



MUSE: Flexible Voiceprint Receptive Fields and Multi-Path Fusion Enhanced Taylor Transformer for U-Net-based Speech Enhancement

Zizhen Lin^{1,*}, Xiaoting Chen^{2,*}, Junyu Wang^{1,*}

¹School of Electronic Information, Sichuan University, China

²School of Software, Yunnan University, China

*Authors contributed equally to this work.

linzizhen17@163.com, 12022219155@mail.ynu.edu.cn, junyu.wang@stu.scu.edu.cn

Abstract

Achieving a balance between lightweight design and high performance remains a challenging task for speech enhancement. In this paper, we introduce Multi-path Enhanced Taylor (MET) Transformer based U-Net for Speech Enhancement (MUSE), a lightweight speech enhancement network built upon the U-Net architecture. Our approach incorporates a novel Multi-path Enhanced Taylor (MET) Transformer block, which integrates Deformable Embedding (DE) to enable flexible receptive fields for voiceprints. The MET Transformer is uniquely designed to fuse Channel and Spatial Attention (CSA) branches, facilitating channel information exchange and addressing spatial attention deficits within the Taylor-Transformer framework. Through extensive experiments conducted on the VoiceBank+DEMAND dataset, we demonstrate that MUSE achieves competitive performance while significantly reducing both training and deployment costs, boasting a mere 0.51M parameters.

Index Terms: speech enhancement, multi-path enhanced taylor transformer, simplified channel attention, deformable embedding, u-net

1. Introduction

Speech enhancement algorithms, also known as speech denoising algorithms, constitute a pivotal task within the realm of speech processing. Their applications extend to improving the quality of recorded audio, enhancing call quality, and augmenting the accuracy of speech recognition systems. In recent years, the flourishing landscape of deep learning has given rise to numerous advanced algorithms in the field of speech enhancement [1, 2, 3, 4]. Noteworthy is the fact that these deep learning algorithms exhibit a remarkable capability to suppress complex and non-stationary noise.

Recently, researchers have explored diverse approaches in the realm of speech enhancement, various domains such as the time domain [5, 6, 7, 8], time-frequency domain [9, 10, 11, 12]. T-F methods apply short-time Fourier transform (STFT), which transform speech signal into time-frequency spectrum. In pursuit of a more comprehensive extraction of magnitude and phase features from speech signals, a variety of multi-domain fusion methods have been applied to Speech Enhancement (SE), by extract feature of magnitude and phase separately [11].

For previous methods, a higher number of channels (64) is often required to achieve peak performance. The popular Two-Stage (TS) [7] structural network in speech enhancement demonstrates strong competitiveness. The TS structure typically performs downsampling only once along the frequency dimension due to the large size of the feature maps. This implies that such structures require greater memory and computational resources, particularly as the length of the speech increases, ne-

cessitating significant GPU memory during both training and inference stages. Training or deploying such speech enhancement models on devices with limited GPU memory presents evident challenges. To mitigate GPU memory consumption for the same length of speech, the simplest approach is to reduce the number of channels to decrease model size. However, for speech enhancement tasks, the model needs to map speech features to a high-dimensional space; insufficient channel numbers often result in decreased ability to discriminate noise features from speech features, potentially leading to significant performance degradation. We retrained the state-of-the-art (SOTA) model MP-SEnet [11] with 16 channels as a baseline, subsequently referred to as MP-SEnet-16. Experimental results indicate that while MP-SEnet-16 significantly reduces GPU memory requirements and parameter count compared to its 64 channels versions, the model's performance also significantly deteriorates. Therefore, for scenarios with low channel counts, we consider continuing to use the TS structure as suboptimal. Further more, for voiceprint information in speech, the feature shape often resembles an elongated crescent, contrasting with normal conv kernels that are fixed in a square shape. Hence, learning features from those specific shaped voiceprint becomes challenging and inefficient.

We believe that the reasons limiting the performance boundaries of small-sized speech enhancement systems can be summarized as follows:

- We posit that the TS-Conformer structure is suboptimal for light weight speech enhancement networks, as the serial architecture fails to establish a robust information flow and hinders the effective integration of low-level and high-level features across different layers when network is shallow. Consequently, this limitation results in inadequate representational capacity of the network. Furthermore, the elevated computational costs impose constraints on the quantity of conformer blocks that can be feasibly employed.
- For idiosyncratic voiceprints, standard convolutions exhibit limited adaptability in their receptive fields, rendering them inefficient in capturing distinctive features inherent in the vocal characteristics.
- While vanilla Transformer-based approaches, such as multi-head self-attention (MSA) [13] have demonstrated improved performance and the ability to capture intricate features, a critical consideration arises regarding the trade off between performance and computational cost. Employing a more efficient MSA mechanism and superior model architecture can be considered.

