


RESEARCH

Open Access



Speech enhancement from fused features based on deep neural network and gated recurrent unit network

Youming Wang^{1,2*} , Jiali Han¹, Tianqi Zhang¹ and Didi Qing¹

*Correspondence:

xautroland@126.com

¹ School of Automation,
Xi'an University of Posts
and Telecommunications,
Xi'an 710121, China

Full list of author information
is available at the end of the
article

Abstract

Speech is easily interfered by external environment in reality, which results in the loss of important features. Deep learning has become a popular speech enhancement method because of its superior potential in solving nonlinear mapping problems for complex features. However, the deficiency of traditional deep learning methods is the weak learning capability of important information from previous time steps and long-term event dependencies between the time-series data. To overcome this problem, we propose a novel speech enhancement method based on the fused features of deep neural networks (DNNs) and gated recurrent unit (GRU). The proposed method uses GRU to reduce the number of parameters of DNNs and acquire the context information of the speech, which improves the enhanced speech quality and intelligibility. Firstly, DNN with multiple hidden layers is used to learn the mapping relationship between the logarithmic power spectrum (LPS) features of noisy speech and clean speech. Secondly, the LPS feature of the deep neural network is fused with the noisy speech as the input of GRU network to compensate the missing context information. Finally, GRU network is performed to learn the mapping relationship between LPS features and log power spectrum features of clean speech spectrum. The proposed model is experimentally compared with traditional speech enhancement models, including DNN, CNN, LSTM and GRU. Experimental results demonstrate that the PESQ, SSNR and STOI of the proposed algorithm are improved by 30.72%, 39.84% and 5.53%, respectively, compared with the noise signal under the condition of matched noise. Under the condition of unmatched noise, the PESQ and STOI of the algorithm are improved by 23.8% and 37.36%, respectively. The advantage of the proposed method is that it uses the key information of features to suppress noise in both matched and unmatched noise cases and the proposed method outperforms other common methods in speech enhancement.

Keywords: Speech enhancement, Deep neural network, Gated recurrent unit, Speech quality

1 Introduction

In the past several decades, speech enhancement has attracted considerable research interest due to the wide application of voice-based solutions for real-world applications. The purpose of speech enhancement is to improve speech quality and intelligibility under

the interfering noise conditions. Recently, the classic noise reduction methods including spectral subtraction (SS), Wiener filtering (WF), hidden Markov model (HMM) and statistical model-based algorithms have been widely studied to remove or attenuate additive noise from noisy speeches [1–4]. Spectral subtraction is one of the typical speech enhancement algorithms proposed to remove environment noise, but the resulting enhanced speech often suffers from annoying musical artifact called musical noise. The Wiener filter is a linear estimator and minimizes the mean-squared error between the original and enhanced speech, which depends on the filter transfer function from sample to sample based on the speech signal statistics. HMMs are doubly stochastic processes or probabilistic functions of Markov chains that model time-series data as the evolution of a hidden state variable through a discrete set of possible values. There are two problems to solve for the traditional HMM, which are the limitation of conditional independence and difficulty of processing segmental features. The performance of conventional methods is generally dependent on the nature of the background noise and the statistical properties of speech, because traditional methods need to estimate power spectrum of noise. However, it is difficult to accurately estimate different types of noise with nonlinear or non-stationary features.

In recent years, deep learning became increasingly popular as a mapping method between the noisy and clean speech signals to accomplish the task of enhancing a desired speech signal. The fully connected structure of multi-layer neuron nodes and the application of nonlinear activation functions enables the deep learning to solve various classification and regression models for the separation of the speech and the noise. Deep learning with multiple nonlinear layers only needs the current observation data and has strong nonlinear mapping and self-learning abilities to learn generalizable features from large amounts of training data. The advantage of deep learning for speech enhancement is that it can remove the noise considerably from the noisy speech, because it makes no assumptions about the statistical properties of the signals and uses a large collection of noise types to generate diverse noisy speech samples for training. Representative deep learning models like convolutional neural networks (CNN), deep neural networks (DNN) and recurrent neural networks (RNN) have been successfully applied into fields like computer vision and natural language processing [5–7]. Recently, deep learning with a large training data set has shown good generalization capabilities to unseen noise types and better performance in both noise reduction and speech distortion over the conventional approaches [8, 9]. In [10], a deep convolutional neural network (CNN) is proposed to improve recognition accuracy for noise robust speech recognition, and it also can reduce word error rate (WER) significantly. A deep auto-encoder (DAE) is introduced to address the mapping relationship of the Mel-frequency power spectra between noisy speech and clean speech, and denoising DAE provided superior speech enhancement performance compared with a minimum mean square error-based speech enhancement [11]. A speech enhancement framework based on the DNN and restricted Boltzmann machine (RBM) is proposed, where RBM is introduced to initialize the multiple-layer deep architecture [12]. Although deep learning methods have achieved great success, the long-term dependencies hidden in time-series data are not considered and utilized in traditional deep learning. Specifically, there are data redundancy, data missing and abnormal data in time-series data. So it is necessary to model long-term dependencies

in time-series data to enlarge the receptive field and discover longer patterns in speech enhancement.

To address the problem, recurrent neural network (RNN) [13] and long short-term memory (LSTM) [14] have been proposed to learn the temporal relations and capture time dependencies of time-series data. RNNs with gated mechanism learn long time sequences via a way that information in nodes of hidden layers will be recycled to achieve time-series memory. Long short-term memory is a typical structure in RNNs, where different gates are used to control the percentage of saving, dropping temporal information and receiving incoming information. Recurrent neural network and long short-term memory have been demonstrated in the applications with sequential data, which can model the relationship between previous frame and current frame to capture the long-term context information [15, 16]. However, LSTM often has the problems of gradient disappearance and gradient explosion. As a variant and improved version of LSTM, GRU can use the previous input of prediction information and maintain a longer-term information dependence, which reduces the number of gate units on the LSTM model and solves the gradient disappearance problem of RNN. Due to its special structure of an update gate and a reset gate, GRU controls the flow of information through learning gates and further controls input and memory of gates; thereby, it saves computer memory and simultaneously captures the dependence of time-series information. In [17], a bitwise GRU network is used for the single-channel source separation task. A GRU-based recurrent neural network method to learn the desired critical band gains over each frequency band is presented in [18]. Recently, a speech emotion recognition model based on Bi-GRU is proposed and shows good recognition accuracy [19]. Gated recurrent units that will result in inessential content are reserved when the unprocessed data are used as input.

