Computers, Materials & Continua
DOI: 10.32604/cmc.2023.027379
Article

Tech Science Press

Speech Enhancement via Mask-Mapping Based Residual Dense Network

Lin Zhou[1,*], Xijin Chen[1], Chaoyan Wu[1], Qiuyue Zhong[1], Xu Cheng[2] and Yibin Tang[3]

[1]School of Information Science and Engineering, Southeast University, Nanjing, 210096, China
[2]Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, FI-90014, Finland
[3]College of IOT Engineering, Hohai University, Changzhou, 213022, China
*Corresponding Author: Lin Zhou. Email: Linzhou@seu.edu.cn
Received: 16 January 2022; Accepted: 06 April 2022

Abstract: Masking-based and spectrum mapping-based methods are the two main algorithms of speech enhancement with deep neural network (DNN). But the mapping-based methods only utilizes the phase of noisy speech, which limits the upper bound of speech enhancement performance. Masking-based methods need to accurately estimate the masking which is still the key problem. Combining the advantages of above two types of methods, this paper proposes the speech enhancement algorithm MM-RDN (masking-mapping residual dense network) based on masking-mapping (MM) and residual dense network (RDN). Using the logarithmic power spectrogram (LPS) of consecutive frames, MM estimates the ideal ratio masking (IRM) matrix of consecutive frames. RDN can make full use of feature maps of all layers. Meanwhile, using the global residual learning to combine the shallow features and deep features, RDN obtains the global dense features from the LPS, thereby improves estimated accuracy of the IRM matrix. Simulations show that the proposed method achieves attractive speech enhancement performance in various acoustic environments. Specifically, in the untrained acoustic test with limited priors, e.g., unmatched signal-to-noise ratio (SNR) and unmatched noise category, MM-RDN can still outperform the existing convolutional recurrent network (CRN) method in the measures of perceptual evaluation of speech quality (PESQ) and other evaluation indexes. It indicates that the proposed algorithm is more generalized in untrained conditions.

Keywords: Mask-mapping-based method; residual dense block; speech enhancement

1 Introduction

Speech enhancement is a fundamental task in speech signal processing, which is widely used in various scenarios, e.g., mobile phone, intelligent vehicles [1] and medical devices [2,3]. It is performed as a front-end signal procedure for automatic speech recognition (ASR), speaker identification, hearing-aid devices and cochlear implant. At present, speech enhancement based on deep learning (DL) is

treated as a supervised learning problem, which can be divided into two categories: spectrum mapping [4] and masking [5], according to the training target.

The masking-based method focuses on separating clean speech from background interference by estimating masking value, which describes the time-frequency relationships of clean speech to noise. Generally, the masking of current time frequency (TF) unit is estimated through features of previous and current frames due to the causal system. The ideal binary masking (IBM) is the commonly used masking, which is firstly adopted in the DL based speech separation [4]. In this work, a pre-trained DNN is used to estimate the IBM on each sub-band. The DNN with support vector machines (DNN-SVM) system demonstrates good generalization [6]. Besides IBM, IRM [7], complex IRM [8], phase-sensitive mask (PSM) [9] and spectral magnitude mask (SMM) [10] are also designed as training targets. In terms of speech quality, ratio masking performs better than binary masking. In 2014, Wang use the DNN to estimate IBM and IRM, which indicates that DNN-based mask estimation method can significantly improve speech enhancement. Overall, the IRM and the SMM are the preferred targets, and the DNN based on ratio masking performs better than unsupervised speech enhancement [11].

The mapping-based method aims to estimate the magnitude spectrogram or temporal representation of clean speech directly from noisy speech, which naturally avoids the masking selection in the masking-based method. Related research indicates that the mapping has superiority to the masking at a low SNR [12]. A deep autoencoder (DAE) is the first proposed algorithm to map the Mel-power spectrum of degraded speech to the clean one [5]. In the later research, log spectral magnitude and log Mel-spectrum are used in DL-based speech separation [13,14]. Also, DNN is exploited in the LPS mapping [15]. Compared with DNN, convolutional neural networks (CNN) can obtain more accurate local features, which can better recover the high-frequency of the speech signal, and improve the quality and intelligibility of the enhanced speech [16,17]. The generative adversarial networks (GANs) learn the nonlinear transformation from noisy speech to clean speech by generating confrontation, which has generalization in untrained conditions [18]. The DNN, GANs or CNN-based speech enhancement rarely consider the temporal characteristics of the speech, which limits the performance of enhancement. With the self-feedback neurons, Recurrent Neural Network (RNN) can process the sequence signals, and achieve better performance on speech enhancement [19]. The optimization of RNN via the back propagation through time (BPTT) has the problem of vanishing and exploding gradients [20], long short-time memory recurrent neural network (LSTM-RNN) is proposed to solve this problem [12], and improves both the speech quality and intelligibility.

As the recent study [12] indicates that masking is advantageous at higher SNRs and mapping is more advantageous at lower SNRs. We combined these two types of speech enhancement, which is denoted as MM based method. The MM method maps LPS [21] to IRM matrix of consecutive frames, not just the IRM of the current frame. The RDN [22] makes full use of features of all layers through local dense connection. Also, using the global residual learning, RDN combines the shallow features and deep features to obtain the global dense features from the LPS spectrogram, thereby improves accuracy of the masking estimation. The proposed MM-RDN speech enhancement outperforms the mapping-based CRN [23] which reached State-of-the-Art (SOTA) level in the enhancement speech.

The outline of the paper is organized as follow. In Section 2, the architecture and the implementation of the proposed method are described in detail. Simulation results and analysis are presented in Section 3. Finally, conclusions are drawn in Section 4.

2 Method Description

The proposed MM-RDN based speech enhancement system is illustrated in Fig. 1. In training, the LPS of consecutive frames is extracted, and treated as the input features for the RDN network. IRM of the corresponding frames composes the two-dimensional IRM matrix, which is used as the training target. The RDN network is trained to establish relationship between the LPS and the IRM matrix. In testing, the RDN outputs the estimated ratio masking (ERM) matrix, which is used to reconstruct the clean speech with the original noisy speech.

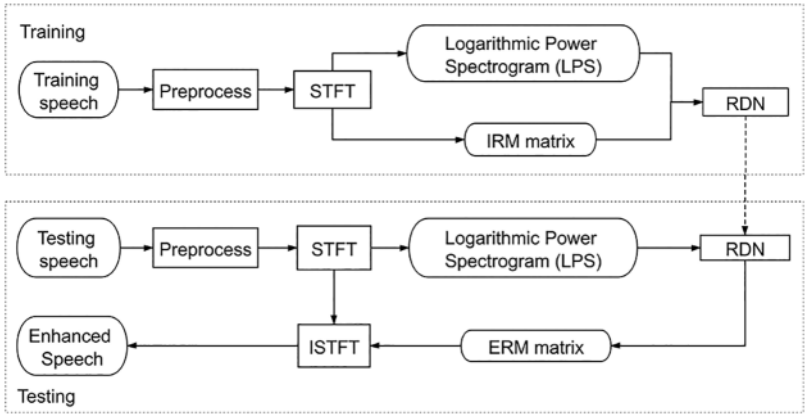

Figure 1: The block diagram of proposed algorithm

The noisy speech signal is formulated as:

$$x(n) = s(n) + v(n) \tag{1}$$

where $x(n)$, s(n) and $v(n)$ denote noisy speech, clean speech and additive noise respectively. n represents the time index.

After framing and windowing, the short-time Fourier transform (STFT) of signal can be written as:

$$X(k,f) = \sum_{m=0}^{M-1} x(k,m) e^{-j\frac{2\pi mf}{M}} \quad f = 0, 1, \cdots, M-1 \tag{2}$$

where $X(k, f)$ is the spectrum of kth frame temporal signal $x(k, m)$. f is frequency bin index, and M is the length of STFT.

