D²Net: A Denoising and Dereverberation Network Based on Two-branch Encoder and Dual-path Transformer

Liusong Wang*†, Wenbing Wei*†, Yadong Chen*†, and Ying Hu*†
* School of Information Science and Engineering, Xinjiang University, Urumqi, China
† Key Laboratory of Signal Detection and Processing in Xinjiang, Urumqi, China
E-mail: wls@stu.xju.edu.cn, huying@xju.edu.cn

***Abstract*—The simultaneous denoising and dereverberation for single-channel mixture speech under the complicated acoustic environment is considered to be a challengeable task. In this paper, we propose a denoising and dereverberation network named as D²Net in which a two-branch encoder (TBE) is designed to extract and selectively fuse features with different granularity. In addition, we design a global-local dual-path transformer (GLDPT) which introduces the local dense synthesizer attention (LDSA) in the dual-path transformer to improve the perception of local information. We evaluated our proposed D²Net and conducted ablation studies on the VoiceBank+DEMAND and WHAMR! datasets. Meanwhile, we chose three types of data in the WHAMR! dataset to verify the ability of the D²Net on the tasks of denoising-only, dereverberation-only, and simultaneous denoising and dereverberation, respectively. Experimental results show that our proposed model outperforms the comparative models, and all achieve better performance on the tasks of simultaneous denoising and dereverberation, dereverberation-only, and denoising-only, while keeping a small number of network parameters.**

***Index Terms*—Speech denoising, Speech dereverberation, Two-branch encoder, Dual-path transformer**

I. Introduction

Speech is easily degraded by background noise and room reverberation, resulting in a significant decline of speech intelligibility and speech recognition performance. Reverberation is the accumulation of multiple reflections of a signal as it travels around the room from source to microphone. Room reverberation can be characterized by the room impulse response (RIR) related to the position of the sound source and microphone [1] [2]. Speech denoising and dereverberation processing is an indispensable front-end task and has been widely studied in speech recognition [3] [4]. In this paper, we focus on the task of single-channel denoising and dereverberation, which is more challengeable.

Generally, the spatial information is extracted and fed into the network of multi-channel dereverberation, such as inter-channel phase differences (IPDs) [5] and inter-channel convolution differences (ICDs) [6]. However, for the task of single-channel dereverberation, the spatial information is not available. There are some researches that exploit cascaded modules to complete denoising and dereverberation [7] [8].

According to the input feature, speech enhancement methods can be divided into two categories: time-domain methods and time-frequency (T-F) domain methods. The time-domain methods estimate the clean waveform directly from the mixture speech in the time domain [9] [10]. The traditional T-F domain methods usually adopt the magnitude spectrogram obtained by the short time Fourier transform (STFT) operation [11]. In this case, the performance is limited because the phase information can not be utilized effectively. Many recent studies have begun to adopt the complex-valued spectrogram, which can be decomposed into the amplitude and phase in polar coordinates or the real and imaginary parts in Cartesian coordinates. The complex-valued spectrogram served as input is verified to improve the performance of the speech denoising network [12].

In recent years, dual-path networks have shown good performance in speech separation [13] [14] and speech enhancement [10]. Wang K et al. introduced transformer [15] into dual-path network structure and proposed a time-domain speech enhancement model: Two-stage transformer based neural network (TSTNN), which greatly improves the performance of the speech enhancement [10]. Several studies reported that dot-product self-attention may not be indispensable to the transformer models. Xu M et al. proposed the local dense synthesizer attention (LDSA) which dispenses with dot products and pairwise interactions, and restricts the attention scope to a local range around the current central frame to reduce the computational complexity and improve the performance [16].

Inspired by the above-mentioned works, we propose a single-channel network for simultaneous denoising and dereverberation named as D²Net, in which a two-branch encoder (TBE) is designed to extract and selectively fuse the different granularity features from two branches. Meanwhile, we design a global-local dual-path transformer (GLDPT) which introduces the LDSA in the dual-path transformer to improve the perception of local information. We evaluated our proposed D²Net and conducted ablation studies on the VoiceBank+DEMAND and WHAMR! datasets.

