Applied Acoustics 157 (2020) 107019

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Speech enhancement based on simple recurrent unit network

Xingyue Cui, Zhe Chen, Fuliang Yin *

School of Information and Communication Engineering, Dalian University of Technology, Dalian 116023, China

ARTICLE INFO

Article history:
Received 6 April 2019
Received in revised form 23 July 2019
Accepted 31 August 2019

Keywords:
Speech enhancement
Deep neural network
Simple recurrent unit
Power spectra

ABSTRACT

Speech enhancement is a crucial and challenging task in many applications. A novel speech enhancement method based on the simple recurrent unit (SRU) is proposed in this paper. First, the log-power spectra of noisy and clean speeches are extracted. Then, the mapping relationship between noisy and clean speech spectra is learned by a multiple-layer stacked SRU network. Finally, the well-trained model is used to predict the corresponding clean speech spectra from the noisy speech spectra and the whole clean speech waveform can be recovered. Compared with the existing algorithms, DNN, LSTM and GRU, the proposed method achieves significant improvements at training speed and has capability to balance the performance and the training time. Experimental results demonstrate the validity and robustness of the proposed method.

1. Introduction

Speech enhancement has attracted considerable attention in both academia and industry communities. For many applications, typically speech recognition, hearing prosthesis and telecommunication, the received speech signals are always corrupted due to background noise, which results in the deterioration of speech quality and intelligibility. Hence, speech enhancement as a powerful speech pre-processing is very necessary.

Some approaches for speech enhancement have been developed. One of the most classic techniques is the spectral subtraction (SS). The core of spectral subtraction is to subtract an estimate of the noise power spectra from the noisy signal spectra to obtain the clean signal [1–3]. Another technique for speech enhancement is the Wiener filtering [4,5], which requires an estimate of the priori signal-to-noise ratio (SNR) [6]. Subspace method proposed by Ephraim and Trees [7] is also a common approach, and its improved versions can be seen in [8,9]. Moreover, the parametric or statistical model based algorithms are also widely used in speech enhancement. Refs. [10,11] presented the model based minimum mean-square error (MMSE) estimators and maximum a posteriori (MAP) estimators, respectively. Ephraim [12] utilized the hidden Markov model (HMM) to model the dynamics of speech and noise processes. Kundu et al. [13] developed an MMSE estimator based on the Gaussian mixture model (GMM) for a clean speech signal. In [14], a Bayesian non-negative matrix factorization (NMF) approach was proposed to obtain the MMSE estimate of the magnitude spectra of clean speech. Considering the interframe correlation, a new linear model for speech spectra estimation was proposed, then extended to multi-frame optimal filters [15]. In [16], an iterative longest matching segment approach was presented to further improve the performance.

Although the above-mentioned algorithms have capability to handle background noise, they are limited in non-stationary noise, and are cumbersome and complicated. Therefore, to cope with non-stationary noise, more recent studies perform speech enhancement in a supervised manner. In 2012, deep learning was introduced for speech separation by Wang and Wang [17,18]. They used a feedforward DNN with restricted Boltzmann machine (RBM) pre-training for subband classification to estimate the ideal binary mask (IBM), and extended it to a two stage DNN [19], where the first stage is trained to estimate the subband IBM and the second stage explicitly incorporates the time–frequency context. In [20], ideal ratio mask (IRM) was used to replace IBM to promote speech intelligibility. Another network for noise reduction is deep autoencoder (DAE). A basic autoencoder (AE) is an unsupervised learning machine, which has a symmetric architecture with one hidden layer. Multiple AEs with greedy layer-wise pretraining can be stacked into a DAE, which subjects to the traditional supervised fine-tuning. In [21], Lu et al. proposed a DAE model that maps from the Mel-frequency power spectra of noisy speech to that of clean speech. Subsequently, in order to deal with non-stationary noise, Xu et al. [22,23] proposed a DNN network with the RBM pretraining. After training, DNN has the capability to estimate the clean speech from a noisy input. On this basis, in

* Corresponding author.
E-mail addresses: xiechoah@mail.dlut.edu.cn (X. Cui), zhechen@dlut.edu.cn (Z. Chen), flyin@dlut.edu.cn (F. Yin).

https://doi.org/10.1016/j.apacoust.2019.107019
0003-682X/

[24], a skip connection strategy was incorporated into DNN, and in [25], a phase-aware speech enhancement algorithm based on DNN was proposed to improve the enhancement performance. In addition, due to its strong regression capability, recurrent neural network (RNN) has also been utilized in speech enhancement. Weninger et al. proposed a RNN-LSTM network to tackle speech-related problems, including speech enhancement [26], speech separation [27] and speech recognition [28]. In [29], Sun et al. proposed an ensemble framework with multiple-target joint learning based on LSTM. Lee et al. [30] used a deep bi-directional LSTM structure to suppress wind noise. Moreover, speech enhancement techniques based on convolutional neural network (CNN) were also developed. Fu et al. [31,32] investigated an SNR-aware CNN model, then extended it to a fully convolutional network for speech enhancement. Qian et al. [33] proposed a very deep CNN to improve recognition accuracy for noise robust speech recognition. In [34], a two-channel beamforming approach based on the concatenation of CNN, LSTM and DNN networks was presented. Besides, there are some other potential methods to deal with speech enhancement. In [35], a deep stacking network (DSN) was proposed to handle the task of speech separation and pitch estimation simultaneously. In [36], speech enhancement at the phoneme level was studied. Recently, Pascual et al. [37] proposed a speech enhancement generative adversarial network (SEGAN), which gives a new perspective in handling enhancement problems. These existing neural network methods have good capability for noise suppression, but suffer from complex structure and long training time.

In order to decrease the training time of deep network while maintaining the performance of noise suppression, a novel SRU-based speech enhancement method is proposed in this paper to deal with non-stationary noise without any pre-training of noise models. The deep network consists of multiple stacked simple recurrent unit (SRU) [38] which is a recurrent architecture that trades off between training speed and performance enhancement. When using the proposed SRU network, the log-power spectra of noisy speech and clean speech are extracted as the input and output of the network respectively, then the network are trained for the sake of learning the nonlinear relationship between noisy speech and clean speech. Experiments are carried out to verify the performance and robustness of the SRU-based speech enhancement network.

The residue sections of this paper are organized as follows. Section 2 introduces the SRU architecture as well as the SRU-based speech enhancement network in detail. Section 3 presents a series of experiments to evaluate the performance of the proposed network under varying conditions. Finally, Section 4 concludes the paper.

2. SRU-based speech enhancement

At present, RNN based networks can achieve speech enhancement [26,29,30], but their computation load is large and the training time is long. Therefore, in this paper, a simplified recurrent architecture named SRU is used to cope with these problems. In this section, we describe the process of speech enhancement based on SRU network. First, the spectral features of speech are employed as network input and output. Then, the architecture and parallel principle of SRU are introduced. Finally, the whole structure of model is given in details, including the overall learning framework and the SRU network training procedure.

2.1. Spectral features

In general, when dealing with speech enhancement problems by means of neural network, log-power spectra of signals are often adopted as the input and output features of network. Given a time domain signal, set the corresponding frame length to N samples with a frame shift of $N/2$ samples. Then, a short-time Fourier transform is applied to each overlapping windowed frame. Here, the log-power spectra at the k-th frequency bin of the n-th frame is denoted as $X(n,k)$, and each frame can be described using a vector $\mathbf{x}(n)$ as follows

$$\mathbf{x}(n) = [X(n,0), X(n,1), \cdots, X(n,N)]^{\top} \tag{1}$$

In order to balance the continuity of speech signals and the data redundancy, the length of input features is expanded to multiple frames, which contains more acoustic context information from adjacent frames. Therefore, the input features for network are renewed as

$$\tilde{\mathbf{x}}(n) = [\mathbf{x}(n-l), \cdots, \mathbf{x}(n), \mathbf{x}(n+1), \cdots, \mathbf{x}(n+l)]^{\top} \tag{2}$$

where l is the number of adjacent frames on each side. Besides, similar to the input of network, the desired output of network is the log-power spectra in the current frame n. Here, it should be pointed out that all the input and output features are extracted from noisy and clean speeches, respectively and they are aligned on each frame.

