# TSMGAN-II: Generative Adversarial Network Based on Two-Stage Mask Transformer and Information Interaction for Speech Enhancement

Lianxin Lin[1[0000-0003-1906-8470]], Yaowen Li[1[0009-0003-3335-3398]], and Haizhou Wang[1,✉ [0000-0003-1197-5906]]

[1] School of Cyber Science and Engineering, Sichuan University, Chengdu, 610207, China
{linlianxin,liyaowen}@stu.scu.edu.cn, whzh.nc@scu.edu.cn

**Abstract.** Speech Enhancement is significantly applied in speech processing, as a foundation for downstream tasks. Nowadays, neural networks are well applied in speech enhancement. However, there remain considerable difficulties for neural networks to improve speech quality. Firstly, existing methods have the problem of speech over-suppression. Because they have not yet taken into account that neural networks influence not only background noise but also clean speech during enhancement. This issue can negatively impact the following tasks. Secondly, striking a balance between model complexity and performance is crucial, especially when deploying on resource-constrained hardware. Existing models often prioritize performance, overlooking the issue of complexity. To solve the problems above, we propose a novel Generative Adversarial Network based on Two-Stage Mask Transformer and Information Interaction (TSMGAN-II), consisting of an attention encoder, a two-stage mask transformer, and a dual-feature decoder with information interaction. It effectively captures and models both amplitude and spectral characteristics within the time-frequency domain. Experiments on the VoiceBank+DEMAND dataset show that our model, with 1.39 million parameters, achieves state-of-the-art performance with PESQ of 3.40 and SSNR of 11.81. Moreover, we also introduce a lightweight model with just 0.59M parameters, achieving 97% of the performance of SOTA models with PESQ of 3.31 and SSNR of 11.53.

**Keywords:** Speech enhancement, generative adversarial network, attention encoder, mask formerblock, information interaction.

## 1 Introduction

### 1.1 Background

In practice, factors such as environmental noise often reduce speech clarity. Hence, speech enhancement aims to filter out such noise for clear, pristine speech signals, aiding future tasks [1]. Neural networks for speech enhancement have become popular compared to traditional static signal processing methods due to their effective capture of the dynamic structure within speech [2].

With the development of information technology, speech has evolved from just a means of human communication to serve as a pivotal input for different devices. The number of applications requiring speech inputs, like speech recognition, has significantly increased. However, noise interference often degrades speech signals, thereby diminishing the performance of these applications [3]. For instance, noise can lead to inaccuracies in the output from speech recognition systems. Addressing this issue demands more than noise elimination, as neural network methods often fail to distinguish between clean speech and background noise, thereby unnecessarily suppressing clean speech as well. Neural networks often confuse clean speech with noise, suppressing both. Thus, understanding speech enhancement requires both noise elimination and clean speech preservation, an aspect often overlooked in current methods.

Moreover, the range of devices making use of enhanced speech is broad and diverse, including smart speakers for voice-controlled home appliances, smartphones integrated with artificial intelligence, etc. These devices utilize enhanced speech to improve their functionality, becoming integral to our daily lives. However, due to cost considerations, device manufacturers often equip these devices with limited hardware capabilities. As a result, most devices fail to utilize advanced neural network methods for speech enhancement fully because these techniques mostly require substantial computational resources [4].

### 1.2 Challenges

In summary, in the field of speech enhancement, the following problems still exist:

Firstly, existing methods tend to over-suppress speech during enhancement, inadvertently reducing noise and target speech. Even though the mild over-suppression may not affect intelligibility or metrics like PESQ, it significantly influences metrics like SSNR and the performance of downstream applications. In brief, addressing over-suppression issues can enhance downstream application performance.

Secondly, the growing model overhead poses a significant challenge. Due to resource constraints, many models aren't feasible for use in real-world applications like voice-controlled devices, limiting speech enhancement adoption and requiring a balance between performance and overhead.

### 1.3 Contributions

As for the above challenges, this paper proposes a new method based on a two-stage mask transformer and information interaction in a generative adversarial network. The following are the contributions of our work:

1. We propose a novel model TSMGAN-II for speech enhancement, which employs our proposed mask formerblock, convolutional attention, and information interaction mechanism. The evaluation results show that it significantly outperforms all approaches of speech enhancement with PESQ of 3.40 and SSNR of 11.81.