In this study, we introduce a novel U-Net based model MUSE for light weight speech enhancement. To ensure a resilient flow of information and impede the efficient integration of features

spanning both low-level and high-level across various layers, the TS-Conformer [11] was omitted in favor of a U-Net [14] architecture that incorporates deformable convolutions. To effectively and flexibly learn voiceprint feature and high-frequency information, concomitantly facilitating inter-channel information exchange, we propose a novel Multi path Enhanced Taylor (MET) attention mechanism. We substitute Taylor-Multi-head Self Attention (T-MSA) for MSA, significantly reducing computational complexity while capturing both global and high-frequency features of the speech signal. To address the inherent insensitivity of T-MSA to channel information [15], we introduce a streamlined channel and spatial attention branch called CSA. The features from these three branches are integrated to form the MET attention, which serves as the core module of the backbone network. Ultimately, amplitude and phase spectra are independently decoded, and the enhanced speech signal is reconstructed through Inverse Short-Time Fourier Transform (ISTFT).

Our model demonstrates remarkable effectiveness. Ultimately, through empirical validation on the VoiceBank-Demand dataset, MUSE achieved outstanding performance with extremely low parameter count of 0.51M and remarkably high utilization of memory (A 8GB GPU is sufficient for training).

2. Proposed method

In this section, we provide a comprehensive exposition of MUSE, which is founded upon a U-Net architectural framework, comprising dense convolution codec, deformable embedding, and MET-Transformer.

2.1. Model architecture

we subject the input signal y to Short-Time Fourier Transform (STFT), yielding the magnitude spectrum $\mathbf{Y}_m \in \mathbb{R}^{T \times F}$ and the phase spectrum $\mathbf{Y}_p \in \mathbb{R}^{T \times F}$. Following the methodology of PHASEN [16], a power-law compression is applied to \mathbf{Y}_m , resulting in the compressed magnitude spectrum $\mathbf{Y}_m^c \in \mathbb{R}^{T \times F}$. The compressed \mathbf{Y}_m^c is concatenated with \mathbf{Y}_p to form $\mathbf{Y}_{in}^c \in \mathbb{R}^{T \times F \times 2}$, which is then fed into a U-Net architecture incorporating the MET-Transformer. For the ultimate decoding layer, independent feature extraction is performed on the magnitude spectra $\hat{\mathbf{X}}_m \in \mathbb{R}^{T \times F}$ and phase spectra $\hat{\mathbf{X}}_p \in \mathbb{R}^{T \times F}$. These individual features are subsequently fused into a spectrogram, and the signal is reconstructed through Inverse Short-Time Fourier Transform (ISTFT).

Following the input encoder layer, the MET encoder-decoder is introduced, with subsequent layers incorporating both upsampling and downsampling operations. Each stage comprises deformable embedding (DE) and MET-Transformer blocks. The DE yields tokens of varying scales for input into the Transformer. Channel expansion $1d, 2d, 3d$ is achieved for both encoders and decoders through downsampling and upsampling modules. In the ultimate layer, analogous to the structure of the initial encoder layer, independent learning is performed for the magnitude and phase spectra.

2.2. Deformable Embedding

In the realm of acoustic signal processing, sound frequencies and features exhibit significant variability. Voiceprints typically possess unique shapes, thus, convolutional receptive fields may capture specific voiceprint features inadequately. In pursuit of a more adaptable receptive field configuration, akin to strategies employed in the field of image processing, we introduce

depthwise separable and deformable convolutions (DSDCN) [15]. Through the application of depthwise convolution and pointwise convolution, we mitigated computational complexity and reduced parameter count, consequently enhancing the efficiency of speech signal processing. We have incorporated Deformable Embedding at the head of each U-Net encoder-decoder, as illustrated by the MET Transformer block in Figure 2. DSDCN concurrently possesses the capability to learn intricate details of voiceprint information at fine scales and large-scale voiceprint features. This implies that DSDCN not only has the capability to capture the shape of voiceprint features but also exhibits the ability to expand the receptive field, similar to the behavior of dilated convolutions. Hardswish activation is employed to harness heightened non-linear characteristics, enabling the network to extract more intricate acoustic features. The application of DSDCN is avoided in the input encoder and magnitude-phase decoder of the network. This strategic exclusion arises from the inherent flexibility of deformable convolutions in achieving adaptable receptive fields through feature offsets, as the use of DCN across the full spectrum of the input and output codec stages is not efficient.