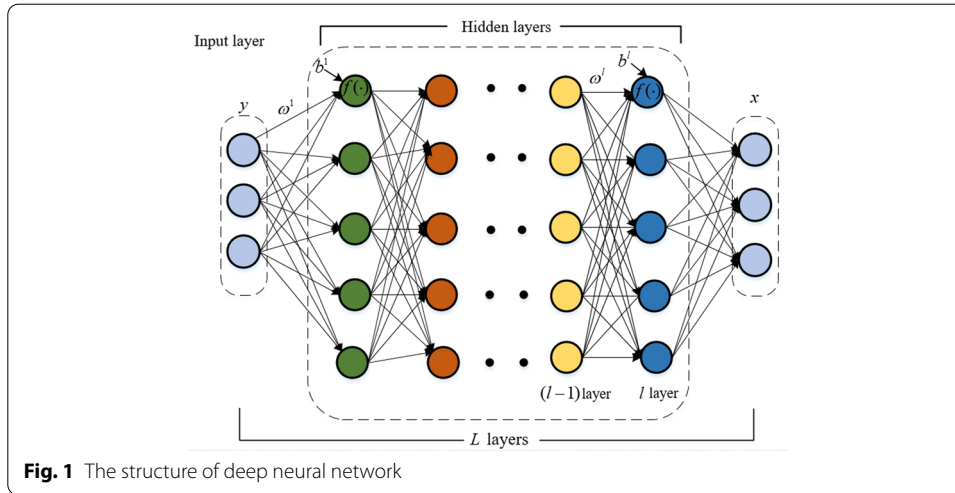
A novel DNN-GRU method is proposed to take advantage of both deep neural network and recurrent neural network to drastically reduce the number of parameters and simultaneously improve speech quality and speech intelligibility in this paper. The DNN with three fully connected layers is employed to establish a mapping function between noisy speech and clean speech. In order to learn context information while decreasing the training time of deep learning, the LPS features from DNN model and noisy speech are fused and learned by a GRU-based speech enhancement method. The proposed DNN-GRU network combines the output of the speech pre-processed by DNN with the features of noisy speech to compensate the lack of context information and improve the enhanced speech quality and intelligibility.

The rest of this paper is organized as follows. Section 2 introduces the DNN and GRU architectures. The DNN-GRU model is introduced in Sect. 3. Experiments are presented in Sect. 4 to evaluate the performance of the proposed algorithm. Finally, the conclusion is given in Sect. 5.

2 Preliminaries

2.1 Deep neural network

Deep neural network is a kind of feed-forward neural network, which contains the input layer, several hidden layers and the output layer [20]. Figure 1 shows the topological



structure of DNN. Deep neural network has the ability to learn some features from multiple layers to ensure that the neural structure can construct a complex mapping function. The nodes between two adjacent layers of DNN are fully connected, and the nodes on the same layer are not connected to each other. As the number of layers and width of the network increases, the characteristics of DNN become more complex and the training time becomes longer.

Deep neural network that generates output vector from input vector is expressed by

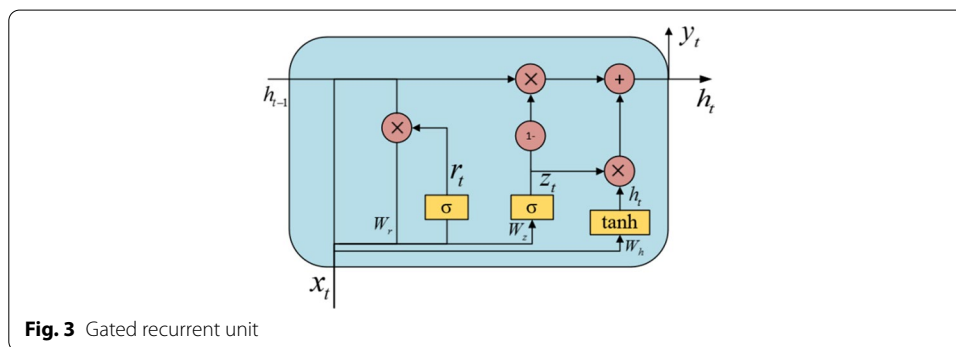
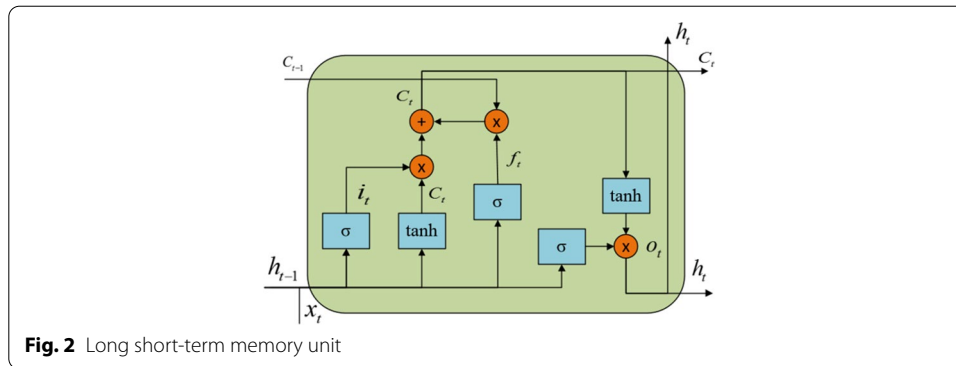
$$\begin{cases} h^1 = f^1(w^1 y + b^1) \\ h^l = f^l(w^l h^{l-1} + b^l) \\ x = f^L(w^L h^L + b^L) \end{cases} \quad (1)$$

where $1 \leq l \leq L$, $h^0 = y$, $h^L = x$. $h^{l-1} \in R^{d_{l-1} \times 1}$ is the d_{l-1} dimensional output vector of $(l-1)$ -th layer, and $h^l \in R^{d_l \times 1}$ is the d_l dimensional output vector of l -th layers. Additionally, $w^l \in R^{d_l \times d_{l-1}}$ and $b^l \in R^{d_l \times 1}$ are the weight matrix, with bias from $(l-1)$ -th hidden layer to the l -th hidden layer, $f^l(\cdot)$ is the activation function on the l -th hidden layer, and L -th layer is the output layer.

Since the nonlinearity of activation functions is crucial for the success of predictive models, the nonlinear activation functions are commonly used to enhance the model accuracy including Sigmoid, Tanh and ReLU. Scaled exponential linear unit (SeLU) has a unique characteristic in the ability to automatically normalize its output toward predefined mean and variance, which can be described by

$$f(x) = \lambda \begin{cases} x & x > 0 \\ ae^x - a & x \leq 0 \end{cases} \quad (2)$$

where λ and α are two fixed parameters, in general, $\lambda = 1.05$ and $\alpha = 1.67$. The SeLU activation function has saturation zone but no dead zone, and the output will be magnified after activation.



2.2 Gated recurrent unit

Gated recurrent unit (GRU) network is regarded as an updated version of LSTM with a simple structure including memory cell and gate units [21]. Long short-term memory and gated recurrent unit are improved versions of RNN, which are considered as powerful schemes for modeling temporal and sequential data and capturing long-term dependencies on datasets. Compared with the RNN, GRU has promising features on the balance between fast computation and capture capability for the mapping relationship among time-series datasets. By introducing gating mechanisms into the architecture, GRUs provide a trained model with consistent memory capable of seizing short-term and long-term dependencies among speech frames effectively.

Figures 2 and 3 depict the structures of LSTM and GRU, respectively. The LSTM has an input, output and forget gate. In the GRU cell, this is handled via an update gate and a reset gate, where the update gate mostly does what in the LSTM is done by the input and forget gate. The main difference is the presence or absence of an output gate, which tells how much of the content is presented to the next layer of the network. Compared with the LSTM network structure, GRU can solve the prediction problem of long interval long delay time series. Gated recurrent unit can outperform LSTM units both in terms of convergence in CPU time and in terms of parameter updates and generalization [22].

As shown in Fig. 3, the reset gate is used to control the degree of ignoring the information of the previous moment and the update gates control whether the status of GRU is updated and how many the gating units are updated. The activation gate h_t of the GRU at time t is a linear interpolation between the previous activation h_{t-1} and the next activation h_t .

