WaveNet MH-SRU: Deep and Wide Multiple-history Simple Recurrent Unit for Speech Recognition

Hengguan Huang and *Brian Mak*

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
{hhuangaj,mak}@cse.ust.hk

Abstract

Recently the high-order multiple-history long short-time memory (MH-LSTM) is proposed for speech recognition. The high-order connections to an MH-LSTM cell come from signals similar to the original one but running with progressively longer time lags. Each time-lagged signal keeps a slightly different history of the input, and the history ensemble helps improve the model's robustness against mis-labeling or mis-alignments in the training targets. When an MH-LSTM cell is unfolded in time, it becomes a deep neural network with wider and wider layers towards the inputs. Although a wider MH-LSTM is more resilient to noisy data, we notice that when it is too wide, it is easily under-trained and the increased history/input contexts are not effectively utilized. In this paper, LSTMs are replaced by the faster simple recurrent units (SRUs), which remove the dependency on the hidden states of an MH-LSTM, and high-order connections are instead directly applied to the inputs of an MH-SRU. The high-order connections are further modeled by a WaveNet block which makes use of dilated causal convolution to provide a wider receptive field more effectively. Experimental results on NTIMIT and CHiME-2 demonstrate the effectiveness of the new WaveNet MH-SRU. For example, on CHiME-2, it achieves 1.8% and 0.6% absolute WER reductions over the SRU and MH-SRU baseline models that have similar number of model parameters.

Index Terms: long short-term memory, recurrent neural network, WaveNet, dilated convolution

1. Introduction and related works

Recurrent neural networks (RNNs) are capable of modeling the temporal dependencies among the observations of a sequence signal [1]. One successful implementation of RNN is the *long short-time memory* (LSTM) [2] which has been successfully applied to diverse machine learning problems such as automatic speech recognition (ASR) [3], language modeling [4], machine translation [5] and computer vision [6].

In ASR, it is beneficial for an acoustic model to capture the interactions among acoustic events in a wide contextual window. However, it is difficult for a simple RNN to learn long-term dependencies in a sequence due to the vanishing gradient problem [7]. LSTM alleviates the problem by maintaining a constant error flow through the LSTM cells [2]. Another RNN variant called *Nonlinear AutoRegressive model with eXogenous inputs* (NARX) network [8] introduces high-order feedback paths between an RNN hidden state and its preceding states so that gradients can flow through additional paths that are multiple time steps apart. Clockwork RNN and multi-timescale LSTM [9, 10, 11] further improve on NARX by grouping the hidden states into modules that run at different clock speeds, thus capturing information at different (usually exponential) time scales. The highway LSTM [12], on the other hand, deals with the problem by setting up highway connections between the current input and the output layer so that the unfolded highway LSTM is equivalent to a very deep feedforward residual networks [13]. The simple recurrent unit (SRU) [14], which may be viewed as a simplified highway LSTM, further speeds up the recurrent network computations by making the gate computation dependent only on the current input.

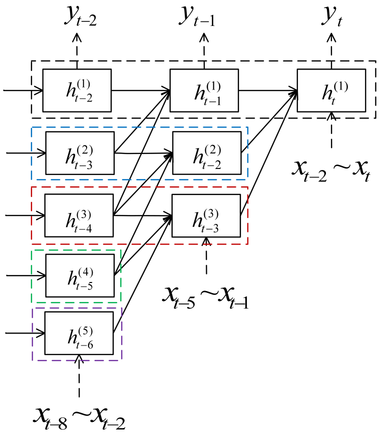

Figure 1: *An MH-LSTM cell unfolded in time.*

In [15], a novel high-order LSTM called *multiple-history LSTM* (MH-LSTM) is proposed to deal with noisy information during acoustic modeling. For example, during the semi-supervised training of acoustic models, most of the training targets are obtained from the recognition results of other (usually inferior) models, and decoding errors — wrong labels or wrong state alignments — are inevitable. The problem gets worse with noisy training speeches which produce more decoding errors. The MH-LSTM introduces high-order connections to each LSTM cell which come from signals similar to the original one except that they run at progressively longer time lags. As a result, each time-lagged signal keeps a slightly different history of the input, and the ensemble of these histories helps smooth out the mis-alignment and mis-labeling noises in the training targets. When an MH-LSTM cell is unfolded in time as in Fig. 1, it is a deep neural network with wider and wider layers at the bottom in terms of history and input contexts.