2.2. Simple recurrent unit

In 2017, Lei et al. [38] proposed the simple recurrent unit (SRU), which is a simplified recurrent architecture to solve natural language tasks. Different from other common structures, for instance, Long Short-term Memory (LSTM) and Gated Recurrent Units (GRU), the majority of the computations in SRU can be easily parallelized since the calculations of each step are less depend on the completed previous ones. And then, the results of parallel computations are combined with the residual computations via a fast recurrent structure. Therefore, SRU has more capability to balance serial and parallelized computations than traditional recurrent units.

As described in [38], a single forget gate is the basic component of SRU. When given an input sequences $\mathbf{x}_t$ at time t, the linear transform $\tilde{\mathbf{x}}_t$ and the forget gate $\mathbf{f}_t$ can be calculated as

$$\tilde{\mathbf{x}}_t = \mathbf{W}\mathbf{x}_t \tag{3}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_f\mathbf{x}_t + \mathbf{b}_f) \tag{4}$$

where $\mathbf{W}$ and $\mathbf{W}_f$ are weight matrices, $\mathbf{b}_f$ is the bias term of forget gate, and $\sigma(\cdot)$ is the sigmoid function. Eq. (3) and Eq. (4) indicate that the computations about $\mathbf{x}_t$ and $\mathbf{f}_t$ only depend on $\mathbf{x}_t$, which makes it possible to parallelize across all time steps. Besides, the forget gate is used to modulate the internal state $\mathbf{c}_t$ as well, which calculates the output state $\mathbf{h}_t$ as follows

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + (1 - \mathbf{f}_t) \odot \tilde{\mathbf{x}}_t \tag{5}$$

$$\mathbf{h}_t = g(\mathbf{c}_t) \tag{6}$$

where $g(\cdot)$ is the nonlinear activation function for producing the output state $\mathbf{h}_t$.

The complete architecture of SRU also includes skip connections and highway connections. Besides, in order to combine the internal state $g(\mathbf{c}_t)$ and the input $\mathbf{x}_t$, a reset gate $\mathbf{r}_t$ similar to the forget gate is also added to compute the output state $\mathbf{h}_t$. The complete architecture of the SRU is described as:

$$\tilde{\mathbf{x}}_t = \mathbf{W}\mathbf{x}_t \tag{7}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_f\mathbf{x}_t + \mathbf{b}_f) \tag{8}$$

$$\mathbf{r}_t = \sigma(\mathbf{W}_r\mathbf{x}_t + \mathbf{b}_r) \tag{9}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + (1 - \mathbf{f}_t) \odot \tilde{\mathbf{x}}_t \tag{10}$$

$$\mathbf{h}_t = \mathbf{r}_t \odot g(\mathbf{c}_t) + (1 - \mathbf{r}_t) \odot \mathbf{x}_t \tag{11}$$

where $\mathbf{W}_r$ and $\mathbf{b}_r$ are weight matrix and bias term of reset gate, respectively.

From Eqs. (7)–(11), it can be found that the SRU completely drops the connection between the gating computations of step t and the states of step t-1, $\mathbf{h}_{t-1}$ and $\mathbf{c}_{t-1}$. Thus, when given a sequence of input vectors $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, the $\{\tilde{\mathbf{x}}_t, \mathbf{f}_t, \mathbf{r}_t\}$ for different $t = 1, 2, \cdots, n$ are independent and computed in parallel. More specifically, two optimizations are allowed for SRU formulation. First, matrix multiplications of all time steps can be batched, that is, the matrix multiplications in Eqs. (7)–(11) are grouped into a single batch, which is achieved as follows

$$\mathbf{U}^\top = \begin{pmatrix} \mathbf{W} \\ \mathbf{W_t} \\ \mathbf{W_f} \end{pmatrix} [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \tag{12}$$

where n is the sequence length, $\mathbf{U}$ is the resulting matrix of size $n \times 3d$ and d is the hidden state size. Second, all point-wise operations can be compiled into a single fused kernel and parallelized in the dimension of the hidden state. This optimization will decrease additional kernel launching latency and the time of data access.

From the above analysis, the computational complexity of SRU is presented as follows. When the input is a mini-batch of b sequences, the computational complexity of one SRU layer is $\mathcal{O}(nbd)$ [39]. In contrast, due to the hidden-to–hidden multiplications (e.g. $\mathbf{W}\mathbf{h}_{t-1}$), the computational complexity of one LSTM layer is $\mathcal{O}(nbd^2)$, and each dimension can not be independently parallelized. Similar to LSTM, the current computation of GRU is also depend on the previous output state $\mathbf{h}_{t-1}$, thus, the computational complexity of one GRU layer is $\mathcal{O}(nbd^2)$ as well. Obviously, SRU has much lower computational complexity than that of LSTM and GRU, which is consistent with the theory.

2.3. SRU-based speech enhancement

As described in Section 2.2, SRU is kind of efficient and fast recurrent network, which not only inherits the advantages of RNN network, but also achieves parallel computation. Thus, it is suitable for speech enhancement.

(1) Overall Learning Framework

Fig. 1 shows the overall procedure based on SRU network, which includes two stages: training and enhancement. It is a supervised process that multiple SRU layers are adopted to learn the mapping from noisy speech features to clean speech features. In the training stage, a SRU-based regression model is trained using a collection of speech data, which are composed of noisy and clean pairs by the log-power spectra features. In the enhancement stage, the noisy speech features are processed by the well-trained SRU-based model to predict the clean speech features, in which the estimated log-power spectra features of the obtained clean speech are defined as $\hat{X}^l(n,k)$. Then the reconstructed spectra $\hat{X}^r(n,k)$ can be calculated as

$$\hat{X}^r(n,k) = \exp\left\{\hat{X}^l(n,k)/2\right\} \exp\left\{j\angle Y^r(n,k)\right\} \tag{13}$$

where $\angle Y^r(n,k)$ denotes the kth phase of the nth frame from the original noisy speech. After above operations, a frame of clean speech is derived by IDFT from the current frame spectra and the whole waveform can be synthesized through the overlap-add method.

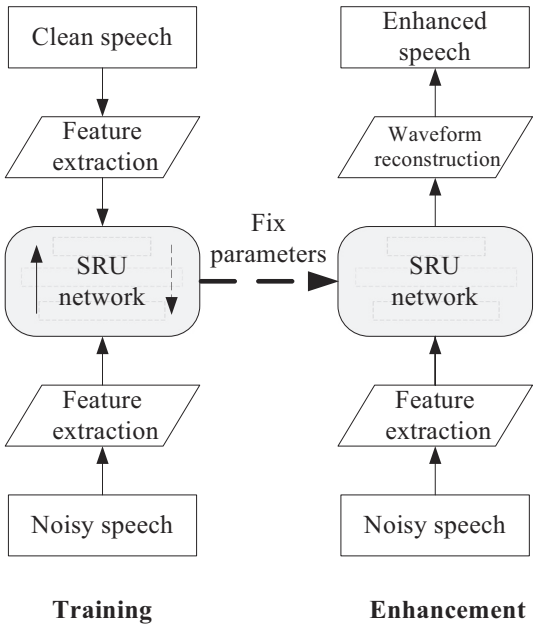

Fig. 1. Block diagram of SRU-based speech enhancement.