The equation of GRU can be described as:

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \quad (3)$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \quad (4)$$

$$h_t \% = \tanh(W_h x_t + U_h r_t \cdot h_{t-1}) \quad (5)$$

$$h_t = (1 - z_t)h_{t-1} + z_t h_t \% \quad (6)$$

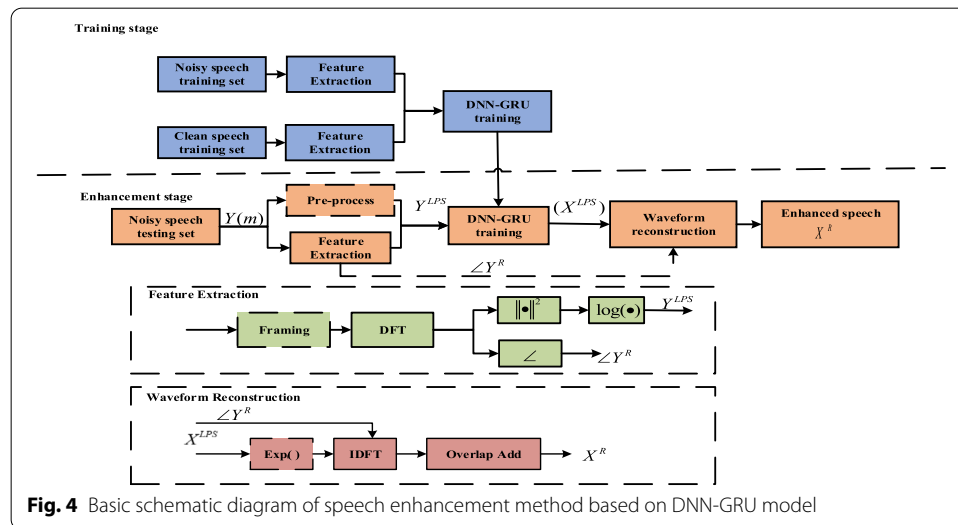
where r_t is the reset gate determining the number of ignored prior information. x_t represents the input of memory unit, z_t is the update gate which determines the number of information input to the next state cell. W_r , W_h and W_z represent weight vectors corresponding to the gates in the memory unit, respectively.

Although GRUs can handle long-term sequential for time-series data, their gate structures can lead to the disregard of important content in a long sequence [23, 24]. Gated recurrent units may lead to poor models where important information from previous time steps and long-term event dependencies is not well addressed during training stage. In this paper, we present an approach that alleviates this problem by introducing a novel DNN-GRU model which is capable of sustaining crucial content in long-term sequential data.

3 Speech enhancement based on the DNN and GRU network

3.1 Overall learning framework

Figure 4 shows the overall procedure based on DNN-GRU model, which includes the training phase and enhancement phase. Before training, a variety of LPS features for noisy speech and clean speech are extracted. In the training phase, two-stage speech enhancement neural network with nonlinearities is adopted, which can learn mapping from noisy speech features to clean speech features. Firstly, LPS features of the noisy speech and clean speech are inputted to a fully connected feed-forward DNN to obtain



the optimal weights, bias and hyper-parameters. Then, the LPS features of DNN pre-processed and noisy speech are combined to compensate the missing time-series information. Lastly, the new LPS speech features and the LPS features of clean speech are used to build the mapping function of GRU network to achieve noise reduction. In the enhancement stage, the noisy speech is sent into the well-trained DNN-GRU model to predict the LPS features of clean speech. The estimated LPS feature is used as waveform recovery to obtain the clean speech. The enhanced speech by the DNN-GRU model is coherent, which guarantees the contextual information of the speech signal and improves the speech intelligibility and quality.

In Fig. 4, $Y(m)$ is the noisy speech, Y^{LPS} is the LPS features of noisy speech, X^{LPS} is the LPS features, X^R is the estimated speech, and $\angle Y^R$ is the phase of speech.

3.2 DNN-GRU model-based training

Clean speech and noise are added to construct noisy speech. The clean speech and noise form voice pair datasets which are divided into training sets and test sets.

$$Y(m) = X(m) + N(m) \quad (7)$$

where $Y(m)$, $X(m)$ and $N(m)$ represent noisy speech, clean speech and noise at time m , respectively.

In the LPS domain, the target values of different frequency bins are predicted independently without any correlation constraint, and can be transformed back to the waveform domain without any information loss. The extraction process of LPS features is as follows.

First, the speech signal is decomposed into 25 ms frames with 10 ms frame shift by pre-processing as shown in Eq. (8). Each frame is smoothed with hamming window.

$$Y_t(n) = \sum_{p=n-L+1}^n y(p)w(n-p) \quad (8)$$

where $Y_t(n)$ is the t -th frame speech signal, and t is the sample point of $Y_t(n)$. L is the frame length, and p denotes the window length. A discrete Fourier transform (DFT) is performed on $Y_t(n)$ to obtain the spectrum of each frame as shown in Eq. (9):

$$Y(t, f) = \sum_{n=0}^{N-1} Y_t(n) e^{-j \frac{2\pi}{N} fn} (f = 0, 1, 2 \dots N-1) \quad (9)$$

$$Y^{\text{LPS}}(t, f) = \log(|Y(t, f)|)^2 \quad (10)$$

where f represents the f -th frequency point at time-frame unit t , and N is the number of DFT points. The LPS features are obtained by logarithmic function which can be compressed as follows:

Pseudocode

Algorithm: DNN-GRU algorithm
1. STEP 1 Constructing training and test sets
2. STEP 2 Feature extraction
Input: $Y(m)$
Output: $Y^{LPS}(t, f)$
3. Step 1: Framing:
4. Step 2: DFT:
5. STEP 3 Local noise reduction by DNN: $X^P(t) = x_{t+k} \big _{k=-\tau}^{\tau} = f^{DNN}(X_t \theta)$
6. STEP 4 Combining
Input: $X^P(t), Y(t)$
Output: $Y^*(t)$
7. Step 1: preparing: $X^P(t) = x_{t+k} \big _{k=-\tau}^{\tau} = f^{DNN}(X_t \theta)$
8. Step 2: preparing: $Y(t) = \{y(t-\tau), y(t-\tau+1), \dots, y(t), \dots, y(t+\tau-1), y(t+\tau)\}$
9. Step 3: feature fusion: $Y^*(t) = (y_{t+k+i}^* \big _{k=-\tau}^{\tau}) \big _{i=-\tau}^{\tau} = X^P(t) \cup Y(t)$
10. STEP 5: Global noise reduction by GRU: $X^R(t) = g^{GRU}(\cdot \eta) = \{x^R(t-\tau), x^R(t-\tau+1), \dots, x^R(t), \dots, x^R(t+\tau-1), x^R(t+\tau)\}$
11. STEP 6: Waveform recovery: $X^R(n, k) = \exp\{X(n, k) / 2\} \exp\{j \angle Y^R(n, k)\}$

3.3 DNN-GRU model

The sequence of the noisy LPS features are used as input of the established DNN-GRU model. The DNN-GRU model for speech enhancement contains 8 layers, which consists of an input layer, three hidden layers of DNN with a sequencing size of 1024–1024–1024, one feature fusion layer with size of 512, two GRU layers and one output layer. To capture the nonlinear variations of data, the SeLU is selected as the activation function in the hidden layers of DNN. The structure of DNN-GRU model is shown in Fig. 5).

