Embedding Encoder-Decoder With Attention Mechanism for Monaural Speech Enhancement

TIAN LAN\textsuperscript{1,2}, WENZHENG YE\textsuperscript{1}, YILAN LYU\textsuperscript{1}, JUNYI ZHANG\textsuperscript{2}, AND QIAO LIU\textsuperscript{1}

\textsuperscript{1}School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610051, China
\textsuperscript{2}Electromagnetic Spectrum Cognition and Management Key Laboratory of Hebei Province, Shijiazhuang 050081, China

Corresponding author: Qiao Liu (qliu@uestc.edu.cn)

This work was supported in part by the National Science of China under Grant U19B2028 and Grant 61772117, in part by the National Science and Technology Commission Innovation Project under Grant 19-H863-01-ZT-003, in part by the CETC Big Data Research Institute under Grant 10-2018039, in part by the Sichuan Hi-Tech Industrialization Program under Grant 2018GFW0150, and in part by the Fundamental Research Funds for the Central Universities under Grant ZYGX2019J077.

**ABSTRACT** The auditory selection framework with attention and memory (ASAM), which has an attention mechanism, embedding generator, generated embedding array, and life-long memory, is used to deal with mixed speech. When ASAM is applied to speech enhancement, the discrepancy between the voice and noise feature memories is huge and the separability of noise and voice is increased. However, ASAM cannot achieve desirable performance in terms of speech enhancement because it fails to utilize the time-frequency dependence of the embedding vectors to generate a corresponding mask unit. This work proposes a novel embedding encoder-decoder (EED), and a convolutional neural network (CNN) is used as decoder. The CNN structure is good at detecting local patterns, which can be exploited to extract correlation embedding data from the embedding array to generate the target spectrogram. This work evaluates a similar ASAM, EED with an LSTM encoder and a CNN decoder (RC-EED), RC-EED with an attention mechanism (RC-AEED), other similar EED structures and baseline models. Experiment results show that RC-EED and RC-AEED networks have good performance on speech enhancement task at low signal-to-noise ratio conditions. In addition, RC-AEED exhibits superior speech enhancement performance over ASAM and achieves better speech quality than do deep recurrent network and convolutional recurrent network.

**INDEX TERMS** Speech enhancement, embedding encoder-decoder, convolutional neural network, attention mechanism, neural network.

**I. INTRODUCTION**

Speech enhancement has attracted considerable research attention for several decades. It aims to remove noise or reverberation from a noisy speech signal and improve the signal’s intelligibility and quality. This challenging task is significant to some real-world applications, such as telephone conferencing, speech recognition systems, hearing aid devices, and conference recording. Obvious progress has been achieved in speech enhancement owing to the introduction of deep learning approaches, which outperform conventional methods, including spectral subtraction [1] and the Wiener filter method [2], which are based on stationary noise assumption. However, deep learning approaches still fail to achieve desirable performance in low signal-to-noise ratio (SNR) conditions.

Recently, several fully convolutional network (FCN)-based speech enhancement algorithms were proposed. The FCN only contains CNN layers and discards the fully connected layer. This structure was proposed by Long et al. [3], who applied it on semantic segmentation. They replaced a densely connected network with the FCN and achieved obvious improvement in pixel accuracy, mean accuracy, and mean intersection over union. In a special aspect, they reinterpreted this phenomenon as a trade-off, that is, the receptive field sizes of the filters are decreased to help extract information at a finer scale. Mao et al. [4] used the FCN structure for image denoising and proposed a symmetric convolution encoder and deconvolution decoder with skipping layers. [5], [6] introduced the FCN structure to speech enhancement and found that the performance of the proposed convolutional encoder-decoder (CED) can exceed that of the structure with RNN and fully connected network.