Logarithmic power spectrum is defined as:

$$X_s(k,f) = 10 \log_{10}[|X(k,f)|^2] \tag{3}$$

According to STFT symmetry, the first $M/2$ logarithmic power spectrum of M/2 consecutive frames are spliced together to obtain a two-dimensional LPS $C(l)$, which is defined as:

$$C(l)=\begin{bmatrix} X_S\left(\frac{M}{2}\cdot l,\frac{M}{2}-1\right) & X_S\left(\frac{M}{2}\cdot l+1,\frac{M}{2}-1\right) & \cdots & X_S\left(\frac{M}{2}\cdot l+\frac{M}{2}-1,\frac{M}{2}-1\right) \\ \vdots & \vdots & \ddots & \vdots \\ X_S\left(\frac{M}{2}\cdot l,1\right) & X_S\left(\frac{M}{2}\cdot l+1,1\right) & \cdots & X_S\left(\frac{M}{2}\cdot l+\frac{M}{2}-1,1\right) \\ X_S\left(\frac{M}{2}\cdot l,0\right) & X_S\left(\frac{M}{2}\cdot l+1,0\right) & \cdots & X_S\left(\frac{M}{2}\cdot l+\frac{M}{2}-1,0\right) \end{bmatrix} \tag{4}$$

The training target is the IRM matrix, which IRM is calculated using the following formula:

$$IRM(k,f)=\left(\frac{S(k,f)^2}{S(k,f)^2+V(k,f)^2}\right)^{\beta} \tag{5}$$

where $S(k, f)$ represents the spectrum of the clean speech $s(n)$ after preprocessing and STFT, and $V(k, f)$ is the spectrum of the noise. The adjustable parameter β is 0.5.

The IRM matrix $R(l)$ corresponding to $C(l)$ is computed as follow:

$$R(l)=\begin{bmatrix} IRM\left(\frac{M}{2}\cdot l,\frac{M}{2}-1\right) & IRM\left(\frac{M}{2}\cdot l+1,\frac{M}{2}-1\right) & \cdots & IRM\left(\frac{M}{2}\cdot l+\frac{M}{2}-1,\frac{M}{2}-1\right) \\ \vdots & \vdots & \ddots & \vdots \\ IRM\left(\frac{M}{2}\cdot l,1\right) & IRM\left(\frac{M}{2}\cdot l+1,1\right) & \cdots & IRM\left(\frac{M}{2}\cdot l+\frac{M}{2}-1,1\right) \\ IRM\left(\frac{M}{2}\cdot l,0\right) & IRM\left(\frac{M}{2}\cdot l+1,0\right) & \cdots & IRM\left(\frac{M}{2}\cdot l+\frac{M}{2}-1,0\right) \end{bmatrix} \tag{6}$$

2.1 Masking Mapping

The masking-based method uses multi-frame features to predict the masking of a certain frame [15,24], as shown in Fig. 2a. The methods are divided into two categories: causal and non-causal one, which use different frames to estimate the masking (as shown by the blue dashed box). Generally speaking, causal speech enhancement is closer to actual application scenarios. The mapping-based method realizes the spectrum-to-spectrum mapping [23,25], as shown in Fig. 2b. This method maps the noisy spectrum directly to its corresponding clean spectrum.

But both of these two methods have the own shortcomings. First of all, the masking method ignores the spectral correlation between the consecutive frames and cannot make full use of the two-dimensional convolution kernel. Secondly, although the spectrum mapping utilizes the two-dimensional information of the spectrogram, the masking can provide the richer information than that of the spectrogram through the comparison of Figs. 2a and 2b.

Based on the above analysis, MM is presented to estimate the masking matrix of multi frames on the LPS, as shown in Fig. 2c. The difference between MM and spectrum mapping is that the training target is no longer the spectrum of clean speech, but the IRM matrix, and the difference between the

masking method and the proposed method is that MM estimates the IRM matrix, rather than the IRM for a single frame.

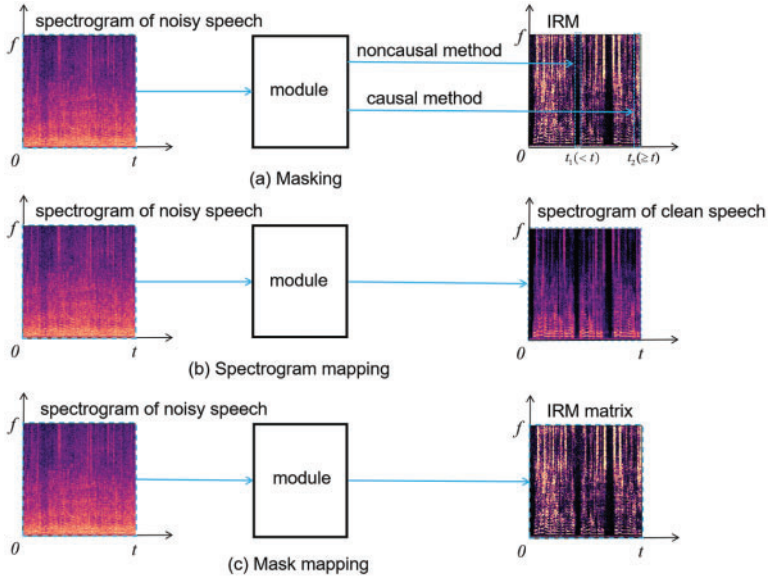

Figure 2: The training target of different method

2.2 Mask-mapping Based on RDN

The structure of proposed MM-RDN is shown in Fig. 3. The network contains down-sampling, dense feature extraction module stacked by residual dense block (RDB) and up-sampling. In Fig. 3, k represents the size of convolution kernel in convolution layer (Conv) and deconvolution layer (Deconv), o is the number of convolutional kernels, and s represents the convolution step. The down-sampling extracts the local and structural features, and also reduces the size of the feature maps by the convolutional kernel. The Conv layer is followed by batch normalization (BN), Dropout [26] and ReLU, which significantly reduces the computation cost and parameters load, and also increases the receptive field. Then distinguishable features are extracted by a stack of 6 RDBs. The up-sampling restores the feature map through convolutional layer with a step size of 1/2. The skip connections between down-sampling and up-sampling provide the combination of local and global feature and avoid the gradient vanishing.

The following sub-section describes the residual block and dense block of RDB in detail.

2.3 Residual Block and Dens Block

The structure of residual block is showing in Fig. 4. The skip connection structure alleviates the problem of gradient vanishing and network degradation, and can well deal with the problems caused by network Deeping.

DenseNet [27] is composed of several dense blocks (DB) as shown in Fig. 5. In DB, there is a skip connection between any two layers, and the input of each layer is the union of outputs of all the previous layers, which means the features learned in certain layer are the input for the all subsequent layers. DenseNet not only alleviates the problem of gradient vanishing, but also realizes the multiplexing of features extracted in the hidden layers.

