WaveCRN: An Efficient Convolutional Recurrent Neural Network for End-to-end Speech Enhancement

Tsun-An Hsieh, Hsin-Min Wang, Xugang Lu, and Yu Tsao, *Member, IEEE*

***Abstract*—Due to the simple design pipeline, end-to-end (E2E) neural models for speech enhancement (SE) have attracted great interest. In order to improve the performance of the E2E model, the locality and temporal sequential properties of speech should be efficiently taken into account when modelling. However, in most current E2E models for SE, these properties are either not fully considered or are too complex to be realized. In this paper, we propose an efficient E2E SE model, termed WaveCRN. In WaveCRN, the speech locality feature is captured by a convolutional neural network (CNN), while the temporal sequential property of the locality feature is modeled by stacked simple recurrent units (SRU). Unlike a conventional temporal sequential model that uses a long short-term memory (LSTM) network, which is difficult to parallelize, SRU can be efficiently parallelized in calculation with even fewer model parameters. In addition, in order to more effectively suppress the noise components in the input noisy speech, we derive a novel restricted feature masking (RFM) approach that performs enhancement on the feature maps in the hidden layers; this is different from the approach that applies the estimated ratio mask on the noisy spectral features, which is commonly used in speech separation methods. Experimental results on speech denoising and compressed speech restoration tasks confirm that with the lightweight architecture of SRU and the feature-mapping-based RFM, WaveCRN performs comparably with other state-of-the-art approaches with notably reduced model complexity and inference time.**

***Index Terms*—Compressed speech restoration, simple recurrent unit, raw waveform speech enhancement, convolutional recurrent neural networks**

I. Introduction

SPEECH related applications, such as automatic speech recognition (ASR), voice communication, and assistive hearing devices, play an important role in modern society. However, most of these applications are not robust when noises are involved, and speech enhancement (SE) [1]–[8] has been used as a fundamental tool in these applications. SE aims to improve the quality and intelligibility of the original speech signal. Traditional SE approaches are derived based on the statistical properties of speech and distortion signals (e.g., Wiener filtering [9]). Although these traditional SE approaches perform well under many conditions, the enhancement performance degrades when the statistical properties are not fulfilled.

The powerful transformation capabilities of deep learning algorithms have enabled revolutionary results for a wide variety of traditional classification/regression problems. In recent years, researchers have tried to incorporate deep learning algorithms into the SE task. Many SE systems are derived to carry out enhancement on the frequency-domain acoustic features, where the speech signals are analyzed and reconstructed by using the short-time Fourier transform (STFT) and inverse STFT, respectively [10]–[14]. Lu *et al.* [3] presented a magnitude spectrogram based enhancement method using a fully connected deep denoising auto-encoder [3]. Fu *et al.* [15] used convolutional neural networks (CNNs) to capture better local information. Weninger *et al.* [16] and Maas *et al.* [17] used recurrent neural networks (RNNs) to enhance speech signals and subsequently increase the robustness of ASR systems. Later on, Xu *et al.* [18] used stacked simple recurrent units (SRU) to build an SE system, which provides comparable denoising performance while less training time as compared to a long short-term memory (LSTM)-based SE system. Some other approaches [19], [20] combine CNN and RNN to capture the spatial and temporal correlations jointly. Although the above-mentioned approaches that carry out enhancement in the frequency domain can already provide outstanding performance, the enhanced speech signals cannot reach perfection due to lack of accurate phase information. To tackle this problem, Fu *et al.* [21] and Williamson *et al.* [22] proposed to adopting complex ratio masking and complex spectral mapping, respectively, to enhance distorted speech. Takahashi *et al.* [23] formulated the phase estimation as a classification problem for source separation. In the meanwhile, some approaches [24]–[28] propose to directly perform enhancement on the raw waveform.