Some new structures that combine LSTMs and CNNs were proposed [7]–[9] and were proven to have promising...
performance in speech enhancement. Tan and Wang [8] developed the convolutional recurrent network (CRN), which inserts two LSTM layers into the middle of a CED network. The structure can be regarded as an extension of the CED in [5]. The experiment results show that CRN outperforms two special LSTM networks on trained and untrained speaker datasets when tested at −2 and −5 dB SNR levels. Zhao et al. [9] proposed the EHNet, which also merges CNN and LSTM components. The convolutional component is leveraged to exploit local patterns in a spectrogram, and the bidirectional recurrent component captures the dynamic correlations between consecutive frames. Under a high SNR condition, EHNet achieves better perceptual evaluation of speech quality (PESQ) [26] and word error ratio (WER) than does an RNN [10] and two DNNs [11], [12]. Ge et al. [7] also developed a similar structure composed of an attention mechanism. In this special structure, an enhanced frame is generated by the surrounding frames of input, and the frames with speech information are given more attention.

This study proposes the application of a unified auditory selection framework by modeling attention and memory [13] and an embedding decoder on speech enhancement, which is an LSTM and CNN combined structure. ASAM is an auditory selection and speech separation framework with attention and life-long memory, which employs the same embedding generation method in deep clustering [14]. It maps each T-F unit in the mixture spectrogram to a high dimensional vector and the mapped vectors form an embedding array. The life-long memory is updated by a memory vector, and the memory vector serves as priori speech feature to extract target voice feature in the embedding array. However, ASAM fails to utilize the time-frequency dependence of the embedding vectors to generate a corresponding mask unit. As shown in Fig. 1, every unit of the frequency-selective attentional filter’s mask [15] is generated by a single embedding from the array and the life-long memory, and the surrounding information is ignored.

Inspired by the successful introduction of CNNs in feature extraction in sentiment classification [16], [17], we develop a similar CNN structure to generate an enhanced spectrum from the embeddings. The CNN is also an FCN, and the enhanced speech spectrogram is directly generated by the last layer of the CNN. The spectrum generator is combined with an LSTM embedding generator to build an recurrent-convolutional embedding encoder-decoder network (RC-EED). CNNs are good at extracting local patterns and can ignore irrelevant data to form better representations [18]. A CNN and LSTM union structure was also used in [8], [9] to extract special local patterns from a spectrum and model the temporal dependencies in a latent space. In the proposed model, the CNN embedding decoder is utilized to detect the local spatial patterns of the embedding array and generate T-F units in the spectrum from the local embedding representations, which can leverage more correlative embeddings to form the spectrum. In addition, we find the similar memory block and attention mechanism in ASAM can be applicant in the RC-EED network and the embedding array can be filtered by a stacked mask. In our experience, the results show that the RC-EED and the RC-AEED have good performance on speech enhancement task in low SNR conditions and them both apparently outperform a similar ASAM model in different noise conditions. Specially, RC-AEED achieves better PESQ than RC-EED.

The rest of this paper is organized as follows. Section II introduces the proposed models. Section III presents the details of the experimental procedure and the analysis of the experimental results. Finally, section IV concludes the paper.

### II. ALGORITHM DESCRIPTION

The RC-EED and RC-AEED are shown in Fig. 2. RC-AEED is the solid part of the figure. With noisy and clean speech spectra as input, the model is composed of two embedding encoders, one memory block, and one embedding decoder.
embedding array, respectively. The clean speech embedding representation is accumulated into a memory vector in the storage region. An attention mechanism then takes the memory vector and the noisy speech embedding array as input to form an ideal radio mask (IRM) [19]–[21]. Afterward, the noisy speech embedding array is pointwise multiplied by a stacked IRM to develop an enhanced speech embedding array. The stacked IRM is a repeated overlay mask. Finally, an embedding decoder transforms the enhanced speech embedding array to the enhanced speech spectrum, which is the solid line part in Fig. 2.

The RC-EED has no attention structure and masking layer. Only an LSTM embedding encoder of noisy speech exists in the RC-EED. The dashed line part of Fig. 2 shows the dataflow between the encoder and the decoder.

A. EMBEDDING ENCODER
The encoder aims to map each T-F unit of the speech spectrum to a high-dimensional embedding. In ASAM, this structure is constructed by LSTM layers and a fully connected layer and then the output of the structure is reshaped to an embedding array. In [22], a spectrogram and a feature map generated from speech wave were concatenated to be a hybrid-domain feature map. Next, the new feature map is entered to a repeated structure of 1-d dilated CNN layers, the repeated number is 4.

