INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

SE-Conformer: Time-Domain Speech Enhancement using Conformer

Eesung Kim, Hyeji Seo

AI R&D Lab, Kakao Enterprise
235, Pangyoyeok-ro, Bundang-gu, Seongnam-si, Gyeonggi-do 13494, Korea
chris.ekim@kakaoenterprise.com, heize.s@kakaoenterprise.com

Abstract

Convolution-augmented transformer (conformer) has recently shown competitive results in speech-domain applications, such as automatic speech recognition, continuous speech separation, and sound event detection. Conformer can capture both the short and long-term temporal sequence information by attending to the whole sequence at once with multi-head self-attention and convolutional neural network. However, the effectiveness of conformer in speech enhancement has not been demonstrated. In this paper, we propose an end-to-end speech enhancement architecture (SE-Conformer), incorporating a convolutional encoder–decoder and conformer, designed to be directly applied to the time-domain signal. We performed evaluations on both the VoiceBank-DEMAND Corpus (VCTK) and Librispeech datasets in terms of objective speech quality metrics. The experimental results show that the proposed model outperforms other competitive baselines in speech enhancement performance.

Index Terms: Time-Domain Speech Enhancement, Conformer

1. Introduction

Speech enhancement (SE) entails the enhancement of clean speech signal from noisy input signals. This enhancement helps improve the intelligibility and quality of speech. As the importance of speech-based interface technologies has increased in recent years, SE has attracted particular attention in applications such as automatic speech recognition (ASR), teleconferencing, and video calls [1]. Traditional SE approaches are mainly statistical-based models, including spectral subtraction, Wiener filter, and the optimally modified log-spectral amplitude (OMLSA) estimator [2]. However, these approaches have shown limited performance in speech with non-stationary noises that are common in real-world environments. To address this limitation, deep learning approaches have been widely investigated in recent years for better SE capability by employing recurrent neural networks (RNNs), convolutional neural networks (CNNs), convolutional recurrent neural networks (CRNNs) [3], and generative adversarial networks (GANs) [4].

One of the mainstream applications of deep learning for SE is algorithms based on the time-frequency (TF) domain, which is computed with a short-time Fourier transform (STFT). This can be further divided into two approaches: mask-based target and mapping-based target. Mask-based approaches estimate the ideal ratio mask from noisy acoustic features, then multiply noisy magnitude spectra, and reconstruct clean speech signals [5]. Mapping-based targets estimate the clean magnitude spectrum from a noisy magnitude [3]. However, these approaches have some difficulties, which cause audible artifacts only by estimating the magnitude spectrum of the waveform while maintaining noisy phase information. To alleviate these obstacles, some researchers have considered incorporating the phase information of the noisy signal to enhance the speech signal [6].

Another stream of SE research is the time-domain end-to-end method, which is designed to directly estimate clean speech waveforms from noisy speech waveforms [7, 8, 9]. This approach benefits the learning of proper representations that are well suited for enhancing clean speech. In [7], the authors built an efficient WaveCRN architecture, which processes a waveform with a one-dimensional CNN module and a stacked simple recurrent unit module. ConvTasnet and its variants [8] leverage a convolutional encoder–decoder with skip connection (CED) structure and a temporal convolutional network (TCN) to learn long-range dependencies of the time-domain signal. Wave-U-Net [10] adapted the U-Net structure, using downsampling and upsampling blocks with skip connections, to extract effective features from various time scales. In [9], the authors introduced a Demucs composed of an effective CED and bidirectional long short-term memory (BLSTM) structure to enable sequence modeling in latent space. RNNs and their variants, however, cause an unstable information propagation problem with large timescales because of their inherently sequential nature [11].

In recent years, transformer [12], which can consider long-term dependency effectively by applying multi-head self-attention (MHSA), perform well in sequence-to-sequence domains, in speech-related applications such as ASR [13], speech enhancement [14, 15], and source separation [16]. In [17], the authors built a transformer with Gaussian-weighted self-attention (T-GSA), whose attention weights are attenuated according to the distance between correlated symbols. Multi-stage SE system (MSSA-TCN) [14] comprises self-attention block followed by stacks of dilated TCN blocks every each stage in TF domain. In [15], the authors employed a U-Net structure with an encoder–decoder, in which each layer comprised a dense block and an attention module.

