2024 Tenth International Conference on Communications and Electronics (ICCE) | 979-8-3503-7979-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCE62051.2024.10634678

DConvT: Deep Convolution-Transformer Network Utilizing Multi-scale Temporal Attention for Speech Enhancement

Hoang Ngoc Chau, Anh Xuan Tran Thi, and Quoc Cuong Nguyen*

School of Electrical and Electronic Engineering, Hanoi University of Science and Technology
Hanoi, 100000, Vietnam
chau.hn222175m@sis.hust.edu.vn, xuan.tranthianh@hust.edu.vn
*Corresponding author: cuong.nguyenquoc@hust.edu.vn

***Abstract*—Transformers have made significant progress in a wide range of speech processing tasks. However, they are not widely explored in real-time speech enhancement due to the complex distribution of noise over speech signal in challenging scenarios. Pure transformers can hardly capture the structure of speech harmonics and how they change over time in the noisy speech spectrogram. To fill the gap, we propose a deep convolution-transformer network (DConvT), in which the multi-scale temporal attention is designed to effectively model different speech structures from multiple temporal scales in a convolutional encoder-decoder architecture. In DConvT, multi-scale temporal information is extracted by dilated convolutions and projected to the query/key-value sequence of the self-attention operation. In this way, the self-attention can perform global modeling on a variety of temporal structures separately. The multi-scale temporal attention results are subsequently merged by concatenation to fuse the information of multiple temporal patterns. In addition, we adopt a convolutional encoder-decoder (CED) architecture to efficiently model and down-sample the frequency dimension of the speech spectrogram, which produces compact sequences for temporal modeling. The experimental results on 2020 Deep Noise Suppression Challenge (DNS20) dataset show that DConvT achieves state-of-the-art performance among existing speech enhancement baselines.**

***Index Terms*—Speech enhancement, deep neural networks, transformer.**

I. Introduction

Speech enhancement (SE) involves retrieving clean speech from a noisy mixture to improve speech perceptual quality and intelligibility for a wide range of applications, such as telecommunication, automatic speech recognition, and hearing aids [1]. Compared to traditional statistical approaches, data-driven approaches using deep neural networks (DNNs) have pushed the performance of SE systems to remarkably further limits. This is mostly due to their robust ability to handle nonstationary noises through nonlinear approximation [2], [3].

Generally, DNN-based approaches function either in the time domain (speech waveform) [4]–[6] or in the time-frequency (T-F) domain [7]–[9], where the latter representation is produced by applying the Short-Time Fourier Transform (STFT) to the speech waveform, resulting a spectrogram. This paper focuses on the T-F domain as the spectrogram representation provides a concise representation of the spectral patterns that are important for speech perception and noise components. For real-time SE, recent studies commonly combine the effectiveness of convolutional neural networks (CNNs) and recurrent neural networks (RNNs). In [7], the authors proposed a complex convolution recurrent network, where the complex convolutional encoder-decoder (CED) extracts hierarchical spectral features with phase-awareness and the encoded high-level features are fed to long short-term memory networks (LSTMs) for frame-by-frame modeling. This design scheme ranked first in the Interspeech 2020 Deep Noise Suppression Challenge (DNS20) and inspired several subsequent works [10]–[12]. The authors in [10], [13] further improved the recurrent module by introducing RNNs along the frequency axis to learn better correlations between frequency bands. For explicit phase estimation, Li et al. [8], [9] decoupled phase and magnitude into separated branches and optimized temporal modeling through temporal convolutional networks (TCNs) [14]. However, although RNNs and TCNs have exhibited their effectiveness in previous works, they lack the capability to explicitly capture global correlations in temporal representation by design.

Transformers are well-known for global context modeling due to their ability to model dependencies between all elements in a sequence simultaneously [15]. However, they are not widely investigated in the real-time SE task and still lag behind the aforementioned temporal modeling methods [3], [16]. The noisy spectrogram possesses structural correlations between frames, which can be taken advantage of by the inductive bias from internal designs of RNNs and TCNs, such as sequential characteristics, locality, and translation-invariance. In contrast, pure transformers can hardly learn these temporal structures, since one STFT frame contains little context information. Several works have proposed transformer-based models for SE [17]–[19]. Nevertheless, these models are evaluated on a small dataset [20] and are noncausal designs for non-real-time SE, which can be computationally expensive.

