IOP Conf. Series: Materials Science and Engineering **719** (2020) 012027 doi:10.1088/1757-899X/719/1/012027

Single Channel Speech Enhancement Algorithm based on BLSTM-DNN Bidirectional Optimized Hybrid Model

Xiaoyue Sun[1], Ruwei Li[1], Tao Li[1] and Dengcai Yang[2]

1. Department of Information and Communication Engineering, University of Beijing Technology, Faculty of Artificial Intelligence, Science building, room 611, Beijing, China;
2. Institute of Science and Technology Development, Beijing University of Technology, Knowledge and Practice building, room 415, Beijing, China.
Email: li_stu_public@163.com

Abstract. The performance of existing speech enhancement algorithms based on deep learning is not ideal in complex noise environment. To improve the problem, a bidirectional optimized hybrid network named BLSTM-DNN is constructed based on bidirectional long-short term memory (BLSTM) network and fully-connected deep neural network (DNN). This structure uses BLSTM to extract high-level information including past and future temporal context of noisy speech. Next, fully-connected DNN fits the high-level information to ideal ratio mask (IRM). Finally, the IRMs estimated by the BLSTM-DNN are used to enhance the noisy speech. Experimental results show that the proposed method can effectively improve the speech quality and intelligibility under unknown noise conditions.

1. Introduction

Recently, with the development of deep learning technology, deep neural networks have been widely used in the field of speech enhancement due to their simple network structure, powerful modeling capabilities and good fitting effects [1]. For example, Wang et al. [2] learn a variety of complementary features under the DNN framework to further improve the performance of speech enhancement. Li et al. use DNN to achieve noise classification and denoising [3]. However, although contextual information can be incorporated into the training of the DNN through frame expansion [2], that cannot be represented by the DNN in the long-term acoustic model. Recurrent neural networks (RNNs) can correctly represent that relationship between the previous frame and the current frame [4]. But because of signal internal structure and the lack of nonlinear activation function, the gradient of RNN will gradually vanish and explode after multi-stage propagation, which makes RNNs difficult to capture the long-term contextual information [5]. Long short-term memory networks may alleviate this problem by introducing a series of gates and the concepts of memory cell to dynamically control the flow of information, which makes the networks have long-term memory [6]. In order to access past and future context information, Graves et al. [7] propose a Bidirectional long-short term memory (BLSTM) networks based on LSTM, which integrate forward and backward contextual information into the model to simulate the timing variation of speech and noise over a longer period of time [8] and improve the speech enhancement performance. However, more network parameters and lower fitting efficiency result in the BLSTM model not easily converging in speech enhancement tasks [9].

In this paper, a novel speech enhancement algorithm named BLSTM-DNN is proposed based on the advantage of BLSTM and DNN to capture the contextual information a while improving the matching efficiency of the networks, which combined the advantages of BLSTM and DNN to jointly optimize them. Bidirectional context information for noisy speech is extracted by BLSTM, which is used by DNN to fit the training target. The effectiveness of the proposed algorithm is verified by experiments from the aspects of enhanced speech quality and intelligibility as well as the waveform and spectrogram.

2. Bidirectional Optimized Speech Enhancement

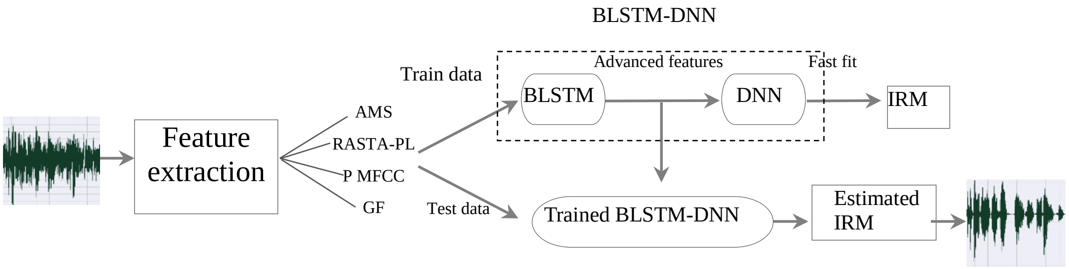


Figure 1. Schematic diagram of the proposed speech enhancement algorithm.