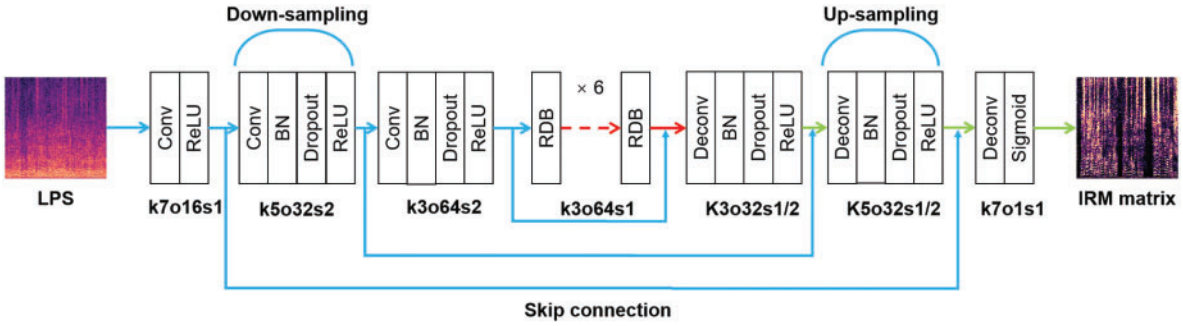

Figure 3: The structure of MM-RDN

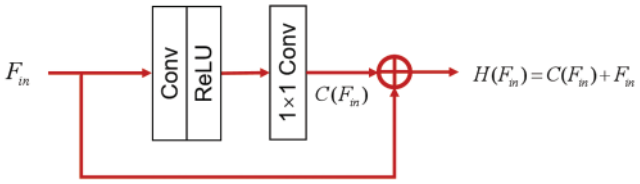

Figure 4: The structure of residual block

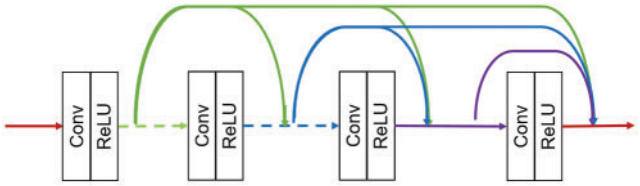

Figure 5: The structure of DB

RDB [22] combines the residual block and dense block, as shown in Fig. 6. RDB can not only obtain the state from the previous RDB, but also make full use of the feature of all layers through the local dense connections. The CM mechanism ensures that the previous RDB output can pass to each layer of the current RDB, which is formed by dense connection, local feature fusion (LFF) and local residual learning (LRL).

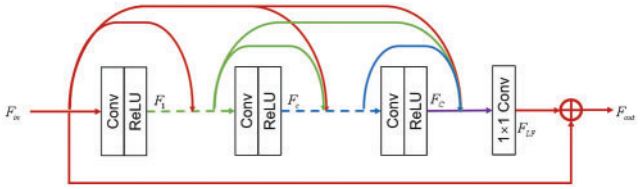

Figure 6: The structure of RDB

For an RDB, F_{in} denotes the input, then the output of cth convolutional layer of RDB F_c can be expressed as:

$$F_c = \sigma\left(W_c\left[F_{in}, F_1, \cdots F_{c-1}\right]\right) \tag{7}$$

where σ denotes ReLU [28] activation function and W_c is the weight of the cth Conv layer. $[\cdot]$ refers to the concatenation of the input.

The input of the last convolutional layer is the local features of all convolutional layers to obtain the local feature fusion F_{LF}, which is formulates as

$$F_{LF} = H_{LFF}\left([F_{in}, F_1, \cdots, F_c, \cdots F_C]\right) \tag{8}$$

where H_{LFF} denotes the 1×1 convolutional layer and the C is the number of the convolutional layers.

The RDB output can be obtained by add the input and the fused local features, which realizes the local residual learning:

$$F_{out} = F_{in} + F_{LF} \tag{9}$$

3 Simulation Setup and Result Analysis

3.1 Simulation Setup

To evaluate the proposed algorithm, clean speech signals are taken from the CHAINS corpus [29]. The dataset consists of recordings of 36 speakers. Four fables' sentences by 9 males and 9 females are used in training while 33 sentences from the TIMIT corpus of 3 males and 3 females are used in testing. The speakers of the training differ from that of the testing. 4 types of noise (babble, factory, pink, white) from the NOISEX-92 database [30] are added to the mentioned utterances at 4 different SNR, i.e., -5, 0, 5 and 10 dB. In addition, 3 untrained types of noise (baccaneer2, leopard, f16) at SNR -5, 0, 5 and 10 dB are used to test the generalization of the algorithms. Besides, untrained SNR of -7.5, -2.5, 2.5, 7.5, 12.5 dB with untrained noises are also added to the testing dataset. The sampling rate is 16 kHz.

To obtain the spectrum, the framing length is 256 with an overlap of 192 samples. After Hamming windowing, 256 points STFT is performed on each frame. As described above, the dimension of LPS is 128×128 representing the log-power spectrum of consecutive frame. In the proposed MM-RDN method, we utilized 2 down-sampling blocks and 2 up-sampling blocks. RDN has 6 RDBs with 3 skip connections in Fig. 3. The probability value of Dropout is 0.5 to increase generalization. Adam optimizer [31] optimizes the network with a learning rate of 0.0002 under the mean square error (MSE) criterion, and the hyper-parameters of momentum decay are set to 0.9 and 0.999 respectively. The model was trained for 10 epochs.

In the simulation, we firstly discuss the effect of the frame length on the performance of the proposed algorithm. The frame length is set to 128, 64, and 32, respectively, and the corresponding algorithms are denoted as MM-RDN128, MM-RDN64, and MM-RDN32.

To evaluate the quality of the enhanced speech, source to distortion radio (SDR) [32], PESQ [33] (from $-$ 0.5 to 4.5), mean opinion score (MOS) prediction of the intrusiveness of background noise (CBAK) (from 1 to 5), extended short-time objective intelligibility (ESTOI) [34] (from 0 to 1) and MOS prediction of the overall effect (COVL) (from 1 to 5) [35] are selected. SDR is used to estimate the overall distortion of the signal. PESQ and ESTOI are parameters for evaluating the speech perceptual quality and intelligibility, respectively. CBAK and COVL are comprehensive indicators related to subjective evaluation. In Section 3.3, the CRN [21] method which reached SOTA level in enhancement speech is compared with the proposed method MM-RDN of the most appropriate window length both in the matched and unmatched environments.

3.2 Effect of Frame Length on Performance of MM-RDN

In this section, the performance of MM-RDN with different frame length is compared in the matched noisy environment, and the results are shown in Tab. 1. For the unmatched environment, the testing dataset has different noise type or different SNR with the training dataset. Tab. 2 is the comparison results for MM-RDN with different frame lengths, in which only the noise types are different in the testing and the training. Tab. 3 presents the results on trained noise type and untrained SNR. Specifically, the SNRs in the testing are −5, 0, 5 and 10 dB, while the SNRs in the training are −7.5, −2.5, 2.5, 7.5 and 12.5 dB. Here, the noise types of testing dataset are the same as that of the training dataset. Tab. 4 displayed the results of MM-RDN with different frame length on untrained noise and untrained SNR.

Table 1: Metrics of noisy and enhanced speech in matched environments for different frame lengths

Model	Noisy			MM-RDN32			MM-RDN64			MM-RDN128		
SNR(dB)	SDR	PESQ	ESTOI	SDR	PESQ	ESTOI	SDR	PESQ	ESTOI	SDR	PESQ	ESTOI
−5	−5.098	1.035	0.266	2.061	1.082	0.378	2.510	1.104	0.399	**3.210**	**1.121**	**0.410**
0	−0.226	1.043	0.414	6.617	1.210	0.558	6.848	1.252	0.580	**7.239**	**1.290**	**0.589**
5	4.732	1.086	0.573	10.637	1.460	0.716	10.761	1.529	0.733	**10.985**	**1.592**	**0.739**
10	9.719	1.202	0.722	14.392	1.863	0.830	14.562	1.954	0.841	**14.671**	**2.050**	**0.844**

Table 2: Metrics of noisy and enhanced speech on unseen noise type for different window lengths

Model	Noisy			MM-RDN32			MM-RDN64			MM-RDN128		
SNR(dB)	SDR	PESQ	ESTOI	SDR	PESQ	ESTOI	SDR	PESQ	ESTOI	SDR	PESQ	ESTOI
−5	−4.178	1.043	0.334	−1.375	1.066	0.377	−1.220	1.068	0.379	**−0.478**	**1.076**	**0.387**
0	0.722	1.075	0.467	4.169	1.135	0.518	4.273	1.137	0.523	**5.137**	**1.161**	**0.533**
5	5.691	1.157	0.608	9.281	1.311	0.660	9.372	1.321	0.670	**9.972**	**1.381**	**0.679**
10	10.681	1.335	0.740	13.780	1.657	0.786	13.936	1.699	0.796	**14.205**	**1.791**	**0.801**