2. We propose an innovative lightweight model with only 0.58M parameters based on TSMGAN-II. Through altering of the encoder, decoder and interaction information mechanism, we significantly reduce model complexity, which achieves performance

closer to SOTA with PESQ of 3.31 and SSNR of 11.53, maintaining the lowest complexity among existing methods.

3. We investigate the structures of the encoder and decoder of speech signals and explore the information interaction during speech enhancement. We also improve the transformer by integrating our mask formerblock. We demonstrate the effectiveness of our proposed model to solve the above problems through comprehensive ablation experiments and others. Additionally, we have made our code[1] publicly available.

## 2 Proposed Method

### 2.1 Method Description

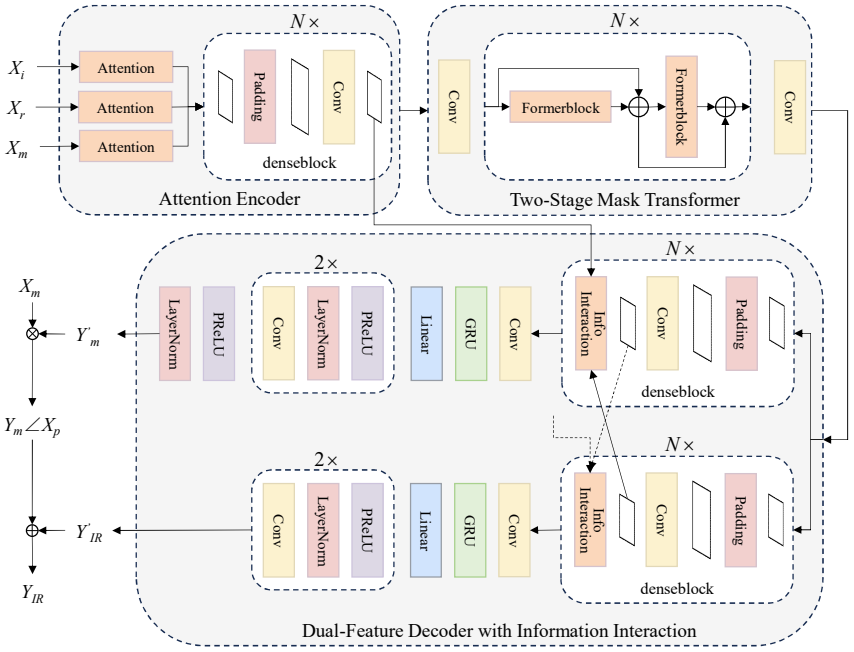


**Fig. 1.** The overall architecture of generator

In our method, the process of generating target speech is:

$$X_o = STFT\left(x(t)\right) \tag{1}$$

$$X = \left|X_o\right|^c e^{jX_p} = X_m e^{jX_p} = X_r + jX_i \tag{2}$$

$$Y = Generator\left(X_m, X_r, X_i\right) \tag{3}$$

$$Y_o = \left|Y\right|^{1/c} e^{jY_p} = Y_m e^{jY_p} = Y_r + jY_i \tag{4}$$

[1] https://github.com/yiyepianzhounc/TSMGAN-II

$$y(t) = iSTFT(Y_o) \tag{5}$$

where $x(t)$, $X_p$, $X_m$, $X_r$, $X_i$ represent the input speech, phase, amplitude, real part and imaginary part of the compressed spectrogram respectively, and Eq. (2) and Eq. (4) represent the power-law compression and decompression.

### 2.2 Generator

We use a generative adversarial network for speech enhancement, shown in Fig. 1. It consists of an attention encoder, a two-stage mask transformer and a dual-feature decoder with information interaction.

Audio features $X_m$, $X_r$, $X_i$ in the time-frequency domain feed the generator network. The dual-feature decoder yields the amplitude mask $Y_m^{'}$ and complex spectrogram $Y_{IR}^{'}$, deriving the enhanced complex spectrogram with the formula [5]:

$$Y_{IR} = Y_m \angle X_p + Y_{IR}^{'} = X_m Y_m^{'} \angle X_p + Y_{IR}^{'} \tag{6}$$

### 2.3 Attention Encoder

The encoder consists of multi-scale convolution attention blocks and a DenseNet.

As shown in Fig. 2, the convolution attention block with different scales (kernel sizes of $[k_s, k_m, k_b]$) captures and extracts short-, mid- and long-term features of $X_m$, $X_r$, $X_i$, each with unique parameters for multi-scale feature extraction.