[15] shows that a wider MH-LSTM network is more resilient to noisy data as it increases its memory capacity to learn the long-term dependencies in speech. However, we notice that when an MH-LSTM is too wide, it is easily under-trained and

the increased history/input contexts are not effectively utilized. To better model with wider history/input context and speed up training, this paper proposes to use WaveNet block [16] and SRU to implement MH-LSTM, and the ensuing model will be called *WaveNet MH-SRU*. That is, the faster SRU is adopted as the building block to construct an MH-SRU by removing the dependency on the hidden states of an MH-LSTM, and high-order connections are directly applied on inputs instead. A WaveNet block is further used to model the high-order input connections, which makes use of dilated causal convolution to provide a larger receptive field, and is able to capture wider feature contexts with fewer layers.

2. The Model

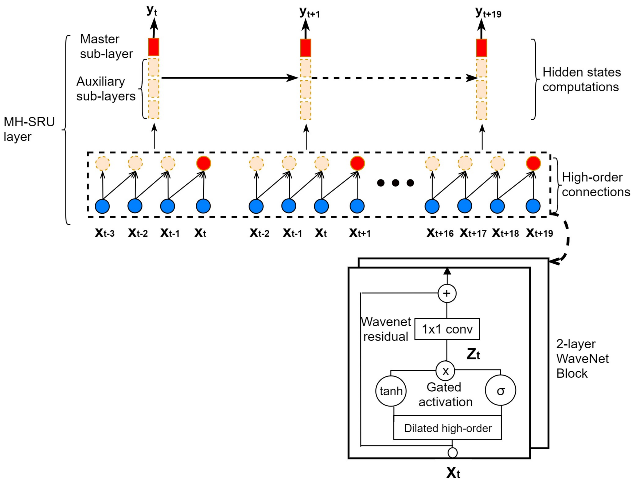

Figure 2: *Architecture of a WaveNet MH-SRU with 4 sub-layers. The original high-order connections in an MH-SRU layer are modeled with a 2-layer WaveNet block.*

2.1. Overview

WaveNet MH-RNN follows the general design of MH-RNN [15] and groups the hidden states into sub-layers, where the signals run with different time lags, with the exception of the top sub-layer, which runs with no time lag. The top sub-layer is also called the *master sub-layer*, and is directly connected to the output layer. The auxiliary sub-layers are arranged in the order of increasing time lags. All of the sub-layers are initialized differently so that each sub-layer represents a different history of the inputs and these histories are used as inputs for the next layer in a deep WaveNet MH-SRU model.

In this paper, the basic building block of an MH-RNN is an SRU. An MH-RNN that is composed of SRU is called MH-SRU. SRU parallelizes most computations of an MH-RNN by removing the dependency on the hidden states, and high-order connections can be directly applied to the inputs of each MH-SRU layer alone. Furthermore, we propose modeling the high-order connections in MH-SRU by a WaveNet block. The dilated causal convolutional layers where the convolution is performed along the time axis of the data in a WaveNet block effectively implement the high-order causal connections in the MH-SRU. In addition, the progressive dilations in a WaveNet block provide a larger receptive field of the input to each sub-layer (including the input layer). For example, a single-layer time-unfolded 2nd-order MH-SRU shown in Fig. 2 has input features from the past two time steps for each sub-layer. In a 2nd-order WaveNet MH-SRU using 2-layer WaveNet blocks having dilation of one for the first layer and dilation of two for the second layer will make use of input features from the past four time steps for each sub-layer.