The proposed system is shown in figure 1. In the training phase, the noisy speech is decomposed into time-frequency units, and the Mel-frequency cepstral coefficients (MFCC), amplitude modulation spectrogram (AMS), relative spectral transform and perceptual linear prediction (RASTA-PLP) and gammatone frequency (GF) are extracted from the time-frequency units as the inputs of BLSTM-DNN [10]. High-level abstraction information of noisy features is extracted by BLSTM, which is used to fit the IRM by DNN. In the enhancement phase, the complementary features of the test noisy speech are input into the trained BLSTM-DNN to estimate IRMs. It is used to weight the noisy speech to synthesize the enhanced speech.

2.1.BLSTM-DNN Hybrid Model Construction

In order to further incorporate past and future context into model to simulate the temporal changes of speech as well as noise in a longer time range and improve the fitting ability of the network, the BLSTM-DNN bidirectional optimized hybrid model is constructed. The topology of the model is shown in figure. 2. The input layer $x = (x_1, x_2, \cdots, x_T)$ is a frame-level complementary feature vector, where $T = 246$ represents the feature vector dimension. The first three layers are BLSTM layers, each of which consists of a forward LSTM layer and a backward LSTM layer. The remaining two layers are fully connected DNN layers.

2.2.BLSTM-based Bidirectional Optimization

Figure 2 shows that the BLSTM layer scans the input sequence in two opposite directions and connects to the same output layer. The output layer is updated by the backward hidden layer from t=T iteration to 1 and the forward hidden layer from t=1 iteration to T, which is used to calculate the forward hidden sequence, the backward hidden sequence and the output sequence. That bidirectional optimization process can further simulate the temporal context between speech and noise. This process is implemented by the following formulas:

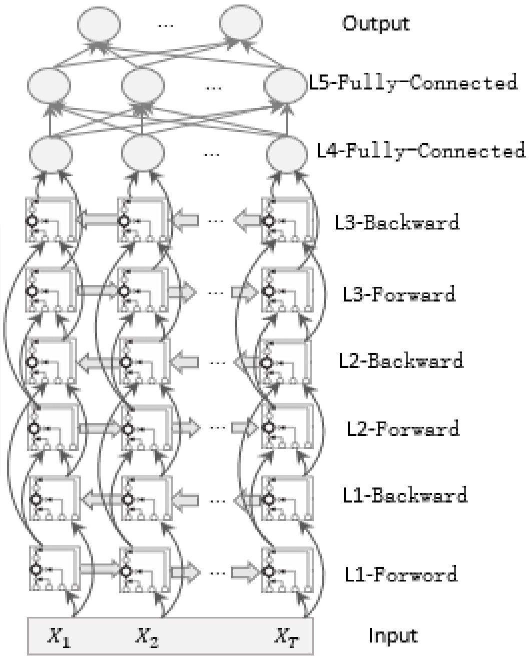
$$\overrightarrow{h_t} = H\left(W_{x\vec{h}} x_t + W_{\vec{h}\vec{h}} h_{t-1} + b_{\vec{h}}\right) \tag{1}$$

$$\overleftarrow{h_t} = H\left(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}\right) \tag{2}$$

$$y_t = W_{\overleftarrow{h}y} \vec{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y \tag{3}$$