(2) SRU Network Training

The architecture of the SRU network is illustrated in Fig. 2. It is a feedforward neural network with many levels of non-linearities, which can represent a highly non-linear regression function that maps noisy speech features to clean speech features. During the training, network parameters are randomly initialized and the input–output features are all normalized to zero mean and unit variance. Then, the back-propagation algorithm with the log hyperbolic cosine (logcosh) objective function, as shown in Eq. (14), is utilized to sufficiently train the SRU network.

$$L = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} \log\left(\cosh\left(\hat{X}(n,k) - X(n,k)\right)\right) \tag{14}$$

where L is the logarithm of the hyperbolic cosine, $\hat{X}(n,k)$ and $X(n,k)$ denote the kth predicted and clean frequency bins of the log-spectral feature at sample frame index n, respectively. In addition, N indicates the mini-batch size and K indicates the size of the log-spectral feature vector, which is 129 as described in Section 2.1.

To better capture the nonlinear variations of data, rectified linear unit (ReLU) and hard sigmoid are used as the activation function and recurrent activation function in hidden layers, respectively. Adam optimizer is performed in mini-batches to update the weights and biases of hidden neurons. Besides, other hyper-parameters, such as the number of layers, the number of hidden neurons and learning rate, are all set according to different conditions.

As described above, if training data is diverse and large enough, the SRU network has the potential of learning the nonlinear relationship between noisy speech and clean speech without any prior knowledge, which is different from traditional model-based algorithms. Besides, compared with DNN and common RNN, SRU has stronger regression and parallel computing capabilities, and therefore in theory, SRU-based speech enhancement has faster training speed and can balance the performance and the training time.

3. Experiments and result discussions

In order to evaluate the performance of the proposed method, several experiments are carried out. The SRU-based speech

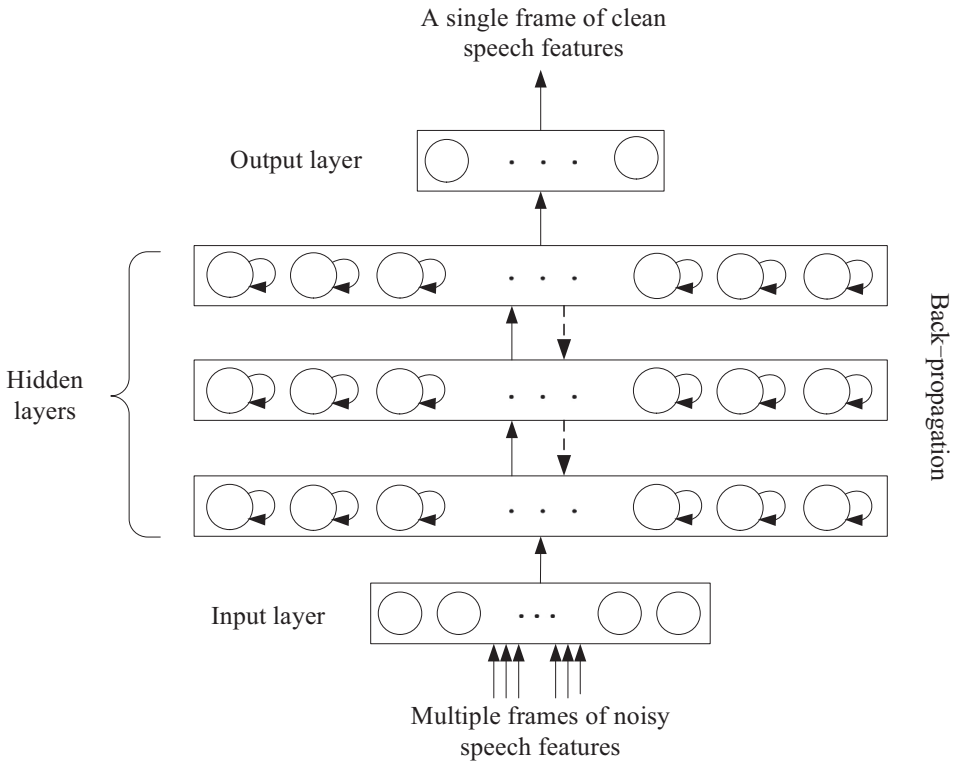

Fig. 2. Architecture of SRU network.

enhancement scheme is first compared with several existing models. Then, the overall performance comparisons on 15 noises among different models are given. Finally, some special cases are discussed to verify the generalization capability of the proposed scheme.

3.1. Simulation setup

During the training phase, 4620 speech from the training set of the TIMIT database are used as clean samples, and the noise samples with 115 noise types, including 100 made by Ohio State University [40] and 15 made by USTC [23], are used for synthesizing noisy speech. More specifically, in our experiments, all the clean speech are mixed with the above-mentioned noises at 6 levels of SNR (−5 dB, 0 dB, 5 dB, 10 dB, 15 dB and 20 dB) to construct a multi-condition stereo training set. However, to balance the performances of networks and their time consumption, only 20 h (27720 samples) are chosen as training subset from the entire training data (about 2720 h).

During the test phase, 158 clean speech are randomly selected from TIMIT test set and 15 types of noises from the NOISEX-92 database [41] are considered to generate noisy speech. It should be pointed out that our work mainly focuses on the evaluation of mismatched noises between training and testing, so all the utterances in the test set are different from those in the training set.

As for signal analysis, all the clean and noisy waveforms are resampled to 8KHz, and the corresponding frame length is set to 256 samples (i.e. 32 ms) with a 128 samples frame shift. Then, a short-time Fourier transform is used to compute the log-power spectral features of signals, and the dimension of the feature vector is 129. Furthermore, to evaluate the performance of the proposed SRU-based scheme, we compare it with three existing networks, namely DNN, LSTM and GRU. Then, four criterions including perceptual evaluation of speech quality (PESQ) [42], segmental SNR (SSNR), short-time objective intelligibility (STOI) [43] and mean objective score (MOS) are used to assess the enhanced speeches from different networks.

In the following experiments, to make a fair comparison, all implementations of network are executed under the same condition. The parameters of all network are set as follows: the number of the adjacent frame is 5, thus the dimension of the network input is $129 \times 11 = 1419$; the maximum number of the epoch is 50000 and the mini-batch size is 1024. Other parameters such as the number of layers and learning rate are set according to different conditions.

3.2. General performance of SRU-based speech enhancement

3.2.1. Evaluation of SRU-based scheme

As we known, RNN networks are suitable for dealing with speech-related issues. Besides, DNN network proposed by Xu [22] is widely used in speech enhancement as well. Thus, in the following experiments, the SRU-based speech enhancement will be compared with three typical models: LSTM, GRU and DNN. All the parameters are set as described in Section 3.1, and the learning rate, the number of the hidden layers and the hidden units for each layer is 0.001, 3 and 1024, respectively. In addition, due to the space limitation, we only display the results of five types of noises, namely White, Volvo, Babble, Machine gun and HF channel. Here, the first two are stationary noises and the last three are non-stationary noises.

Table 1 presents the time consuming of four training models. From Table 1, it can be claimed that the implementation of 3-layer SRU model (SRU_3 for short) is significantly faster than other three models, and the time required is almost half that of LSTM model. But for performance, as shown in Table 2, the PESQ score of 3-layer SRU is worse than general RNN models, only better than

Table 1
Time consumption of training five models.

Algorithm	DNN	LSTM	GRU	SRU_3	SRU_4
Training time(s)	690	19672	15217	9752	13526

the DNN model. Thus, to further balance the performance and time consumption, the depth of SRU model is extended to four layers.