Two embedding encoders are adopted in the embedding encoder-decoder models. One is similar to the LSTM block in ASAM, and the other has a similar structure to that in [22]. The LSTM encoder is made up of two LSTM layers and a fully connected layer. The structure of the CNN encoder in [22] is also followed to build the CNN encoder in comparison, but the repeated number is reduced to 3. As show in the Fig. 2, the CNN encoder is only used to deal with noisy spectrum. The EED with CNN encoder and CNN decoder is named CC-EED. The encoding process can be expressed as follows:

\[
\begin{align*}
A &= EE(X) \quad (1) \\
L &= FC(NN(X)) \quad (2) \\
A &= \text{reshape}(L) \quad (3)
\end{align*}
\]

where \(EE(\cdot)\) represents an embedding encoder network, input \(X\) represents the speech spectrum corrupted with noise, and \(A \in \mathbb{R}^{T \times F \times E}\) is the generated embedding array. Formulas (2) and (3) denote the intermediate process of (1), where \(NN(\cdot)\) represents the LSTM or CNN structure; \(FC(\cdot)\) refers to a fully connected layer; \(\text{reshape}(\cdot)\) is the reshape operation; and \(L \in \mathbb{R}^{T \times FE}\). The embedding representation of clean speech is only generated by an LSTM embedding encoder, and an embedding array of clean speech is used to generate a feature vector:

\[
a = \frac{\sum_{0 \leq t \leq T} \sum_{0 \leq f \leq F} AC_{t,f}}{TF} \quad (4)
\]

where \(AC\) represents an embedding array of clean speech, \(a\) represents the embedding representation of clean speech. Note that the generation process of the memory is slightly different from that in [13], of which is simpler, but we don’t find any apparent influence on the experiment results.

B. MEMORY BLOCK
A memory block with a \(1 \times E\) shape is developed in the proposed model. In every minibatch, embedding representations of the clean speech of different speakers are saved to the memory block. The operation producing the memory vector is the same as the method in [13], but the memory vector stores the general feature of different voices. The memory update formula can be expressed as

\[
m' = \frac{a + m}{||a + m||} \quad (5)
\]

where \(m \in \mathbb{R}^{E}\) represents the memory vector, and \(a\) is the embedding representation of clean speech. Note that the \(m\) is initialized to a zero vector.

C. ATTENTION MECHANISM AND MASK
The attention mechanism aims to calculate the mask by the noisy speech spectrum embedding array and the memory vector. Every element is related to the embedding at the corresponding position and the memory vector. The method to compute the attention weight is given by the following equation:

\[
\alpha_{t,f} = \text{sigmoid}(m \cdot A_{t,f}) \quad (6)
\]

where \(A_{t,f}\) represents any embedding in the noisy speech spectrum embedding array, and \(\alpha\) can be regarded as a T-F mask. In the ASAM, the mask is directly used to filter the enhanced spectrum from the mixture. In the proposed model, the mask is utilized to update the noisy embedding array. The mask has to be copied at first and then the same masks are stacked together to match the shape of the embedding array.

D. EMBEDDING DECODER
An enhanced speech spectrogram is reconstructed from the embedding array in this block. The process can be regarded as an inverse process of embedding generation. To overcome the problem of using a single embedding in ASAM, we propose to adopt a CNN structure to extract the features of local correlation embeddings, as illustrated in Fig. 3. Given that the
energy in the spectrogram is continuous along the time and frequency dimension, we suppose that the embedding array generated from the spectrogram has a similar property and the CNN is appropriate for exploiting rich local patterns. One way to construct an embedding decoder is using fully connected layers, but the structure of the neural network could introduce significant irrelevant information when each T-F unit is generated. Multiple structures of CNN are designed to be the embedding decoder, which can be expressed by the following formula:

\[ Y = ED(A') \]  

where \( ED(\cdot) \) represents the embedding decoder, \( Y \) represents the enhanced speech spectrogram, \( A' \) represents an embedding array.