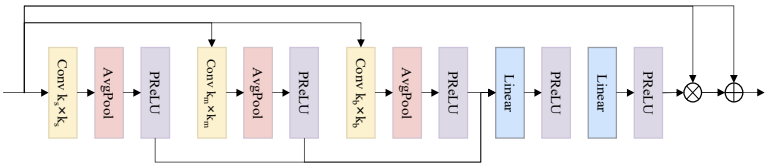


**Fig. 2.** The overall architecture of the convolution attention block

The purified outputs from the convolutional attention blocks are merged to obtain $X \in R^{B\times 3\times T\times F}$. An 1×1 convolution increases channels to $C$, forming the input $X \in R^{B\times C\times T\times F}$ for DenseNet [6] composed of four dense blocks. In addition, the output of each dense block will be saved for processing by the dual-feature decoder.

### 2.4 Two-Stage Mask Transformer

There are 2D convolution layers preceding and following the two-stage mask transformer to halve and restore the input features' channel respectively. Inspired by TSTNN [7], we use a dual-path transformer with our proposed mask formerblock to handle temporal and frequency features, that is, separately processing $X_t \in R^{B\times C/2\times T\times F}$ and $X_f \in R^{B\times C/2\times F\times T}$. As shown in Fig. 3, the structure of our mask formerblock uses a GRU, a GELU and an FC layer for feature extraction along the corresponding

dimension and employs a mask from the multi-head attention layer to eliminate inconsistent speech features.

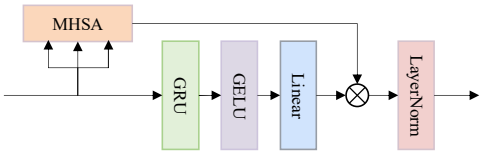


**Fig. 3.** The overall architecture of the formerblock based on multi-head attention

## 2.5 Dual-Feature Decoder with Information Interaction

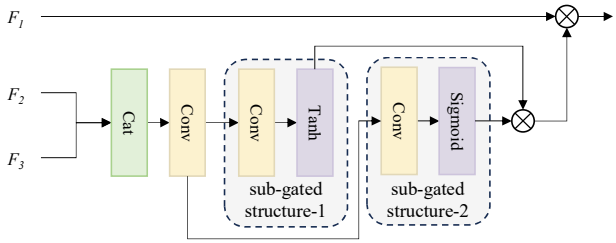


**Fig. 4.** The overall architecture of the gated structure in information interaction module

The dual-feature decoder, primarily comprising an amplitude and a complex spectrum decoder, outputs predicted amplitude mask and complex spectrum respectively. Each decoder includes a DenseNet, a gated structure for information interaction shown in Fig. 4 and an output decoding layer. Within the DenseNet, the gated structure generates an output mask through two sub-gated structures by using the merged outputs from another decoder and the encoder's related dense block, with logic shown below:

$$Gate1 = Tanh\left(Conv\left(MergedVector\right)\right) \tag{7}$$

$$Gate2 = Sigmoid\left(Conv\left(MergedVector\right)\right) \tag{8}$$

$$X_F^{'} = X_F \times Mask \ = X_F \times Gate1 \times Gate2 \tag{9}$$

After the DenseNet, using the output decoding layer, the amplitude decoder outputs the amplitude mask $Y_m^{'} \in R^{B\times 1\times T\times F}$, and the complex spectrum decoder outputs the complex spectrum $Y_{IR}^{'} \in R^{B\times 2\times T\times F}$.

## 2.6 Lightweight Model

To minimize the model's complexity, we introduced some modifications to our model. We directly employ output as input for the subsequent dense block, differentiated in Fig. 5. This approach effectively reduces the input dimensionality for each dense block, significantly conserving model parameters. Additionally, in the dual-feature decoder, only sub-gated structure-1 is retained in the gated structure.