Where $\vec{h} = (\overrightarrow{h_1}, \cdots, \overrightarrow{h_t}, \cdots, \overrightarrow{h_T})$ denotes the forward hidden layer sequence. $\overleftarrow{h} = (\overrightarrow{h_1}, \cdots, \overrightarrow{h_t}, \cdots, \overrightarrow{h_T})$ denotes the backward hidden layer sequences. $y^l = (y_1, \cdots, y_t, \cdots, y_T)$ indicates the output sequence after bidirectional optimization of the lth (l=1, 2, 3) BLSTM layer. H denotes the activation function of hidden state. W_{xh}, W_{hh} and W_{hy} represent the weight matrix of the input layer-hidden layer, the hidden layer-hidden layer and the hidden layer-output layer, respectively. $b_{\vec{h}}$, $b_{\overleftarrow{h}}$ and b_y represent the bias of the forward and backward hidden sequences and the offset of the output vector, respectively. $t = (1,2,\cdots,T)$ indicates the time index in the iterative process.

Figure 2. Topology diagram of the BLSTM-DNN



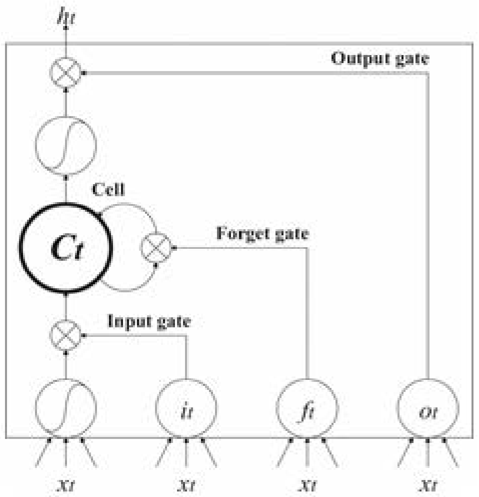


Figure 3. Block diagram of a single memory module in BLSTM

The forward hidden sequence $\vec{h}$ and the backward hidden sequence $\overleftarrow{h}$ for the BLSTM layer are obtained by optimizing the cell state c_t in each memory module (shown in figure 3) as follows:

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

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

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

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_0) \tag{7}$$

$$h_t = o_t \cdot tanh(c_t) \tag{8}$$

Where i_t, f_t, o_t, c_t represent the input gate, the forgetting gate, the output gate and the cell, respectively. The dimension of them is the same as the hidden layer vector h_t. W_{hi} denotes the hidden layer-input gate weight matrix. W_{xo} denotes the input layer-output layer weight matrix. c_t is the cell state at time t. h_t is the hidden layer state at timet, and $\sigma(\cdot)$ is the sigmoid function.

Table 1. Test results of PESQ

Noise	SNR (dB)	DNN	BLSTM	BLSTM-DNN
Babble	10	3.026	3.442	3.483
	5	2.765	3.036	3.188
	0	2.425	2.816	2.841
	-5	1.997	2.467	2.498
White	10	2.832	3.019	3.106
	5	2.659	2.778	2.821
	0	2.316	2.420	2.445
	-5	1.913	2.041	2.058

2.3.DNN-based Fitting Process

The formula for network iteration of the fully-connected DNN layers ($l = 4{,}5$) is as follows:

$$y^l = \varphi\left(W^l y^{l-1} + b^l\right) \tag{9}$$

Where y^l, W^l and b^lrepresent the output of the activation units, weight matrix and the bias of lth DNN layer, respectively. φ represents the ReLU activation function. The function of fully connected DNN layers is efficiently fit the training target through gradient descent and error backpropagation.