### III. EVALUATIONS AND COMPARISONS

#### A. DATASET

This study selects TIMIT corpus and Noisex92 [23] as the speech and noise datasets, respectively. Noisex92 is a commonly used noise dataset. A total of 1984 utterances are randomly extracted from the TIMIT corpus and 5 noises are used in the training phase. The voices are corrupted with five types of noise at different SNRs, \([-5, -4, -3, -2, -1, 0]\) dB. The noises include Babble, Factory1, Destroyerops, F16, and White. Other different 300 utterances of seen and unseen speakers in TIMIT are randomly selected to construct the test set. Three unseen noises (i.e., Factory2, Leopard, and Hfchannel) and three seen noises (i.e., Babble, Factory1, and F16) from the NOISEX-92 corpus are adopted for testing. All waveforms are sampled at 16 kHz.

#### B. EXPERIMENT SETUP

The short-time Fourier transform (STFT) and a Hamming window with 512 and an overlap interval with 256 are used to extract the speech features from the waveform. We then obtain a speech spectrum with a shape of \(235 \times 257 \) and feed it to the neural network.

Three models are chosen as the reference approaches, including deep recurrent network (DRN) [24], CRN [8] and ASAM. Four proposed models with similar structure, such as RC-EED, convolutional-convolutional EED (CC-EED), RC-AEED and recurrent-fully_connected AEED (RF-AEED) are also evaluated. Note that DRN and CRN are completely different from the proposed RC-EED or RC-AEED, while ASAM, RC-EED, CC-EED, RF-AEED and RC-AEED have similar structures.

The ASAM baseline is similar to that in [13], the two embedding encoders of which consist of two unidirectional LSTM layers and a fully connected layer. In the embedding encoder of the noisy speech, each LSTM layer has 784 cells, and the fully connected layer has 257 units. An LSTM layer with 120 cells is used in the other embedding encoder. Two different embedding arrays of size \(235 \times 257 \times 40\) are then generated. The embedding array produced by clean speech forms a memory vector of size 40, which is subsequently stored in a memory block. The shape of the memory block is the same as the proposed method, and the memory block stores the speech features of all speakers.

RC-EED and CC-EED both have an embedding encoder-decoder structure. The embedding generator of the RC-EED is the same as that of the ASAM baseline. However, in the CC-EED, the generator is replaced by a similar CNN structure as that in [22], and the embedding decoders of the two models are two CNN layers with a kernel size of \(3 \times 3\). The first layer has 40 input channels and 10 output channels, while the second layer has 10 input channels and 1 output channel. The decoder of RC-AEED has the same structure parameters as above. The RF-AEED, which replaces the CNN embedding decoder with a fully connected network, is built to verify the effectiveness of the CNN decoder. The fully connected layers have 2048 and 257 units respectively. During the test phase, the two AEED models directly use the memory vector generated in training phase, and the memory vector will not be updated again.

Four different decoders in the RC-AEED model are also developed, and the performances of all the decoders are evaluated, as shown in Table 7.

The models are trained with the Adam optimizer at a learning rate of 0.001. The mean squared error in the time-frequency domain is used as the objective function. The models with a minibatch size of 32 are trained on the utterance level.

#### C. RESULTS AND ANALYSIS

Short-term objective intelligibility (STOI) [25], PESQ, and scale-invariant signal-to-distortion ratio (SI-SDR) [27] are used as evaluation metrics in the experiment. Tables 1, 2 present the experiment results on three seen noises and untrained speakers. Generally, ASAM cannot achieve satisfactory performance in speech enhancement. DRN, CRN, and the proposed methods have better performance than the ASAM in nearly all conditions. In the \(-10\) dB SNR condition, most of the STOI and PESQ metrics of the DRN are inferior to those of the CRN and the RC-AEED, indicating that CRN and RC-AEED have better stability in the low SNR environment. Table 2 shows that the proposed RC-AEED achieves the best PESQ among the seven models in all the noise conditions. When Babble is used as the test noise, RC-AEED generates 0.33, 0.56, 0.70, and 0.69 PESQ gains over the unprocessed mixtures in \(-10, -5, 0,\) and 5 dB SNR conditions. In Table 1, RC-AEED and CRN have the highest STOI scores. RC-AEED exhibits better performance when tested on F16 noise, while CRN has greater STOI gain when tested on Babble. Specifically, in the \(-10\) dB SNR case, RC-AEED exhibits the best performance in terms of the STOI and PESQ metrics.