2.3. MET-Transformer block

In order to comprehensively learn the entire spectral characteristics of the speech signal, we prioritized efficiency in the design of the new Multi-path Enhanced Taylor (MET) Transformer block. MET is composed of three paths: The first path incorporates Taylor multi-head attention (T-MSA). In comparison to MSA, T-MSA attenuates attention across channels, emphasizing global self-attention, and excelling in handling high-frequency information. The second path, comprised of pooling and convolution linearly, is designed to compensate for channel information exchange in T-MSA. The third path, a combination of pointwise and depthwise convolutions, is responsible for learning spatially-invariant features and maintaining the stability of information flow. In order to streamline the description, we designate the second and third paths as the Channel and Spatial Attention (CSA) branch. Inspired by the 'Simple Gate' in NAFnet [17], we perform element-wise multiplication of these three branches, imparting a certain degree of non-linearity to our module. In a manner akin to the structure of the Transformer, we sum the MET with the residual connection [18], subsequently linearly connecting it to the FFN module after layer normalization.

T-MSA Branch. For vanilla Transformer, we have the following formula:

$$V' = \text{Softmax} \left(\frac{Q^T K}{\sqrt{D}} \right) V^T \quad (1)$$

As the computation of softmax involves the natural logarithm of the exponentiated values e^x :

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_{i=1}^n e^{x_i}} \in (0, 1) \quad (2)$$

To approximate (1) We can write a generalized function as:

$$V'_i = \frac{\sum_{j=1}^N f(Q_i, K_j) V_j}{\sum_{j=1}^N f(Q_i, K_j)} \quad (3)$$

where the matrix with ' i ' as the subscript represents the vector of the i th row of a given matrix, and ' $f(\cdot)$ ' denotes any similarity function, Equation (3) reduces to Equation (1) by

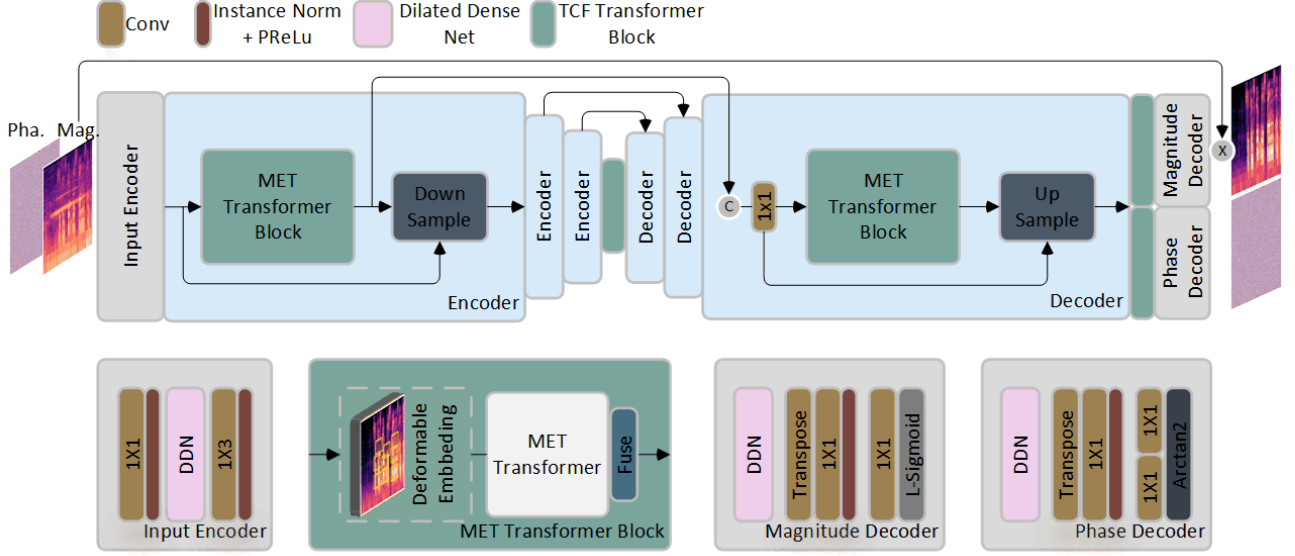


Figure 1: The overall architecture of MUSE.