The training time of 4-layer SRU model (SRU_4 for short) is showed in Table 1, and the PESQ score on five noises at different SNRs are displayed in Table 2. From Table 1, it can be stated that although the training time increases with the depth of the SRU model increasing, it is still much smaller than LSTM and GRU models. As for performance, SRU_4 has an advantage in dealing with stationary noise, for example, in white noise case, the average PESQ over six different SNR levels (from −5 dB to 20 dB) improves from 2.02 to 2.66, which ranks first among five models. For non-stationary noise, LSTM outperforms other models, and the average PESQ improves from 2.44 to 2.82. SRU_4 ranks second, with a gap of 0.02 compared with LSTM. Following are GRU and SRU_3, which have third best and fourth rank. Besides, by looking in details, with the input SNR decreasing from 20 dB to −5 dB, the improvement of PESQ increases gradually, such as White, Volvo and Machine gun noises. But for Babble and HF channel noises, the best improvement appear at 5 dB.

Furthermore, to analyze the SRU model comprehensively, Figs. 3–7 present the SSNR results on the above-mentioned noises at different SNRs, respectively. In Figs. 3–7, it is clear that when the SNR of input speech is low(−5 dB), the output speech obtained by the LSTM has the highest SSNR, and the SRU_4 just ranks second. When the input SNR is 0 dB, the performance of SRU_4 is slightly inferior to LSTM but is still competitive. For other input SNRs, i.e. 5 dB, 10 dB, 15 dB and 20 dB, the performance of SRU_4 are far

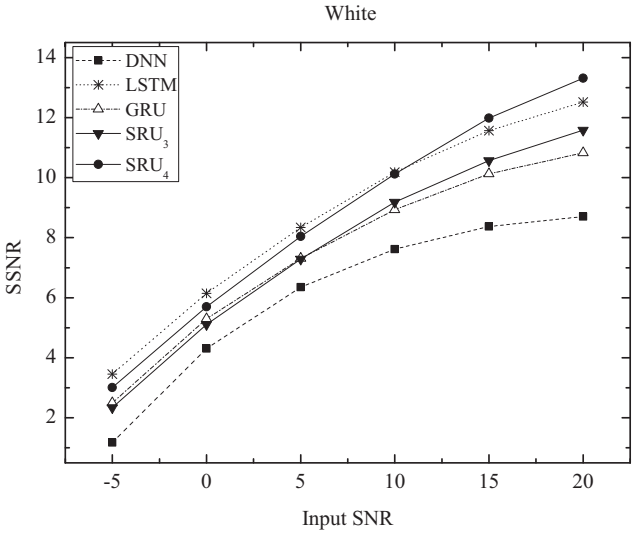

Fig. 3. SSNR results of White noise for different models.

superior to other four models, which reveals that SRU_4 is suitable for more scenarios and has better applicability.

Unlike aforementioned speech quality evaluation measures, STOI is a representative approach to evaluate speech intelligibility [43]. It computes the correlation between temporal envelopes of the clean and processed speeches in short-time segments as an intelligibility indicator and has been verified to have a high correlation with speech intelligibility of human listeners. Thus, STOI is an important index to assess speech enhancement algorithms as well.

Table 2
PESQ comparison on the test set at different input SNRs of unseen noise environments.

Noise types	SNR	Unproc.	DNN	LSTM	GRU	SRU_3	SRU_4
White	20 dB	2.84	3.12	**3.28**	3.23	3.17	**3.28**
	15 dB	2.50	2.92	3.05	3.02	2.95	**3.06**
	10 dB	2.15	2.68	2.80	2.77	2.72	**2.81**
	5 dB	1.82	2.38	2.52	2.49	2.47	**2.54**
	0 dB	1.51	2.06	2.23	2.19	2.20	**2.27**
	−5 dB	1.28	1.69	1.92	1.81	1.87	**1.97**
Volvo	20 dB	4.29	3.47	3.87	3.72	3.70	**3.89**
	15 dB	4.12	3.44	3.83	3.69	3.65	**3.84**
	10 dB	3.85	3.40	**3.76**	3.63	3.57	**3.76**
	5 dB	3.52	3.33	**3.66**	3.55	3.46	3.64
	0 dB	3.18	3.23	**3.52**	3.42	3.29	3.46
	−5 dB	2.86	3.07	**3.33**	3.26	3.05	3.22
Babble	20 dB	3.19	3.25	**3.47**	3.38	3.34	**3.47**
	15 dB	2.87	3.06	**3.25**	3.18	3.13	3.23
	10 dB	2.55	2.79	**2.98**	2.92	2.86	2.95
	5 dB	2.22	2.44	**2.63**	2.60	2.56	2.62
	0 dB	1.91	2.00	2.19	2.20	2.17	**2.22**
	−5 dB	1.58	1.55	1.69	1.76	1.72	**1.77**
Machine gun	20 dB	3.59	3.37	**3.70**	3.59	3.58	**3.70**
	15 dB	3.36	3.28	**3.59**	3.49	3.47	3.58
	10 dB	3.11	3.16	**3.46**	3.37	3.34	3.44
	5 dB	2.82	3.00	**3.32**	3.23	3.20	3.29
	0 dB	2.50	2.81	**3.16**	3.05	3.04	3.12
	−5 dB	2.11	2.54	**2.94**	2.80	2.80	2.89
HF channel	20 dB	2.82	3.04	**3.18**	3.14	3.07	3.14
	15 dB	2.48	2.78	**2.87**	**2.87**	2.80	2.84
	10 dB	2.15	2.52	2.54	**2.58**	2.49	2.51
	5 dB	1.83	2.24	2.22	**2.27**	2.17	2.18
	0 dB	1.56	1.93	1.92	**1.95**	1.86	1.86
	−5 dB	1.35	1.60	**1.64**	1.58	1.57	1.60

The best performance are highlighted in bold.

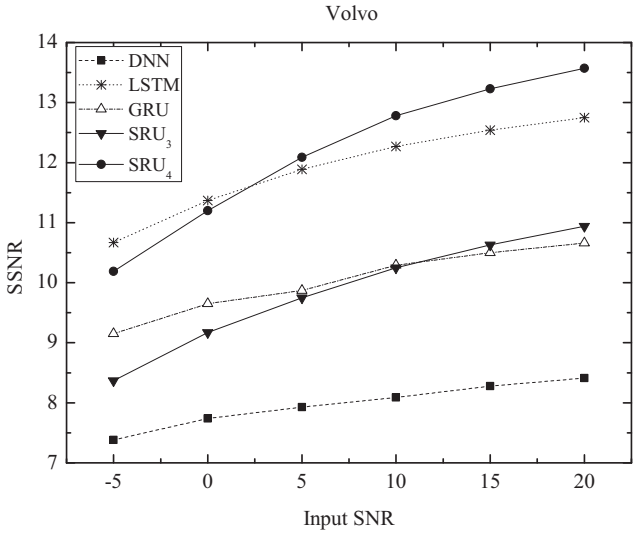

Fig. 4. SSNR results of Volvo noise for different models.

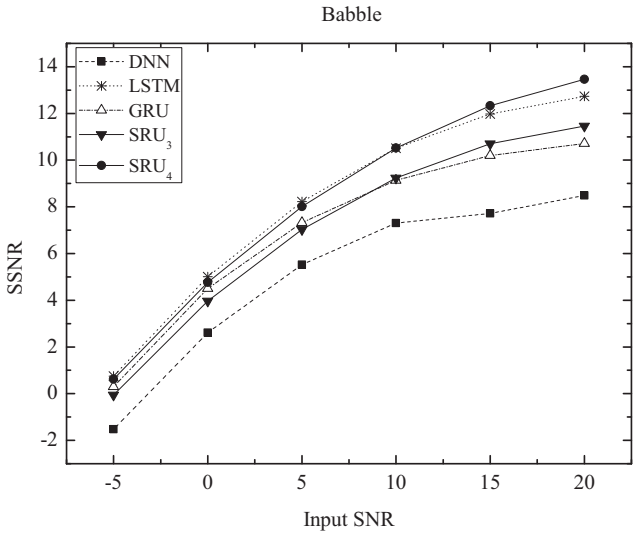

Fig. 5. SSNR results of Babble noise for different models.