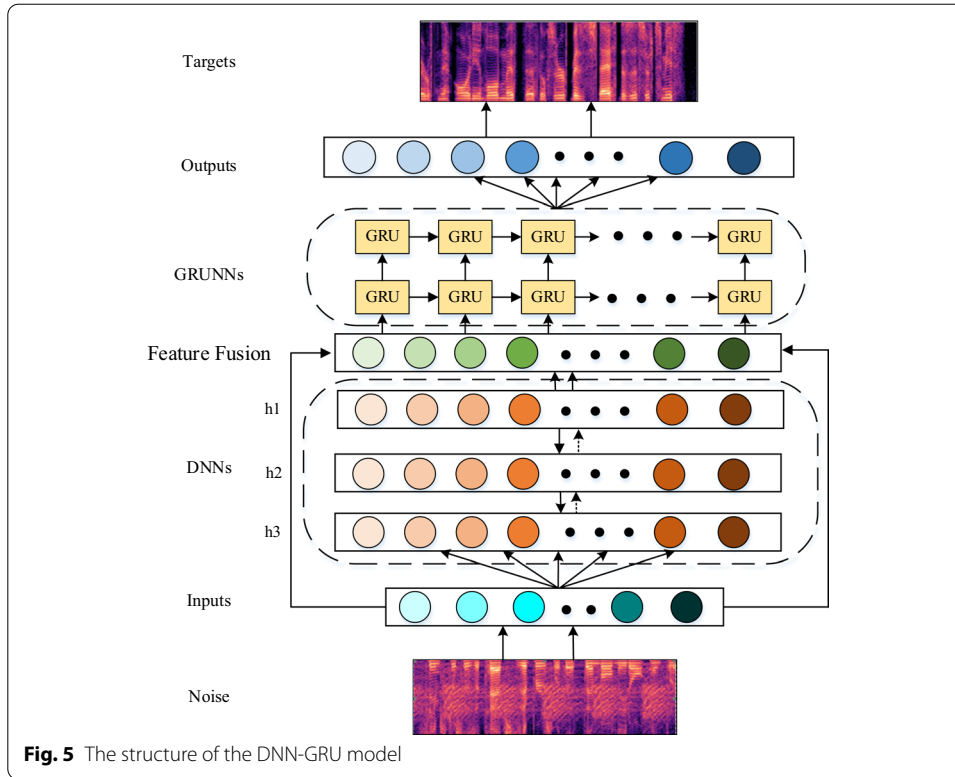
Firstly, a DNN with three hidden layers is typically used to learn the mapping between the local LPS features of noisy speech and clean speech to estimate the clean LPS features from the noisy ones in the first stage.

$$Y(t) = \{y(t-\tau), y(t-\tau+1), \dots, y(t), \dots, y(t+\tau)\} \quad (11)$$

$$X^P(t) = x_{t+k} \big|_{k=-\tau}^{\tau} = f^{DNN}(X_t | \theta), \tau \in (1, X^R(t)) \quad (12)$$

where $Y_t \in R^N$ denotes the noisy LPS vector, $\{x_{t+k}\}_{k=-\tau}^{\tau} \in R^N$ is the enhancement LPS vectors, k is the front-end frames, and $f^{DNN}(Y_t | \theta)$ means the DNN-based function that directly maps the noisy LPS features to clean ones, with DNN parameter set to θ .

The standard back-propagation (BP) algorithm has the ability to address dropout regularization. The DNN training adopts dropout regularization to overcome over-fitting, which randomly discards the neurons with a certain probability to prevent complex correlation among hidden neurons. The mini-batch stochastic gradient descent is a simple



but effective method; it also is used to solve the problem of the over-fitting in a large scale of deep network widely. The dropout rate is set as 0.25 in this paper. In the training stage, a linear activation function is used for the output layer. The number of iterations of the standard BP algorithm is 100. The mean squared error (MSE) is used as the loss function, which minimizes the error between the predicted and noisy speech features.

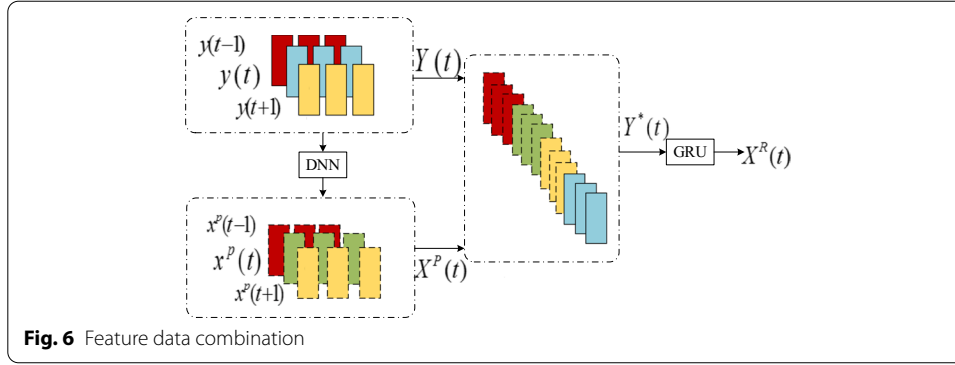
$$MES = \sqrt{\frac{\sum_{i=1}^L (X^{\text{LPS}}(t) - X^R(t))^2}{L}} \quad (13)$$

where L is the total number of samples, $X^{\text{LPS}}(t)$ denotes t -th clean LPS features, and $X^R(t)$ represent the predicted LPS features.

Adam optimizer is used to update the weights and biases of hidden neurons in mini-batches. Furthermore, the rest of hyper-parameters including learning rate, the number of layers and hidden neurons depends on different conditions. As described above, if training data is diverse and large enough, the DNN-GRU model has the potential to learn the nonlinear relationship between noisy speech and clean speech without any prior knowledge.

Secondly, to capture the effective contextual information in features, the layer of feature fusion is adopted. As shown in Fig. 6, DNN-GRU has a cascade architecture consisting of a prior NN (DNN) and a posterior NN (GRU-NN) for the first and second stage of DNN-GRU.

In Fig. 6, $x^p(t-1)$, $x^p(t)$ and $x^p(t+1)$ are the LPS features of three frames after the first stage of DNN, respectively. $y(t-1)$, $y(t)$ and $y(t+1)$ are the LPS feature of noisy



ones. $Y(t)$ and $X^P(t)$ are added and expanded in the form of Fig. 6, forming $Y^*(t)$. Input the $Y^*(t)$ into the GRU network for the second stage.

Since the noisy speech contains the time-series information, the combined features are expected from the LPS features of noisy and the LPS features of DNN processing. The new feature frames are combined with the noisy speech frame as follows:

$$Y^*(t) = (y_{t+k+i}^*|_{k=-\tau}^{\tau})|_{i=-\tau}^{\tau} = X^P(t) \cup Y(t) \quad (14)$$

where $X^P(t)$ includes all base predictions for $x^p(t) \in R^N$, and $Y^*(t)$ containing 128 LPS vectors is input into the GRU network. i is the front-end frames of noisy speech.