In this paper, we propose an end-to-end raw waveform-mapping-based SE method using a convolutional recurrent neural network, termed WaveCRN. Two tasks are used to test the proposed WaveCRN SE model: (1) speech denoising and (2) compressed speech restoration. For speech denoising, we evaluate our method on an open-source dataset [29] and obtain state-of-the-art PESQ (perceptual evaluation of speech quality) scores [30] using a relatively simple architecture and L1-loss function. For compressed speech restoration, evaluated on the TIMIT database [31], the proposed WaveCRN model recovers extremely compressed speech (compressing speech samples from 16-bit to 2-bit) with a notable relative STOI (short-time objective intelligibility) [32] improvement of 75.51% (from 0.49 to 0.86).

II. Related Works

In this section, we review existing raw waveform based SE approaches. Several studies have shown that the phase

Thanks to MOST of Taiwan for funding.

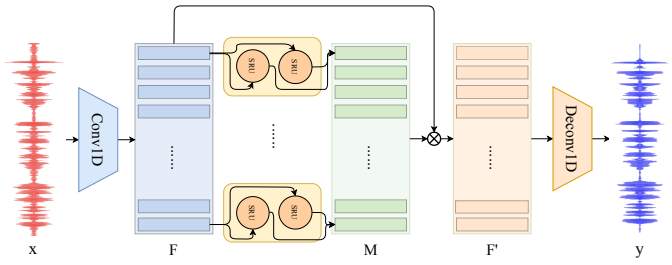


Fig. 1. The architecture of the proposed WaveCRN model. For local feature extraction, a 1D CNN maps the noisy audio **x** into a 2D feature map **F**. Bi-SRU then encodes **F** into an restricted feature mask (RFM) **M**, which is element-wisely multiplied by **F** to generate a masked feature map **F'**. Finally, a transposed 1D convolution layer recovers the enhanced waveform **y** from **F'**.

information is important when converting spectral features to waveforms. A class of studies [22], [23] conducted phase-aware SE using a complex ratio mask (cRM) to jointly reconstruct magnitude and phase. Wang *et al.* [33] proposed using a convolutional recurrent neural network (CRN) for real-time noise/speaker-independent SE and [34] in a close-talk scenario. The performance of these approaches are superior to previous works based on the magnitude spectrogram and ideal ratio mask (IRM). In the field of ASR, researchers have found that using a raw waveform input can achieve lower word error rates than using hand-crafted features [35], [36]. For the SE task, fully convolutional network (FCN) has been popularly used to perform waveform-mapping directly [25], [37]–[40]. Compared to a fully connected architecture, FCN retains better local information and thus can more accurately model the high-frequency-components of speech signals. More recently, Pandey *et al.* proposed to use a temporal convolutional neural network (TCNN) to more precisely characterize temporal features and perform SE in the time domain [26].

III. Methodology

In this section, we describe the details of our SE system. The architecture is a fully differentiable end-to-end neural network that does not require pre-processing and handcrafted features. We leverage the advantages of CNN and RNN to model spatial and temporal information. The overall architecture of the proposed WaveCRN is shown in Fig. 1.

A. 1D Convolutional Input Module

Most of previous deep-learning-based SE approaches use log-power-spectrum (LPS) as input. Therefore, pre-processing is required to convert the raw waveform into LPS features, which are then fed into the deep-learning model. Then, the phase information of the noisy speech is used to reconstruct the enhanced waveform. To perform time-domain SE, we design a light-weighted 1D CNN input module to substitute the STFT processing. Benefited by the nature of neural networks, the CNN module is fully trainable. An input noisy audio $\mathbf{X}$ ($\mathbf{X} \in R^{N\times 1\times L}$) is convolved with a two-dimensional tensor $\mathbf{W}$ ($\mathbf{W} \in R^{C\times K}$) to extract the feature map $\mathbf{F} \in R^{N\times C\times T}$, where N, C, K, T, L are the batch size, number of channels, kernel size, time steps, and audio length, respectively. Notably, to reduce the sequence length for computational efficiency, we set the convolution stride to half the size of the kernel, so the length of $\mathbf{F}$ is reduced from L to $T = 2L/K + 1$.