The remainder of this paper is organized as follows. The structure of D²Net is introduced in Section II, and datasets and experiment setup in Section III. Section IV describes the ablation studies and comparisons of denoising and dereverberation performance with other models in two datasets. The

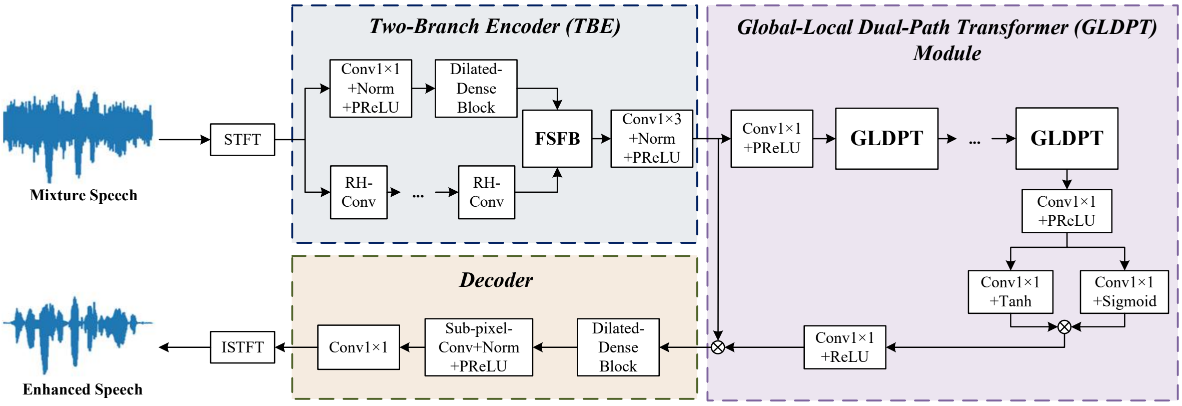

Fig. 1. Diagram of the D²Net. ⊗ represents the element-wise multiplication.

conclusions are drawn in Section V.

II. Methodology

Our proposed D²Net following an encoder-decoder structure is composed of a two-branch encoder (TBE), a global-local dual-path transformer (GLDPT) module, and a decoder. The GLDPT module is used to estimate the mask of target speech. Fig. 1 shows the overall structure of D²Net. *Norm* denotes the switchable normalization operation [17].

A. Two-Branch Encoder

The two-branch encoder (TBE) consists of the upper and lower branches of feature extract layers and a feature selective fusion block (FSFB). The mixture waveform is converted into spectrogram by the STFT operation. The input of the TBE, a 3-D tensor $X \in \mathbb{R}^{2\times T\times F}$, is formed by the stacking of the real and imaginary parts of the complex-valued spectrogram.

The upper branch of feature extract layer consists of a convolutional layer and a dilated-dense block [18]. The channel number is increased to 64 by the operation of convolution with the kernel size of 1×1. The dilated-dense block utilizes 4 dilation convolution layers to continuously expand receptive fields to aggregate context at different resolutions. The lower branch consists of multiple residual hybrid convolution (RH-Conv) blocks [19]. The RH-Conv block makes better use of time-frequency dependence to extract more fine-grained features as shown in Fig. 2.

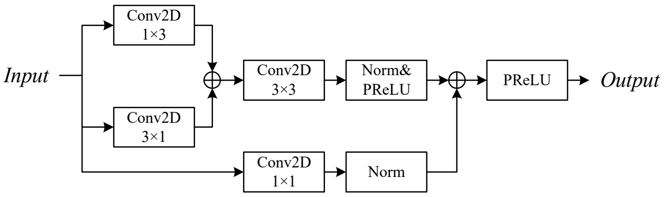

Fig. 2. Diagram of the residual hybrid convolution (RH-Conv) block. ⊕ represents the element-wise addition.