In this paper, we propose a novel deep convolution-transformer network (DConvT) to take advantage of both

979-8-3503-7979-2/24/$31.00 ©2024 IEEE

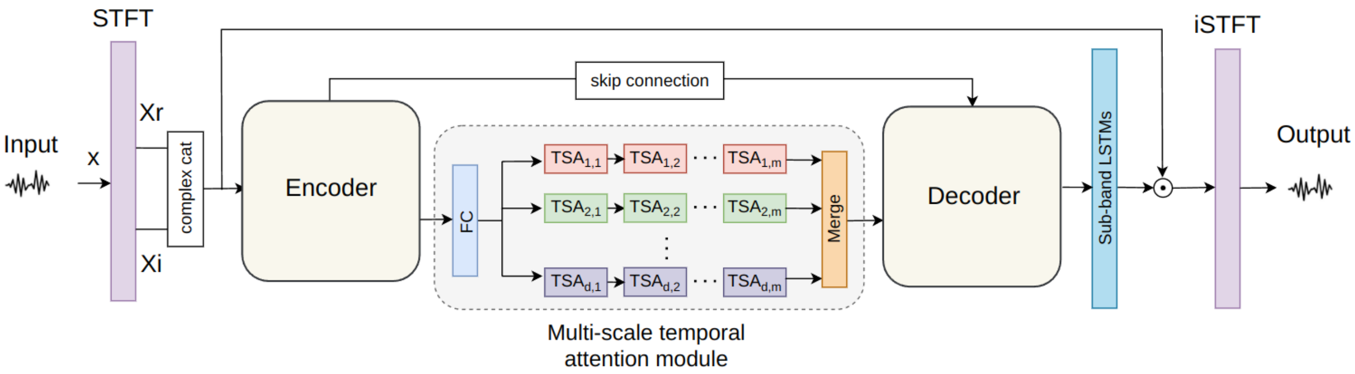

Fig. 1. System flowchart of proposed DConvT. ⊙ represents element-wise complex multiplication

CNNs and transformers. We first build the temporal self-attention block to extract the global correlations in temporal representation. For better mining of structural information, causal dilated convolutions are introduced in the query/key-value projection of temporal attention blocks. This enhances the self-attention operation from frame-to-frame modeling to temporal relations modeling, since a series of close elements has more informative structures than one single element. To learn more diverse temporal information, we propose the multi-scale temporal attention module, which separately captures various temporal structures through multiple attention branches with different convolution dilation rates. The results from multiple branches are subsequently merged to obtain comprehensive information integration. The CED architecture is adopted due to its efficiency in extracting hierarchical spectral features and producing a compact sequence for temporal modeling. Furthermore, a sub-band LSTMs module is introduced after the CED to remove redundant noises at the sub-band level. Our experimental results on DNS20 dataset show that DConvT consistently outperforms other advanced baselines, indicating the effectiveness of the proposed method.

The rest of this paper is structured as follows. In section II, we provide a detailed presentation of the proposed methodology. Section III presents information on experiments, including datasets, model and training configurations, and experimental results. Section IV concludes this paper.

II. Proposed Method

A. Problem formulation

The noisy speech signal in the STFT domain can be formulated as:

$$\mathbf{X}(t,f) = \mathbf{S}(t,f) + \mathbf{N}(t,f), \tag{1}$$

where $\mathbf{X}(t,f), \mathbf{S}(t,f)$ and $\mathbf{N}(t,f)$ denote the complex-valued time-frequency bin of noisy mixture, the clean speech received at the microphone and the noise, respectively, at time index t and frequency index f. The objective of the SE task is to separate $\mathbf{S}(t,f)$ apart from $\mathbf{X}(t,f)$.

B. Overview of the network architecture

Fig. 1 depicts the proposed DConvT network architecture. The goal of our network is to extract the clean speech from a noisy mixture $\mathbf{x}$ through its STFT representation $\mathbf{X} \in \mathbb{C}^{F\times T}$, where $F \times T$ denotes the spectrogram resolution. For phase reconstruction, DConvT estimates the complex ratio mask (CRM) [21] $\mathbf{M} \in \mathbb{C}^{F\times T}$. The estimated speech spectrogram $\hat{\mathbf{S}}$ is formulated as: $\hat{\mathbf{S}} = \mathbf{M} \odot \mathbf{X}$, here $\odot$ represents the element-wise complex multiplication operation. The mask estimation process is built on a CED architecture. Specifically, the encoder contains a series of encoder layers, which learn high-level spectral features and reduce the spectral dimension. Each encoder layer is comprised of a complex 2D convolution, a complex batch normalization, and a PReLU [22] activation function. Given a complex input $\mathbf{U} = \mathbf{U}_i + j\mathbf{U}_i$ and a complex kernel $\mathbf{L} = \mathbf{L}_i + j\mathbf{L}_i$, a complex 2D convolution can be formulated as:

$$\mathbf{U}_{out} = (\mathbf{U}_r * \mathbf{L}_r - \mathbf{U}_i * \mathbf{L}_i) + j(\mathbf{U}_r * \mathbf{L}_i + \mathbf{U}_i * \mathbf{L}_r) \tag{2}$$

The encoded representation is passed through the proposed multi-scale temporal attention module and then reconstructed back to the original resolution by the decoder, which is similar to the encoder with complex transposed 2D convolutions. For redundant noise removal, a sub-band LSTMs module is introduced to estimate the CRM.

C. Multi-scale temporal attention module

For temporal modeling, we propose the use of dilated convolutions as a novel query/key-value projection in our temporal self-attention blocks (TSA), as shown in Fig. 2. Particularly, the encoder output $\mathbf{F} \in \mathbb{R}^{C\times T}$ is projected to the query matrix $\mathbf{Q} \in \mathbb{R}^{H\times T}$, the key matrix $\mathbf{K} \in \mathbb{R}^{H\times T}$, and the value matrix $\mathbf{K} \in \mathbb{R}^{H\times T}$ as follows:

$$\mathbf{Q} = \mathbf{W}_Q * \mathbf{F}, \mathbf{K} = \mathbf{W}_K * \mathbf{F}, \mathbf{V} = \mathbf{W}_V * \mathbf{F} \tag{3}$$

where $\mathbf{W}_Q, \mathbf{W}_K$, and $\mathbf{W}_V$ are learnable weights of different dilated convolution projections (DConv). In traditional transformers, the query/key-value projection is a linear transformation of the input, which can be understood as a point-wise convolution operation (PConv). We extend the PConv to DConv

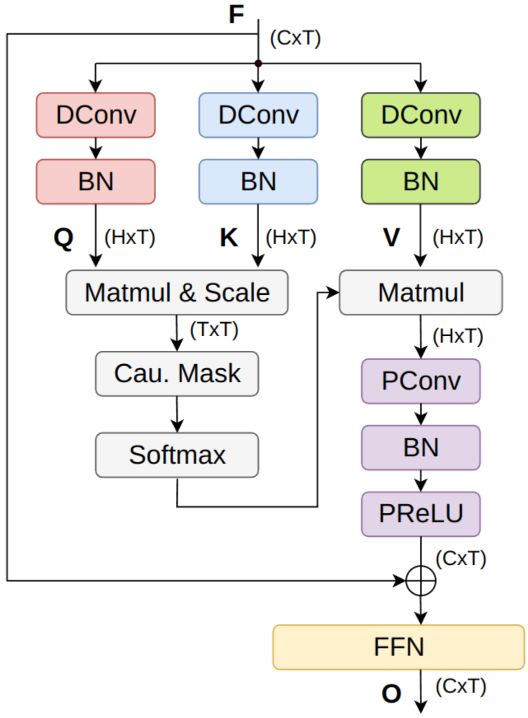

Fig. 2. Schematic diagram of the temporal self-attention block (TSA)

with a kernel of size k and a dilation rate $2^{i-1}, i \in \{1, 2, ..., d\}$. Such dilated convolution filters allow the query/key-value projection to aggregate structural information and temporal relations. Batch normalization is included after each DConv layer. Then, the attention score matrix $\mathbf{P} \in \mathbb{R}^{T \times T}$ is calculated through scaled dot-product as follows:

$$\mathbf{P} = Softmax(CausalMask(\frac{\mathbf{QK^T}}{\sqrt{H}})). \tag{4}$$

Here a triangular mask is applied to the scaled dot-product result to ensure the causality for real-time SE. The value matrix $\mathbf{V}$ is multiplied with the score matrix $\mathbf{P}$ to produce the attention result. After projecting the feature back to the original channel C, the TSA block output $\mathbf{O} \in \mathbb{R}^{C \times T}$ can be obtained through residual connection and a feed-forward network (FFN), which enriches the feature diversity [15].