2.2. Multiple-history simple recurrent unit (MH-SRU)

When LSTM is used to implement MH-RNN, it runs slowly with the computation bottleneck at the gates. SRU [14] simplifies the architecture of LSTM and increases its running speed by dropping the connections between its internal hidden states so that the majority of computation for each step can be done in parallel. In this paper, we accelerate the computation of our previous MH-LSTM by replacing the LSTMs with the simpler and faster SRUs. Below are the updating formulas of the m-th sub-layer of a p-th order MH-SRU which runs with n histories where $1 \leq m \leq n$ and the master sub-layer has the index 1.

$$\left[\hat{\mathbf{r}}_t^{(m)}, \hat{\mathbf{f}}_t^{(m)}, \hat{\mathbf{c}}_t^{(m)}\right] = \sum_{k=0}^{p-1} \mathbf{W}_{x(k)} \mathbf{x}_{t-k-m+1} + \mathbf{b} \quad (1)$$

$$\mathbf{r}_t^{(m)} = \sigma(\hat{\mathbf{r}}_t^{(m)}) \quad (2)$$

$$\mathbf{f}_t^{(m)} = \sigma(\hat{\mathbf{f}}_t^{(m)}) \quad (3)$$

$$\mathbf{c}_t^{(m)} = \mathbf{f}_t^{(m)} \odot \mathbf{c}_{t-1}^{(m)} + (1 - \mathbf{f}_t^{(m)}) \odot \hat{\mathbf{c}}_t^{(m)} \quad (4)$$

$$\mathbf{h}_t^{(m)} = \mathbf{r}_t^{(m)} \odot tanh(\mathbf{c}_t^{(m)}) + (1 - \mathbf{r}_t^{(m)}) \odot \mathbf{x}_t \quad (5)$$

$$\mathbf{y}_t = softmax(\mathbf{W}_y \mathbf{h}_t^{(n)}) \quad (6)$$

where $\mathbf{x}_t$ is the input at time t; $\mathbf{r}_t^{(m)}$, $\mathbf{f}_t^{(m)}$ and $\mathbf{c}_t^{(m)}$ are the reset gate output, the forget gate output and the memory cell output, respectively; $\mathbf{W}_{x(k)}$ and $\mathbf{W}_y$ are the weight matrices; $\mathbf{b}$ is the gate bias vector; $\mathbf{h}_t^{(m)}$ is the hidden state output; $\mathbf{y}_t$ is the softmax output with respect to the context-dependent HMM state label at time t; any quantity with a ˆ is the activation value of the quantity before an activation function is applied; $\odot$ is the element-wise multiplication operation; σ is the logistic sigmoid function.

2.3. WaveNet MH-SRU

WaveNet MH-SRU, as an extension to MH-SRU, aims at better capturing more information from input sequence or multiple histories. One key component of the WaveNet MH-SRU is the dilated high-order connections. It effectively implements a high-order MH-SRU. It is inspired by the dilated causal convolution used by WaveNet. Another advantage of this component is that it can expand the receptive field of each node exponentially while the number of model parameters only grows linearly. In our paper, there are only 2 layers in each WaveNet block and the dilation factors are one and two for the two layers.

Another component of WaveNet MH-SRU is the gated activation function [17]. We use it to produce the dilated high-order outputs from the WaveNet blocks. To be specific, the intermediate output of each layer $\mathbf{Z}_t$ of a stacked WaveNet block in WaveNet MH-SRU is given by

$$\mathbf{Z}_t = \tanh(\mathbf{W}_{xf} * \mathbf{X}_t) \odot \sigma(\mathbf{W}_{xg} * \mathbf{X}_t) \;, \quad (7)$$

where $\mathbf{X}_t = [\mathbf{x}_{t-n+1}, \mathbf{x}_{t-n+2}, ..., \mathbf{x}_t]$ is the input feature sequence; $\mathbf{W}_{xf}$ and $\mathbf{W}_{xg}$ are the convolutional weights for the normal filters and the gates, respectively; $*$ denotes the convolution operator which is performed along the time axis and $\odot$ denotes an element-wise multiplication operator. This activation function produces a weighted output vector by gating the convolution output and is found helpful for noisy inputs because it allows $\mathbf{Z}_t$ to adapt to the input data based on the learned gating values.