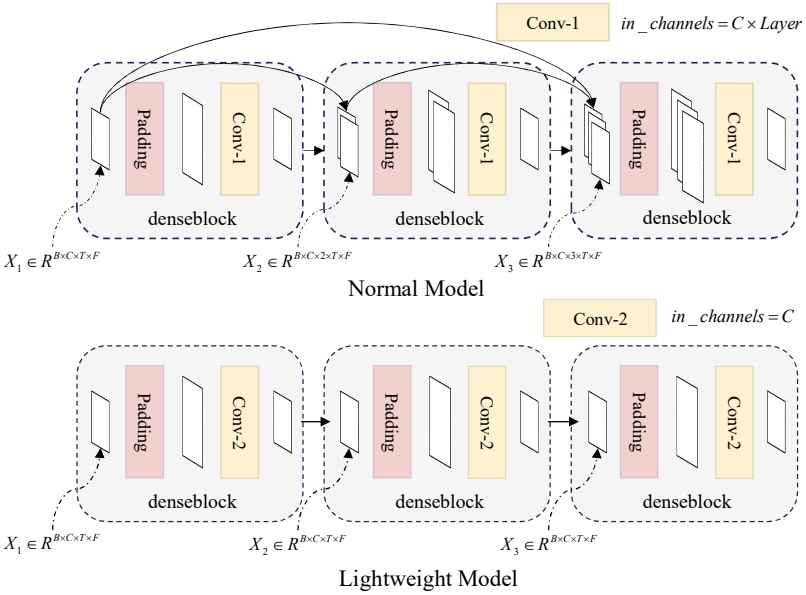


**Fig. 5.** The overall architecture of DenseNet for standard and lightweight model

Our subsequent baseline experiment can affirm the worthiness of this modification.

### 2.7 Discriminator

Adopting CMAGN's [5] discriminator structure, we use a stepped design with a convolution layer with kernels size $4\times 4$, LayerNorm and PReLU. Following that, two linear layers and an adaptive sigmoid are used to process the global average pooling output into a normalized score of two audio amplitudes.

### 2.8 Loss Function

**Loss Function of the Discriminator.**
Following MetricGAN's [8] method, we derive the loss from normalized PESQ scores of two audios falling within $[0,1]$ and employ distorted-clean speech pairs, clean-clean speech pairs and enhanced-clean speech pairs for loss, as shown below [9]:

$$Loss_1 = E_{x,y}\left[\left(D(x,y)-Q(x,y)\right)^2\right] \tag{10}$$

$$Loss_2 = E_{x,y}\left[D(y,y)-1\right] \tag{11}$$

$$Loss_3 = E_{x,y}\left[\left(D(G(x),y)-Q(G(x),y)\right)^2\right] \tag{12}$$

$$Loss = Loss_1 + Loss_2 + Loss_3 \tag{13}$$

Here, $Q$, $D$ and $G$ represents the PESQ score, discriminator, generator, respectively.

**Loss Function of the Generator.**
We utilize the outputs of the discriminator for loss functions. We also use amplitude, complex spectrum and temporal features to calculate the loss, as shown below:

$$L_{Mag} = E_{X_m,\widehat{X_m}}\left[\left\|X_m - \widehat{X_m}\right\|^2\right] \tag{14}$$

$$L_{RI} = E_{X_r,\widehat{X_r}}\left[\left\|X_r - \widehat{X_r}\right\|^2\right] + E_{X_i,\widehat{X_i}}\left[\left\|X_i - \widehat{X_i}\right\|^2\right] \tag{15}$$

$$L_{Time} = E_{X_T,\widehat{X_T}}\left[\left\|X_T - \widehat{X_T}\right\|^2\right] \tag{16}$$

$$L_{GAN} = E_{X_m,\widehat{X_m}}\left[\left\|Discriminator\left(X_m - \widehat{X_m}\right) - 1\right\|^2\right] \tag{17}$$

$$Loss = \alpha_1 L_{Mag} + \alpha_2 L_{RI} + \alpha_3 L_{Time} + \alpha_4 L_{GAN} \tag{18}$$

Here, $X$, $\hat{X}$, $X_m$, $X_r$, $X_i$, $X_T$ represent the distorted audio, enhanced audio, amplitude, real and imaginary spectrum and temporal features of the audio, respectively.

# 3 Experiment

## 3.1 Experimental Setup

**Dataset.**
For evaluation, we use the VoiceBank+DEMAND dataset[2] with 11,572 training utterances and 824 test utterances. The data contains 10 noisy types, with training set SNR from 0 to 15dB increments, and test set SNR from 2.5 to 17.5dB also in 5db steps.