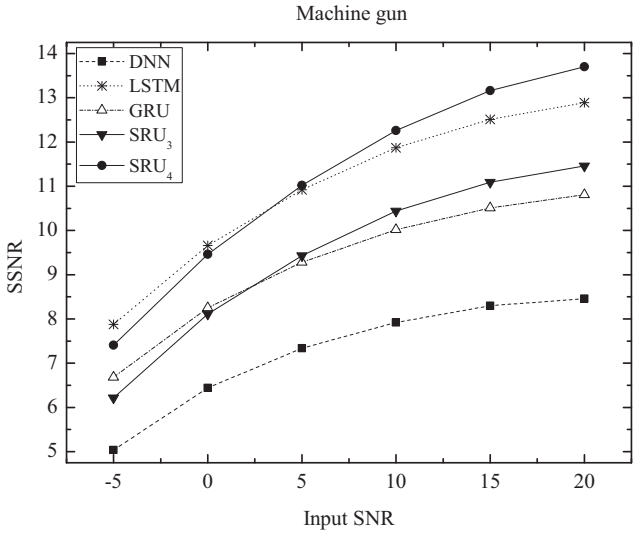

Fig. 6. SSNR results of Machine gun noise for different models.

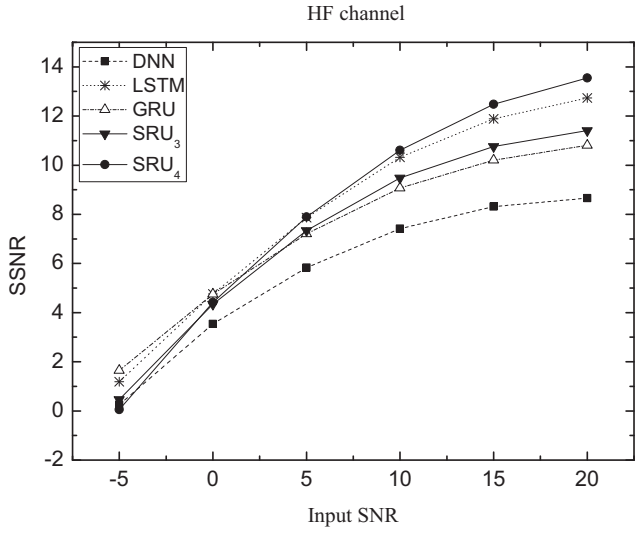

Fig. 7. SSNR results of HF channel noise for different models.

The results of STOI over five types of noises are shown in Figs. 8–12. For the stationary noise environments, as illustrated in Figs. 8 and 9, the LSTM perform better than other models in low SNR conditions ranging from −5 dB to 0 dB, while in high SNR conditions ranging from 5 dB to 20 dB, the performance of SRU_4 are comparable to those of the LSTM, jointly ranking the first place. Figs. 10–12 show the STOI results of non-stationary noise environments. Generally, when the input SNR is higher than 5 dB, the enhanced speeches from SRU_4 and LSTM have the same intelligibility, which outperform than those of GRU, SRU_3 and DNN. However, when the input SNR is less than or equal to 5 dB, the outcomes of five models are slightly different on three noise types. For Babble, the enhanced speech obtained by GRU has the highest STOI under −5 dB SNR input, SRU_4 and LSTM jointly ranking the second place. For Machine gun, LSTM always have the best performance and following is SRU_4. For HF channel, with −5 dB and 0 dB SNR inputs, the enhanced speech from LSTM have the best intelligibility while those from SRU_4, GRU and SRU_3 have the same STOI. It brings us a conclusion that for low SNR inputs, these models are good at dealing with different dynamic noises.

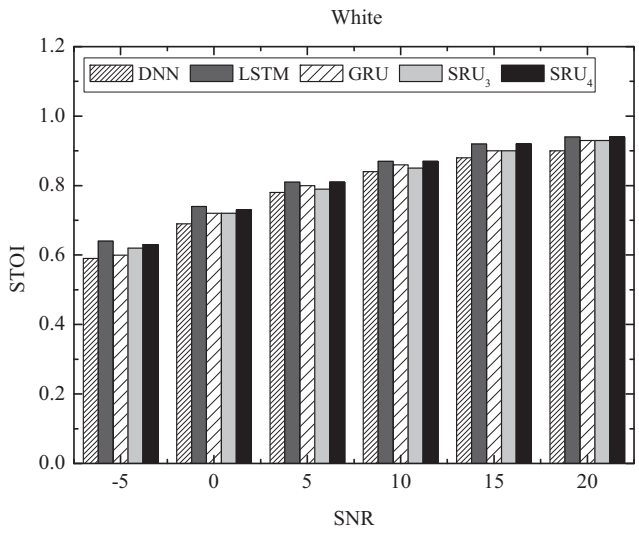

Fig. 8. STOI results of White noise for different models.

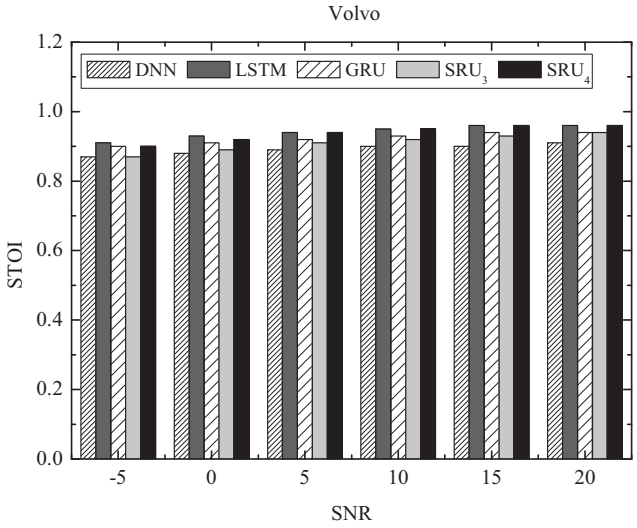

Fig. 9. STOI results of Volvo noise for different models.

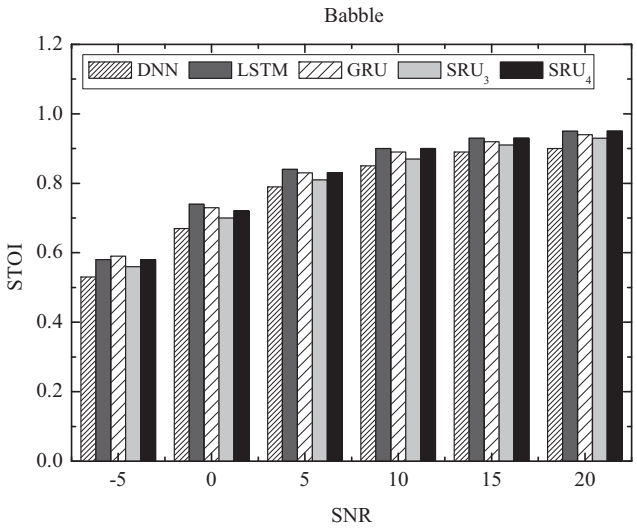

Fig. 10. STOI results of Babble noise for different models.