A WaveNet MH-SRU stacks a number of dilated convolutional layers in each WaveNet block which effectively increase the receptive field of the model. As a result, during training, for the same number of truncated BPTT time steps, the structure of a WaveNet MH-SRU can be much deeper than that of a conventional LSTM because the total number of layers in the time-unfolded neural network is the product of the number of BPTT time steps and the number of dilated convolutional layers in each WaveNet block. This poses a greater challenge to learning the model due to the gradient vanishing problem. A common solution is the introduction of residual connections in a very deep neural networks [13] so that the gradients can be directly passed back to lower layers. In our WaveNet MH-SRU, there are two types of residual connections. The first type is the WaveNet residuals inside a WaveNet block as shown in Fig. 2. The other type is the highway connections between the input and output layers of each WaveNet MH-SRU layer as shown in Eqn. (5).

3. Experiments

The proposed WaveNet MH-SRU was evaluated on two noisy corpora: phoneme recognition on the small-vocabulary NTIMIT [18], and word recognition on the medium-vocabulary CHiME-2 [19].

3.1. Data sets

3.1.1. The NTIMIT speech corpus

The NTIMIT database was collected by transmitting all 6300 original TIMIT [20] utterances over various channels in the NYNEX telephone network [18]. All utterances are time-aligned with the original clean TIMIT utterances. Therefore, the data preparation procedure is the same as that of TIMIT. It is worth noting that there are two major differences between TIMIT and NTIMIT data sets: (i) NTIMIT is noisier than TIMIT with a signal-to-noise ratio (SNR) of 25dB vs. TIMIT's 40dB; (ii) NTIMIT consists of narrowband speech while TIMIT consists of wideband speech because the spectral energy above 3.5 kHz in the original TIMIT utterances are greatly reduced due to the telephone channel. Phoneme recognition was performed using Viterbi decoding with a phone bigram language model estimated from the TIMIT training transcriptions using the Kaldi toolkit [21]. The phoneme recognition performance will be reported in terms of phoneme error rate (PER).

3.1.2. The CHiME-2 speech corpus

The CHiME-2 corpus is a medium-vocabulary corpus, which was generated by convolving clean Wall Street Journal (WSJ0) utterances with binaural room impulse responses, and adding real background noise at SNRs in the range of [-6, 9]dB. The training set contains 7138 simulated noisy utterances from 83 speakers. The transcriptions are the same as those of the original WSJ0 training set. The development and test sets contain 2460 and 1980 simulated noisy utterances spoken by 10 and 8 speakers, respectively. The WSJ0 text corpus is used to train a trigram language model, which consists of 37M words from 1.6M sentences. The speech recognition performance will be reported in terms of word error rate (WER).

3.2. Feature extraction and model training procedure

Different acoustic hidden Markov models (HMM) were built with states being modeled by Gaussian-mixture model (GMM), LSTM RNN, MH-LSTM RNN, MH-SRU RNN or WaveNet MH-SRU RNN.

The GMM-HMM models for both NTIMIT and CHiME-2 were trained with 39-dimensional fMLLR-adapted MFCC features using the standard Kaldi recipes for the tasks. There were 1940 and 1928 context-dependent tied-states for NTIMIT and CHiME-2, respectively.

All neural network acoustic models of NTIMIT were trained with 39-dimensional MFCC features, whereas all neural network acoustic models of CHiME-2 were trained with 40-dimensional mel-filterbank coefficients without their derivatives. Per-speaker mean and variance normalizations were performed on the inputs to all neural network models. The GMM-HMMs were used to derive the state targets for the subsequent DNN and RNN training through forced alignment. For CHiME-2, the state targets for subsequent RNNs training were further obtained by aligning the noisy training data with its DNN acoustic model through the iterative procedure outlined in [22].

Based on the optimal configurations of LSTM/MH-LSTM found in [15], we have the following basic setups: all basic LSTM models had 3 hidden layers with 512 nodes per layer, and the inputs consisted of 5 contextual frames. We define the model order as the number of high-order connections to each hidden node of any MH-RNN model, and the model order of all MH-RNN models (MH-LSTM, MH-SRU, each layer of WaveNet block of WaveNet MH-SRU) was set to 5 in this paper. For all RNN models other than the basic LSTMs, the number of hidden layers and the number of histories were varied from 3–12 and 21–101, respectively, to find the best setting for each task. In the end, all the MH-RNN models had 256 hidden nodes in each sub-layer.