Table 3: Metrics of noisy and enhanced speech on unseen SNR for different window lengths

Model	Noisy			MM-RDN32			MM-RDN64			MM-RDN128		
SNR(dB)	SDR	PESQ	ESTOI	SDR	PESQ	ESTOI	SDR	PESQ	ESTOI	SDR	PESQ	ESTOI
−7.5	−7.457	1.048	0.203	−0.524	1.054	0.294	0.053	1.066	0.309	**0.971**	**1.077**	**0.322**
−2.5	−2.680	1.039	0.337	4.433	1.133	0.469	4.758	1.164	0.490	**5.281**	**1.191**	**0.502**
2.5	2.247	1.059	0.493	8.676	1.317	0.642	8.828	1.372	0.662	**9.128**	**1.422**	**0.670**
7.5	7.224	1.131	0.650	12.540	1.643	0.779	12.661	1.722	0.793	**12.833**	**1.806**	**0.798**
12.5	12.217	1.308	0.785	16.180	2.118	0.870	16.448	2.218	0.879	**16.505**	**2.317**	**0.880**

Table 4: Metrics of noisy and enhanced speech on unseen noise type and unseen SNR

Model	Noisy			MM-RDN32			MM-RDN64			MM-RDN128		
SNR(dB)	SDR	PESQ	ESTOI	SDR	PESQ	ESTOI	SDR	PESQ	ESTOI	SDR	PESQ	ESTOI
−7.5	−6.568	1.038	0.274	−4.268	1.052	0.312	−4.123	1.054	0.313	**−3.535**	**1.060**	**0.318**
−2.5	−1.742	1.055	0.399	1.448	1.091	0.447	1.565	1.093	0.449	**2.418**	**1.107**	**0.458**
2.5	3.202	1.107	0.573	6.786	1.205	0.589	6.888	1.210	0.597	**7.643**	**1.249**	**0.607**
7.5	8.184	1.231	0.676	11.613	1.459	0.727	11.719	1.482	0.737	**12.142**	**1.561**	**0.745**
12.5	13.179	1.472	0.797	15.778	1.901	0.837	16.039	1.962	0.845	**16.150**	**2.065**	**0.848**

In order to clearly display the performance of each indicator of the algorithm, we consolidate the data of Tabs. 1–4 and draw the line chart. The three algorithm performance data in the situation of matched noise type are combined based on the contents of Tabs. 1 and 3 and the comparative incremental results of each parameter are shown in Figs. 7–9. For the situation of unseen noise type are combined based on the contents of Tabs. 2 and 4 and the comparative incremental results of each parameter are shown in Figs. 10–12.

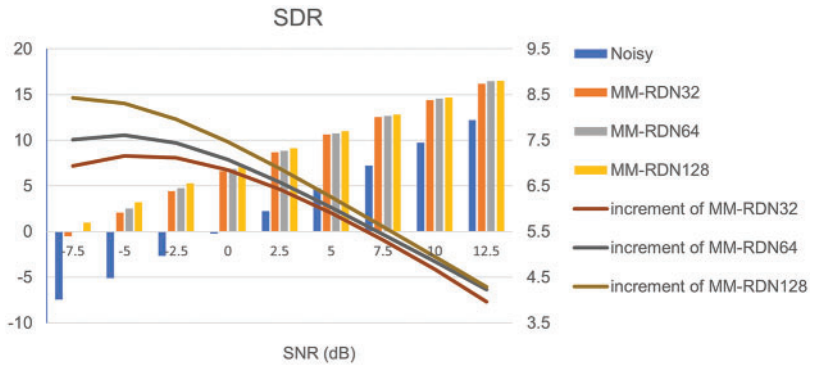

Figure 7: Comparation on SDR between MM-RDN on matched noise for different window lengths

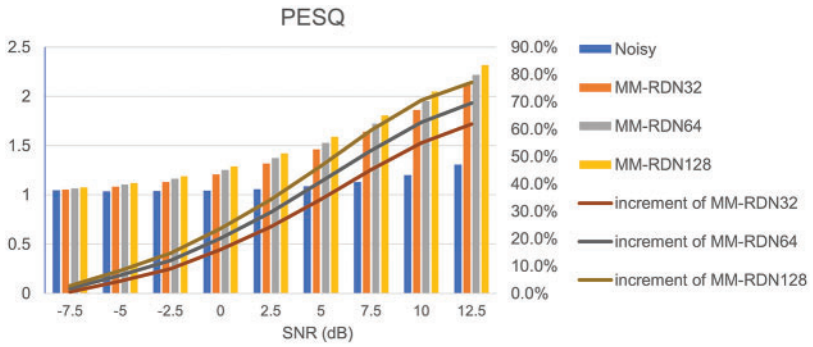

Figure 8: Comparation on PESQ between MM-RDN on matched noise for different window lengths

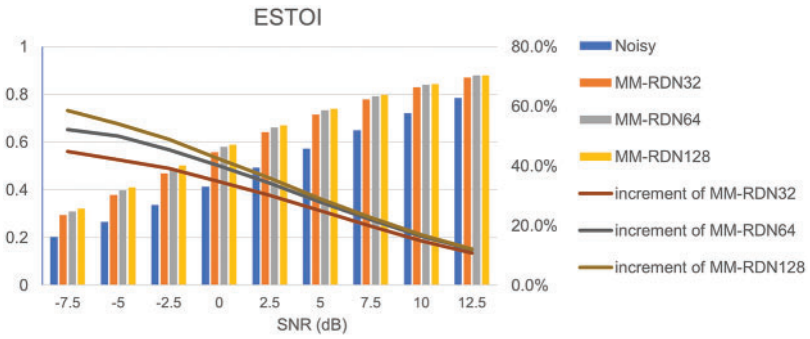

Figure 9: Comparation on ESTOI between MM-RDN on matched noise for different window lengths

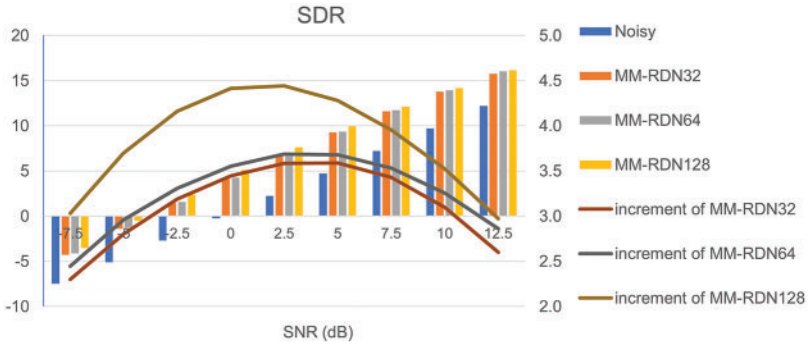

Figure 10: Comparation on SDR between MM-RDN on unseen noise for different window lengths

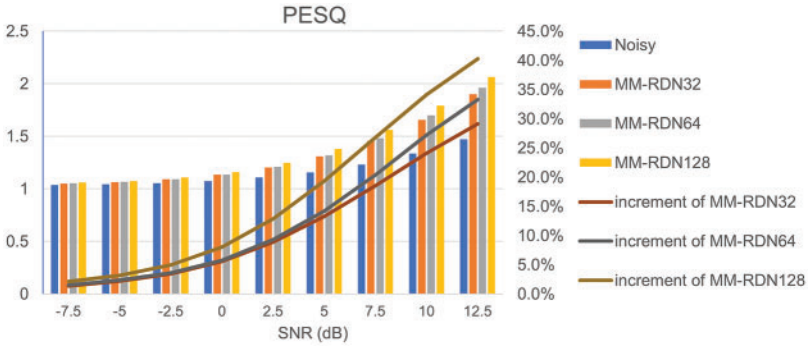

Figure 11: Comparation on PESQ between MM-RDN on unseen noise for different window lengths