The new LPS features of time instance t_k, t_{k-1}, L, t_{k-n} (where k is the current time instance and n is the number of prior frames) are fed into the GRU network with two GRU layers. The first GRU layer has 1024 cells, which encode the input and pass its hidden state to the second GRU layer, which has 512 cells. The two GRU layers are used to establish the mapping from the new feature to the training target features to achieve the whole frames speech enhancement, and meanwhile preserving the contextual information of speech. The GRU network output $x^R(t)$ is the estimated $X^R(t)$.

$$\begin{aligned} X^R(t) &= g^{\text{GRU}}(\cdot|\eta) \\ &= \{x^R(t-\tau), x^R(t-\tau+1), L, x^R(t+\tau)\} \end{aligned} \quad (15)$$

where $g^{\text{GRU}}(\cdot|\eta)$ means the GRU network-based function that directly maps the new features $Y^*(t)$ to clean ones, with GRU network parameter set to η .

3.4 DNN-GRU model-based enhancement

Firstly, the noisy speech is pre-processed in the enhancement stage to obtain a satisfactory enhancement effect. Secondly, the LPS features of noisy speech are extracted and fed into the well-trained DNN-GRU model as test data. To fully display the complementarity of a target set and reduce the impact of network misestimating on enhanced speech, we adopt the estimated LPS to reconstruct enhanced waveform.

Through the DNN-GRU model testing, the estimated LPS feature of the obtained clean speech is defined as $X^{\text{LPS}}(n, k)$. Lastly, the reconstructed spectra $X^R(n, k)$ can be calculated as

$$X^R(n, k) = \exp \{X(n, k)/2\} \exp \{j\angle Y^R(n, k)\} \quad (16)$$

where $\angle Y^R(n, k)$ denotes the k -th phase of the n -th frame from the original noisy speech. After above operations, a frame of clean speech is derived by inverse discrete Fourier transform (IDFT) from the current frame spectra and the whole waveform can be reconstructed.

4 Experiment and result discussion

4.1 Experimental setup

The proposed DNN-GRU model includes training stage and enhancement stage. In the training stage, a fully connected feed-forward DNN-GRU model is used to establish the mapping function of input–output pairs. The trained model can predict the clean speech from corresponding noisy speech. In the enhancement stage, based on the results of the DNN-GRU testing and the online estimated pitch period, the IDFT is utilized to obtain enhanced speech.

During the training stage, 100 speeches from the TIMIT database are used as clean speech, and the 160 noise types of noise samples are randomly selected from Nonspeech and Noise-15 database. The clean speeches are mixed with the noises at 6 levels of signal noise ratio (SNR) to form a noisy set. The noise SNRs are -5 dB, 0 dB, 5 dB, 10 dB, 15 dB and 20 dB, respectively. During the test stage, 40 speeches are randomly selected in the TIMIT test database, and 6 types of noises including Pink, White, Battle, Factory, F16 and Destroy noises are selected from the NOISEX-92 database to form noisy speeches.

4.2 Performance measurement

Three evaluation criteria are used to evaluate the enhanced speech quality, including the perceptual evaluation of speech quality (PESQ) [25], segmental SNR (SSNR) [26] and short-time objective intelligibility (STOI) [27].

4.2.1 PESQ

The PESQ reflects the perceptual quality of the enhanced speech. The PESQ scored from -0.5 to 4.5 , and the PESQ is positively related to the perceptual quality of speech. The PESQ value on six noises in various SNR conditions is presented in Table 1. It can be observed that DNN-GRU model has a superior noise reduction performance. Specifically, the PESQ value of DNN-GRU model is higher than that of the other four models at different SNR levels for White, Factory, F16 and Destroy noises. But for Pink and Battle noises, the PESQ of the proposed model is slightly lower than DNN at 20 dB SNR level. It can be concluded that DNN-GRU model can obtain better speech perceptual quality in variety of environments. Since the proposed framework is compatible with DNN and GRU, it has good performance than single network when processing the different SNRs conditions.

4.2.2 SSNR

Since the speech signal is a short and smooth signal, the SNR values will vary at different times which is changed slowly. The SSNR commonly is used in practical applications to

Table 1 PESQ comparison on the test set at different input SNRs of unseen noise environments

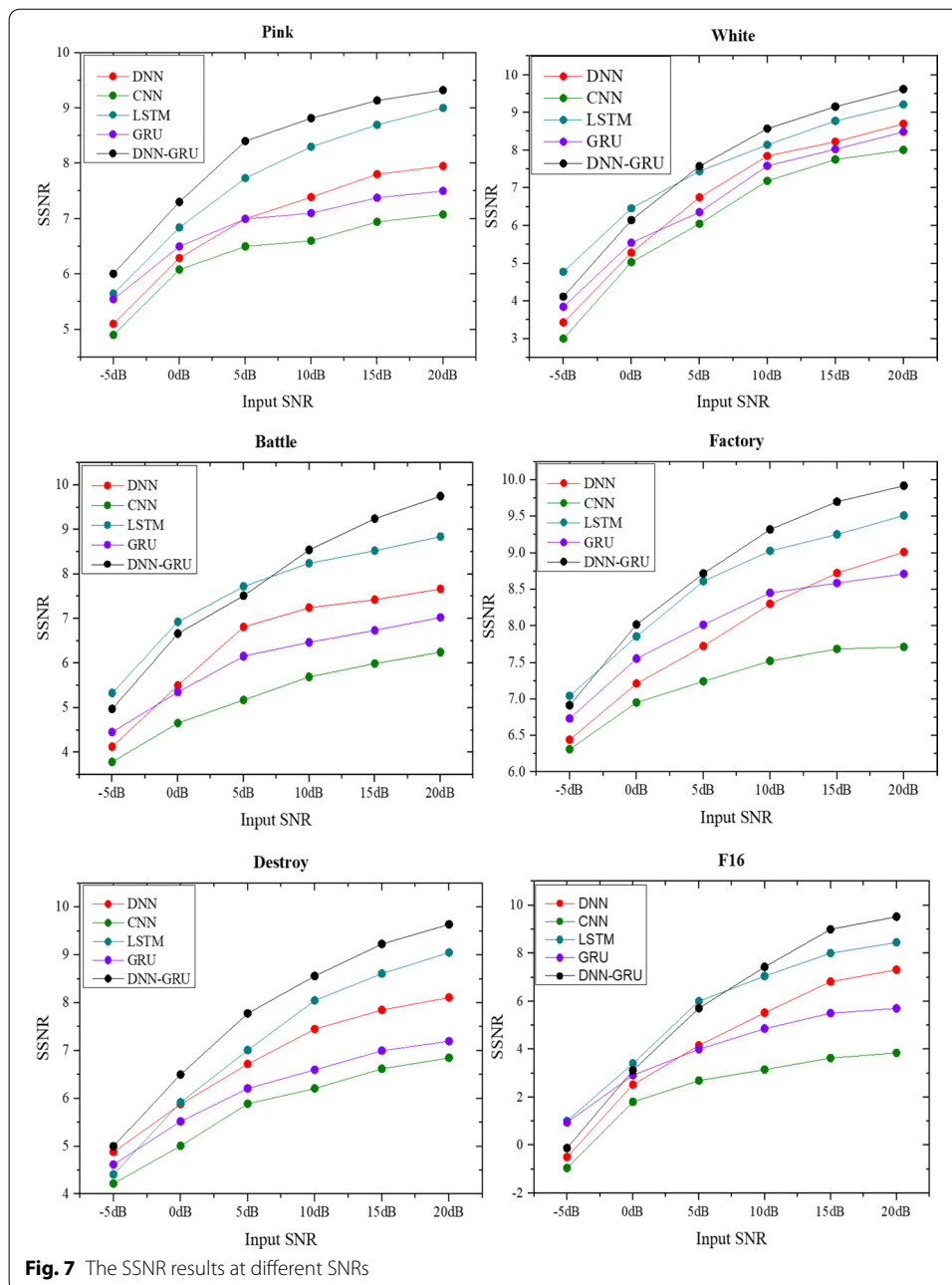
Noise	SNR (dB)	Unprocessed	DNN	CNN	LSTM	GRU	DNN-GRU
Pink	20	3.059	3.330	2.478	3.253	3.181	3.528
	15	2.710	3.075	2.956	3.0151	2.951	3.314
	10	2.358	2.792	2.144	2.724	2.664	2.963
	5	1.962	2.491	2.228	2.411	2.324	2.790
	0	1.606	2.118	1.826	2.049	1.940	2.462
	−5	1.292	1.708	1.371	1.675	1.515	2.114
White	20	2.791	3.156	2.381	3.101	3.080	3.271
	15	2.443	2.921	2.922	2.889	2.847	2.966
	10	2.102	2.659	1.953	2.618	2.605	2.837
	5	1.720	2.356	2.145	2.345	2.287	2.674
	0	1.424	2.057	1.693	2.026	1.957	2.386
	−5	1.200	1.613	1.270	1.671	1.597	1.975
Battle	20	3.152	3.362	2.550	3.339	3.241	3.353
	15	2.831	3.160	2.974	3.103	3.009	3.253
	10	2.503	2.880	2.265	2.818	2.723	2.880
	5	2.126	2.549	2.188	2.522	2.406	2.655
	0	1.796	2.205	1.844	2.156	2.062	2.332
	−5	1.471	1.829	1.446	1.739	1.693	2.175
Factory	20	3.240	3.436	3.396	3.346	3.269	3.642
	15	2.907	3.240	2.970	3.146	3.059	3.423
	10	2.572	2.992	2.289	2.879	2.810	3.229
	5	2.039	2.563	2.209	2.463	2.368	2.937
	0	1.683	2.221	1.811	2.107	1.983	2.592
	−5	1.370	1.797	1.400	1.725	1.593	2.285
F16	20	3.117	3.416	2.527	3.368	3.253	3.587
	15	2.429	2.943	3.013	3.145	3.137	3.308
	10	2.440	2.958	2.265	2.850	2.757	3.169
	5	2.065	2.676	2.292	2.557	2.429	2.882
	0	1.726	2.321	1.906	2.182	2.066	2.538
	−5	1.424	1.965	1.491	1.844	1.744	2.419
Destroy	20	3.191	3.346	2.546	3.375	3.260	3.602
	15	2.878	2.223	2.967	3.174	3.047	3.440
	10	2.547	2.971	2.290	2.898	2.774	3.276
	5	2.193	2.738	2.273	2.635	2.487	2.950
	0	1.828	2.391	1.933	2.205	2.128	2.535
	−5	1.468	1.995	1.533	1.773	1.708	2.393