More recently, convolution-augmented transformer (conformer) have demonstrated superior performance over transformer-based approaches [18] in the areas of ASR, continuous speech separation [19], and sound event detection and separation in domestic environments [20]. In this paper, we propose a conformer-based time-domain speech enhancement (SE-Conformer) that applies a conformer to the effective CED structure. Inspired by [18, 9], we assumed that the conformer block is more appropriate for CED-based structures because of its enhanced ability to reflect the local and global temporal context dependencies by attending the entire sequence in the latent representation. We evaluated our model on both open-source small VoiceBank-DEMAND Corpus (VCTK) and large Librispeech datasets. We demonstrate that the proposed model performs better than the baselines in terms of objective speech quality metrics.

http://dx.doi.org/10.21437/Interspeech.2021-2207

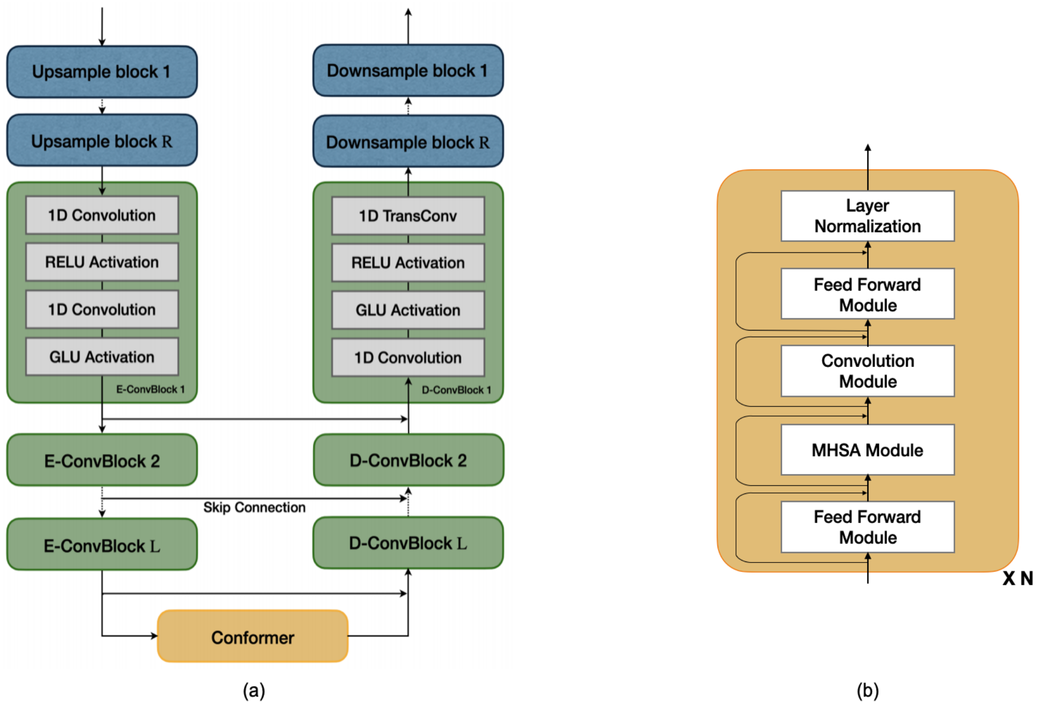

Figure 1: *Block diagram of (a) Overview of proposed architecture and (b) The conformer architecture.*

2. Method

The goal of SE is to estimate the clean speech signal $\hat{\mathbf{x}} \in \mathbb{R}^T$ from the noisy waveform $\mathbf{x} \in \mathbb{R}^T$ contaminated by noise, where T is the number of audio samples. The overall structure is shown in Figure 1 (a). The proposed model consists of a multi-layer CED structure and conformer block. The convolutional encoders perform upsampling and convolution blocks sequentially on the waveform signal to obtain the corresponding latent representations. These representations are applied with conformer block that capture the local context and global context dependencies to model the model sequence information. The decoder performs the convolution blocks with downsampling to reconstruct the time-domain signal from a latent representation estimated through the conformer block. For training objectives, similar to the previous work [9], we simultaneously used L1 and multi-resolution STFT loss [21] as follows:

$$L_{total}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{T}[||\mathbf{x} - \hat{\mathbf{x}}||_1 + \sum_{m=1}^{M} L_{stft}^{(m)}(\mathbf{x}, \hat{\mathbf{x}})] \qquad (1)$$

where the multi-resolution STFT loss is the sum of the STFT losses, which is the sum of the spectral convergence (sc) and magnitude (mag) loss, represented as follows:

$$\begin{aligned} L_{stft}^{(m)}(\mathbf{x}, \hat{\mathbf{x}}) &= L_{sc}^{(m)}(\mathbf{x}, \hat{\mathbf{x}}) + L_{mag}^{(m)}(\mathbf{x}, \hat{\mathbf{x}}) \\ L_{sc}^{(m)}(\mathbf{x}, \hat{\mathbf{x}}) &= \frac{|||STFT^{(m)}(\mathbf{x})| - |STFT^{(m)}(\hat{\mathbf{x}})|||_F}{|||STFT(\mathbf{x})|||_F} \qquad (2) \\ L_{mag}^{(m)}(\mathbf{x}, \hat{\mathbf{x}}) &= \frac{1}{T}||\log|STFT^{(m)}(\mathbf{x})| - \log|STFT^{(m)}(\hat{\mathbf{x}})|||_1 \end{aligned}$$

where $|| \cdot ||_F$ and $|| \cdot ||_1$ denote the Frobenius and L1 norms, respectively. M and $|STFT^{(m)}(\cdot)|$ denote the number of resolution parameter sets for STFT and the magnitude of STFT with the mth analysis parameter set, respectively.

2.1. Encoder

The encoder takes the mixture noise-corrupted waveform as an input $\mathbf{x} \in \mathbb{R}^T$ and learns a latent representation using the R upsampling blocks and the L stack of convolutional blocks (E-ConvBlock). The upsampling block doubled the time resolution using sinc interpolation. Each E-ConvBlock, numbered in order from 1 to l, is composed of two convolutions: the first convolution has a kernel size K_1, stride S_1, H_{l-1} input channels, H_l output channels, followed by rectified linear unit (ReLU) activation, and second convolution with kernel size K_2, stride S_2, H_l input channels, and $2H_l$ output channels, followed by gated linear units (GLU) [22].

2.2. Conformer-based sequence modeling in latent space

Recent studies [18, 19] show that the former is more effective in temporal sequence modeling in terms of improving ASR and source separation performance. The conformer can model local context information by inserting a depth-wise convolution into a transformer, which is effective in global context information modeling. Thereafter, we assume that the conformer is relatively more suited for sequence modeling than BLSTM or transformer in latent space for SE. A diagram of the conformer architecture is shown in Figure 1 (b). The conformer is the N stack of a conformer block, which consists of several modules, including macaron-like feedforward modules, convolution module (ConvBlock), MHSA module, and layer normalization. The feedforward modules are employed as a macaron-structure network, which consists of a linear layer with swish activation [26] and dropout, followed by a second linear layer. The ConvBlock starts with a pointwise convolution and a GLU, followed by a 1-

Table 1: *Comparison of performances of the proposed model in terms of PESQ, CSIG, CBAK, COVL, and STOI scores with the previously reported scores of baselines, based on the VCTK dataset.*

Model	Domain	PESQ	CSIG	CBAK	COVL	STOI
Noisy	-	1.97	3.35	2.44	2.63	0.92
HiFi-GAN [4]	Frequency	2.94	4.07	3.07	3.49	-
DeepMMSE [23]	Frequency	2.95	4.31	3.46	3.64	0.94
MHSA-SPK [24]	Frequency	2.99	4.15	3.42	3.57	-
MSSA-TCN [14]	Frequency	3.02	4.29	3.50	3.67	0.94
RDL-Net [25]	Frequency	3.02	4.38	3.43	3.72	0.94
T-GSA [17]	Frequency	3.06	4.18	**3.59**	3.62	-
Wave-U-Net [10]	Time	2.40	3.52	3.24	2.96	-
DEMUCS [9]	Time	3.07	4.31	3.40	3.63	0.95
SE-Conformer (Proposed)	Time	**3.13**	**4.45**	3.55	**3.82**	0.95

D depth-wise convolution layer with a Batchnorm, Swish activation, and pointwise convolution. We used the MHSA without relative positional embedding [27]. The prenorm residual unit [28] applies to all modules, including the feedforward, MHSA, and ConvBlock modules. With the latent representation derived by encoders, described in Section 2.1., the conformer blocks are performed and finally taken Sigmoid activation function to smooth the output value of the conformer blocks.