B. Temporal Encoder

We adopt a bidirectional SRU (Bi-SRU) to capture the temporal correlation of the feature maps extracted by the input module in both directions. For one feature map $\mathbf{f} \in R^{C\times T}$, it can be formulated as a sequence $\mathbf{H} = [h_1, h_2, ..., h_T]$, $h_t \in R^C$, and then passed to the recurrent feature extractor. The hidden state extracted in both directions are concatenated as $\hat{h}_t = [\vec{h}_t, \overleftarrow{h}_t]$. An affine transform is used to ensure that the dimensions of input and output feature maps are the same.

C. Restricted Feature Mask

The restricted optimal ratio mask (ORM) has been widely used in SE and speech separation tasks [41]. For our task, an alternative restricted ORM, called the restricted feature mask (RFM) $\mathbf{M} \in R^{N\times C\times T}$, where all the elements are in the range of -1 to 1, is applied to mask the feature map $\mathbf{F}$ as:

$$\mathbf{F}' = \mathbf{M} \circ \mathbf{F}. \tag{1}$$

$\mathbf{F}'$ is the masked feature map estimated by element-wisely multiplying the mask $\mathbf{M}$ and the feature map $\mathbf{F}$ for waveform generation. The main difference between the restricted ORM and RFM is that the former is applied in the time-frequency domain while the latter transforms the feature map, rather than directly applied in the time-frequency domain.

D. Waveform Generation

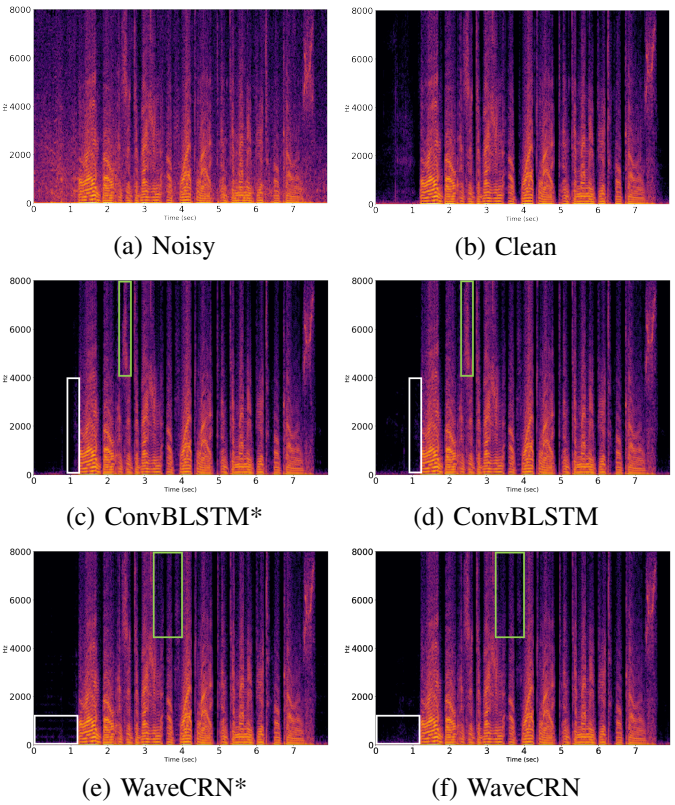
As described in Section 3.1, the sequence length is reduced from L to T due to the stride in the convolution process. Length restoration is essential to generate an output waveform of the same length as the input. Given the input length, output length, stride, and padding as L_{in}, L_{out}, S, and P, the relation of L_{in} and L_{out} can be formulated as:

$$L_{out} = (L_{in} - 1) \times S - 2 \times P + (K - 1) + 1. \tag{2}$$

Let $L_{in} = T$, $S = K/2$, $P = K/2$, we have $L_{out} = L$. That is, the input and output lengths are guaranteed to be the same.