Table 2. Test results of STOI

Noise	SNR (dB)	DNN	BLSTM	BLSTM-DNN
Babble	10	0.944	0.9509	0.9544
	5	0.891	0.9004	0.9155
	0	0.806	0.8112	0.8452
	-5	0.655	0.7336	0.7376
White	10	0.957	0.9580	0.9597
	5	0.921	0.9216	0.9216
	0	0.862	0.8701	0.8715
	-5	0.788	0.7957	0.8016

Table 3. Test results of LSD

Noise	SNR (dB)	DNN	BLSTM	BLSTM-DNN
Babble	10	6.312	4.065	4.062
	5	7.319	4.567	4.543
	0	8.952	6.792	6.514
	-5	11.634	7.540	6.929
White	10	7.821	4.239	4.184
	5	8.932	4.948	4.283
	0	10.271	5.418	5.405
	-5	11.700	6.413	5.552

3. Experimental Results and Analysis

3.1.Experimental Data

288 clean utterances of the experiments come from the NTT corpus. 100 noise types are from TIMIT database in training stage to improve the generalization capacity of unseen environments. This paper selects 200 utterances, which are corrupted with each noise type at 10dB,5dB, 0dB and -5dB to build training set. The remaining 88 utterances are used to construct the test set for each combination of noise types and SNR levels. Babble, white and factory2 from the NOISEX-92 corpus are used to evaluate on unseen noise types. All the signals are sampled at 16kHz rate. The frame length and shift are 320 and 160 samples, respectively. In order to verify the effectiveness of the proposed algorithm, we select on complementary features speech enhancement model based on DNN [2] as the first contrast algorithm, and complementary features speech enhancement model based on BLSTM as the second contrast algorithm. Both DNN and BLSTM model have 4 hidden layers with 1024 nodes for each layer. The Dropout ratio is 0.2, and the number of iterations is set to 50.

3.2.Speech Enhancement Performance Comparison

3.2.1.*Objective performance evaluation.* The test results of Segmental Signal to Noise Ratio (SegSNR), Evaluation of Speech Quality (PESQ), Log-Spectral Distortion (LSD) and Short-time Objective Intelligibility (STOI) are shown in table 1, table 2, table3 and table4, respectively. It can be seen from the tables that the objective test results of the BLSTM-DNN model are better than the BLSTM and

DNN models. Because BLSTM can further incorporate bidirectional temporal context and simulate the temporal variation of speech and noise in a longer time, and makes the model have the high-efficiency fitting ability that the BLSTM model does not.

Table 4. Test results of SegSNR

Noise	SNR (dB)	DNN	BLSTM	BLSTM-DNN
Babble	10	0.864	2.272	2.377
	5	-1.951	0.622	1.967
	0	-6.409	-4.275	1.262
	-5	-14.845	-10.063	-8.509
White	10	3.503	3.942	4.411
	5	1.167	1.354	2.208
	0	-0.870	-0.982	0.074
	-5	-3.237	-2.510	-2.420

3.2.2. Comparison of waveform and spectrogram. In order to compare the speech enhancement effects of DNN, BLSTM and BLSTM-DNN visually, figure 4 shows the waveform and spectrogram of the proposed algorithm and the contrast algorithm with factory2 noise at SNR = -5dB. We can know from the spectrogram that the speech enhanced by BLSTM-DNN is more similar with pure speech, because BLSTM layers could build long-term dependence of noisy speech to better preserve the temporal context information. Besides, the speech enhanced with BLSTM-DNN has less background noise, especially in the non-speech segment. That indicates that the fully connected DNN could efficiently fit speech information processed by BLSTM layers to the training target to achieve better denoising performance. The experimental results show that BLSTM-DNN has better denoising performance than BLSTM and DNN, which is consistent with the objective test results.

4. Conclusion

The proposed algorithm can bi-directionally model the temporal context information of noisy speech, and simulate the timing changes of speech and noise in a longer period through BLSTM. In addition, the high-level information extracted by the BLSTM quickly fits the IRM by using the efficient fitting ability of the DNN. It ensures that the deep neural network has a good fitting ability while modeling the temporal relationship of the speech. Compared with the comparison algorithm, the proposed algorithm improves the quality and intelligibility in unseen noise conditions.