Tables 3, 4 show the experiment results on three unseen noises and untrained speakers. The DRN, CRN, and proposed methods have better noise generalization than ASAM. RC-AEED achieves the best PESQ score in most noise conditions. The STOI score of RC-AEED is only inferior
TABLE 1. STOI comparisons based on seen noises between various models.

| Target | Babble | Factory1 | F16 |
|--------|--------|----------|-----|
|        | −10 dB | −5 dB    | 0 dB | 5 dB | −10 dB | −5 dB | 0 dB | 5 dB | −10 dB | −5 dB | 0 dB | 5 dB |
| Noisy  | 0.447  | 0.556    | 0.675 | 0.783 | 0.429  | 0.531 | 0.653 | 0.770 | 0.452  | 0.559 | 0.685 | 0.801 |
| DRN    | 0.463  | 0.642    | 0.785 | 0.871 | 0.526  | 0.692 | 0.801 | 0.876 | 0.520  | 0.682 | 0.807 | 0.884 |
| CRN    | 0.509  | 0.679    | 0.799 | 0.867 | 0.570  | 0.720 | 0.815 | 0.873 | 0.514  | 0.668 | 0.794 | 0.871 |
| ASAM   | 0.443  | 0.597    | 0.739 | 0.823 | 0.492  | 0.644 | 0.753 | 0.826 | 0.491  | 0.641 | 0.757 | 0.838 |
| RC-EED | 0.492  | 0.672    | 0.796 | 0.865 | 0.580  | 0.724 | 0.813 | 0.868 | 0.556  | 0.706 | 0.813 | 0.877 |
| RF-AEED| 0.475  | 0.644    | 0.776 | 0.847 | 0.560  | 0.705 | 0.799 | 0.854 | 0.532  | 0.692 | 0.801 | 0.862 |
| RC-AEED| 0.497  | 0.677    | 0.799 | 0.866 | **0.582** | 0.721 | 0.813 | 0.867 | **0.556** | 0.714 | **0.819** | 0.878 |

The generation of memory vector of ASAM here is the same as RC-AEED. We also evaluated the ASAM structure using original method of [13], and didn’t find it had any apparent influence on the experiment results.

TABLE 2. PESQ comparisons based on seen noises between various models.

| Target | Babble | Factory1 | F16 |
|--------|--------|----------|-----|
|        | −10 dB | −5 dB    | 0 dB | 5 dB | −10 dB | −5 dB | 0 dB | 5 dB | −10 dB | −5 dB | 0 dB | 5 dB |
| Noisy  | 0.88   | 1.17     | 1.53 | 1.92 | 0.83   | 1.07  | 1.42 | 1.81 | 0.88   | 1.16  | 1.50 | 1.89 |
| DRN    | 1.07   | 1.63     | 2.14 | 2.51 | 1.21   | 1.83  | 2.28 | 2.62 | 1.42   | 1.88  | 2.27 | 2.58 |
| CRN    | 1.11   | 1.65     | 2.10 | 2.45 | 1.31   | 1.82  | 2.23 | 2.52 | 1.36   | 1.68  | 2.10 | 2.46 |
| ASAM   | 0.91   | 1.30     | 1.80 | 2.20 | 0.96   | 1.45  | 1.90 | 2.24 | 0.99   | 1.43  | 1.87 | 2.25 |
| RC-EED | 1.13   | 1.67     | 2.18 | 2.58 | 1.43   | 1.95  | 2.37 | **2.69** | 1.38   | 1.86  | 2.28 | 2.64 |
| RF-AEED| 1.01   | 1.54     | 2.03 | 2.38 | 1.28   | 1.78  | 2.19 | 2.47 | 1.20   | 1.70  | 2.14 | 2.46 |
| RC-AEED| **1.21** | **1.73** | **2.23** | **2.61** | **1.46** | **1.97** | **2.39** | **2.69** | **1.46** | **1.93** | **2.34** | **2.68** |

TABLE 3. STOI comparisons based on unseen noises between various models.