reflect the performance measurement of enhanced speech, which is defined to evaluate the performance of noise reduction by

$$\text{SSNR} = \frac{10}{M} \sum_{m=1}^{M-1} \log_{10} \frac{\sum_{n=N_m}^{N_m+N-1} x^2(n)}{\sum_{n=N_m}^{N_m+N-1} (x(n) - \hat{x}(n))^2} \quad (17)$$

where m is the frame index, M is the total number of frames, N_m , and N denote the minimum length and total length of the frame, respectively. $x(n)$ represents the clean speech, and $\hat{x}(n)$ denotes the enhanced speech.

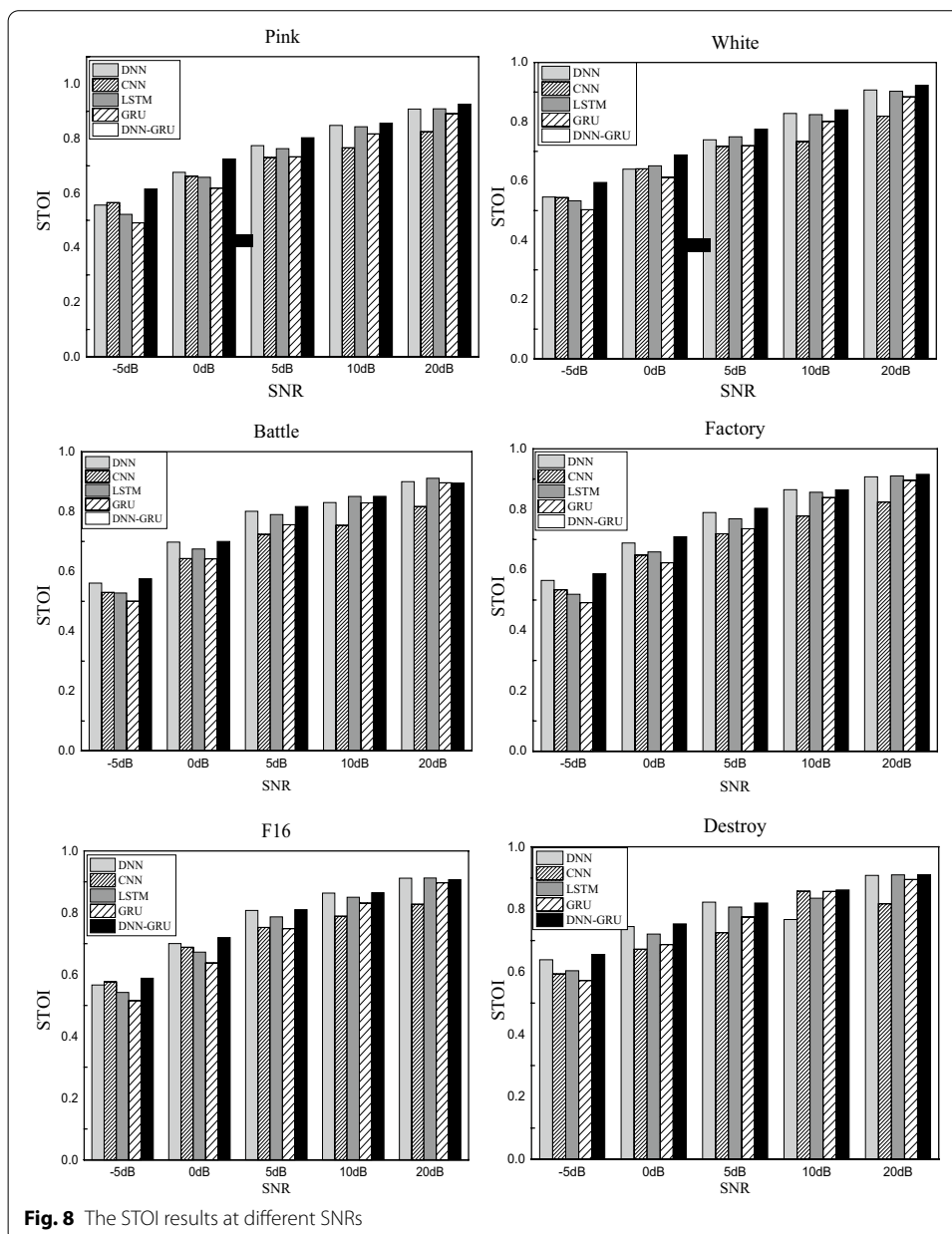
Figure 7 presents the SSNR results at different SNRs. It can be seen that when the input SNR is from 5 to 20 dB, the SSNR of the DNN-GRU model is better than that of the other reference models. It can be inferred that the DNN-GRU model has good noise reduction ability. Under -5 dB and 0 dB conditions, the results of five models are obviously different. Specifically, the LSTM model has excellent results in White, Battle and F16 noises, but the DNN-GRU is still very competitive. For other noise conditions such as Pink and Destroy, the DNN-GRU always has superior SSNR scores. Overall, although the performance of the DNN-GRU model is slightly inferior under lower SNR conditions, the DNN-GRU model is better than other models in



most cases, which verifies our proposed model DNN-GRU has good speech quality and intelligibility.