It can be seen that the evaluation index of MM-RDN at all SNR is better than that of noisy speech, which means the MM-RDN can effectively improve the quality and intelligibility. In addition, MM-RDN128 obtains the best results, which indicates that the longer frame length can get better performance. Since the proposed algorithm uses a convolutional network to extract the high-level features of LPS, the longer the frame length, the better the convolution operation can capture the long-term and short-term correlations of speech features. Similarly, the frame length cannot be increased indefinitely. Due to the short-term stability of the speech, the frame length is too long to destroy the correlation between the LPS of adjacent frames, and the convolution operation cannot extract accurate high-level features. From Fig. 7, the incremental trend of SDR is consistent. When the frame length

is 128 and 64, as the SNR increases, SDR increment relative to the original noisy speech gradually decreases. This shows that increasing the frame length can more effectively improve the SDR in a low SNR environment. However, when the frame length is 32, the frame length is too small and the corresponding speech duration is too short to provide enough information for network learning, which limits the performance improvement of the MM-RDN algorithm. Also, Fig. 8 shows that MM-RDN can effectively improve the PESQ of enhanced speech. With the increase of SNR, the increase of PESQ of the algorithm also increases, indicating that MM-RDN can better improve the speech quality in the case of high SNR. Moreover, the longer the frame length, the more obvious the improvement of PESQ, which indicates that the frame length affects the improvement of the algorithm in speech quality. For the ESTOI in Fig. 9, the long frame length can improve the speech intelligibility. In general, under the matched noise type, MM-RDN still shows a certain generalization to the noise SNR.

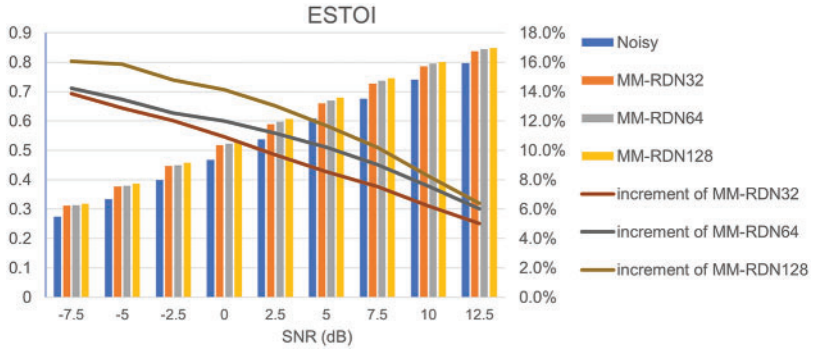

Figure 12: Comparation on ESTOI between MM-RDN on unseen noise for different window lengths

As shown in Figs. 10–12, the proposed algorithm can still effectively improve the perceptual quality of speech under unseen noise environment. Compared with Figs. 8, 9, 11 and 12 shows that the change trend of the algorithm for PESQ and ESTOI is consistent with the matched noise situation. Compared with the results of Fig. 7, when the noise is unseen, from the Fig. 10, the SDR increment of the enhanced speech does not always decrease, but first rises and then falls. The overall incremental data is lower the matching noise case. The results indicate that it is more difficult to improve the SDR at a low SNR unseen acoustic environment. It shows that the algorithm has limited generalization to unseen environment. At the same time, when the frame length is increased from 64 to 128, the SDR is improved more obvious than when the frame length is increased from 32 to 64, which means that the increase in frame length can compensate for the algorithm's impact on performance under unseen acoustic environments. Also, MM-RDN128 with longest frame length maintains the best performance, which indicates that the algorithm performance is related to frame length.MM-RDN can effectively improve the SDR, PSEQ and ESTOI of noisy speech, and the longer the frame length, the more obvious the SDR, PESQ and ESTOI improvement. Moreover, the performance gap between different frame lengths also increases. The results indicate that MM-RDN has certain generalization to noise type.

In general, through the above simulations, the following conclusions can be drawn: 1. MM-RDN can effectively improve the speech quality and intelligibility in different environments. 2. The increase of the frame length has a positive effect on the performance improvement of the proposed algorithm. 3 Increasing the length of the frame can more significantly improve the SDR and ESTOI at low SNR, and improve PESQ at high SNR.

Therefore, 128 is selected as the frame length of MM-RDN in the algorithm comparison in the following section, where MM-RD128 is denoted as MM-RDN.

3.3 Simulation of MM-RDN and Other Model Results and Analysis

Firstly, the performance of CRN and MM-RDN are compared in the matched noisy environment, and the results are shown in Tab. 5. Here noisy speech means unprocessed speech. For the unmatched environment, the testing dataset has different noise type or different SNR with the training dataset. Tab. 6 is the comparison results for MM-RDN and CRN, in which only the noise types are different in the testing and the training. Tab. 7 presents the results on trained noise type and untrained SNR. Specifically, the SNRs in the testing are −5, 0, 5 and 10 dB, while the SNRs in the training are −7.5, −2.5, 2.5, 7.5 and 12.5 dB, while the noise types are the same. Tab. 8 displayed the results on untrained noise and untrained SNR.

Table 5: Metrics of noisy and enhanced speech in matched environments for different algorithm

Model	Noisy				CRN				MM-RDN			
SNR(dB)	PESQ	CBAK	ESTOI	COVL	PESQ	CBAK	ESTOI	COVL	PESQ	CBAK	ESTOI	COVL
−5	1.035	1.185	0.266	1.089	1.086	1.518	0.367	1.264	**1.121**	**1.579**	**0.410**	**1.365**
0	1.043	1.461	0.414	1.187	1.212	1.926	0.524	1.596	**1.290**	**2.007**	**0.589**	**1.718**
5	1.086	1.825	0.573	1.383	1.457	2.371	0.689	2.016	**1.592**	**2.471**	**0.739**	**2.152**
10	1.202	2.243	0.722	1.707	1.850	2.862	0.811	2.513	**2.050**	**2.983**	**0.844**	**2.683**

Table 6: Metrics of noisy and enhanced speech on unseen noise type

Model	Noisy				CRN				MM-RDN			
SNR (dB)	PESQ	CBAK	ESTOI	COVL	PESQ	CBAK	ESTOI	COVL	PESQ	CBAK	ESTOI	COVL
−5	1.037	1.197	0.275	1.048	1.062	1.329	0.323	1.043	**1.073**	**1.317**	**0.339**	**1.020**
0	1.051	1.481	0.416	1.167	1.129	1.751	0.473	1.291	**1.143**	**1.757**	**0.494**	**1.278**
5	1.091	1.844	0.566	1.388	1.284	2.196	0.625	1.699	**1.332**	**2.234**	**0.648**	**1.736**
10	1.203	2.259	0.709	1.711	1.591	2.673	0.759	2.192	**1.701**	**2.745**	**0.780**	**2.281**

Table 7: Metrics of noisy and enhanced speech on unseen SNR

Model	Noisy				CRN				MM-RDN			
SNR (dB)	PESQ	CBAK	ESTOI	COVL	PESQ	CBAK	ESTOI	COVL	PESQ	CBAK	ESTOI	COVL
−7.5	1.048	1.108	0.203	1.067	1.058	1.347	0.266	1.154	**1.077**	**1.388**	**0.322**	**1.232**
−2.5	1.039	1.308	0.337	1.130	1.136	1.714	0.434	1.415	**1.191**	**1.787**	**0.502**	**1.529**
2.5	1.059	1.636	0.493	1.270	1.371	2.143	0.611	1.796	**1.422**	**2.234**	**0.670**	**1.924**