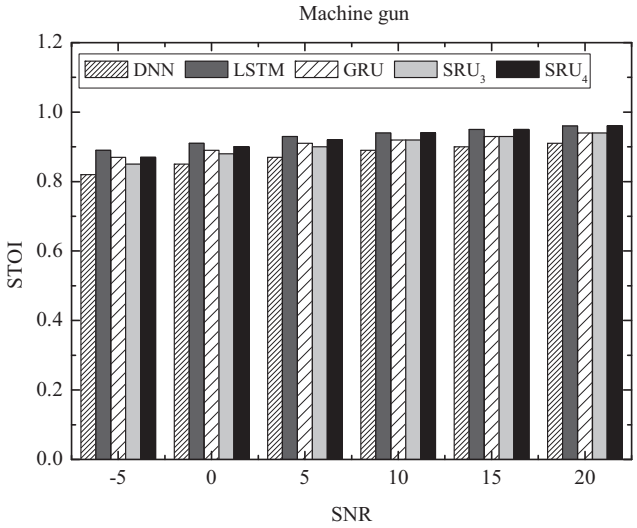

Fig. 11. STOI results of Machine gun noise for different models.

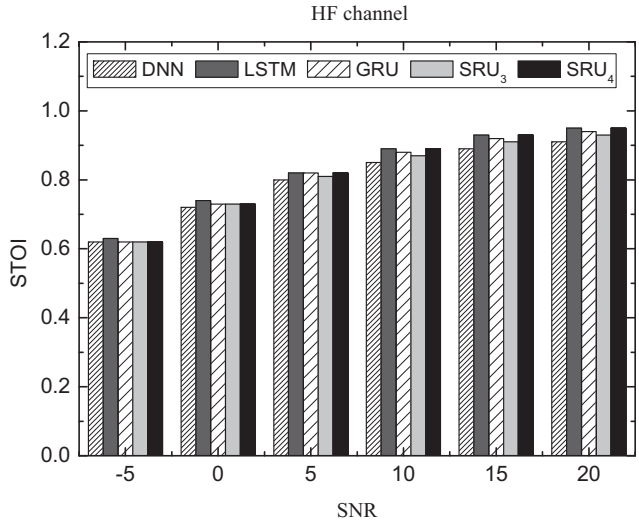

Fig. 12. STOI results of HF channel noise for different models.

3.2.2. Evaluation of the model depth

As above mentioned, the performance of the 4-layer SRU-based model is far superior to that of the 3-layer one and its consumption of training time is smaller than that of conventional RNN networks. That is to say, the depth of the network has a large impact on performance. Therefore, in this experiment, the effect of different depths on the training time and enhanced results are taken into consideration.

The parameters of this experiment are set as described in Section 3.1, and the number of the hidden layers is increased from 3 to 5, with 1024 hidden units for each layer. Besides, to obtain the best performance of different models, the learning rate is changed with the number of layers increasing.

Before the experiment, Table 3 presents the total number of parameters involved in the training stage among DNN, LSTM, GRU, SRU_3 and SRU_4 models. From Table 3, it is clear that in RNN networks, SRU_3 has the least parameter size, while LSTM has the largest one. Moreover, although the layer of SRU is extended to 4, the number of its parameters is still much smaller than those of 3-layer LSTM and 3-layer GRU, which indicates that the comparison for performance and training time of SRU_4, $LSTM_3$ and GRU_3 models are fair and meaningful.

Tables 4 and 5 list the training time and average PESQ results (5 dB input SNR) of DNN, LSTM, GRU and SRU models under different layers. It can be observed that first, training RNN networks does take more time than training DNN network, which is consistent with our perception. Second, combined with performance, deeper neural network architectures have a better regression capability to some extent, indicating that more detail features can be learned with a deeper hidden layer. But in turn, when the layer is increased to 5, the performance of models are worse or equal to that of 4-layer models. This may because the 5-layer model is too large for experimental dataset, which leads to overfitting. Third, by observing the results of RNN networks in details, it is interesting to note that the training time of SRU_4 is 1.45 times faster than $LSTM_3$, but its performance is still competitive to that of $LSTM_3$. For 4-layer models, although the performance of the $LSTM_4$ and GRU_4 are 0.1

Table 3
The number of parameters among five models.

	DNN	LSTM	GRU	SRU_3	SRU_4
Total params.	3.69 M	21.64 M	16.27 M	6.96 M	10.11 M

Table 4
Time consumption of network training at different layers.

	Time(s)			
	DNN	LSTM	GRU	SRU
Layer-3	690	19672	15217	9752
Layer-4	808	26931	20707	13526
Layer-5	915	35339	27126	17628

Table 5
Average PESQ results among four network at different layers on fifteen unseen noise.

	PESQ			
	DNN	LSTM	GRU	SRU
Layer-3	2.62	2.77	2.74	2.68
Layer-4	2.68	2.77	2.78	2.76
Layer-5	2.67	2.73	2.78	2.75

and 0.2 better than that of the SRU_4, the implementation of them are 2 and 1.53 times as fast as SRU_4, respectively. In addition, with the number of layers deepening, the time consuming of training RNN-based models are increased disproportionately. For instance, the difference of training time between $LSTM_4$ and $LSTM_3$ is 7259s, but it is extended to 8408s when comparing that between $LSTM_5$ and $LSTM_4$. Similarly, the difference in training time for SRU is increased from 3774s to 4102s, but it is still significantly smaller than that of LSTM. Therefore, it can be stated that SRU is more suitable to construct very deep networks. In other words, the deeper the network, the less the time required for training SRU compared with convention RNN networks.

From above analyses, it can be concluded that the performance of the proposed SRU scheme is comparable to LSTM, and superior to GRU and DNN to some extent. It does balance the training time and model performance, which has the potential to dispose more complex problems.

3.3. Overall performance evaluation

In order to get a representative result, the whole 15 noise types are tested among DNN, LSTM, GRU and SRU models at different SNRs. The configurations of all models are set as described in Section 3.1, and the learning rate is 0.001 and the number of hidden layers is 3 (4 for SRU), with 1024 hidden units for each layer. Table 6 presents the results of average PESQ for overall performance. From it, we can note that the overall performance of SRU_4 is as outstanding as that of LSTM with an improvement from 2.45 to 2.86. Following are GRU and SRU_3, with an improvement of 0.36 and 0.31, respectively. Moreover, when the input SNR decreases from 20 dB to −5 dB, the improvement of PESQ increases from 0.23 to 0.51 and the best improvement appears at 0 dB.

In addition, the average STOI results to represent the intelligibility of the enhanced speech are also listed in Table 7. It can be claimed that the STOI of DNN is slightly worse than that of the noisy speech, whereas the STOI of RNN based networks are all better than that of noisy speech. The enhanced speeches from SRU_4 and LSTM have the best intelligibility with an average STOI increment from 0.82 to 0.85, and GRU and SRU_3 obtained 0.02 and 0.01 STOI improvement, respectively. Besides, for the overall intelligent of speech, we are more concerned about low SNR conditions. Thus, although with the high input SNR(20 dB), the intelligibility of all models decrease, while SRU_4 and LSTM achieve an 0.07 STOI improvement with low input SNR(−5 dB).

Table 6
Average PESQ results among five models on the whole fifteen noises at different SNRs.

SNR	Unproc.	DNN	LSTM	GRU	SRU_3	SRU_4
20 dB	3.25	3.25	3.48	3.40	3.35	3.48
15 dB	2.93	3.09	3.27	3.21	3.16	3.27
10 dB	2.61	2.88	3.04	2.99	2.93	3.03
5 dB	2.29	2.62	2.77	2.74	2.68	2.76
0 dB	1.96	2.31	2.47	2.44	2.39	2.47
−5 dB	1.67	1.94	2.13	2.09	2.06	2.13
Avg	2.45	2.68	2.86	2.81	2.76	2.86

Table 7
Average STOI results among five models on the whole fifteen noises at different SNRs.