E. Model Structure Overview

In summary, as shown in Fig. 1, our model leverages the benefits of CNN and RNN. Given a noisy speech utterance, for local feature extraction, a 1D CNN maps the noisy audio **x** into a 2D feature map **F**. Bi-SRU then encodes **F** into an RFM **M**, which is element-wisely multiplied by **F** to generate a masked feature map **F'**. Finally, a transposed 1D convolution layer is used to recover the enhanced waveform **y** from **F'**.



(a) Noisy (b) Clean (c) ConvBLSTM* (d) ConvBLSTM (e) WaveCRN* (f) WaveCRN

Fig. 2. Magnitude spectrograms of noisy, clean and enhanced speech by ConvBLSTM, ConvBLSTM*, WaveCRN, and WaveCRN*, where models marked with * generate enhanced speech directly without the RFM.

F. Comparing LSTM and SRU

In [42], the SRU has been confirmed to provide comparable performance while better parallelization than LSTM. LSTM recursively encodes sequence similarity with sequential gates. However, the dependency on hidden states leads to slow training and inference. In contrast, all gates in the SRU depend on the input of the corresponding time, and the temporal correlation is captured by adding a highway connection between the recurrent layers. Therefore, the gates in the SRU are computed simultaneously. Furthermore, replacing the matrix multiplication with the Hadamard product while computing the state vectors speeds up the forward and backward pass calculation. The above advantages make SRUs very suitable to be fundamental blocks to build an SE system.

IV. Experiments

This section presents information of the datasets and our experimental setup including the hyper-parameters and the model architecture. Experimental results and analyses will be discussed by quantitative results and the visualization of the enhanced speech for each model.

A. Datasets

1) Speech Denoising: For the speech denoising task, an open-source dataset [29] was used, which incorporates the voice bank corpus [43] and DEMAND [44]. In the voice bank corpus, 28 out of 30 speakers are used for training and the remaining speakers are used for testing. For the training set, the clean speech is combined with 10 types of noises with 4 SNR conditions (0, 5, 10, and 15 dB), while 5 types of unseen noises are mixed with the clean speech under 4 different SNR conditions (2.5, 7.5, 12.5, and 17.5) for the testing set.

2) Compressed (2-bit) Speech Restoration: For the compressed speech restoration task, we used the TIMIT corpus [31]. The original speech samples were recorded in 16 kHz and in 16-bit format. In this set of experiemtns, we compressed each sample to a 2-bit format, and thus each compressed sample was represented by -1, 0, or +1. In this way, we successfully save 87.5% of bits and accordingly reduce the data transmission and storage requirement. We believe this compression scheme is potentially applicable to real-world IoT scenarios. Expressing the original speech as $\hat{y}$ and the compressed speech as $\mathrm{sgn}(\hat{y})$, the optimization process becomes:

$$\min_{\theta} \|\hat{y} - g_{\theta}(\mathrm{sgn}(\hat{y}))\|_1, \tag{3}$$

where g_θ denotes the SE process.

TABLE I
Results of the speech denoising task. A higher score indicates better performance. The bold values indicate the corresponding best performance. Models marked with * generate enhanced speech directly without the RFM.