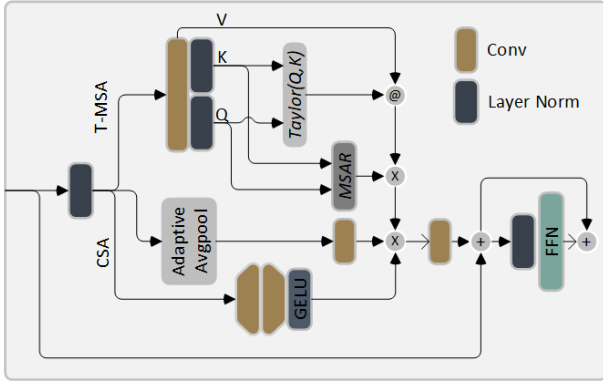


Figure 2: MET Transformer.

setting $f(Q_i, K_j) = \exp\left(\frac{Q_i^T K_j}{\sqrt{D}}\right)$. To elucidate this transition, the application of the Taylor formula [19] for a first-order expansion at 0 allows us to write $Taylor(Q_i, K_j) = 1 + \frac{Q_i^T K_j}{\sqrt{D}} + o\left(\frac{Q_i^T K_j}{\sqrt{D}}\right) \approx \exp\left(\frac{Q_i^T K_j}{\sqrt{D}}\right)$ and reformulate Equation (3) as follows:

$$V'_i = \frac{\sum_{j=1}^N (1 + Q_i^T K_j + o(Q_i^T K_j)) V_j^T}{\sum_{j=1}^N (1 + Q_i^T K_j + o(Q_i^T K_j))} \quad (4)$$

The computational complexity pertaining to both the MSA and the T-MSA, with respect to an spectrum comprising $t \times f$ patches, is explicated as follows:

$$\Omega(\text{MSA}) = 4tfD^2 + 2t^2f^2D \quad (5)$$

$$\Omega(\text{T-MSA}) = 18tfD + 2tfD^2 \quad (6)$$

Generally speaking, D is significantly smaller than tf . From this, we can conclude that T-MSA can significantly reduce computational complexity while approaching MSA representation, especially as tf increases, the reduction in complexity becomes more pronounced.

CSA Branch. Differing from MSA, T-MSA accentuates global attention while attenuating channel attention [15]. We designed channel and spatial attention (CSA) branch. To guide the network towards inter-channel feature relationships and emphasize crucial channel information, we employ adaptive average pooling to calculate weights between channels [20]. In an effort to reduce network complexity and parameter count, we adopt a strategy inspired by NAFnet [17], retaining only the essential components of channel attention and directly employing 1×1 convolution operations for inter-channel information exchange. Notably, the element-wise multiplication operation of 3 branches introduces nonlinearity to MET like Simple Gate [17]. Within the spatial attention [21] branch, we directly utilize pointwise and 3×3 depthwise convolutions [22] with GELU activation, incurring minimal computational cost and maintaining a low parameter count. This approach captures spatial information at different scales, aiding the network in focusing on crucial regions in both magnitude and phase spectra [16].

2.4. Dense Convolution Codec

In the input-output encoder, we borrowed the Codec from MP-SNet [11]. The Dilated Dense-Net structure comprises four dilated convolutional blocks with dense residual connections [23], each block having dilation factors set to 1, 2, 4, 8. In dilated convolutions, our primary objective is to augment the receptive field while preserving the kernel size and layer counts, ensuring the efficient capture of globally and intricately featured information.

3. EXPERIMENTS

3.1. Datasets

In our study, we employed the VoiceBank+DEMAND [29] dataset, which provides a comprehensive collection of high-fidelity utterances, both clean and mixed. The training set consists of 11,572 utterances, totaling 9.4 hours, delivered by 28 distinct speakers. Conversely, the test set comprises 824 utterances, amounting to 0.6 hours, articulated by 2 speakers not

Table 1: Comparison with other methods on VoiceBank+DEMAND dataset. “-” denotes the result is not provided in the original paper.