To model different temporal structures, we introduce the multi-scale temporal attention module (MSTA), as shown in Fig. 1. Each branch of MSTA applies a different dilated convolution filter for the key/query-value projection to capture diverse structures in temporal representation. Given the output from the last TSA block of i-th branch is $\mathbf{O}_{i,m}$, the merge operation can be formulated as follows:

$$\mathbf{Y} = Linear(concat(\mathbf{O}_{i,m}, i \in \{1, 2, ..., d\})) \tag{5}$$

where $\mathbf{Y}$ is the temporal module output, which is subsequently fed into the decoder.

D. Sub-band LSTMs module

To reduce the redundant noise at sub-band level of the original spectrogram, we utilize the sequential bias advantage of LSTMs and introduce a sub-band LSTMs module. The module performs frequency-wise sequential modeling and produces the CRM as follows:

$$\mathbf{M} = MLP(LSTM(Norm(\mathbf{I}))) \tag{6}$$

where $\mathbf{I} \in \mathbb{R}^{C \times F \times T}$ is the decoder output and $LSTM$ is a two-layer LSTMs operating on sub-band time sequences. MLP is the final multilayer perceptron that produces the complex output mask $\mathbf{M} \in \mathbb{C}^{F \times T}$.

E. Joint loss function

We first adopt the power-law compressed spectrum loss function [23], which can be defined as:

$$\mathcal{L}_{spec} = (1-\alpha)\sum_{t,f}||S|^c - |\hat{S}|^c| + \alpha\sum_{t,f}||S|^c e^{j\varphi_S} - |\hat{S}|^c e^{j\varphi_S}|, \tag{7}$$

where S and $\hat{S}$ are the STFT representation of the clean and reference speech signals, respectively. We set the hyperparameters $\alpha = 0.3$ and $c = 0.3$, as investigated in previous studies [23].

The popular objective metric Scale-Invariant Signal-to-Noise Ratio (SI-SNR) [24] is also considered:

$$\begin{cases} s_{target} = \dfrac{\langle \hat{s}, s \rangle \cdot s}{\|s\|_2^2}; \\ e_{noise} = \hat{s} - s; \\ \mathcal{L}_{SI-SNR}(s, \hat{s}) = 10log_{10}(\dfrac{\|s_{target}\|_2^2}{\|e_{noise}\|_2^2}), \end{cases} \tag{8}$$

here $\hat{s}$ and s are the estimated and reference speech signals in the time-domain, respectively. $\|\cdot\|_2$ is Euclidean norm and $\langle \cdot, \cdot \rangle$ represents the dot product of two vectors.

The joint loss function utilized for the system is:

$$\mathcal{L}(s, \hat{s}) = \mathcal{L}_{Spec} + \mathcal{L}_{SI-SNR}(s, \hat{s}). \tag{9}$$

III. Experiments

A. Dataset

1) DNS 2020 Challenge dataset: We employ the DNS-2020 dataset [3], which is a popular large-scale dataset for monaural real-time SE. This dataset comprises 500 hours of clean voice recordings obtained from a total of 2150 speakers. The noise set includes around 180 hours of video from 150 different classes. For the training set, we generate around 1000 hours of 6-second noisy-clean pairs sampled at 16 kHz with Signal-to-Noise Ratio (SNR) varies randomly from -5 dB to 20 dB, followed by a 5-hour validation set. Finally, the system is evaluated on the non-blind synthetic test set provided by the challenge.

B. Experimental Configuration

A 32 ms Hanning window with 50% overlap and 512 STFT points is employed, yielding a total of 257 spectral dimensions.

TABLE I
OBJECTIVE RESULTS OF DIFFERENT SYSTEM

System	Param.(M)	PESQ$_{WB}$	PESQ$_{NB}$	STOI(%)	SI-SDR
Unprocessed	-	1.58	2.45	91.52	9.07
DConvT	11.88	**3.20**	**3.64**	**97.35**	**19.55**
w/o. Dilation	9.91	3.15	3.61	97.18	19.31
w/o. Multi-branch	4.15	3.15	3.60	97.12	19.26
w/o. MSTA	2.50	3.07	3.55	96.97	19.20
w/o. Sub-band LSTMs	11.80	3.10	3.58	96.98	18.30