The features of the upper and lower branches, F_{b1}, $F_{b2} \in \mathbb{R}^{64\times T\times F}$, are fed into the FSFB. Inspired by [20], the architecture of designed FSFB is shown in Fig. 3. The F_{b1} and F_{b2} are concatenated along with the channel dimension, and undergo a linear layer (*Linear*) and sigmoid activation operation (*Sigmoid*) to generate a gated parameter *w* for F_{b1} and thus *1-w* for F_{b2}. The whole procedure can be formulated as follows:

$$w = Sigmoid\left(Linear\left(Concat\left[F_{b1}, F_{b2}\right]\right)\right) \tag{1}$$

$$F = ReLU((w \otimes F_{b1} + (1-w) \otimes F_{b2}) + F_{b1} + F_{b2}) \tag{2}$$

where $F \in \mathbb{R}^{64\times T\times F}$ is the output of the FSFB. The fusion scheme of the FSFB combines each other's features to preserve key cues and discard inessential information [20]. After that, the frequency dimension is down-sampled through a convolutional layer with the kernel size of 1×3 and stride of 1×2.

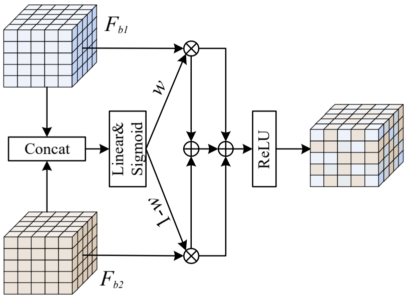

Fig. 3. Diagram of the feature selective fusion block (FSFB) architecture. *w* denotes a weighted factor.

B. Global-Local Dual-Path Transformer (GLDPT) Module

The GLDPT module consists of multiple GLDPTs and convolutional layers. A GLDPT consists of right-and-left two improved transformers as shown in Fig. 4.

In order to reduce the computational complexity, we use a convolutional layer with the kernel size of 1×1 followed by a PReLU operation to halve the channel number. For each improved transformer in Fig. 4, the local dense synthesizer attention (LDSA) [16] block is introduced after the multi-head self-attention (MHSA) block. Compared with self-attention, the current frame in the LDSA can only be restricted to interact with its neighboring frames so that the LDSA has the ability to extract fine-grained local features [16]. Thus, the improved transformer in the GLDPT is more efficient than the original transformer in modeling the global and local dependencies of speech sequences. The global and local contextual information along with the time and frequency dimensions can be captured and extracted by swapping the time and frequency dimensions.

The input Y is firstly fed into the MHSA block.

$$Y' = LN\Big(MHSA\Big(Y\Big) + Y\Big) \tag{3}$$

where Y' denotes the output through the residual connection and layer normalization (*LN*) after the MHSA block. Then the Y' is passed through the LDSA block.

$$Y'' = LN\left(LDSA\left(Y'\right) + Y'\right) \tag{4}$$

where Y'' denotes the output through the residual connection and layer normalization after the LDSA block. In the feed-forward network (FFN), we adopt a GRU layer (*GRU*) as the first layer to capture the long-term context dependencies for both past and future frames of the time series [10]. The output of the FFN is as follows:

$$Y_{ffn} = Linear\left(ReLU\left(GRU\left(Y''\right)\right)\right) \tag{5}$$

The output of an improved transformer in the GLDPT is as follows:

$$Y_{out} = LN\left(Y_{ffn} + Y''\right) \tag{6}$$

The first improved transformer is followed by the group normalization operation (*GN*) and residual connection. Then, the feature is fed into the next improved transformer after a permutation operation by swapping the time and frequency dimensions.

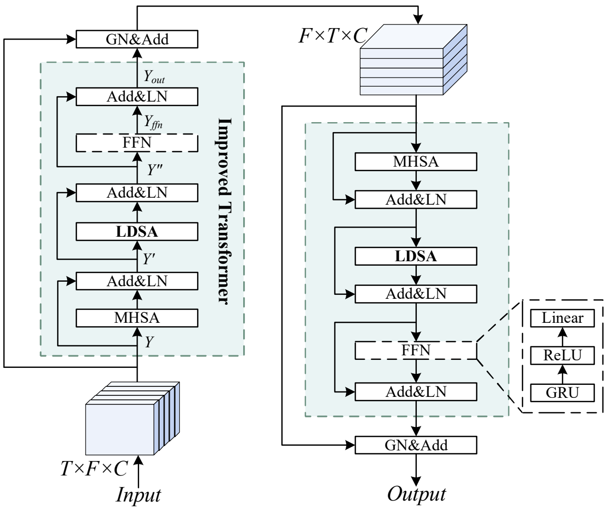

Fig. 4. Diagram of the global-local dual-path transformer (GLDPT) architecture. *Add* represents $\oplus$.