(Continued)

Table 7: Continued

Model	Noisy				CRN				MM-RDN			
SNR (dB)	PESQ	CBAK	ESTOI	COVL	PESQ	CBAK	ESTOI	COVL	PESQ	CBAK	ESTOI	COVL
7.5	1.131	2.026	0.650	1.531	1.634	2.610	0.755	2.254	**1.806**	**2.722**	**0.798**	**2.409**
12.5	1.308	2.476	0.785	1.914	2.095	3.119	0.855	2.786	**2.317**	**3.250**	**0.880**	**2.968**

Table 8: Metrics of noisy and enhanced speech on unseen noise type and unseen SNR

Model	Noisy				CRN				MM-RDN			
SNR (dB)	PESQ	CBAK	ESTOI	COVL	PESQ	CBAK	ESTOI	COVL	PESQ	CBAK	ESTOI	COVL
−7.5	1.035	1.107	0.214	1.020	1.051	1.167	0.254	1.011	**1.061**	**1.151**	**0.266**	**1.002**
−2.5	1.042	1.322	0.343	1.096	1.085	1.534	0.396	1.137	**1.097**	**1.529**	**0.415**	**1.111**
2.5	1.065	1.656	0.490	1.263	1.189	1.969	0.549	1.481	**1.218**	**1.993**	**0.573**	**1.492**
7.5	1.135	2.044	0.639	1.536	1.414	2.429	0.695	1.933	**1.493**	**2.484**	**0.719**	**1.998**
12.5	1.302	2.490	0.772	1.914	1.811	2.929	0.815	2.470	**1.961**	**3.020**	**0.833**	**2.590**

Consolidate the data of Tabs. 5–8 and draw the line chart. The algorithm performance data in the situation of matched noise type are combined based on the contents of Tabs. 5 and 7 and the comparative incremental results of each parameter are shown in Figs. 13–16. For the situation of unseen noise type are combined based on the contents of Tabs. 6 and 8 and the comparative incremental results of each parameter are shown in Figs. 17–20.

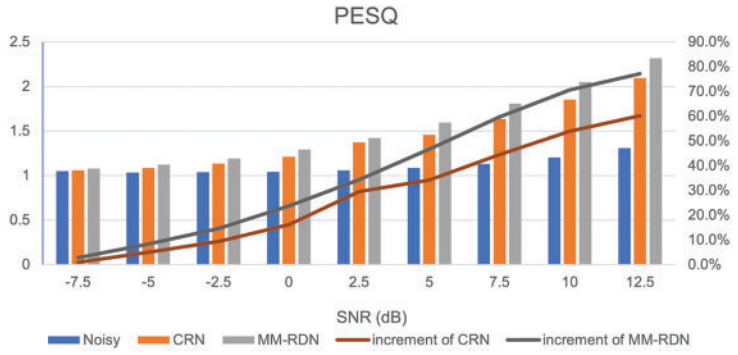

Figure 13: Comparation on PESQ between MM-RDN and CRN on matched noise

It can be seen that MM-RDN gets the best scores at all the SNR on all the metrics, which means it not only effectively reduces the noise, but also improves the perceptual quality and intelligibility of enhanced speech. Additionally, for MM-RDN, Figs. 13–14 shows that the perceptual quality is improved more obviously under high SNR, and Figs. 15 and 16 shows that the intelligibility is improved more obviously under low SNR condition. From Figs. 13–16, the performance of MM-RDN is stable. With matching noise types, the metrics maintain the similar trend for different SNR which demonstrates the robustness and generalization of the propose algorithm to the SNR.