TABLE II
COMPARISON WITH OTHER ADVANCED BASELINES ON THE DNS20 CHALLENGE NON-BLIND TEST SET

System	Year	Param.(M)	PESQ$_{WB}$	PESQ$_{NB}$	STOI(%)	SI-SDR
Unprocessed	-	-	1.58	2.45	91.52	9.07
DCCRN [7]	2020	3.67	-	3.27	-	-
PoCoNet [26]	2020	50	2.75	-	-	-
DCCRN+ [27]	2021	3.3	-	3.33	-	-
CTS-Net [9]	2021	4.35	2.94	3.42	96.66	17.99
FullSubNet+ [28]	2022	8.67	2.98	3.50	96.69	18.34
FS-CANet [29]	2022	4.21	3.02	3.51	96.74	18.08
FRNet [30]	2022	7.52	3.14	3.52	96.91	18.75
GaGNet [8]	2022	5.94	3.17	3.56	97.13	18.91
Inter-SubNet [31]	2023	2.29	3.00	3.50	96.61	18.05
CTFUNet [32]	2023	6.1	3.18	**3.64**	97.17	18.66
DConvT (ours)	2024	11.88	**3.20**	**3.64**	**97.35**	**19.55**

AdamW [25] optimization is adopted using an initial learning rate of 0.001. The learning rate is halved if the validation result does not increase after 2 consecutive epochs.

The encoder and decoder's channel numbers are $\{32, 64, 128, 128, 256, 256\}$. In both the encoder and decoder layers, the 2D complex convolution employs a kernel size of (5, 2) and a stride size of (2, 1). The channel C of the MSTA module is 256. The number of TA blocks in one branch is $m = 5$ and the number of parallel branches is $d = 4$. In a TSA block, the channel dimension is squeezed to $H = 64$ and the FFN for feature transformation is 2 consecutive feedforward networks with an inverted bottleneck factor of 2. The channel number of the sub-band LSTMs module is 64.

C. Evaluation Metrics

In this study, we adopt four popular speech evaluation metrics, including Perceptual Evaluation of Speech Quality (PESQ) [33], Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [24], Short-Time Objective Intelligibility (STOI) [34], and Extended Short-Time Objective Intelligibility (ESTOI) [35].

D. Experimental Results

1) Ablation study: We first conduct ablation experiments to examine the contribution of each sub-module to the overall performance of DConvT. The ablation results evaluated on DNS20 non-blind synthetic test set are shown in Table I. After replacing dilated convolution projection with point-wise convolutions (w/o. Dilation), the SE performance decreased in all four metrics (e.g. 3.20 to 3.15 in PESQ$_{WB}$). By changing the query/key-value projection of all branches to a standard linear transformation, the TSA structures in each branch become similar, resulting in the possibility of different branches having the same functionality. Interestingly, the performance of using one temporal attention branch (w/o. Multi-branch) is relatively similar to removing dilated convolutions, which can confirm the above assumption. Moreover, removing MSTA leads to considerable degradation in performance (e.g. 3.20 to 3.07 in PESQ$_{WB}$ and 3.64 to 3.55 in PESQ$_{NB}$), demonstrating the effectiveness of the proposed temporal modeling module. The sub-band LSTMs module further improves the overall performance.

2) Comparison with other advanced baselines: Table II shows the comparison results on the DNS20 challenge non-blind synthetic test set. One can observe that our proposed method significantly outperforms other state-of-the-art real-time SE networks. Compared to decoupled structures for separately modeling magnitude and complex spectrum such as GaGNet and FRNet, DConvT using only complex spectrum shows 0.06 and 0.03 improvements in PESQ$_{WB}$, respectively. With transformer-based temporal modeling in the bottleneck of CED, DConvT surpasses CTFUNet, a multi-resolution TCNs integrated with time-frequency self-attention model, in PESQ$_{WB}$, STOI, and SI-SDR metrics. In addition, we assess the real-time factor (RTF) of DConvT on an Intel Core i5 quad-core CPU with single thread. DConvT has an RTF of 0.46, which fully meets the DNS 2020 Challenge criteria for real-time processing. The computational complexity of DConvT is about 5.98G multiply-accumulate operations per second (MACs).

IV. Conclusion

In this paper, we propose a monaural speech enhancement model for real-time applications, named as DConvT. Inspired by the inductive bias of convolutions and the global modeling capability of transformers, we re-design the self-attention operation by improving the query/key-value projection with dilated convolutions. A multi-branch approach is employed to guide the network learn various acoustic structures through different dilation factors. The multi-scale information of temporal representation are then aggregated by a concat operation. While the multi-scale temporal attention module updates the frequency information globally, a sub-band LSTMs module is integrated after the CED for frequency-wise feature refinement. Experimental results conducted on DNS20 dataset show that DConvT not only outperforms other state-of-the-art approaches but also satisfies the real-time processing requirement, which demonstrates the superiority of our proposed approach. In the future, we will focus on optimizing the query/key-value projection filters and extending the network to wider-band spectrum modeling.