| Target | Factory2 | Leopard | Hfchannel |
|--------|----------|---------|-----------|
|        | −10 dB   | −5 dB   | 0 dB      | 5 dB     | −10 dB   | −5 dB   | 0 dB   | 5 dB     |
| Noisy  | 0.541    | 0.649   | 0.752    | 0.838   | 0.713    | 0.765   | 0.805  | 0.842   |
| DRN    | 0.645    | 0.771   | 0.851    | **0.902** | 0.569    | 0.760   | **0.858** | **0.898** |
| CRN    | **0.684** | **0.792** | **0.855** | 0.892   | **0.724** | **0.803** | 0.850  | 0.879   |
| ASAM   | 0.575    | 0.708   | 0.797    | 0.853   | 0.642    | 0.723   | 0.806  | 0.847   |
| RC-EED | 0.677    | 0.784   | 0.847    | 0.886   | 0.681    | 0.788   | 0.846  | 0.877   |
| RF-AEED| 0.651    | 0.762   | 0.833    | 0.854   | 0.679    | 0.773   | 0.829  | 0.865   |
| RC-AEED| 0.680    | 0.787   | 0.848    | 0.886   | 0.718    | 0.800   | 0.847  | 0.876   |

TABLE 4. PESQ comparisons based on unseen noises between various models.

| Target | Factory2 | Leopard | Hfchannel |
|--------|----------|---------|-----------|
|        | −10 dB   | −5 dB   | 0 dB      | 5 dB     | −10 dB   | −5 dB   | 0 dB   | 5 dB     |
| Noisy  | 0.99     | 1.32    | 1.71     | 2.11     | 1.31     | 1.71    | 2.09   | 2.44     |
| DRN    | 1.65     | 2.14    | 2.50     | 2.80     | 1.23     | 2.01    | 2.58   | **2.90** |
| CRN    | 1.66     | 2.10    | 2.42     | 2.66     | 1.77     | 2.20    | 2.48   | 2.68     |
| ASAM   | 1.21     | 1.70    | 2.12     | 2.43     | 1.52     | 1.87    | 2.31   | 2.61     |
| RC-EED | 1.79     | 2.24    | 2.58     | **2.82** | 1.87     | 2.31    | 2.61   | 2.84     |
| RF-AEED| 1.58     | 2.06    | 2.38     | 2.58     | 1.78     | 2.14    | 2.41   | 2.61     |
| RC-AEED| **1.80** | **2.26** | **2.59** | **2.82** | **1.95** | **2.33** | **2.62** | **2.83** |

The generation of memory vector of ASAM here is the same as RC-AEED. We also evaluated the ASAM structure using original method of [13], and didn’t find it had any apparent influence on the experiment results.

to those of CRN and RF-AEED in the −10 and −5 dB SNR conditions. RF-AEED has a higher STOI score on the Hfchannel noise when tested in low SNR conditions, but it cannot derive consistent performance on the other noises. Specifically, when Leopard noise is tested, only RC-AEED and CRN have slight STOI improvements in the −10 dB SNR condition, whereas the others lose some objective intelligibility of the speech.
Table 5 shows the mean experimental results of SI-SDR on the seen and unseen noises. There is no distinct difference in SI-SDR metrics among the DRN, CRN, and proposed methods. Some tendencies can be observed from the five tables. RC-AEED with an attention mechanism obtains higher PESQ and STOI scores than the RC-EED. This outcome means that a proper attention mechanism can boost the performance of the RC-EED, and the stacked masking layer can filter some interference in the embedding array, which is broadly analogous to the T-F masking method. Meanwhile, the performance of RF-AEED is inferior to that of RC-AEED in most cases. This result is probably because the CNN can extract local pattern of embedding array more efficiently [18] and the fully connected layers introduce a large amount of irrelevant information that affects the formation of the spectrum.

Table 6 shows the performance of the DRN, CRN, and proposed methods on both trained and untrained speakers in the $-10$, $-5$, $0$, $5$ dB SNR conditions, and all the seen and unseen noises are used. Generally, a similar tendency is observed between the experiment results of the trained and untrained speakers. RC-AEED and RC-EED both have good PESQ scores on the trained and untrained speaker dataset.

We evaluated the metric results of CC-EED and RC-EED on the seen and unseen noises. We find CC-EED has poor generalization ability. In preliminary experiments, we observed CC-EED has poor generalization, it cannot perform well and even degrades the speech in the unseen noises condition. Fig. 4 only shows the results on the seen noise. We can find CC-EED has poor performance in unseen SNR condition, and RC-EED consistently outperforms CC-EED in all metrics.