In the generalization experiment [10], we constructed three test sets to evaluate the model performance for varying noise conditions. The first set, designed for noise environment matching, comprised 800 clean utterances from the VoiceBank+DEMAND test set, combined with 4 unseen noises (DLIVING, OMEETING, SPSQUARE, TBUS noise) from the DEMAND dataset. The second set, for noise environment mismatch, includes the same 800 clean utterances, mixed with two unseen noises (factory1 and babble noise) from the NOISEX-92 dataset[3]. Lastly, the channel environment mismatch test set contains 800 clean utterances from TIMIT[4], mixed with two unseen noises (factory1 and babble noise). Each dataset is mixed at SNRs of 0, 6, 12 and 18 dB with 200 utterances per SNR level.

All audios are resampled to 16kHz and the duration of each sample is 2 seconds.

**Implementation.**
In the experiments, we use a window length of 25ms and a hop size of 6.25ms Hamming window for STFT and iSTFT. The power coefficient $c$ for audio power

[2] https://datashare.ed.ac.uk/handle/10283/2791
[3] http://spib.linse.ufsc.br/noise.html
[4] https://catalog.ldc.upenn.edu/LDC93S1

compression is set to 0.3 [9]. The DenseNet and transformer layers are set to 4. The channel size $C$ for the standard and lightweight models are 64, 48. The kernel size $[k_s, k_m, k_b]$ of the attention convolution blocks is set to $[3,5,10]$. For the generator's loss, the parameters are set to $\alpha_1 = 0.9, \alpha_2 = 0.1, \alpha_3 = 0.2, \alpha_4 = 0.05$. We use the Adam optimizer for the generator and discriminator for 100 epochs. The initial learning rate of the generator is 0.0004 and is decreased by 2% every two epochs. The learning rate of the discriminator is kept twice as large as that of the generator.

**Evaluation Metrics.**
To objectively assess our model and measure enhanced speech quality, we use metrics: PESQ [11] (-0.5 to 4.5), SSNR [12] (-10 to 35), CSIG [13], CBAK [13], and COVL [13] (each 1 to 5). For all metrics, higher values indicate better performance.

### 3.2 Baseline

We compared the standard and lightweight models we proposed with other excellent methods on the VoiceBank+DEMAND dataset, as shown in Table. 1.

**Table 1.** Comparison with other methods on the VoiceBank+DEMAND dataset.

| Method | Para(M) | SSNR | PESQ | CSIG | CBAK | COVL |
|---|---|---|---|---|---|---|
| Noisy | N/A | 1.68 | 1.97 | 3.34 | 2.44 | 2.63 |
| TSTNN (2021) [7] | 0.92 | 9.70 | 2.96 | 4.10 | 3.77 | 3.52 |
| SE-Conformer (2021) [14] | N/A | N/A | 3.13 | 4.45 | 3.55 | 3.82 |
| DB-AIAT (2022) [15] | 2.81 | 10.79 | 3.31 | <u>4.61</u> | 3.75 | 3.96 |
| DeepFilterNet2 (2022) [16] | 2.31 | N/A | 3.08 | 4.30 | 3.40 | 3.70 |
| PCS (2022) [17] | N/A | N/A | <u>3.35</u> | 4.43 | N/A | 3.92 |
| MetricGAN-OKD (2023)[18] | 1.89 | N/A | 3.24 | 4.23 | 3.07 | 3.73 |
| NHS-SM-SE (2023) [19] | 0.92 | 9.33 | 3.29 | **4.65** | 3.67 | 3.99 |
| CCFNet+ (2023) [10] | <u>0.62</u> | 10.03 | 3.03 | 4.27 | 3.55 | 3.61 |
| E-CDNN-SRU (2023) [20] | 1.22 | 10.35 | 3.25 | N/A | N/A | N/A |
| **Ours-Lightweight** | **0.58** | <u>11.53</u> | 3.31 | 4.50 | <u>3.81</u> | 3.95 |
| **Ours** | 1.39 | **11.81** | **3.40** | 4.54 | **3.88** | **4.03** |

The results show that our standard model surpasses the majority of other methods, achieving the highest ranking in SSNR, PESQ, CBAK, and COVL. As mentioned above, the issue of speech over-suppression was unaddressed by most methods, resulting in low SSNR. Notably, our standard model not only achieves a significantly higher SSNR of 11.81 but also maintains superior performance across other metrics with PESQ of 3.40. Moreover, our lightweight model shows near-optimal metrics with SSNR of 11.53, PESQ of 3.31, and the fewest parameters. In summary, both our standard and lightweight models deliver superior performance with low complexity.