Method	Architecture	Parameters	PESQ	CSIG	CBAK	COVL	STOI
Noisy	-	-	1.97	3.35	2.44	2.63	0.91
SEGAN [5]	U-Net	43.18M	2.16	3.48	2.94	2.80	0.92
DEMUCS [8]	U-Net	33.53M	3.07	4.31	3.40	3.63	0.95
SE-Conformer [24]	U-Net	-	3.13	4.45	3.55	3.82	0.95
MetricGAN+ [25]	LSTM	-	3.15	4.14	3.16	3.64	-
TSTNN [7]	Two-Stage Transformer	0.92M	2.96	4.33	3.53	3.67	0.95
DB-AIAT [10]	Two-Stage Transformer	2.81M	3.31	4.61	3.75	3.96	-
DPT-FSNet [26]	Two-Stage Transformer	0.88M	3.33	4.58	3.72	4.00	0.96
PHASEN [16]	Two-Stream DNN	20.9M	2.99	4.18	3.45	3.50	0.95
MetricGAN-OKDv2 [27]	LSTM	0.82M	3.12	4.27	3.16	3.71	0.95
MANNER [6]	U-Net	24.07M	3.21	4.53	3.65	3.91	0.95
MANNER-S-5.3GF [28]	U-Net	0.90M	3.06	4.42	3.58	3.77	0.95
MUSE(Ours)	U-Net	0.51M	3.37	4.63	3.80	4.10	0.95

represented in the training data. Notably, the noise profiles in the test set, derived from recorded DEMAND datasets and synthetic sources, diverge from those present in the training set. Signal-to-noise ratios vary between the datasets, with the test set featuring ratios of 0dB, 5dB, 10dB, and 15dB, while the training set encompasses ratios of 2.5dB, 7.5dB, 12.5dB, and 17.5dB.

3.2. Model setup

Throughout the training process, speech data was uniformly segmented into 30700 points, employing an FFT size of 510, a window length of 510, a hop length of 100, and a sampling rate of 16000¹. The training configuration encompassed a batch size of 2, a learning rate set to 0.0005 with a decay factor of 0.99, dense channels is 16, all the models were trained using the AdamW [30] optimizer until 100 epochs. During training, a single 8GB RTX 3070ti GPU was utilized.

3.3. Evaluation metrics

We have selected a suite of widely accepted metrics to assess the quality of denoised speech. These metrics include the Perceptual Evaluation of Speech Quality (PESQ) [31], which provides scores ranging from -0.5 to 4.5, the Segmental Signal-to-Noise Ratio (SSNR), and Composite Mean Opinion Score (MOS) metrics as outlined in literature. The MOS metrics encompass the MOS prediction of signal distortion (CSIG), MOS prediction of background noise intrusiveness (CBAK), and MOS prediction of overall effect (COVL), each scored within a range of 1 to 5. Furthermore, Speech Transmission Index (STOI) [32] is employed, offering scores from 0 to 1 to evaluate speech intelligibility, where higher values indicate superior performance across all metrics under consideration.

3.4. Result

Comparative analysis of objective metrics on the VoiceBank+DEMAND dataset, as presented in Table 1. Due to MUSE is based on the U-Net, we selected 5 U-Net based method including SEGAN, DEMUCS, SE-Conformer, MANNER, and MANNER-S-5.3GF. Furthermore, we specifically focused on selecting lightweight models with a parameter count less than 1M for comparison, including TSTNN, DPT-FSNet,

MetricGAN-OKDv2, and MANNER-S-5.3GF. Even with the inherent high parameter count of the U-Net architecture, MUSE achieved a parameter count of 0.51M, which is lower than all baseline models. Additionally, apart from slightly lower performance in STOI compared to DPT-FSNet, MUSE outperforms baseline models in terms of PESQ, CSIG, CBAK, and COVL metrics.

Table 2: **Ablation Study:** CSA means Channel and Spatial Attention; T-MSA means Taylor-Multi-head Self Attention; DE means Deformable Embedding. We retrained the MP-SEnet with 16 dense channels as our baseline namely MP-SEnet-16.