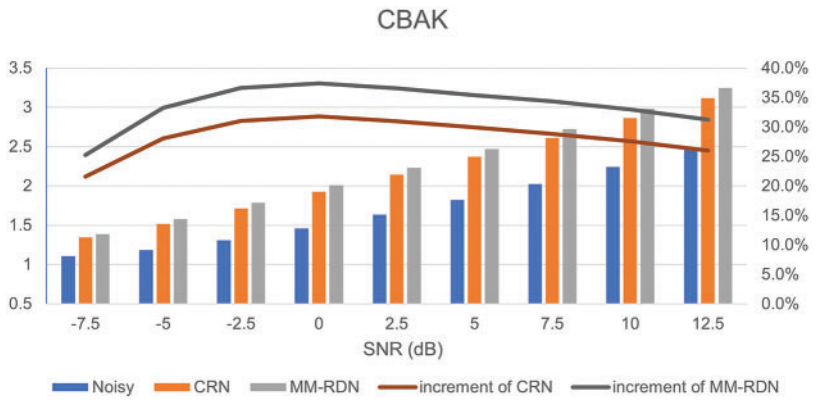

Figure 14: Comparation on CBAK between MM-RDN and CRN on matched noise

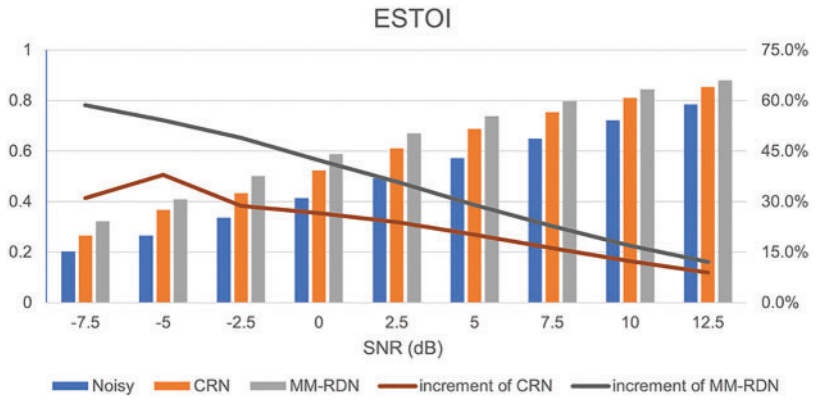

Figure 15: Comparation on ESTOI between MM-RDN and CRN on matched noise

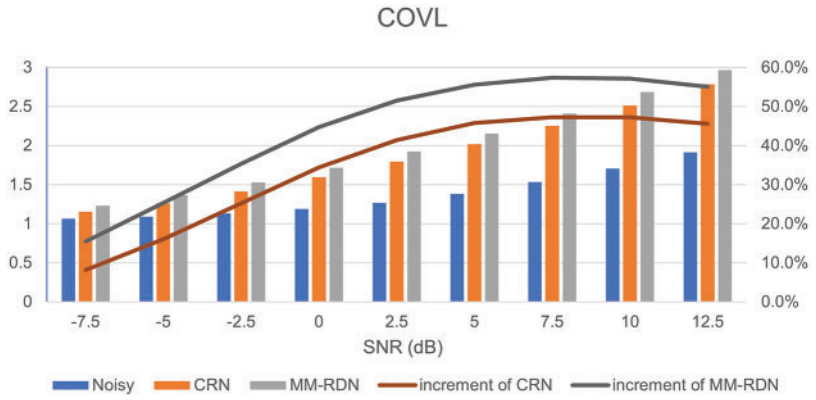

Figure 16: Comparation on COVL between MM-RDN and CRN on matched noise

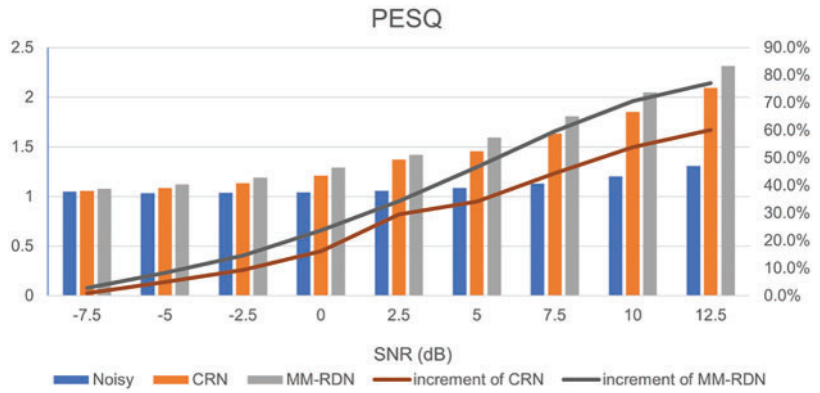

Figure 17: Comparation on PESQ between MM-RDN and CRN on unseen noise

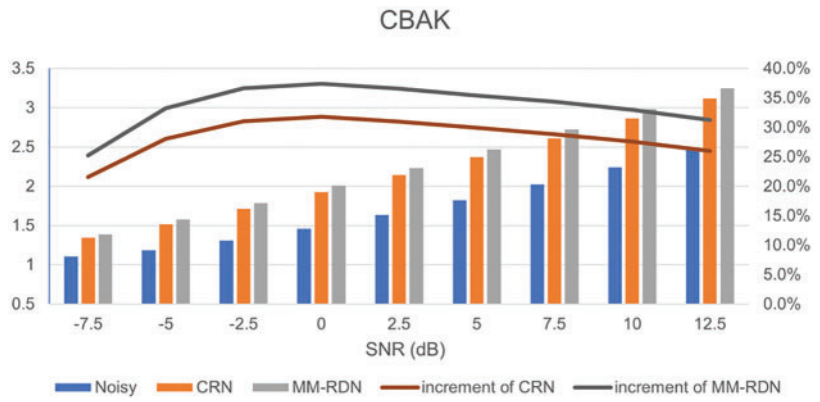

Figure 18: Comparation on CBAK between MM-RDN and CRN on unseen noise

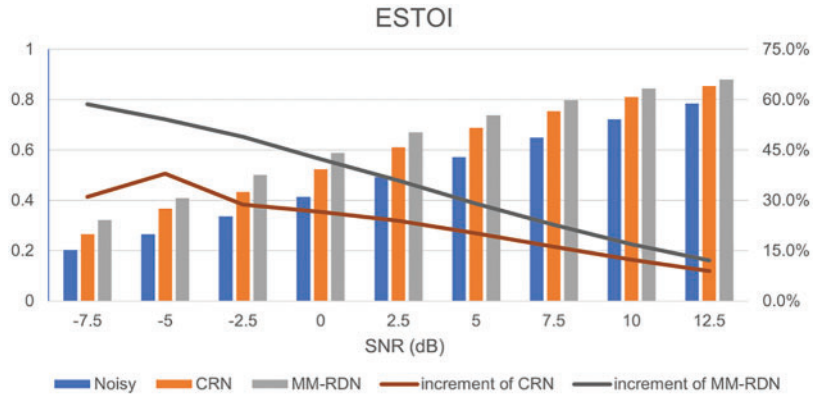

Figure 19: Comparation on ESTOI between MM-RDN and CRN on unseen noise

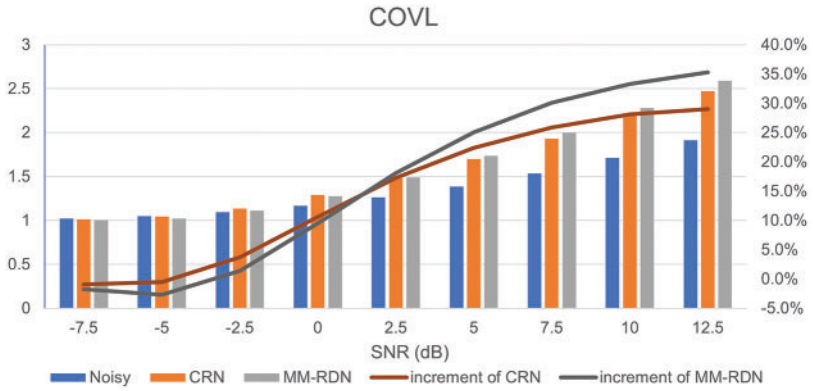

Figure 20: Comparation on COVL between MM-RDN and CRN on unseen noise

As shown in Fig. 17–20, the performance metrics of MM-RDN are superior to that of CRN, including the quality and intelligibility. At low SNR, MM-RDN is slightly inferior to CRN in COVL, but MM-RDN is significantly superior when SNR is greater than zero. Also, PESQ and COVL of MM-RDN increases faster than CBAK and ESTOI, which means the perceptual of speech is significantly improved. The results show that MM-RDN has certain generalization in the overall effect to noise type.

Thus, the proposed method achieved the best results, which demonstrate the robustness and generalization to the SNR and noise type. In addition to the performance improvement, it should be noted that the number of parameters in MM-RDN is about 27.99% of that of CRN. As is known to all, a small network would avoid overfitting and assure the generalization of a model. Compared with CRN, MM-RDN not only has the better performance, but also reduces the computation cost and parameters load.

In terms of the signal quality and intelligibility, MM-RDN is more robust and generalizable to noise type and SNR than CRN. It suggested that the proposed algorithm has better performance than the masking-based and spectrum mapping based method. Figs. 21 and 22 shows the waveform and amplitude spectrum of noisy speech and the enhanced speech by MM-RDN.