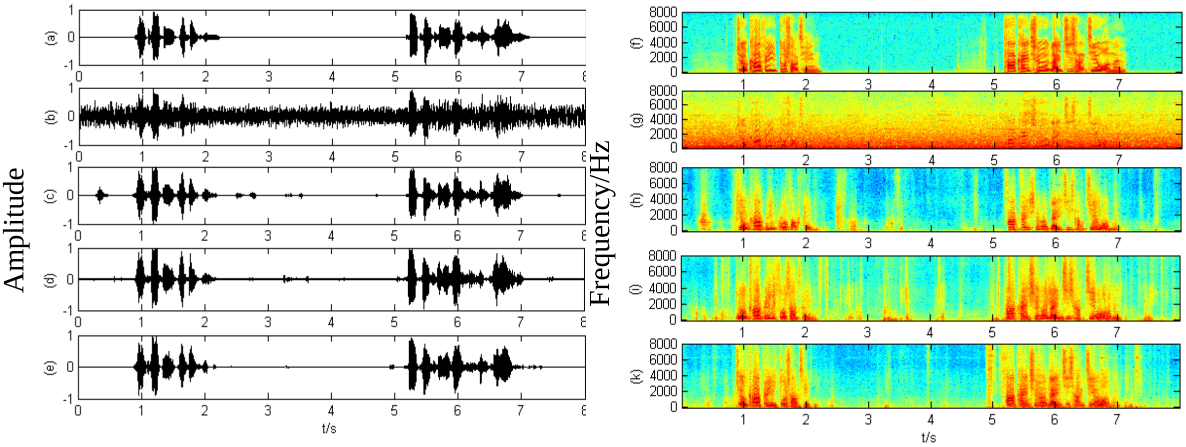


Figure 4. Speech enhancement effect samples with -5dB factory2 noise
(a),(f)Waveform and spectrogram of clean speech (b),(g)waveform and spectrogram of noisy speech(c),(h) waveform and spectrogram of speech with DNN (d),(i) waveform and spectrogram of speech with BLSTM (e),(j) waveform and spectrogram of enhanced speech with BLSTM-DNN

5. Acknowledgments

This work was supported by Beijing Municipal Natural Science Foundation Cooperation Beijing Education Committee (No. 201910005007) and the National Natural Science Foundation of China (No. 61871007).

6. Reference

[1] Li R, Sun X, Liu Y, et al. Multi-resolution auditory cepstral coefficient and adaptive mask for speech enhancement with deep neural network[J]. EURASIP Journal on Advances in Signal Processing, 2019, 2019(1): 22.
[2] Wang Y, Narayanan A, Wang D L. On training targets for supervised speech separation[J]. IEEE/ACM transactions on audio, speech, and language processing, 2014, 22(12): 1849-1858.
[3] Li R, Liu Y, Shi Y, et al. ILMSAF based speech enhancement with DNN and noise classification[J]. Speech Communication, 2016, 85: 53-70.
[4] Mehri S, Kumar K, Gulrajani I, et al. SampleRNN: An unconditional end-to-end neural audio generation model[J]. arXiv preprint arXiv:1612.07837, 2016.
[5] Tang Z, Shi Y, Wang D, et al. Memory visualization for gated recurrent neural networks in speech recognition[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017: 2736-2740.
[6] Sun L, Du J, Dai L R, et al. Multiple-target deep learning for LSTM-RNN based speech enhancement[C]//2017 Hands-free Speech Communications and Microphone Arrays (HSCMA). IEEE, 2017: 136-140.
[7] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5-6): 602-610.
[8] Cheng G, Huang L, Sun J, et al. Bidirectional LSTM with Extended Input Context[C]//2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE, 2018: 364-368.
[9] Wang Q, Du J, Dai L R, et al. A multiobjective learning and ensembling approach to high-performance speech enhancement with compact neural network architectures[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 2018, 26(7): 1181-1193.
[10] Chen J, Wang Y, Wang D L. A feature study for classification-based speech separation at low signal-to-noise ratios[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1993-2002.