As shown in Fig. 1, after the last GLDPT, a convolutional layer with the kernel size of 1×1 followed by a PReLU operation is used to restore the channel number to 64. Then the feature map is fed into two convolutional layers with the kernel size of 1×1 followed by a nonlinearity activation operation, respectively. The outputs are multiplied together and fed into a convolutional layer with the kernel size of 1×1 followed by a ReLU operation to get the mask of target speech.

Finally, the estimated mask and the input of the GLDPT module are multiplied to serve as the input of the decoder.

C. Decoder

The decoder is composed of a dilated-dense block and a sub-pixel convolutional layer [21]. The structure of the dilated-dense block is consistent with that in the TBE. After a convolutional layer with the kernel size of 1×1, the channel number is reduced to 2 that two channels represent the real and imaginary parts of spectrogram of estimated speech, respectively. And finally, the waveform of estimated target speech is restructured by the inverse STFT (ISTFT) operation.

D. Loss Function

The loss function combines both time-domain and time-frequency domain losses. The loss function is as follows:

$$Loss = \alpha * Loss_T + \beta * Loss_{TF} \tag{7}$$

where α and β are tunable parameters and set to 0.4 and 0.6 in this paper, respectively. $Loss_T$ is the mean square error (MSE) loss:

$$Loss_T = \frac{1}{N}\sum_{n=0}^{N-1}\left(x_n - \tilde{x}_n\right) \tag{8}$$

where x and $\tilde{x}$ are the sample of the clean speech and enhanced speech, respectively. N denotes the number of samples in the waveform. $Loss_{TF}$ is the L1 loss:

$$\begin{aligned} Loss_{TF} =& \frac{1}{TF}\sum_{t=0}^{T-1}\sum_{f=0}^{F-1}\Big|\Big(\big|X_r(t,f)\big| - \big|\tilde{X}_r(t,f)\big|\Big) \\ &+ \Big(\big|X_i(t,f)\big| - \big|\tilde{X}_i(t,f)\big|\Big)\Big| \end{aligned} \tag{9}$$

where X and $\tilde{X}$ are the spectrogram of the clean speech and enhanced speech, respectively. r and i denote the real and imaginary parts of complex-valued spectrogram. T and F are the number of frames and frequency bins, respectively.

III. Experiments

A. Datasets and Evaluation Metrics

We evaluated the denoising performance of our proposed D²Net on the VoiceBank+DEMAND dataset which contains pre-mixed noisy speech and its paired clean speech. The clean set is selected from the VoiceBank corpus [24], where the training set contains 11,572 utterances and the test set contains 872 utterances. For the pre-mixed noisy speech, the training set contains 40 different noise conditions with 10 types of noises (8 from the DEMAND database [25] and 2 artificially generated) at 4 signal noise ratios (SNRs) from 0 dB to 15 dB at 5 dB intervals, and the test set contains 20 different noise conditions with 5 types of unseen noises from the DEMAND database at 4 SNRs from 2.5 dB to 17.5 dB at 5 dB intervals.