4.2.3 STOI

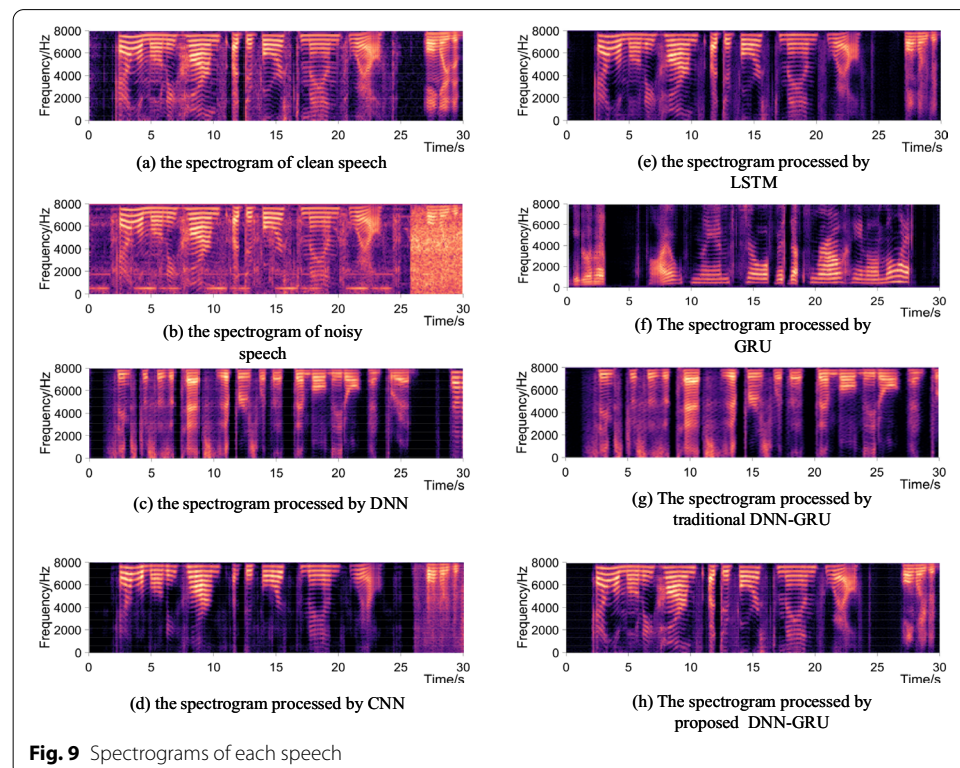
The STOI is a speech intelligibility indicator, which indicates the correlation between temporal envelopes of the clean speech and enhanced speech in short-time segments. The value range of STOI is between 0 and 1, and the larger STOI value denotes the better the speech intelligibility. Figure 8 shows the results of STOI under the six different noise environments. Even though the proposed model has a little decline compared with LSTM at 20 dB under the Battle noise environment, the performance of STOI is better



than reference models generally. Specifically, the DNN-GRU model performs better than other models at low SNR conditions ranging from -5 to 5 dB. In high SNR conditions ranging from 5 to 20 dB, DNN, LSTM and DNN-GRU have excellent noise reduction performance. These phenomena are caused by the superimposition of sine waves, and the reconstructed speech will reduce the intelligibility of the speech in a way. So it can be summarized that these models have good capabilities for the lower SNR conditions.

4.3 Spectrogram comparison

In order to reflect visually differences between the proposed model and the other models, the comparison of the spectrograms of the five representative models with the Pink noise at 5 dB input SNR level is shown in Fig. 9. Its x-label represents time and takes values from 0 to 30 s. Its y-axis represents the frequency and takes values from 0 to 8000 Hz. It can be observed that the DNN and CNN have good noise reduction effects for single-frame speech signals, but they do not have the ability to process time-series signal, so there is a clear fault phenomenon. In addition, the LSTM and GRU have powerful processing capability to correct the front-end frames of speech, but their abilities of noise reduction are relatively poor. Figure 9g–h shows the enhanced speech resulting of traditional two-stage neural network speech enhancement and the proposed DNN-GRU model, respectively. Due to the new features fused with the features between the original signal and processing signal by DNN, the single-frame signal processing capability of DNN and the context information maintaining of sequence signals by the RNN are observed. Thus, the proposed speech enhancement model guarantees a good noise reduction effect and ensures the coherence of speech signal. Compared with the



traditional two-stage neural network model, the speech reconstructed by the proposed method is more complete because it can retain more spectrum details and has superior noise reduction effect.

Simultaneously, Table 2 lists the total number of parameters included in the training stage of DNN, CNN, LSTM, GRU and DNN-GRU. It is clearly seen that among LSTM, GRU and DNN-GRU, the DNN-GRU has the least parameter size and LSTM has the largest size. Compared with the DNN and CNN, more parameters need to be obtained for DNN-GRU network, but it can achieve feature fusion and maintain the continuity of speech signals.

4.4 PESQ and STOI results under mismatch input SNRs

To verify the capability of the DNN-GRU model for speech enhancement with multiple noises, we select four types mismatch noises including 17 dB, 8 dB, 2 dB and -7 dB as input SNRs. The proposed DNN-GRU model is also compared with the reference model including DNN, CNN, LSTM and GRU. The average PESQ results of each model are described in Table 3. The learning rate of DNN-GRU model is selected as 0.0001; the rest of the parameters are consistent with the previous experiment.

In Table 3, it can be seen that despite networks using mismatched SNRs for input, the proposed DNN-GRU model still has a superior performance compared with other models. The DNN-GRU model improves PESQ value by 0.567 improvement, and the performance of GRU is slightly lower than LSTM. Furthermore, Table 4 lists the average STOI results. The enhanced speech using DNN-GRU model also has the best

Table 2 The number of parameters among five models

	DNN	CNN	LSTM	GRU	DNN-GRU
Total parameters	6.05 M	3.68 M	18.09 M	16.38 M	12.17 M

Table 3 Average PESQ results among five models under mismatch input SNRs

SNR	Unprocessed	DNN	CNN	LSTM	GRU	DNN-GRU
17 dB	3.032	3.172	3.069	3.292	3.266	3.587
8 dB	2.405	2.568	2.574	2.778	2.682	2.893
2 dB	1.973	2.156	1.987	2.329	2.237	2.489
-7 dB	1.418	1.725	1.525	1.883	1.863	1.928
Average	2.207	2.405	2.289	2.571	2.512	2.774

Table 4 Average STOI results among five models under mismatch input SNRs

SNR	Unprocessed	DNN	CNN	LSTM	GRU	DNN-GRU
17 dB	0.856	0.857	0.853	0.868	0.868	0.898
8 dB	0.781	0.765	0.743	0.824	0.795	0.832
2 dB	0.739	0.763	0.690	0.720	0.696	0.752
-7 dB	0.546	0.553	0.549	0.510	0.491	0.565
Average	0.727	0.734	0.730	0.731	0.712	0.762

performance, and the STOI result is similar to the matching SNRs result. According to the existing results, it can be inferred that the mismatched SNRs signal is combined with the sentence, and the STOI index represents the average value of the sentence. Compared with the reference speech enhancement models, the DNN-GRU model speech enhancement is considered to be more capable of suppressing non-stationary noise more and denoising less residual noise. Therefore, it can be summarized that the DNN-GRU model can achieve superior performance for the mismatched SNRs, and it has satisfied adaptability and robustness.