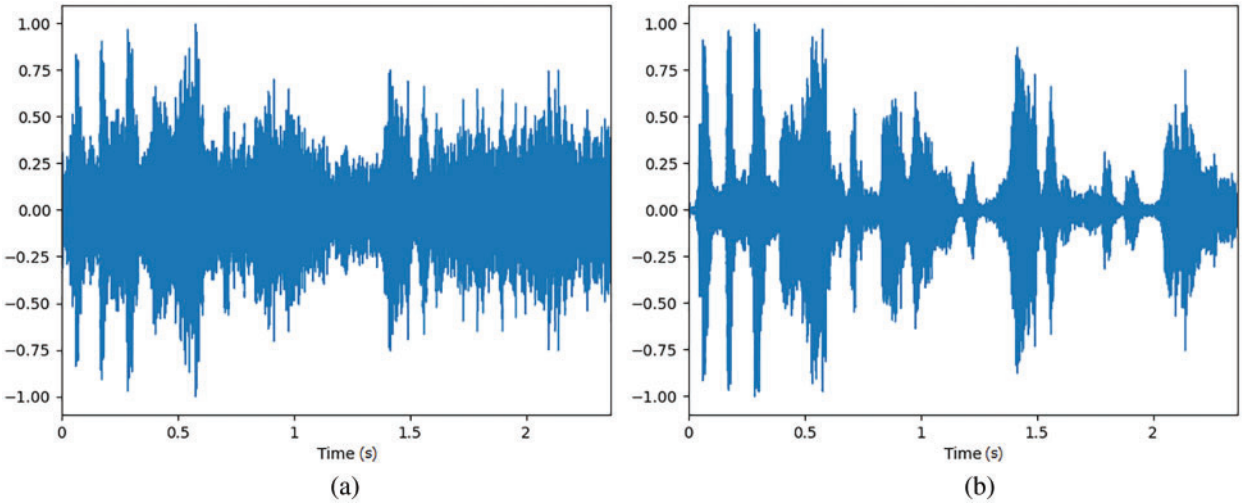

Figure 21: Waveform of (a) noisy speech and (b) enhanced speech by MM-RDN

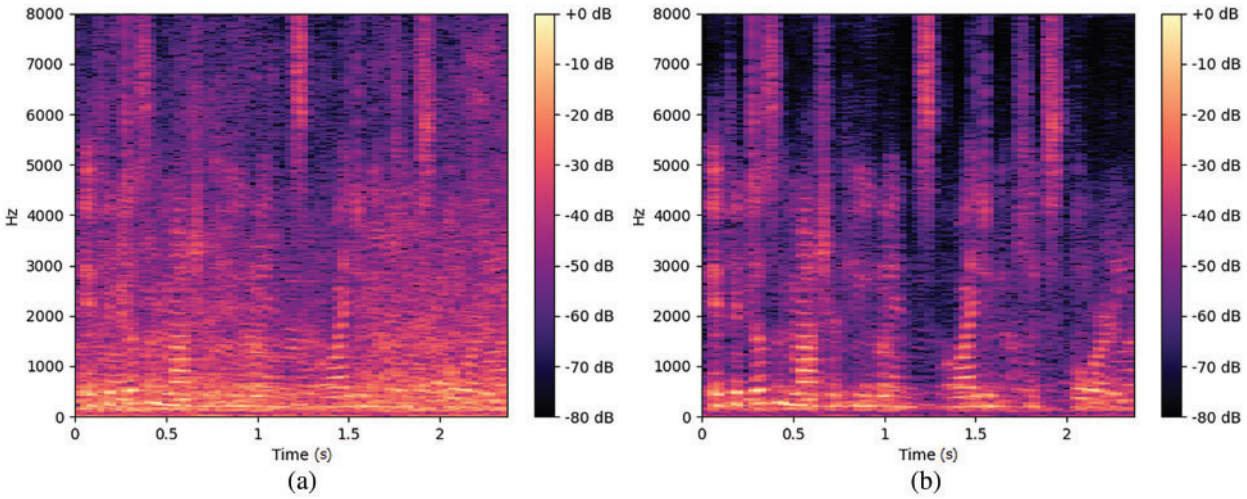

Figure 22: Amplitude spectrum of (a) noisy speech and (b) enhanced speech by MM-RDN

4 Conclusions

We introduce a MM based speech enhancement method using RDN and LPS of speech. The proposed MM-RDN reduces the network parameters load and avoids over-fitting through dense connection layers, LFF and LRL. At the same time, the method makes full use of the two-dimensional inter-frame information of LPS and the prior information of IRM, thus effectively improve perceptual quality and speech intelligibility of enhance speech, and also has generalization to the noise. Although the algorithm in this paper uses the two-dimensional inter-frame information, it is not enough to mine the timing characteristics of speech. In the future, the network may be further improved by means of timing convolution.

Funding Statement: This work is supported by the National Key Research and Development Program of China under Grant 2020YFC2004003 and Grant 2020YFC2004002, and the National Nature Science Foundation of China (NSFC) under Grant No. 61571106.

Conflicts of Interest: We declare that they have no conflicts of interest to report regarding the present study.

References

[1] P. Yan, J. Zou, Z. Li and X. Yang, "Infrared and visible image fusion based on nsst and rdn," *Intelligent Automation & Soft Computing*, vol. 28, no. 1, pp. 213–225, 2021.

[2] M. O. El-Habbak, M. A. Abdelalim, H. N. Mohamed, M. H. Abd-Elaty, A. M. Hammouda *et al.,* "Enhancing Parkinson's disease diagnosis accuracy through speech signal algorithm modeling," *Computers, Materials & Continua*, vol. 70, no. 2, pp. 2953–2969, 2022.

[3] G. Jyoshna, M. Zia and L. Koteswararao, "An efficient reference free adaptive learning process for speech enhancement applications," *Computers, Materials & Continua*, vol. 70, no. 2, pp. 3067–3080, 2022.

[4] Y. X. Wang and D. L. Wang, "Boosting classification-based speech separation using temporal dynamics," in *13th Annual Conf. of the Int. Speech Communication Association 2012*, Portland, OR, USA, pp. 1526–1529, 2012.

[5] X. G. Lu, Y. Tsao, S. Matsuda and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *14th Annual Conf. of the Int. Speech Communication Association 2013*, Lyon, France, pp. 436–440, 2013.
[6] Y. X. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.
[7] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Workshop on Speech Separation by Humans and Machines*, Montreal, Canada, pp. 181–197, 2005.
[8] D. S. Williamson, Y. X. Wang and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.
[9] H. Erdogan, J. R. Hershey, S. Watanabe and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, South Brisbane, QLD, Australia, pp. 708–712, 2015.
[10] Y. X. Wang, A. Narayanan and D. L. Wang, "On training targets for supervised speech separation," *IEEE-ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.
[11] D. L. Wang and J. T. Chen, "Supervised speech separation based on deep learning: An overview," in *IEEE-ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.
[12] L. Sun, J. Du, L. R. Dai and C. H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Conf. on Hands-Free Communications and Microphone Arrays*, San Francisco, CA, USA, pp. 136–140, 2017.
[13] P. S. Huang, M. Kim, M. Hasegawa-Johnson and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE-ACM Transactions on Audio Speech and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.
[14] Y. Xu, J. Du, L. R. Dai and C. H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.
[15] Y. Xu, J. Du, L. R. Dai and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.
[16] T. Kounovsky and J. Malek, "Single channel speech enhancement using convolutional neural network," in *IEEE Int. Workshop of Electronics Control Measurement Signals and Their Application to Mechatronics*, Donostia-San, Spain, pp. 1–5, 2017.
[17] Mustaqeem and S. Kwon, "1D-Cnn: Speech emotion recognition system using a stacked network with dilated cnn features," *Computers, Materials & Continua*, vol. 67, no. 3, pp. 4039–4059, 2021.
[18] L. Zhou, Q. Y. Zhong, T. Y. Wang, S. Y. Lu and H. M. Hu, "Speech enhancement via residual dense generative adversarial network," *Computer Systems Science and Engineering*, vol. 38, no. 3, pp. 279–289, 2021.
[19] L. C. Yann, Y. Bengio and H. Geoffrey, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[20] R. Pascanu, T. Mikolov and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Int. Conf. on Machine Learning*, Atlanta, GA, USA, pp. 2347–2355, 2013.
[21] F. Weninger, H. Erdogan, S. Watanabe, E., Vincent, J. Le Roux *et al.,* "Speech enhancement with lstm recurrent neural networks and its application to noise-robust ASR," in *Int. Conf. on Latent Variable Analysis and Signal Separation*, Tech Univ Liberec, Liberec, Czech Republic, pp. 91–99, 2015.
[22] Y. L. Zhang, Y. P. Tian, Y. Kong, B. N. Zhong and Y. Fu, "Residual dense network for image super-resolution," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 2472–2481, 2018.
[23] K. Tan and D. L. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *19th Annual Conf. of the Int. Speech Communication Association 2018*, Hyderabad, India, pp. 3229–3233, 2018.
[24] S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," in *18th Annual Conf. of the Int. Speech Communication Association 2017*, Stockholm, Sweden, pp. 1993–1997, 2017.

[25] H. Zhao, S. Zarar, I. Tashev and C. H. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Calgary, Canada, pp. 2401–2405, 2018.
[26] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever and R. R. Salakhutdinov, *Improving Neural Networks by Preventing co-Adaptation of Feature Detectors*, 2012. [Online]. Available: https://arxiv.org/abs/1207.0580.
[27] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 2261–2269, 2017.
[28] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, pp. 234–241, 2017.
[29] C. Fred, G. Marco, L. Thomas and J. Simko, "The chains corpus: Characterizing individual speakers," in *SPECOM-2006*, St Petersburg, Russia, pp. 431–435, 2006.
[30] A. Varga and H. J. M. Steeneken, "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 Communication*, vol. 12, no. 3, pp. 247–251, 1993.
[31] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. on Learning Representations*, San Diego, CA, USA, pp. 1–9, 2015.
[32] E. Vincent, R. Gribonval and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.
[33] ITU-T P.862.2, "Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs," *Telecommunication Standardization Sector of ITU*, 2007.
[34] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE-ACM Transactions on Audio Speech and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.
[35] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio Speech and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.