To more clearly observe the performance of our model, we use one audio from the test set to separately display the clean, distorted and enhanced speech via Waveform,

Logarithmic spectrogram and Mel spectrogram, as shown in Fig. 6-8. It can be clearly observed from each figure that the clean speech and enhanced speech are nearly identical. Our model effectively reduces noise while largely retaining clean speech.

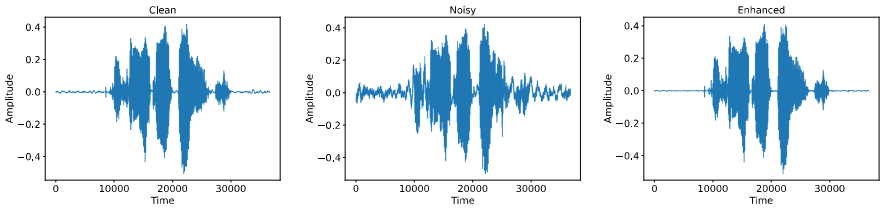


**Fig. 6.** Waveform of clean, noisy and enhanced speech.

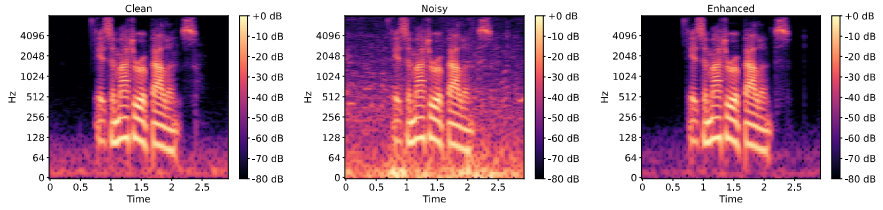


**Fig. 7.** Logarithmic spectrogram of clean, noisy and enhanced speech.

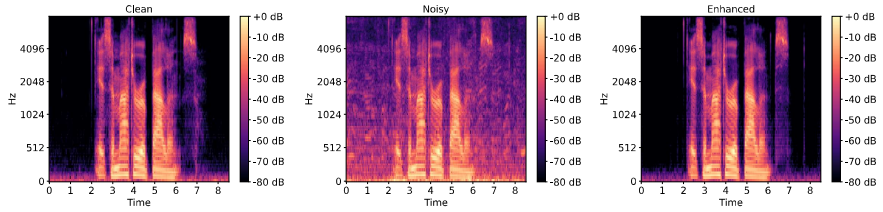


**Fig. 8.** Mel spectrogram of clean, noisy and enhanced speech.

### 3.3 Ablation Experiment

The following groups about the information interaction mechanism, the attention block and the mask formerblock are designed to explore the influence of different sources of information in the information interaction mechanism, different blocks in the two-stage mask transformer and the attention block, with result shown in Table. 2.

1. Complete model: TSMGAN-II
2. Without the information interaction mechanism: II-None.
3. The information interaction mechanism from the encoder: II-E.
4. The information interaction mechanism from another decoder: II-D.
5. Without the attention block: AB-None.
6. Transformer with block substituted the block form TSTSNN: TS-1.
7. Transformer with block substituted the block form SE-Conformer: TS-2.

**Table 2.** The result of the ablation experiment.

| Method | SSNR | PESQ | CSIG | CBAK | COVL |
|---|---|---|---|---|---|
| TSMGAN-II | **11.81** | **3.40** | **4.54** | **3.88** | **4.03** |
| II-None | 11.73 | 3.36 | 4.50 | 3.85 | 3.98 |
| II-D | 11.63 | 3.30 | 4.48 | 3.79 | 3.94 |
| II-E | 11.79 | 3.34 | 4.50 | 3.86 | 3.97 |
| AB-None | 11.71 | 3.36 | 4.47 | 3.85 | 3.97 |
| TS-1 | 11.58 | 3.27 | 4.47 | 3.80 | 3.92 |
| TS-2 | 11.61 | 3.33 | 4.50 | 3.83 | 3.97 |

For the information interaction mechanism, encoder information greatly boosts metrics such as SSNR and CBAK. On the contrary, additional decoder information alone reduces performance. Utilizing both encoder and decoder information together yields noticeable benefits, improving SSNR by 0.08 and PESQ by 0.04. For the mask formerblock, it effectively boosts all metrics and notably improves the model's SSNR indicator, achieving an increase of 0.20 and 0.23 compared with the block introduced in TSTNN and SE-Conformer. Hence, our mask former block excels in preserving target speech and suppressing noise. For the attention block, it improves significantly on SSNR, which demonstrates its filtering effect on input speech.