Model	PESQ	CSIG	CBAK	COVL
MUSE	3.37	4.63	3.80	4.10
w/o CSA	3.29	4.62	3.76	4.04
w/o T-MSA	3.27	4.62	3.76	4.02
w/o DE	3.34	4.63	3.78	4.08
MP-SEnet-16 (baseline)	3.21	4.54	3.72	3.95

Ablation study compared the scores of multiple objective metrics on the VoiceBank+DEMAND dataset, as shown in Table 2. “w/o” denotes the removal of a certain component. We experimented on the major breakthroughs of MUSE, including the CSA branch, T-MSA branch, and Deformable Embedding in our proposed MET Transformer. MP-SEnet-16 represents a reproduced result of MP-SEnet [11] with only 16 dense channels, which serves as our original baseline model.

4. Conclusions

In this paper, we proposed a lightweight speech enhancement network named MUSE, achieved competitive performance with only 0.51M parameters. We firstly apply Deformable Embedding to flexibly adapt the shape of voiceprint. We propose a novel multi path enhanced Taylor-Transformer (MET) Transformer, utilizing pooling and convolutional branches to achieve channel information exchange and spatial attention missing in light weighted Taylor-Transformer. MUSE reduces training and deployment costs greatly. In the future, we will explore how to make MUSE into a real-time speech enhancement system, enabling it to have a broader range of applications.

¹<https://github.com/huaidanquede/MUSE-Speech-Enhancement>

5. References

- [1] A. E. Bulut and K. Koishida, "Low-latency single channel speech enhancement using u-net convolutional neural networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6214–6218.
- [2] D. N. Tran and K. Koishida, "Single-channel speech enhancement by subspace affinity minimization," in *INTERSPEECH*, 2020, pp. 2447–2451.
- [3] S. Wisdom, J. R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, and R. A. Saurous, "Differentiable consistency constraints for improved deep speech enhancement," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 900–904.
- [4] E. Moliner and V. Välimäki, "A two-stage u-net for high-fidelity denoising of historical recordings," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 841–845.
- [5] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *Interspeech 2017*, 2017.
- [6] H. J. Park, B. H. Kang, W. Shin, J. S. Kim, and S. W. Han, "Manner: Multi-view attention network for noise erasure," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7842–7846.
- [7] 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.
- [8] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Demucs: Deep extractor for music sources with extra unlabeled data remixed," *arXiv preprint arXiv:1909.01174*, 2019.
- [9] J. Wang, "Efficient Encoder-Decoder and Dual-Path Conformer for Comprehensive Feature Learning in Speech Enhancement," in *Proc. INTERSPEECH 2023*, 2023, pp. 2853–2857.
- [10] G. Yu *et al.*, "Dual-branch attention-in-attention transformer for single-channel speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022, pp. 7847–7851.
- [11] Y.-X. Lu, Y. Ai, and Z.-H. Ling, "MP-SENet: A Speech Enhancement Model with Parallel Denoising of Magnitude and Phase Spectra," in *Proc. INTERSPEECH 2023*, 2023, pp. 3834–3838.
- [12] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based Metric GAN for Speech Enhancement," in *Proc. Interspeech 2022*, 2022, pp. 936–940.
- [13] Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 181–185.
- [14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.
- [15] Y. Qiu, K. Zhang, C. Wang, W. Luo, H. Li, and Z. Jin, "Mbtaylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 802–12 813.
- [16] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9458–9465.
- [17] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *European Conference on Computer Vision*. Springer, 2022, pp. 17–33.
- [18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [19] B. Taylor, *Methodus incrementorum directa & inversa*. Inny, 1717.
- [20] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 534–11 542.
- [21] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, "An empirical study of spatial attention mechanisms in deep networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6688–6697.
- [22] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
- [23] A. Pandey and D. L. Wang, "Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 6629–6633.
- [24] E. Kim and H. Seo, "Se-conformer: Time-domain speech enhancement using conformer," in *Interspeech*, 2021, pp. 2736–2740.
- [25] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "Metricgan+: An improved version of metricgan for speech enhancement," *arXiv e-prints*, pp. arXiv–2104, 2021.
- [26] F. Dang, H. Chen, and P. Zhang, "Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022, pp. 6857–6861.
- [27] W. Shin, B. H. Lee, J. S. Kim, H. J. Park, and S. W. Han, "Metricgan-okd: Multi-metric optimization of metricgan via on-line knowledge distillation for speech enhancement," 2023.
- [28] W. Shin, H. J. Park, J. S. Kim, B. H. Lee, and S. W. Han, "Multi-view attention transfer for efficient speech enhancement," *arXiv preprint arXiv:2208.10367*, 2022.
- [29] 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.
- [30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [31] 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.
- [32] 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.