SNR	Unproc.	DNN	LSTM	GRU	SRU_3	SRU_4
20 dB	0.97	0.91	0.95	0.94	0.93	0.95
15 dB	0.93	0.89	0.93	0.92	0.91	0.93
10 dB	0.88	0.86	0.90	0.89	0.88	0.90
5 dB	0.81	0.82	0.86	0.85	0.83	0.85
0 dB	0.72	0.74	0.78	0.78	0.76	0.78
−5 dB	0.61	0.65	0.68	0.68	0.66	0.68
Avg	0.82	0.81	0.85	0.84	0.83	0.85

Although the objective evaluation measures PESQ and STOI have given useful results of human perception, MOS test, which is a typical subject measure, can provide a further evaluation on the real usability of the proposed method by the end users. In MOS test, 60 enhanced utterances processing by five models are randomly selected and 10 listeners, which equally distributed in males and females, are conducted to evaluate the distortion level in the speech. Fig. 13 lists the scores of MOS test. From it, we can state that most listeners prefer the enhanced speeches based on LSTM and SRU_4 models, which is consistent with objective experimental results. Besides, in some cases, such as 0 dB and 10 dB input SNRs, the enhancement results of DNN are not satisfactory, but on the whole, all networks have achieved noise reduction and improved the speech quality.

3.4. Evaluation of other situations

To verify the robustness of the SRU-based model, two special cases are taken into consideration in this subsection. First, we test the model performance under mismatch SNRs, then the condition of multiple noises mixing is employed in the evaluation.

3.4.1. Mismatched SNRs

In this experiment, we choose −8 dB, 2 dB and 7 dB input SNRs which do not appear in the training set to assess the performance

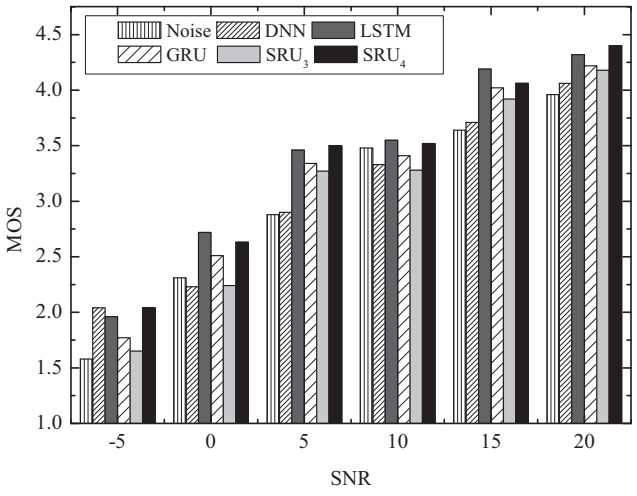

Fig. 13. MOS results of different models for different SNRs.

of SRU model. DNN, LSTM and GRU are still utilized as comparison algorithms. Here, the configurations of five models are set as described in Section 3.1, and the learning rate is 0.001 and the number of hidden layers is 3 (4 for SRU), with 1024 hidden units for each layer.

Table 8 lists the average PESQ results of 15 noise types under mismatched input SNRs. From Table 8, it is clear that although networks are fed with speech under mismatched SNRs, they still lead to good performance, which are basically consistent with those of matched SNRs. LSTM performs the best with a 0.46 PESQ improvement, and SRU_4 ranks second with a gap of 0.01. Besides, the average STOI results are presented in Table 9 as well. It is interesting to note that STOI results are also similar to those of the matching SNR conditions. According to our analysis, these may because when synthesizing noisy speech, it is implemented on the sentence, i.e. noise is added to the sentence while SNR represents the average result in one sentence. In fact, there are various SNRs on each frame in one sentence, and they are not limited to a fixed value but within a certain range.

3.4.2. Multiple noises mixed

In this experiment, the speech from the test set are mixed with multiple types of noises in order to test the capability of the SRU model to suppress multiple noises. Here, 5 types of noises from the Noizeus-92 database, including White, Babble, Buccaneer1, Factory2 and Leopard, are first selected, then a total of 15 types of noises are employed to evaluate the performance of SRU-based network. Moreover, −8 dB, 0 dB, 7 dB and 15 dB are selected as input SNRs, which contain match and mismatch conditions. Similar to the experiment of mismatched SNRs, the SRU model is also compared with DNN, LSTM and GRU models, and the parameters for all models are set as mismatched SNRs experiment to get a representative result.

Table 10 displays the average PESQ results under 5 types of noises. From Table 10, it can be seen that SRU_4 outperforms other networks with an absolute 0.47 improvement compared with noisy speech. LSTM is slightly inferior to SRU_4 with a gap of 0.01. Following are GRU and SRU_3, the improvement of PESQ are 0.43 and 0.34, respectively. In addition, the average PESQ results under 15 types of noises are presented in Table 11. In general, the effect of the enhanced speech are basically as same as the results of 5 types of noises mixed. The PESQ of the enhanced speeches obtained from DNN, LSTM, GRU, SRU_3 and SRU_4 models are increased by 0.33, 0.47, 0.43, 0.33 and 0.46, respectively. However, there are still some differences among them. In Table 11, the best performance emerges in LSTM model, while SRU_4 has similar performance but ranks the second. Besides, the performance of the DNN model is improved, which almost catches up with SRU_3, jointly ranking the third.

From above discussion, it can be summarized that for the cases of mismatched SNRs and multiple noises mixed, the proposed scheme displays good capability of noise suppression, which implies that the SRU-based network has better adaptability and robustness.

Table 8
Average PESQ results among five models under mismatch input SNRs.

SNR	Unproc.	DNN	LSTM	GRU	SRU_3	SRU_4
7 dB	2.42	2.73	2.88	2.84	2.78	2.87
2 dB	2.09	2.44	2.60	2.57	2.51	2.59
−8 dB	1.51	1.72	1.92	1.86	1.84	1.92
Avg	2.01	2.30	2.47	2.42	2.38	2.46

Table 9
Average STOI results among five models under mismatch input SNRs.

SNR	Unproc.	DNN	LSTM	GRU	SRU_3	SRU_4
7 dB	0.85	0.84	0.88	0.87	0.87	0.87
2 dB	0.76	0.78	0.82	0.81	0.79	0.82
−8 dB	0.56	0.58	0.61	0.60	0.59	0.60
Avg	0.72	0.73	0.77	0.76	0.75	0.76

Table 10
Average PESQ results among five models under five type noises interference.

SNR	Unproc.	DNN	LSTM	GRU	SRU_3	SRU_4
15 dB	2.87	3.09	3.26	3.2	3.14	3.27
7 dB	2.33	2.70	2.83	2.81	2.71	2.83
0 dB	1.84	2.21	2.37	2.36	2.26	2.39
−8 dB	1.33	1.49	1.73	1.71	1.61	1.76
Avg	2.09	2.37	2.55	2.52	2.43	2.56

Table 11
Average PESQ results among five models under fifteen types noise interference.

SNR	Unproc.	DNN	LSTM	GRU	SRU_3	SRU_4
15 dB	2.74	3.03	3.17	3.12	3.02	3.16
7 dB	2.19	2.61	2.72	2.70	2.56	2.70
0 dB	1.70	2.11	2.25	2.23	2.12	2.24
−8 dB	1.25	1.44	1.61	1.53	1.51	1.60
Avg	1.97	2.30	2.44	2.40	2.30	2.43

4. Conclusion

In order to reduce the time consumption of training while achieving good quality and intelligibility of the enhanced speech, a SRU-based speech enhancement method is proposed in this paper. The log-power spectra of noisy and clean speech pairs are first extracted as the input and output of the network from a large training data set. Then multiple SRUs are stacked to train log-power spectra pairs, aiming to estimate the complicated nonlinear mapping from noisy speech to clean speech. Ultimately, the well-trained SRU has the capability to estimate the spectra of clean speech corresponding to the noisy input and reconstruct a clean speech waveform. Experimental results show that whether in stationary or non-stationary noise environments, the proposed method not only has the shorter training time, but has superior performance than DNN and GRU, which is comparable to LSTM. Furthermore, under mismatched SNRs and multiple noises mixed cases, the SRU-based scheme still presents good denoising capability. So it can be concluded that the proposed network has better adaptability and robustness.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Nos. 61771091, 61871066), National High Technology Research and Development Program (863 Program) of China (No. 2015AA016306), Natural Science Foundation of Liaoning Province of China (No. 20170540159), and Fundamental Research Funds for the Central Universities of China (No. DUT17LAB04).