2.3. Decoder

The decoder is the inverse of the encoder. The decoder takes the output of the conformer blocks and sequentially performs L convolutional blocks (D-ConvBlock) and R downsample blocks to estimate the clean waveform $\hat{\mathbf{x}} \in \mathbb{R}^T$. The l-th D-ConvBlock starts with a convolution with kernel size K_2, stride S_2, input channels $2H_l$, output channels H_l, and GLU activation, followed by a transposed convolution with kernel size K_1 and stride S_1, input channels H_l, and output channels H_{l-1}, and ReLU activation. The downsampling blocks halved the time resolution by maintaining the number of features.

3. Experiments

3.1. Datasets

To verify the effectiveness of the proposed model, we used two datasets: the VCTK [29] and a Librispeech [30]. The VCTK dataset, which has been used in various recent studies [10, 23, 24, 14, 25, 17, 9] contains clear and noisy audio data. The clean data set consists of 30 speakers, selected by the VoiceBank corpus, of which 28 are reserved for the training set and 2 are reserved for the test set. The noisy training dataset was created with a total of 40 noise conditions with two artificial and eight real noise types, obtained from the DEMAND dataset [31] with various signal-to-noise ratios (SNRs) of 15, 10, 5, and 0 dB. Each of the 28 speakers used 40 noise conditions, and the total time was approximately 10 h. For the test data, the test set was generated with a total of 20 noise conditions using various SNRs randomly selected from 2.5, 7.5, 12.5, and 17.5 dB and five real noise types from the DEMAND dataset. Each of the two speakers used 20 noise conditions, and the number of utterances was 786. The test set conditions were different from the training set because the test set used different speaker and noise conditions.

To obtain a large dataset, we generated a simulated dataset based on Librispeech [30]. For the training set, we randomly selected 50 hours of clean speech audio samples from the *train-clean-100*. The training set used randomly selected noise and music sample sequences from the DEMAND [31] and MUSAN [32] noise dataset, with an SNR between -10 and 10 dB. The DEMAND was divided into six categories, and we used two of the three types of noise in each category. As for test data, clean speech are 500 utterances from randomly selected using the *test-clean* and background noises *babble* from the NOISEX-92 dataset [33], *river*, and *restaurant* from the DEMAND dataset, which were not used during the training. The SNR was set to be $\{-5, 0, 5, 10, 15\}$ dB for each noise type in addition to the clean speech. All utterances are sampled at 16 kHz.

3.2. Evaluation metrics

We evaluated the models in terms of subjective and objective scales, and compared them. For composite objective measures, the CSIG score is a signal distortion mean opinion score, the CBAK score measures background intrusiveness, and the COVL score measures speech quality. For an objective measure of naturalness, the perceptual evaluation of speech quality (PESQ) measure is adopted [34], as this measure is highly correlated with subjective quality [35]. The short-time objective intelligibility (STOI) score is used to measure the intelligibility gain by processing the noisy mixture with reference to the clean [36].

3.3. Experimental Setup

For the proposed model, the CED structure takes $R = 2$, $L = 4$, $H_0 = 48$, $K_1 = 8$, $S_1 = 4$, $K_2 = 1$, and $S_2 = 1$. In the middle part of the CED structure, we used two conformer layers ($N = 2$) with four attention heads, 256 attention dimensions, and 256 FFN dimensions without relative positional encodings. We used BLSTM and transformer instead of the conformer as our baseline models. BLSTM has three BLSTM layers with 1024 input dimensions and 512 hidden dimensions. For the training model, we trained the models for 500 epochs on both the VTCK and Librispeech datasets. We use the L1 loss and STFT loss, described in Section 2, with a weight of 0.5, between the predicted and ground truth clean speech waveforms. We used the AdamW optimizer [38] with a learning rate of 1e-4. For multi-resolution STFT loss, M is set to 3 with different FFT sizes, window sizes, and frame shift parameters such as $\in \{512, 50, 240\}$, $\in \{1024, 120, 600\}$, and $\in \{2048, 240, 1200\}$, respectively. For every experiment, we applied a random shift, $Remix$, and $BandMask$ augmentation on-the-fly during the training of the models, as in [9].