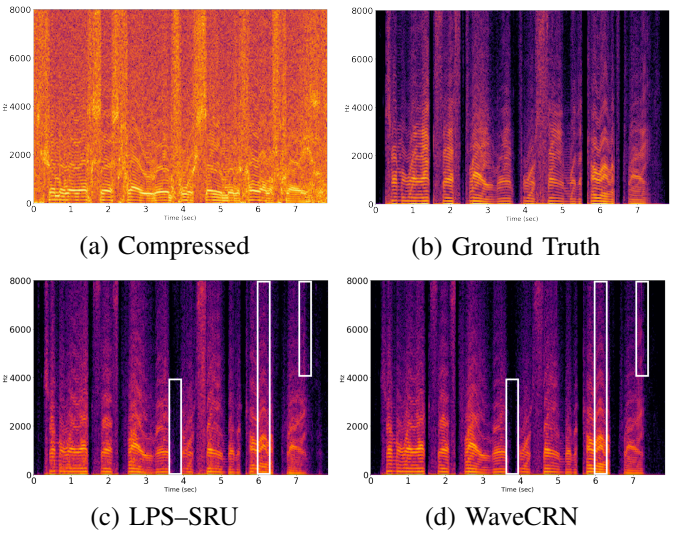
Model	PESQ	CSIG	CBAK	COVL	SSNR
Noisy	1.97	3.35	2.44	2.63	1.68
Wiener	2.22	3.23	2.68	2.67	5.07
SEGAN [38]	2.16	3.48	2.94	2.80	7.73
Wavenet [39]	-	3.62	3.23	2.98	-
Wave-U-Net [45]	2.62	3.91	3.35	3.27	10.05
ConvBLSTM*	2.39	3.19	3.08	2.76	8.78
WaveCRN*	2.46	3.43	3.04	2.89	8.43
ConvBLSTM	2.54	3.83	3.25	3.18	9.33
WaveCRN	**2.64**	**3.94**	**3.37**	**3.29**	**10.26**

B. Model Architecture

In the input module, we extract local features with a 1D convolutional layer, which contains 256 channels with 6ms kernel and 3ms stride. To ensure the recovered audio length to be the same as that of the input, we reflectively pad the input sequence at both sides so that the input length is divisible by the stride size. For the temporal encoder, a Bi-SRU is used. Corresponding to the features extracted from the previous stage, the size of the hidden state is set to 256 with 6 stacks, and each hidden state is transformed to half of its dimension. Next, all hidden states are concatenated together as a mask and element-wisely multiplied by the feature map generated from the first stage. Finally, in the waveform generation step, a transposed convolutional layer maps the 2D feature map into a 1D sequence, which is then passed through a hyperbolic tangent activation function to output the final predicted waveform.

C. Experimental Results and Analyses

1) Speech Denoising: For the speech denoising task, we used five evaluation metrics: **CSIG** that reveals the signal



(a) Compressed (b) Ground Truth

(c) LPS–SRU (d) WaveCRN

Fig. 3. Magnitude spectrograms of original, compressed, and restored speech by LPS–SRU and WaveCRN.

distortion mean opinion score, **CBAK** that represents the background intrusiveness, **COVL** that reveals the speech quality, **SSNR** that shows the segmental SNR, and **PESQ**, as a standard speech quality measure. In addition to Wiener filtering and SEGAN, we listed several well-known SE approaches that use the same L1 loss. A comparative system that combines CNN and BLSTM (termed ConvBLSTM) was also implemented, where SRU in Fig.1 was replaced by LSTM. The combination of CNN and LSTM to process speech signals has been widely investigated [27], [33], [34]. Here in this study, we intend to show that SRU can yield better performance than LSTM for waveform-mapping-based SE in terms of both denoising capability and computation efficiency.

As shown in Table I, WaveCRN performs the best in terms of perceptual and signal-level evaluation metrics. We further investigate the effect of RFM and list the results of ConvBLSTM and WaveCRN without RFM as ConvBLSTM* and WaveCRN*. We can note that RFM can effectviely enhance the denoising capability for both ConvBLSTM and WaveCRN.

Next, we visually investigate the magnitude spectrograms of noisy and clean speech, and the enhanced speech by ConvBLSTM and WaveCRN with and without RFM in Fig. 2. By observing the green-block regions in Fig. 2 (c) and (d) and Fig. 2 (e) and (f), we note that RFM produces less distortions in high-frequency regions. Next, by comparing the white-block regions in Fig. 2 (c) and (d), and Fig. 2 (e) and (f), RFM enables both WaveCRN and ConvBLSTM to more effectively remove noise components. Finally, by comparing Fig. 2 (d) and (f), we can see that WaveCRN has better denoising capability and preserves more consonant information as compared to ConvBLSTM.