References

[1] Weiss M, Aschkensy E, Parson T. Study and the development of the INTEL techniques for improving speech intelligibility. Technical Report NSC-FR/4023. Northvale, USA: Nicolet Scientific Corporation; 1974.
[2] Lockwood P, Boudy J. Experiments with a non-linear spectral subtractor (NSS), hidden Markov models and the projections, for robust recognition in cars. Speech Comm 1992;11(2):215–28.

[3] Kamath S, Loizou P. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In: IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Orlando, USA. 4164–4164.
[4] Lim JS, Oppenheim AV. Enhancement and bandwidth compression of noisy speech. Proc IEEE 1979;67(12):1586–604.
[5] Hu Y, Loizou P. Incorporating psycho-acoustical model in frequency domain speech enhancement. IEEE Signal Process Lett 2004;11(2):270–3.
[6] Scalart P, Filho JV. Speech enhancement based on a priori signal to noise estimation. In: IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP) Atlanta, USA. p. 629–32.
[7] Ephraim Y, Van Trees HL. A signal subspace approach for speech enhancement. IEEE Trans Speech Audio Process 1995;3(4):251–66.
[8] Jensen J, Heusdens R. Improved subspace-based single-channel speech enhancement using generalized super-gaussian priors. IEEE Trans Audio Speech Lang Process 2007;15(3):862–72.
[9] Yang C, Wang J, Wang J, Wu C, Chang K. Design and implementation of subspace-based speech enhancement under in-car noisy environments. IEEE Trans Veh Technol 2008;57(3):1466–79.
[10] Cohen I, Gannot S. Spectral enhancement methods. In: Springer handbook of speech processing. Springer; 2008. p. 873–902.
[11] Lotter T, Vary P. Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model. EURASIP J Appl Signal Process 2005;2005(7):1–17.
[12] Ephraim YA. Bayesian estimation approach for speech enhancement using hidden Markov models. IEEE Trans Signal Process 1992;40(4):725–35.
[13] Kundu A, Chatterjee S, Sreenivas TV. GMM based Bayesian approach to speech enhancement in signal/transform domain. In: IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Las Vegas USA. p. 4893–6.
[14] Mohammadiha N, Smaragdis P, Leijon A. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. IEEE Trans Audio Speech Lang Process 2013;21(10):2140–51.
[15] Huang Y, Benesty J. A multi-frame approach to the frequency domain single-channel noise reduction problem. IEEE Trans Audio, Speech, Lang Process 2012;20(4):1256–69.
[16] Ming J, Crookes D. An iterative longest matching segment approach to speech enhancement with additive noise and channel distortion. Comput Speech Lang 2014;28(6):1269–86.
[17] Wang Y, Wang D. Boosting classification based speech separation using temporal dynamics. In: INTERSPEECH, Portland, USA. p. 1528–31.
[18] Wang Y, Wang D. Towards scaling up classification-based speech separation. IEEE Trans Audio Speech Lang Process 2013;21(7):1381–90.
[19] Healy EW, Yoho SE, Wang Y, Wang D. An algorithm to improve speech recognition in noise for hearing-impaired listeners. J Acoust Soc Am 2013;134(4):3029–38.
[20] Wang Y, Narayanan A, Wang D. On training targets for supervised speech separation. IEEE/ACM Trans Audio Speech Lang Process 2014;22(12):1849–58.
[21] Lu X, Tsao Y, Matsuda S, Hori C. Speech enhancement based on deep denoising autoencoder. In: INTERSPEECH, Lyon, France. p. 555–9.
[22] Xu Y, Du J, Dai L, Lee C. An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process Lett 2014;21(1):65–8.
[23] Xu Y, Du J, Dai L, Lee C. A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans Audio Speech Lang Process 2015;23(1):7–19.
[24] Tu M, Zhang X. Speech enhancement based on deep neural networks with skip connections. In: IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), New Orleans, USA. p. 5565–9.
[25] Zheng N, Zhang X. Phase-aware speech enhancement based on deep neural networks. IEEE/ACM Trans Audio Speech Lang Process 2019;27(1):63–76.
[26] Weninger F, Erdogan H, Watanabe S, Vincent E, Le Roux J, Hershey JR, Schuller B. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In: International Conference on Latent Variable Analysis and Signal Separation, Liberec, Czech Republic; 2015. p. 91–9.
[27] Weninger F, Eyben F, Schuller B. Single-channel speech separation with memory-enhanced recurrent neural networks. In: IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Florence, Italy. p. 3709–13.
[28] Weninger F, Hershey JR, Le Roux J, Schuller B. Discriminatively trained recurrent neural networks for single-channel speech separation. In: IEEE Global Conference on Signal and Information Processing (GlobalSIP) Atlanta, USA. p. 577–81.
[29] Sun L, Du J, Dai L, Lee C. Multiple-target deep learning for LSTM-RNN based speech enhancement. In: Hands-free Speech Communications and Microphone Arrays (HSCMA), San Francisco, USA; 2017. p. 136–40.
[30] Lee J, Kim K, Shabestary T, Kang H. Deep bi-directional long short-term memory based speech enhancement for wind noise reduction. In: Hands-free Speech Communications and Microphone Arrays (HSCMA), San Francisco, USA; 2017. p. 41–5.
[31] Fu SW, Tsao Y, and Lu X. SNR-aware convolutional neural network modeling for speech enhancement. In: INTERSPEECH, San Francisco, USA; 2016. p. 3678-772.
[32] Fu SW, Tsao Y, Lu X, Kawai H. Raw waveform-based speech enhancement by fully convolutional networks. In arXiv: 1703.02205v3; 2017.
[33] Qian Y, Bi M, Tan T, Yu K. Very deep convolutional neural networks for noise robust speech recognition. IEEE Trans Audio Speech Lang Process 2016;24(12):2263–76.
[34] Liu Y, Ganguly A, Kamath K, Kristjansson T. Neural network based time-frequency masking and steering vector estimation for two-channel MVDR beamforming. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada. p. 6717–21.
[35] Zhang X, Zhang H, Nie S, Gao G, Liu W. A pairwise algorithm using the deep stacking network for speech separation and pitch estimation. IEEE Trans Audio Speech Lang Process 2016;24(6):1066–78.
[36] Wang Z, Zhao W, Wang D. Phoneme-specific speech separation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China. p. 146–50.
[37] Pascual S, Bonafonte A, Serra J. SEGAN: Speech enhancement generative adversarial network. In: INTERSPEECH, Stockholm, Sweden; 2017. p. 3642–6.
[38] Lei T, Zhang Y. Training RNNs as fast as CNNs. In arXiv: 1709.02755v2; 2017.
[39] Lei T, Zhang Y, Sida IW, Hui D, Yoav A. Simple recurrent units for highly parallelizable recurrence. In arXiv: 1709.02755v5; 2018.
[40] Hu G. 100 nonspeech environmental sounds; 2004 [Online]. Available:http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html.
[41] Varga A, Steeneken HJM. Assessment for automatic speech recognition: II. NOISEX- 92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun 1993;12(3):247–51.
[42] ITU-T, Rec. P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. International Telecommun Union-Telecommun Standardization Sector; 2001.
[43] Taal CH, Hendriks RC, Heusdens R, Jensen J. A short-time objective intelligibility measure for time-frequency weighted noisy speech, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Dallas, USA; 2010. p. 4214–7.