All RNN training codes were developed using Theano by ourselves. All RNNs were trained by optimizing the categorical cross entropies using BPTT and SGD. The learning rate for LSTM/SRU and MH-LSTM/MH-SRU/WaveNet MH-SRU models was initially set to 0.25 and 0.1, respectively, and it decayed after each iteration until it went below 10^{-6}.

Table 1: *NTIMIT phoneme error rate (PER %). L: number of layers; N: number of nodes per layer; H: number of histories; P: number of model parameters.*

Model	L	N	H	P	PER %
LSTM	3	512	1	10M	31.2
MH-LSTM	3	256	1	6M	30.1
SRU	3	512	1	5M	30.6
SRU	6	512	1	7M	30.4
SRU	9	512	1	10M	30.0
MH-SRU	6	256	21	4M	31.2
MH-SRU	9	256	21	6M	29.6
WaveNet MH-SRU	6	256	101	13M	**29.1**

3.3. Results

3.3.1. NTIMIT

Table 1 shows the NTIMIT phoneme recognition performance of the baseline models and the new WaveNet MH-SRU models, and the effectiveness of using multiple histories in the new models. Firstly, it can be seen that a 3-layer MH-LSTM with 21 histories outperforms an LSTM by 1.1% absolute. SRUs show better performance with increasing number of hidden layers and a 9-layer SRU with only single history achieves similar performance of a 3-layer MH-LSTM. We can also find that MH-SRU requires a deeper structure to achieve better performance than MH-LSTM, and it outperforms MH-LSTM or SRUs for the same number of model parameters. Finally, if more model parameters can be utilized, a 6-layer WaveNet MH-SRU gives the best phoneme recognition performance with 6 histories. If we compare the various models with approximately 10M parameters, we find that a 6-layer WaveNet MH-SRU outperforms a 9-layer MH-SRU by 0.5% absolute, a 9-layer SRU by 0.9% absolute, and a 3-layer LSTM by 2.1% absolute, respectively.

Table 2: *Effect of the number of histories H on NTIMIT phoneme recognition (PER %).*

H	MH-SRU	WaveNet MH-SRU
21	**29.6**	29.7
61	30.0	29.6
101	30.2	**29.1**

To validate the effect of the number of histories on MH-SRU and WaveNet MH-SRU, we selected two models with around 10M parameters, namely a 9-layer MH-SRU and a 6-layer WaveNet MH-SRU, and checked their NTIMIT performance by varying their number of histories from 21 to 101. As can be seen from Table 2, the performance of MH-SRU constantly degrades as the number of histories increases, whereas the performance of WaveNet MH-SRU improves steadily. These results suggest that as the topology of the MH-SRU is more complex with longer feature contexts and more histories, the model was under-trained with the limited amount of data in NTIMIT. On the other hand, the WaveNet MH-SRU can learn more useful information from a bigger feature context and a larger number of histories probably due to its use of dilated convolutions.

Table 3: *Average WER (%) on the CHiME-2 test set. L: number of layers; N: number of nodes per layer; H: number of histories; P: number of model parameters.*

Model	L	N	H	P	WER %
DNN	7	2048	0	30M	29.2
RDNN [23]	7	2048	0	-	27.7
LSTM	3	512	1	10M	27.8
SRU	12	512	1	12M	27.6
SRU	12	1024	1	43M	26.6
MH-SRU	9	256	21	7M	26.4
WaveNet MH-SRU	6	256	101	13M	**25.8**

Table 4: *Detailed WER (%) on each noisy condition of the CHiME-2 test set.*

Model	-6 Db	-3 Db	0 Db	3 Db	6 Db	9 Db
DNN	48.3	37.9	29.4	23.9	18.8	16.7
LSTM	45.1	34.8	28.8	22.3	19.2	16.7
SRU	42.6	**33.0**	27.7	22.0	18.4	15.8
MH-SRU	42.2	33.2	27.8	22.2	18.2	14.9
WaveNet MH-SRU	**41.5**	**33.1**	**26.8**	**21.8**	**16.9**	**14.8**