2) Compressed Speech Restoration: For the compressed speech restoration task, we applied WaveCRN to transform the compressed speech to the uncompressed speech. For comparison, we implemented another SRU-based system, termed LPS–SRU. In LPS–SRU, the SRU structure was identical to the one used in WaveCRN, but the input was the LPS, where the STFT and inverse STFT were used for speech analysis and reconstruction, respectively. The performance was evaluated in terms of the PESQ and STOI scores. From Table II, we can see that WaveCRN/LPS–SRU improves the PESQ score from 1.39 to 2.41/1.97, and the STOI score from 0.49 to 0.86/0.79. Both WaveCRN and LPS–SRU achieve significant improvements, while WaveCRN clearly outperforms LPS–SRU.

TABLE II
THE RESULTS OF THE COMPRESSED SPEECH RESTORATION TASK.

Model	PESQ	STOI
Compressed	1.39	0.49
LPS–SRU	1.97	0.79
WaveCRN	**2.41**	**0.86**

TABLE III
A COMPARISON OF EXECUTION TIME AND NUMBER OF PARAMETERS OF WAVECRN AND CONVBLSTM.

Model	Time (sec)	#parameters (K)
ConvBLSTM	58.039	9093
WaveCRN	**2.289**	**4655**

We further visually investigate the resulting amplitude spectrograms. From Fig. 3 (a) and (b), when the speech samples are compressed to a 2-bit format, the speech quality is notably reduced. By using WaveCRN and LPS–SRU, the restored speech presents a clearer structure, as shown in Fig. 3 (c), and (d). Moreover, the white-block regions show that WaveCRN can restore speech patterns more effectively than LPS–SRU without losing phase information.

Next, we compare WaveCRN and ConvBLSTM in terms of inference time and model complexity. The first column in Table III shows the execution time of the forward pass, and the second column presents the number of parameters. Under the same hyper-parameter setting (number of layers, dimension of hidden states, channel number, etc.), the execution speed of WaveCRN is 25.36 times faster, and the number of parameters is only 51%, as compared to ConvBLSTM .

V. CONCLUSION

In this paper, we proposed the WaveCRN-based E2E SE. By taking the advantage of CNN and SRU, WaveCRN uses a bi-directional architecture to model the temporal correlation of the extracted features. Experimental results of speech denoising and compressed speech restoration tasks show that the proposed WaveCRN has outstanding denoising capability and computational efficiency as compared to related works that use L1 loss. In summary, the contributions of this study are fourfold: (a) WaveCRN is the first work that combines SRU with CNN to perform E2E SE; (b) a novel RFM approach is derived to directly transform noisy features to enhanced features; (c) the SRU model is relatively simple, but yields comparable performance, as compared with other state-of-the-art SE models using the same L1 loss; (d) a new and practical application (i.e., compressed speech restoration) was designed and tested, and promising results were obtained using the proposed WaveCRN model.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Inc., 2nd edition, 2013.

[2] M. Kolbk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM TASLP*, vol. 25, no. 1, pp. 153–167, 2017.

[3] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder.," in *Proc. Interspeech*, 2013.

[4] B. Xia and C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Communication*, vol. 60, pp. 13–29, 2014.

[5] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[6] Z. Meng, J. Li, and Y. Gong, "Adversarial feature-mapping for speech enhancement," in *Proc. Interspeech*, 2017.

[7] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *Proc. ICASSP*, 2018.

[8] L. Chai, J. Du, Q.-F. Liu, and C.-H. Lee, "Using generalized gaussian distributions to improve regression error modeling for deep learning-based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 1919–1931, 2019.

[9] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *ICASSP*, 1996.