TABLE I
EVALUATION RESULTS AND ABLATION ANALYSIS ON THE VOICEBANK+DEMAND DATASET

Model	Para.(Million)	PESQ	STOI	SI-SNR	CSIG	CBAK	COVL
Noisy	-	1.97	0.92	8.45	3.34	2.45	2.63
SEGAN [22] ,2017	97.47	2.16	0.93	-	3.48	2.94	2.80
Wave U-net [9] ,2018	10.00	2.40	-	-	3.52	3.24	2.96
MetricGAN [11] ,2019	1.86	2.86	-	-	3.99	3.18	3.42
DCCRN*[12] ,2020	3.67	2.57	0.94	19.13	3.93	2.90	3.21
DEMUCS(Small) [23] ,2020	18.90	2.93	0.95	-	4.22	3.25	3.52
TSTNN [10] ,2021	**0.92**	2.96	0.95	18.82	4.31	**3.49**	3.66
D^{2}Net	1.13	**3.27**	**0.96**	**19.78**	**4.63**	3.18	**3.92**
-TBE(i)	0.95	3.19	0.96	19.28	4.60	2.59	3.90
-FSFB(ii)	1.12	3.22	0.96	19.45	4.43	3.08	3.77
-GLDPT(iii)	1.10	3.23	0.96	19.51	4.59	3.17	3.94
-TBE-GLDPT(iv)	0.92	3.11	0.95	19.05	4.15	2.79	3.51

* represents the results of the model obtained by our reproduction.

TABLE II
EVALUATION RESULTS AND ABLATION ANALYSIS ON THE WHAMR! DATASET

	Noise+Reverb			*Reverb*			*Noise*		
Model	PESQ	STOI	SI-SNR	PESQ	STOI	SI-SNR	PESQ	STOI	SI-SNR
Mixed	1.11	0.73	-2.73	2.16	0.91	4.38	1.11	0.76	-0.99
DCCRN*	1.59	0.88	5.20	2.59	0.95	7.51	1.66	0.90	9.03
TSTNN*	1.91	0.91	2.89	2.66	0.95	3.56	1.94	0.93	4.17
D^{2}Net	**2.51**	**0.95**	**10.25**	**3.68**	**0.99**	**15.64**	**2.48**	**0.95**	**11.89**
-TBE(i)	2.42	0.95	9.80	3.52	0.99	15.00	2.41	0.95	11.35
-FSFB(ii)	2.49	0.95	10.05	3.67	0.99	15.57	2.45	0.95	11.65
-GLDPT(iii)	2.46	0.95	9.79	3.64	0.99	15.09	2.42	0.95	11.34
-TBE-GLDPT(iv)	2.31	0.94	8.66	3.35	0.98	13.19	2.30	0.94	10.66

* represents the results of the model obtained by our reproduction.

We evaluated the denoising and dereverberation performance of the D^{2}Net on the WHAMR! dataset [26] which is based on the WHAM! dataset [27] with the addition of synthetic room impulse responses (RIRs). Reverberation times are chosen to approximate domestic and classroom environments. The WHAMR! dataset is a noise- and reverberation-augmented version of the wsj0-2mix dataset [28]. The noise in the dataset comes from different urban environments in the San Francisco Bay Area was mixed with clean speaker speech by an SNR of random selection between -6 and +3 dB. The WHAMR! dataset contains many types of utterances where only the noisy and reverberant speech, reverberant speech, and noisy speech were used in the experiments. The training, validation, and test sets contain 20000, 5000, and 3000 utterances, respectively. During the training phase, only the noisy and reverberant speech was selected as the training set. While, during the test phase, three types of data were all evaluated in order to verify the denoising, dereverberation, and simultaneous denoising and dereverberation ability of the D^{2}Net.

In terms of evaluation metrics indexes, we choose perceptual evaluation of speech quality (PESQ) [29], short-time objective intelligibility (STOI) [30], and scale-invariant source-to-noise ratio (SI-SNR) [31]. We also adopt the three most commonly used metrics in the VoiceBank+DEMAND dataset, which are CSIG [32] for signal distortion, CBAK [32] for noise distortion evaluation, and COVL [32] for overall quality evaluation to evaluate the performance of the D^{2}Net. For the experiment of the WHAMR! dataset, only the results of PESQ, STOI, and SI-SNR are shown in Table II.