Table 2: *Comparison of PESQ and STOI scores of the proposed model and baselines in various noisy environment on Librispeech dataset.*

Model		Noisy		DNS-Base [37]		Demucs [9]		SE-Conformer	
Noise	**SNR**	**PESQ**	**STOI**	**PESQ**	**STOI**	**PESQ**	**STOI**	**PESQ**	**STOI**
Babble	-5dB	1.11	0.66	1.23	0.70	1.50	0.82	**1.65**	**0.83**
	0dB	1.19	0.77	1.48	0.82	1.92	0.90	**2.11**	**0.91**
	5dB	1.40	0.86	1.86	0.90	2.36	0.94	**2.54**	**0.95**
	10dB	1.75	0.92	2.29	0.94	2.76	0.96	**2.91**	**0.97**
	15dB	2.25	0.96	2.71	0.96	3.07	0.98	**3.22**	**0.98**
River	-5dB	1.23	0.80	1.81	0.86	2.05	0.90	**2.24**	**0.91**
	0dB	1.45	0.87	2.17	0.91	2.37	0.94	**2.59**	**0.94**
	5dB	1.78	0.92	2.55	0.95	2.66	0.96	**2.92**	**0.96**
	10dB	2.23	0.96	2.92	0.96	3.98	0.97	**3.22**	**0.98**
	15dB	2.78	0.98	3.23	0.98	3.33	0.98	**3.50**	**0.99**
Restaurant	-5dB	1.09	0.63	1.21	0.68	1.38	0.78	**1.53**	**0.80**
	0dB	1.15	0.75	1.43	0.80	1.68	0.88	**1.95**	**0.89**
	5dB	1.31	0.84	1.76	0.88	1.99	0.93	**2.38**	**0.93**
	10dB	1.62	0.91	2.16	0.93	2.33	0.95	**2.75**	**0.96**
	15dB	2.07	0.95	2.59	0.96	2.70	0.97	**3.06**	**0.97**
Average		1.63	0.85	2.09	0.88	2.34	0.92	**2.57**	**0.93**

3.4. Experimental Results

3.4.1. Comparison with previous methods

Table 1 shows a comparison of the metric scores of the proposed model with those of other architectures, such as DeepMMSE [23], HiFi-GAN [4], MHSA-SPK [24], MSSA-TCN [14], RDL-NET [25], T-GSA [17], Wave-U-Net [10], and Demucs [9]. As all these architectures use the same VCTK dataset and the same metrics that we used to train and test our model, we compared the metric scores of our model with the number of baselines reported in the original published works. We can see that our method outperforms baselines. The results indicated that the proposed method preserves better speech quality. To demonstrate the generalization of the proposed model's improvements over baselines, we experimented on a larger Librispeech dataset, as shown in Table 2. In this experiment, we used the DNS-baseline [37] and Demucs [9] as baselines. For the proposed model, the average PESQ and STOI scores under all conditions were 2.57 and 0.93, whereas those for the previous best Demucs model were 2.34 and 0.92, respectively. We found that the proposed model tended to perform better than the baselines in all noisy environments. We can confirm that the proposed model could achieve better speech quality in various background noise for the simulated large data.

Table 3: *Comparison of Objective measures of CED structure with the conformer and other models using the VCTK dataset.*

Model	PESQ	CSIG	CBAK	COVL
CED	2.63	4.07	3.24	3.35
CED + Transformer	2.89	4.30	3.43	3.61
CED + BLSTM [9]	2.94	4.34	3.45	3.67
CED + Conformer	**3.13**	**4.45**	**3.55**	**3.82**

3.4.2. BLSTM/Transformer/Conformer Block

We conducted experiments to compare the performance of the conformer with that of the transformer and BLSTM [9]. For a fair comparison, we used the same number of parameters as the proposed CED structure. As shown in Table 3, applying the conformer yields superior performance over the transformer and BLSTM. This reinforces our idea that conformer using convolution with a transformer is capable of modeling a sequence incorporating both local and global context information.