Acknowledgement

This research is funded by Hanoi University of Science and Technology under grant number T2023-PC-039.

References

[1] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Information Fusion*, p. 101869, 2023.

[2] C. Zheng, H. Zhang, W. Liu, X. Luo, A. Li, X. Li, and B. C. Moore, "Sixty years of frequency-domain monaural speech enhancement: From traditional to deep learning methods," *Trends in Hearing*, vol. 27, p. 23312165231209913, 2023.

[3] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results," in *Proc. Interspeech 2020*, 2020, pp. 2492–2496.

[4] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Proc. Interspeech 2017*, 2017, pp. 3642–3646.

[5] A. Défossez, G. Synnaeve, and Y. Adi, "Real Time Speech Enhancement in the Waveform Domain," in *Proc. Interspeech 2020*, 2020, pp. 3291–3295.

[6] A. Pandey and D. Wang, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6875–6879.

[7] 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," in *Proc. Interspeech 2020*, 2020, pp. 2472–2476.

[8] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Applied Acoustics*, vol. 187, p. 108499, 2022.

[9] A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1829–1843, 2021.

[10] S. Zhang, Y. Kong, S. Lv, Y. Hu, and L. Xie, "F-T-LSTM Based Complex Network for Joint Acoustic Echo Cancellation and Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 4758–4762.

[11] S. Zhao, B. Ma, K. N. Watcharasupat, and W.-S. Gan, "Frcrn: Boosting feature representation using frequency recurrence for monaural speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9281–9285.

[12] S. Lv, Y. Fu, M. Xing, J. Sun, L. Xie, J. Huang, Y. Wang, and T. Yu, "S-dccrn: Super wide band dccrn with learnable complex feature for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7767–7771.

[13] X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-Path Convolution Recurrent Network for Single Channel Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 2811–2815.

[14] 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.

[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] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, A. Ju, M. Zohourian, M. Tang, M. Golestaneh *et al.*, "Icassp 2023 deep noise suppression challenge," *IEEE Open Journal of Signal Processing*, 2024.

[17] 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.

[18] S. Zhao and B. Ma, "D2former: A fully complex dual-path dual-decoder conformer network using joint complex masking and complex spectral mapping for monaural speech enhancement," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[19] G. Yu, A. Li, C. Zheng, Y. Guo, Y. Wang, and H. Wang, "Dual-branch attention-in-attention transformer for single-channel speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7847–7851.

[20] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech." in *SSW*, 2016, pp. 146–152.

[21] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[23] S. Braun and I. Tashev, "Data augmentation and loss normalization for deep noise suppression," in *International Conference on Speech and Computer*. Springer, 2020, pp. 79–86.

[24] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[26] U. Isik, R. Giri, N. Phansalkar, J.-M. Valin, K. Helwani, and A. Krishnaswamy, "Poconet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss," *arXiv preprint arXiv:2008.04470*, 2020.

[27] S. Lv, Y. Hu, S. Zhang, and L. Xie, "Dccrn+: Channel-wise subband dccrn with snr estimation for speech enhancement," *arXiv preprint arXiv:2106.08672*, 2021.

[28] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, "Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7857–7861.

[29] J. Chen, W. Rao, Z. Wang, Z. Wu, Y. Wang, T. Yu, S. Shang, and H. Meng, "Speech Enhancement with Fullband-Subband Cross-Attention Network," in *Proc. Interspeech 2022*, 2022, pp. 976–980.

[30] A. Li, C. Zheng, G. Yu, J. Cai, and X. Li, "Filtering and refining: A collaborative-style framework for single-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2156–2172, 2022.

[31] J. Chen, W. Rao, Z. Wang, J. Lin, Z. Wu, Y. Wang, S. Shang, and H. Meng, "Inter-subnet: Speech enhancement with subband interaction," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[32] S. Xu, Z. Zhang, and M. Wang, "Channel and temporal-frequency attention unet for monaural speech enhancement," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, p. 30, 2023.

[33] 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 *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[34] 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.

[35] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.