As shown in Table 7, we evaluate the performance of different CNN-based embedding decoders at the same noise condition as the evaluation in Table 6. The hyperparameters of the CNN structure are given in $(input\text{channel} / output\text{channel})$ format, and the kernel sizes are $3 \times 3$. In general, the model with a deeper CNN based decoder can achieve higher scores on STOI, PESQ. 3 layers’ decoder gets the best PESQ, and 4 layers’ decoder gets the best STOI.

Fig. 5 illustrates a spectrogram of clean utterance and three spectrograms of the corresponding noisy speech with seen noise enhanced by different methods. Some black dotted boxes are marked in the spectrograms. The spectrogram generated by the CRN has less noise components. However, it also suppresses some of the time-frequency energy of the speech signal by mistake. This problem is mitigated in the proposed methods, and clearer voice features can be found in the right two spectrograms. Furthermore, the RC-AEED generates a better spectrogram.

Table 8 shows the parameter efficiency comparison of RC-AEED, RC-EED and CRN. The decoders of
FIGURE 5. Spectrograms of a clean utterance and the utterance with noise enhanced by different methods. The unprocessed mixture is generated by mixing an utterance of an untrained speaker with factory noise at −5 dB.

TABLE 6. Model comparisons in terms of mean STOI and PESQ scores on trained and untrained speakers.

| Target | Trained Speaker | Untrained Speaker |
|--------|-----------------|-------------------|
|        | STOI | PESQ | STOI | PESQ |
| Noisy  | 0.672 | 1.12 | 0.654 | 1.09 |
| DRN    | 0.766 | 2.12 | 0.755 | 2.10 |
| CRN    | **0.779** | 2.15 | **0.770** | 2.12 |
| ASAM   | 0.767 | 2.08 | 0.751 | 2.07 |
| RC-EED | 0.776 | 2.32 | 0.756 | **2.27** |
| RF-AEED| 0.766 | 1.95 | 0.750 | 2.09 |
| RC-AEED| 0.778 | **2.35** | 0.760 | 2.23 |

TABLE 7. Metric comparisons between the proposed RC-AEED with different embedding decoder structures.

| Embedding Decoder | STOI | PESQ |
|-------------------|------|------|
| L                  |      |      |
| 1                  |      |      |
| 2                  |      |      |
| 3                  |      |      |
| 4                  |      |      |

| Embedding Decoder | STOI | PESQ |
|-------------------|------|------|
| 1                  | 0.756 | 2.22 |
| 2                  | 0.760 | 2.23 |
| 3                  | 0.759 | **2.32** |
| 4                  | **0.763** | 2.27 |

RC-AEED and RC-EED are both two-layers’ CNN structure. We find the proposed models have the same magnitude, and RC-EED has fewer parameters than CRN.

TABLE 8. Trainable parameter comparison of different models.

| Model   | Parameter |
|---------|-----------|
| RC-EED  | 15,962,945 |
| RC-AEED | 19,062,332 |
| CRN [8] | 17,580,000 |

IV. CONCLUSION
Multiple embedding encoder-decoder structures, RC-EED, CC-EED and AEEDs, are proposed in this study. The RC-EED and RC-AEED get good performance on speech
enhancement task at low SNR conditions and apparently outperform a similar ASAM model at different noise conditions. The comparison of the RC-EED and CC-EED shows that the RC-EED with an LSTM-based embedding encoder outperforms the CNN-based encoder-decoder. The RC-AEED is also compared to an AEED with a fully connected decoder to evaluate the efficiency of the CNN decoder. The DRN and the CRN are compared with the proposed methods in the experiment, and the results show the RC-AEED achieves the best PESQ scores in almost all seen and unseen noise conditions. In lower SNR conditions, namely, −10 and −5 dB, the RC-AEED has consistent performance on speech enhancement. In lower SNR conditions, phase distortion has more impact on the quality of enhanced speech. In the future, we will explore to deal with the phase distortion problem, and combine the RC-EED structure and the target complex spectrum.