[10] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM TASLP*, vol. 23, no. 1, pp. 7–19, 2015.

[11] F. Xie and D. Van Compernolle, "A family of MLP based nonlinear spectral estimators for noise reduction," in *Proc. ICASSP*, 1994.

[12] S. Wang, K. Li, Z. Huang, S. M. Siniscalchi, and C.-H. Lee, "A transfer learning and progressive stacking approach to reducing deep model sizes with an application to speech enhancement," in *Proc. ICASSP*, 2017.

[13] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Proc. Interspeech*, 2014.

[14] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," in *Proc. HSCMA*, 2017.

[15] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement," in *Proc. Interspeech*, 2016.

[16] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. LVA/ICA*, 2015.

[17] 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. Interspeech*, 2012.

[18] X. Cui, Z. Chen, and F. Yin, "Speech enhancement based on simple recurrent unit network," *Applied Acoustics*, vol. 157, pp. 107019, 2020.

[19] H. Zhao, S. Zarar, I. Tashev, and C.-H. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *Proc. ICASSP*, 2018.

[20] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM TASLP*, vol. 28, pp. 380–390, 2019.

[21] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. MLSP*, 2017.

[22] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM TASLP*, vol. 25, no. 7, pp. 1492–1501, 2017.

[23] N. Takahashi, P. Agrawal, N. Goswami, and Y. Mitsufuji, "PhaseNet: Discretized phase modeling with deep neural networks for audio source separation," in *Proc. Interspeech*, 2018.

[24] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proc. APSIPA ASC*, 2017.

[25] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Proc. Interspeech*, 2015.

[26] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. Interspeech*, 2019.

[27] J. Li, H. Zhang, X. Zhang, and C. Li, "Single channel speech enhancement using temporal convolutional recurrent neural networks," in *Proc. APSIPA ASC*, 2019.

[28] M. Kolbæk, Z.-H. Tran, Sø. H. Jensen, and J. Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 825–838, 2020.

[29] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proc. SSW*, 2016.

[30] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001.

[31] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus cd-rom. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report*, vol. 93, 1993.

[32] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of timefrequency weighted noisy speech," *IEEE/ACM TASLP*, vol. 19, no. 7, pp. 2125–2136, 2011.

[33] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proc. Interspeech*, 2018.

[34] K. Tan, X. Zhang, and D. Wang, "Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios," in *Proc. ICASSP*, 2019.

[35] D. Palaz, M. Magimai-Doss, and R. Collobert, "Analysis of CNN-based speech recognition system using raw speech as input," in *Proc. Interspeech*, 2015.

[36] P. Golik, Z. Tüske, R. Schlüter, and H. Ney, "Convolutional neural networks for acoustic modeling of raw time signal in LVCSR," in *Proc. Interspeech*, 2015.

[37] S.-W. Fu, T. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM TASLP*, vol. 26, no. 9, pp. 1570–1584, 2018.

[38] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017.

[39] D. Rethage, J. Pons, and X. Serra, "A Wavenet for speech denoising," in *Proc. Interspeech*, 2017.

[40] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florłncio, and M. Hasegawa-Johnson, "Speech enhancement using Bayesian Wavenet," in *Proc. Interspeech*, 2017.

[41] S. Liang, W. Liu, W. Jiang, and W. Xue, "The optimal ratio time-frequency mask for speech separation in terms of the signal-to-noise ratio," *The Journal of JASA*, vol. 134, no. 5, pp. EL452–EL458, 2013.

[42] T. Lei, Y. Zhang, S. I. Wang, H. Dai, and Y. Artzi, "Simple recurrent units for highly parallelizable recurrence," in *Proc. EMNLP*, 2018.

[43] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. O-COCOSDA/CASLRE*, 2013.

[44] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," 2013.

[45] R. Giri, U. Isik, and A. Krishnaswamy, "Attention Wave-U-Net for speech enhancement," in *Proc. WASPAA*, 2019.