3.3.2. CHiME-2

The proposed WaveNet MH-SRU was also evaluated and compared with other models on the noisy CHiME-2 task. Their average recognition performances over test data of different SNRs are shown in Table 3. Again we see that their performances are ranked in the following order:

$$\text{WaveNet MH-SRU} > \text{MH-SRU} > \text{SRU} > \text{LSTM} > \text{DNN}.$$

Specifically, our baseline LSTM and SRU both having 512 hidden nodes per layer and around 10M parameters give similar performance. These results are comparable to the result from [23] which used recurrent deep neural network (RDNN). If we increase the number of hidden nodes in SRU to 1024, the SRU performs better than the baseline LSTM by 1% absolute and approaches the performance of the MH-SRU at the expense of much more model parameters — 6 times more parameters. The new WaveNet MH-SRU performs the best with similar number of model parameters. In summary, for all models with approximately 10M model parameters, the WaveNet MH-SRU outperforms LSTM, SRU and MH-SRU by 2% and 1.8% and 0.6% absolute, respectively.

Table 4 reports the detailed WERs for each test dataset of different SNRs in CHiME-2. The results confirm that WaveNet MH-SRU performs better than all other models at (basically) all SNRs. It again suggests that the effective utilization of the larger number of histories by the WaveNet MH-SRU lends itself to the model's robustness against noises.

4. Conclusions and future work

We introduce a novel model called *WaveNet multiple-history SRU* (WaveNet MH-SRU), which is an extension to MH-LSTM using WaveNet block and SRU. Through a series of carefully designed experiments using NTIMIT and CHiME-2, we show that WaveNet MH-SRU outperforms SRU and MH-SRU by effectively utilizing more histories from the input signal. It is worth noting that although WaveNet MH-SRU makes use of multiple histories, its number of model parameters is comparable to other RNN models. However, it runs more slowly due to the more possible paths in the model. In the future, we will investigate how to speed up its computations.

5. Acknowledgments

The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKUST16215816). We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

6. References

[1] L. V. Fausett, *Fundamentals of neural networks*. Prentice-Hall, 1994.

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

[3] A. Graves, A. r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2013, pp. 6645–6649.

[4] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Proceedings of Interspeech*, 2010, pp. 1045–1048.

[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1409.0473

[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2015.

[7] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[8] T. Lin, B. G. Horne, P. Tino, and C. L. Giles, "Learning long-term dependencies in NARX recurrent neural networks," *IEEE Transactions on Neural Networks*, vol. 7, no. 6, pp. 1329–1338, 1996.

[9] S. El Hihi and Y. Bengio, "Hierarchical recurrent neural networks for long-term dependencies," in *Proceedings of Advances in Neural Information Processing Systems*, 1996, pp. 493–499.

[10] J. Koutník, K. Greff, F. Gomez, and J. Schmidhuber, "A clockwork RNN," in *Proceedings of the International Conference on Machine Learning*, 2014, pp. 1863–1871.

[11] P. Liu, X. Qiu, X. Chen, S. Wu, and X. Huang, "Multi-timescale long short-term memory neural network for modelling sentences and documents," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Lisbon*, 2015, pp. 2326–2335.

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

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[14] T. Lei and Y. Zhang, "Training RNNs as fast as CNNs," *arXiv preprint arXiv:1709.02755*, 2017.

[15] H. Huang and B. Mak, "To improve the robustness of LSTM-RNN acoustic models using higher-order feedback from multiple histories," in *Proceedings of Interspeech*, 2017.

[16] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[17] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves *et al.*, "Conditional image generation with pixelCNN decoders," in *Advances in Neural Information Processing Systems*, 2016, pp. 4790–4798.

[18] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1990, pp. 109–112.

[19] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 126–130.

[20] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, August 1990.

[21] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2011.

[22] 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.

[23] C. Weng, D. Yu, S. Watanabe, and B.-H. F. Juang, "Recurrent deep neural networks for robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2014, pp. 5532–5536.