B. Experimental Setup

The samples are kept with 16 kHz by downsampling, and then randomly sliced into the segment with the length of 4 seconds. For the STFT and ISTFT operations, the window length and hop size are set to 25 ms and 6.25 ms, respectively, and the FFT length is 512. 5 RH-Conv blocks with the channel number of feature maps of 16, 32, 48, 64, and 64, respectively, are equipped in the TBE. There are 4 GLDPTs in the GLDPT module. We use Adam [33] as the optimizer and a gradient clipping with maximum L2-norm of 5 to avoid gradient explosion. During the training phase, the D^{2}Net and comparative models are all trained for 100 epochs. A dynamic strategy [15] is used to adjust the learning rate as follows:

$$lr = \begin{cases} 0.2 \cdot 64^{-0.5} \cdot n \cdot 4000^{-1.5}, & n \leq nwarmup \\ 4e^{-4} \cdot 0.98^{\lfloor \frac{epoch}{2} \rceil}, & n > nwarmup \end{cases} \tag{10}$$

where n is the number of steps.

IV. RESULTS

A. Results on the VoiceBank+DEMAND dataset

For the denoising-only task, we compared our proposed D^{2}Net with the other six models on the VoiceBank+DEMAND dataset. As can be seen from Table I, the D^{2}Net achieves good performance and significant improvements over the other six models, especially on the PESQ and SI-SNR. Compared with the time-domain dual-path transformer model TSTNN [10], the D²Net improves by 0.31 and 0.96 in PESQ and SI-SNR, respectively.

B. Results on the WHAMR! dataset

We evaluated the performance of our proposed D²Net on three types of data in the WHAMR! dataset: noisy and reverberant speech (*Noise+Reverb*), reverberant speech (*Reverb*), and noisy speech (*Noise*), the results as shown in Table II. The D²Net achieves the best performance on the tasks of simultaneous denoising and dereverberation, dereverberation-only, and denoising-only. Compared with the TSTNN, the results of the D²Net achieves considerable improvements on three tasks, especially on the task of dereverberation-only where the STOI score reaches 0.99.

C. Ablation analysis

We also conducted ablation studies on two datasets to verify the effectiveness of TBE, FSFB, and GLDPT in the D²Net. For Table I and Table II, (i) denotes without the TBE that is the lower branch of feature extract layer and FSFB are removed. (ii) denotes without the FSFB that means adopting the element-wise addition method instead of the FSFB. (iii) denotes without the GLDPT that means the LDSA in the GLDPT is removed keeping with the two-stage transformer used in the TSTNN. (iv) denotes that the D²Net preserves the settings in (i) and (iii) simultaneously.

The results of (i) show that the TBE improves the performance in the PESQ score by 0.08 and 0.09 on the VoiceBank+DEMAND and WHAMR! datasets, respectively. This shows that the TBE can effectively extract different granularity features from different branches and fuse them to improve the performance of the network of denoising and dereverberation. The results of (ii) show that the FSFB improves the performance in the PESQ score by 0.05 and 0.02 on the VoiceBank+DEMAND and WHAMR! datasets, respectively. It verifies the FSFB can effectively preserve key cues and discard inessential information. According to the results of (iii), we can find that introducing the LDSA improves the performance in the PESQ score by 0.04 and 0.05 on the VoiceBank+DEMAND and WHAMR! datasets, respectively. These prove the effectiveness of the LDSA on local information extraction. The enhanced speech demos are available online[1].

V. Conclusions

In this paper, we propose a simultaneous denoising and dereverberation network for single-channel mixture speech named D²Net which is based on two-branch encoder (TBE) and dual-path transformer. The TBE we designed can effectively extract different granularity features and utilize the feature selective fusion block (FSFB) to preserve key cues and discard inessential information. Meanwhile, the local dense synthesizer attention (LDSA) is introduced in the global-local dual-path transformer (GLDPT) to capture the fine-grained local features.

We evaluated our proposed D²Net and conducted ablation studies on the VoiceBank+DEMAND and WHAMR! datasets. In addition, we chose three types of data in the WHAMR! dataset to verify the ability of the D²Net on the tasks of denoising-only, dereverberation-only, and simultaneous denoising and dereverberation, respectively. Experimental results show that our proposed D²Net outperforms the comparative models, and all achieve better performance on the tasks of simultaneous denoising and dereverberation, dereverberation-only, and denoising-only, while keeping with a small number of network parameters.