REFERENCES

[1] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.
[2] P. Scalart and J. V. Filho, “Speech enhancement based on a priori signal to noise estimation,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Conf., vol. 2, 1996, pp. 629–632.
[3] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, Jun. 2015, pp. 3431–3440.
[4] X.-J. Mao, C. Shen, and Y.-B. Yang, “Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections,” presented at the Annu. Conf. Neural Inf. Process. Syst., Barcelona, Spain, Dec. 2016.
[5] S. E. Eskimez, P. Soufleris, Z. Duan, and W. Heinzelman, “Front-end speech enhancement for commercial speaker verification systems,” Speech Commun., vol. 99, pp. 101–113, May 2018.
[6] S. R. Park and J. W. Lee, “A fully convolutional neural network for speech enhancement,” in Proc. 18th Annua. Conf. Int. Speech Commun. Assoc., Stockholm, Sweden. Grenoble, France: ISCA, Aug. 2017, pp. 1993–1997.
[7] M. Ge, L. Wang, N. Li, H. Shi, J. Dang, and X. Li, “Environment-dependent attention-driven recurrent convolutional neural network for robust speech enhancement,” in Proc. 26th Annua. Conf. Int. Speech Commun. Assoc., Graz, Austria, 2019, pp. 3135–3137.
[8] K. Tan and D. Wang, “A convolutional recurrent neural network for real-time speech enhancement,” in Proc. 19th Annua. Conf. Int. Speech Commun. Assoc., Hyderabad, India. Grenoble, France: ISCA, 2018, pp. 3229–3233.
[9] H. Zhao, S. Zarar, I. Tashev, and C.-H. Lee, “Convolutional-recurrent neural networks for speech enhancement,” in Proc. 43rd IEEE Int. Conf. Acoust., Speech Signal Process., Calgary, AB, Canada. Piscataway, NJ, USA: IEEE, Apr. 2018, pp. 2401–2405.
[10] A. L. Maas, Q. V. Le, T. M. O’Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, “Recurrent neural networks for noise reduction in robust ASR,” in Proc. 13th Annua. Conf. Int. Speech Commun. Assoc., Portland, OR, USA: Grenoble, France: ISCA, 2012, pp. 22–25.
[11] S. Mirsamadi and I. Tashev, “Causal speech enhancement combining data-driven learning and suppression rule estimation,” in Proc. Interspeech, Sep. 2016, pp. 2870–2874.
[12] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “An experimental study on speech enhancement based on deep neural networks,” IEEE Signal Process. Lett., vol. 21, no. 1, pp. 65–68, Jan. 2014.
[13] J. Xu, L. Shi, G. Liu, X. Chen, and B. Xu, “Modeling attention and memory for audioity selection in a cocktail party environment,” in Proc. 32nd AAAI Conf. Artif. Intell., New Orleans, LA, USA. Menlo Park, CA, USA: AAAI, 2018, pp. 2564–2571.
[14] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in Proc. 41st IEEE Int. Conf. Acoust., Speech Signal Process., Shanghai, China. Piscataway, NJ, USA: IEEE, Mar. 2016, pp. 31–35.
[15] S. Da Costa, W. van der Zwaag, L. M. Miller, S. Clarke, and M. Saenz, “Tuning in to sound: Frequency-selective attentional filter in human primary auditory cortex,” J. Neurosci., vol. 33, no. 5, pp. 1858–1863, Jan. 2013.
[16] B. Huang and K. M. Carley, “Parameterized convolutional neural networks for aspect level sentiment classification,” 2019, arXiv:1909.06276. [Online]. Available: http://arxiv.org/abs/1909.06276

TIAN LAN received the Ph.D. degree in computer science from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2009. He is currently an Associate Professor with the School of Information and Software Engineering, UESTC. His current research interests include medical image processing, speech enhancement, and natural language processing.

WENZHENG YE is currently pursuing the master’s degree in software engineering with UESTC. His research interests include speech enhancement, speech recognition, and machine learning.
YILAN LYU was born in 1997. She is currently pursuing the master’s degree. Her research interests include speech enhancement and speech separation.

JUNYI ZHANG received the Ph.D. degree in computer science from the Beijing University of Posts and Telecommunications, in 2012. His current research interests include digital signal processing, machine learning, and data mining.

QIAO LIU received the Ph.D. degree in computer science from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2010. He is currently a Full Professor with the School of Information and Software Engineering, UESTC. His current research interests include natural language processing, machine learning, and data mining.