolkit," Technical report, Tech. Rep. MSR, Microsoft Research, 2014, 2014. research. microsoft. com/apps/pubs, Tech. Rep., 2014.