Acknowledgment

This work is supported by the Multi-lingual Infomation Technology Research Center of Xinjiang (ZDI145-21), National Natural Science Foundation of China (NSFC) (U1903213) and Tianshan Innovation Team Plan Project of Xinjiang (202101642).

References

[1] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.

[2] O. Hazrati, J. Lee, and P. C. Loizou, "Blind binary masking for reverberation suppression in cochlear implants," *The Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1607–1614, 2013.

[3] R. Gomez, T. Kawahara, and K. Nakadai, "Optimized wavelet-domain filtering under noisy and reverberant conditions," *APSIPA Transactions on Signal and Information Processing*, vol. 4, 2015.

[4] H. Shi, L. Wang, S. Li, C. Fan, J. Dang, and T. Kawahara, "Spectrograms fusion-based end-to-end robust automatic speech recognition," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 438–442.

[5] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "On end-to-end multi-channel time domain speech separation in reverberant environments," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6389–6393.

[6] R. Gu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Enhancing end-to-end multi-channel speech separation via spatial feature learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7319–7323.

[7] Y. Fu, J. Wu, Y. Hu, M. Xing, and L. Xie, "Desnet: A multi-channel network for simultaneous speech dereverberation, enhancement and separation," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 857–864.

[8] A. Li, W. Liu, X. Luo, G. Yu, C. Zheng, and X. Li, "A simultaneous denoising and dereverberation framework with target decoupling," *Proc. Interspeech 2021*, pp. 2801–2805, 2021.

[9] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[10] K. Wang, B. He, and W.-P. Zhu, "Tstnn: Two-stage transformer based neural network for speech enhancement in the time domain," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7098–7102.

[11] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2031–2041.

[12] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *Proc. Interspeech 2020*, pp. 2472–2476, 2020.

[1] https://wangliusong.github.io/D2-Demo/

[13] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[14] J. Chen, Q. Mao, and D. Liu, “Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation,” *Proc. Interspeech 2020*, pp. 2642–2646, 2020.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[16] M. Xu, S. Li, and X.-L. Zhang, “Transformer-based end-to-end speech recognition with local dense synthesizer attention,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5899–5903.

[17] P. Luo, R. Zhang, J. Ren, Z. Peng, and J. Li, “Switchable normalization for learning-to-normalize deep representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 712–728, 2021.

[18] A. Pandey and D. Wang, “Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6629–6633.

[19] Y. Hu, X. Zhu, Y. Li, H. Huang, and L. He, “A multi-grained based attention network for semi-supervised sound event detection,” in *Interspeech*, 2022.

[20] Y. Hu, Y. Chen, W. Yang, L. He, and H. Huang, “Hierarchic temporal convolutional network with cross-domain encoder for music source separation,” *IEEE Signal Processing Letters*, vol. 29, pp. 1517–1521, 2022.

[21] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[22] S. Pascual, A. Bonafonte, and J. Serrà, “Segan: Speech enhancement generative adversarial network,” *Proc. Interspeech 2017*, pp. 3642–3646, 2017.

[23] A. Défossez, G. Synnaeve, and Y. Adi, “Real time speech enhancement in the waveform domain,” *Proc. Interspeech 2020*, pp. 3291–3295, 2020.

[24] C. Veaux, J. Yamagishi, and S. King, “The voice bank corpus: Design, collection and data analysis of a large regional accent speech database,” in *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*. IEEE, 2013, pp. 1–4.

[25] 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.

[26] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, “Whamr!: Noisy and reverberant single-channel speech separation,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 696–700.

[27] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, “Wham!: Extending speech separation to noisy environments,” *Proc. Interspeech 2019*, pp. 1368–1372, 2019.

[28] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[29] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[31] Y. Luo and N. Mesgarani, “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[32] 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.

[33] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR (Poster)*, 2015.