5 Conclusions

This paper proposes a novel speech enhancement strategy based on a novel DNN-GRU model to improve the quality and intelligibility of the enhanced speech. The fully connected DNN is used to learn the complex mapping function between clean speech and noisy speech LPS features. The corresponding predicted clean speech is fused with noisy speech as the input of the GRU network, which can retain the time-series context information of the speech signals. The DNN-GRU model is designed to estimate the spectra of clean speech corresponding to the noisy input and reconstruct a clean speech waveform. The spectrogram and experimental results showed that the proposed model performed superior on the metrics PESQ, SSNR and STOI in various noise environments compared with the traditional speech enhancement models, including DNN, CNN, LSTM and GRU. The experimental results under different mismatch input SNRs and mixed noises indicated that the proposed model had good features of adaptability and robustness. Therefore, it can be concluded that the proposed DNN-GRU model maintains excellent denoising capability and has good speech quality and intelligibility.

Abbreviations

DNN: Deep neural network; DNNs: Deep neural networks; LPS: Logarithmic power spectrum; GRU: Gated recurrent unit; PESQ: Perceptual evaluation of speech quality; STOI: Short-time objective intelligibility; SS: Spectral subtraction; WF: Wiener filtering; HMM: Hidden Markov model; CNN: Convolutional neural network; CNNs: Convolutional neural networks; RNN: Recurrent neural network; RNNs: Recurrent neural networks; WER: Word error rate; DAE: Deep auto-encoder; RBM: Restricted Boltzmann machine; LSTM: Long short-term memory; DNN-GRU: Deep neural network average Gated recurrent unit; SNR: Signal to noise ratio; SSNR: Segmental signal to noise ratio; SNRs: Signal to noise ratios.

Acknowledgements

Not applicable.

Authors' contributions

YW proposed the framework of the whole algorithm; performed the simulations, analysis and interpretation of the results. JH and TZ have participated in the conception and design of this research. JH and DQ drafted and revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the Key Research and Development Program of Shaanxi Province of China (2020SF-377, 2019GY-086). It was also supported by the graduate student innovation fund of Xi'an University of Post and Telecommunications (CXJLD202003).

Availability of data and materials

Please contact the authors for data requests.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

The manuscript does not contain any individual person's data in any form (including individual details, images or videos), and therefore, the consent to publish is not applicable to this article.

Competing interests

The authors declare that they have no competing interests.

Author details

¹School of Automation, Xi'an University of Posts and Telecommunications, Xi'an 710121, China. ²Xi'an Key Laboratory of Advanced Control and Intelligent Process (ACIP), Xi'an 710121, China.

Received: 25 May 2021 Accepted: 13 October 2021

Published online: 24 October 2021

References

1. P.C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd edn. (CRC Press, Cambridge, 2013)
2. C. Valentiniibotinhao, J. Yamagishi, S. King, Evaluating speech intelligibility enhancement for HMM-based synthetic speech in noise (2012)
3. H.N. Moritz, T. Roux, Triggered attention for end-to-end speech recognition. In: *Icassp IEEE International Conference on Acoustics* (IEEE, 2019).
4. T.V. Sreenivas, P. Rao, Pitch extraction from corrupted harmonics of the power spectrum. *J Acoust Soc Am* **65**(1), 223–228 (1979)
5. C. Fdlwa, Vanessa Aparecida de Moraes Weber b e, C. Gvm, et al. Recognition of Pantaneira cattle breed using computer vision and convolutional neural networks-ScienceDirect. *Comput. Electron. Agric.* 175.
6. Analysis of DNN speech signal enhancement for robust speaker recognition (2018)
7. P.P. Barman et al., A RNN based approach for next word prediction in assamese phonetic transcription. *Procedia Comput Sci* **143**, 117–123 (2018)
8. A. Nicolson, K.K. Paliwal, Deep learning for minimum mean-square error approaches to speech enhancement. *Speech Commun* **111**, 44–55 (2019)
9. A. Adeel, M. Gogate, A. Hussain, Contextual audio-visual switching for speech enhancement in real-world environments (2018)
10. Y. Qian, M. Bi, T. Tian et al., Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Trans Audio Speech Lang Process* **24**(12), 2263–2276 (2017)
11. X. G. Lu, Y. Tsao, S. Matsuda et al. Speech enhancement based on deep denoising autoencoder (2013)
12. Y. Xu, J. Du, L.R. Dai et al., An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Process. Lett.* **21**(1), 65–68 (2013)
13. F. Weninger, H. Erdogan, 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), pp. 91–94
14. J. Lee, K. Kim, T. Shabestary, H. Kang. 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), pp. 41–50
15. F. Weninger, J. R. Hershey, J. Le Roux, B. Schuller. Discriminatively trained recurrent neural networks for single-channel speech separation, in: *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on* (IEEE, 2014), pp. 577–581
16. F. Weninger, F. Eyben, B. Schulle. Single-channel speech separation with memory-enhanced recurrent neural networks, in: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2014), pp. 3709–3713
17. Y. Xu, J. Du, L. Dai, C. Lee, A regression approach to speech enhancement based on deep neural networks. *IEEE-ACM Trans. Audio Speech Lang Process* **23**, 7–19 (2015)
18. J. M. Valin, A hybrid DSP/deep learning approach to real-time full-band speech enhancement, in: *IEEE 20th International Workshop on Multimedia Signal Processing* (2018), pp. 1–5
19. Z. Zhu, W. Dai, Y. Hu, Speech emotion recognition model based on Bi-GRU and focal loss-ScienceDirect. *Pattern Recognit Lett* **140**, 358–365 (2020)
20. A.W. Rix, M.P. Hollier, A.P. Hekstra, Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment Part 1—time-delay compensation. *J. Audio Eng. Soc.* **50**, 755–764 (2002)
21. Z. Zhao, W. Chen, X. Wu, LSTM network: a deep learning approach for short-term traffic forecast. *Intell. Transp. Syst. IET* **11**, 68–75 (2017)
22. J. Chung, C. Gulcehre, K. H. Cho, Empirical evaluation of gated recurrent neural networks on sequence modeling (2014)
23. U.S. Bhalla, Dendrites, deep learning, and sequences in the hippocampus. *Hippocampus* **29**, 239–251 (2017)
24. W. Stephen, L. Sijia, S. Sunghwan, Modeling asynchronous event sequences with RNNs. *J. Biomed. Inform.* **83**, 167–177 (2018)
25. 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, in: *International Telecommun Union-Telecommun Standardization Sector* (2001)
26. C. H. Taal, R. C. Hendriks, R. Heusdens. A short-time objective intelligibility measure for time-frequency weighted noisy speech, in: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Dallas, USA, 2010), pp. 4214–7

27. S. Kim, M. Maity, M. Kim. Incremental binarization on recurrent neural networks for single-channel source separation (2019). pp. 376–380

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)