3.4.3. The effect of the components of Conformer

We conducted several experiments to verify the effectiveness of several components in the conformer, including the Macaron FFN and ConvBlock. The results are listed in Table 4. Notably, we can see a significant drop in performance, when ConvBlock is removed from the conformer block. This can be interpreted as ConvBlock being an important factor in capturing local context information based on the transformer model, which captures global context information.

Table 4: *Disentangling Conformer. Comparison of SE-Conformer structure without Macaron FFN, ConvBlock and both of them.*

Model	PESQ	CSIG	CBAK	COVL
SE-Conformer	**3.13**	**4.45**	**3.55**	**3.82**
− Macaron FFN	**3.13**	**4.45**	**3.55**	**3.82**
− ConvBlock	2.88	4.28	3.41	3.60
− ConvBlock & Macaron FFN	2.89	4.30	3.43	3.61

4. Conclusions

In this paper, we proposed a SE-Conformer, which is an end-to-end speech enhancement method based on CED and conformer architecture. It is suitable for sequence modeling by attending the entire sequence at once with self-attention and CNN in latent space. The evaluation results showed the proposed model outperforms other competitive baselines on the VCTK and Librispeech large datasets in terms of standard objective speech quality metrics. In future work, we plan to extend this mechanism to consider real-time processing algorithms.

5. References

[1] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2006.

[2] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[3] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2019.

[4] J. Su, Z. Jin, and A. Finkelstein, "Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," in *INTERSPEECH*, 2020.

[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] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations (ICLR)*, 2018.

[7] T.-A. Hsieh, H.-M. Wang, X. Lu, and Y. Tsao, "Wavecrn: An efficient convolutional recurrent neural network for end-to-end speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 2149–2153, 2020.

[8] V. Kishore, N. Tiwari, and P. Paramasivam, "Improved speech enhancement using tcn with multiple encoder-decoder layers," in *INTERSPEECH*, 2020, pp. 4531–4535.

[9] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *INTERSPEECH*, 2020.

[10] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," in *INTERSPEECH*, 2018.

[11] G. Kerg, B. Kanuparthi, A. G. ALIAS PARTH GOYAL, K. Goyette, Y. Bengio, and G. Lajoie, "Untangling tradeoffs between recurrence and self-attention in artificial neural networks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[13] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7829–7833.

[14] J. Lin, A. J. van Wijngaarden, K.-C. Wang, and M. C. Smith, "Speech enhancement using multi-stage self-attentive temporal convolutional networks," *arXiv:2102.12078*, 2021.

[15] A. Pandey and D. Wang, "Dense cnn with self-attention for time-domain speech enhancement," *arXiv:2009.01941*, 2020.

[16] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[17] J. Kim, M. El-Khamy, and J. Lee, "T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6649–6653.

[18] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *INTERSPEECH*, 2020.

[19] S. Chen, Y. Wu, Z. Chen, J. Li, C. Wang, S. Liu, and M. Zhou, "Continuous speech separation with conformer," *arXiv:2008.05773*, 2020.

[20] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Convolution augmented transformer for semi-supervised sound event detection," in *Proceedings of Workshop Detection Classification Acoustic Scenes Events (DCASE)*, 2020.

[21] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203.

[22] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International Conference on Machine Learning (ICML)*, 2017, pp. 933–941.

[23] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "Deepmmse: A deep learning approach to mmse-based noise power spectral density estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, 2020.

[24] Y. Koizumi, K. Yaiabe, M. Delcroix, Y. Maxuxama, and D. Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[25] M. Nikzad, A. Nicolson, Y. Gao, J. Zhou, K. K. Paliwal, and F. Shang, "Deep residual-dense lattice network for speech enhancement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8552–8559.

[26] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv:1710.05941*, 2017.

[27] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[28] T. Q. Nguyen and J. Salazar, "Transformers without tears: Improving the normalization of self-attention," *The International Workshop on Spoken Language Translation (IWSLT)*, 2019.

[29] C. Valentini-Botinhao *et al.*, "Noisy speech database for training speech enhancement algorithms and tts models," 2017.

[30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[31] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, no. 1. Acoustical Society of America, 2013, p. 035081.

[32] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," 2015.

[33] A. Varga and H. J. 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.

[34] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2001.

[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, 2007.

[36] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010.

[37] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Icassp 2021 deep noise suppression challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[38] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2018.