### 3.4 General Experiment

**Table 3.** Comparison results of different models in matched noise environments

| **Method** | **PESQ** | | | | | **SSNR** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| SNR(dB) | 0 | 6 | 12 | 18 | Avg | 0 | 6 | 12 | 18 | Avg |
| Noisy | 1.15 | 1.35 | 1.72 | 2.30 | 1.63 | -4.10 | -1.26 | 2.89 | 8.07 | 1.40 |
| TSTNN | 1.99 | 2.48 | 2.87 | 3.24 | 2.65 | 5.07 | 6.40 | 8.74 | 11.78 | 7.92 |
| OKD | 2.07 | 2.56 | 2.99 | 3.34 | 2.74 | -1.75 | -0.58 | 1.29 | 3.46 | 0.61 |
| **Ours** | **2.26** | **2.78** | **3.30** | **3.68** | **3.01** | **5.92** | **7.09** | **8.93** | **12.12** | **8.52** |

**Table 4.** Comparison results of different models in mismatched noise environments

| **Method** | **PESQ** | | | | | **SSNR** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| SNR(dB) | 0 | 6 | 12 | 18 | Avg | 0 | 6 | 12 | 18 | Avg |
| Noisy | 1.14 | 1.24 | 1.36 | 1.76 | 1.38 | -3.47 | -0.65 | 2.50 | 7.47 | 1.46 |
| TSTNN | 1.83 | 2.28 | 2.43 | 2.90 | 2.36 | 4.17 | 6.81 | 8.09 | 11.28 | 7.59 |
| OKD | **2.09** | 2.31 | 2.57 | 2.99 | 2.49 | -2.67 | -1.00 | 3.19 | 5.46 | 1.25 |
| **Ours** | 2.05 | **2.38** | **2.67** | **3.17** | **2.57** | **4.56** | **7.44** | **10.50** | **13.73** | **9.06** |

To evaluate our model's effectiveness under different noise conditions, we test it on the test sets mentioned above, compared with MetricGAN-OKD [18] and TSTNN [7].

Table. 3-5 shows our model's top rankings in PESQ and SSNR on three datasets. In contrast, TSTNN and MetricGAN-OKD exhibit significant performance degradation.

This highlights our model's robustness and effectiveness in real-world scenarios. Specifically, our model surpasses TSTNN and MetricGAN-OKD on the SSNR metric, attesting to our methodology's validity for speech over-suppression. Some methods fail to address speech over-suppression such as MetricGAN-OKD, resulting in inferior SSNR performance compared to the original speech signal.

**Table 5.** Comparison results of different models in mismatched channel environments

| **Method** | **PESQ** | | | | | **SSNR** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| SNR(dB) | 0 | 6 | 12 | 18 | Avg | 0 | 6 | 12 | 18 | Avg |
| Noisy | 1.07 | 1.21 | 1.52 | 2.10 | 1.48 | -3.41 | 0.39 | 4.78 | 9.73 | 2.87 |
| TSTNN | 1.15 | 1.43 | 1.92 | 2.35 | 1.71 | 1.10 | 3.99 | 6.84 | 9.27 | 5.30 |
| OKD | 1.24 | 1.23 | 1.41 | 1.58 | 1.37 | -5.28 | -2.79 | -0.60 | -0.88 | -2.39 |
| **Ours** | **1.41** | **1.98** | **2.65** | **3.18** | **2.31** | **3.45** | **6.72** | **9.90** | **12.65** | **8.18** |

## 4 Conclusion

In this paper, we proposed a generative adversarial network-based a on two-stage transformer and information interaction for efficient speech enhancement. It extracts, models and reconstructs speech features in the time-frequency domain from multiple perspectives. We employ an information interaction mechanism to connect the encoder and decoder for better interflow. Our novel mask formerblock boosts speech enhancement abilities. The experiments show that our model outperforms most models with SSNR of 11.81 and PESQ of 3.40, effectively addressing the problem of model complexity and speech over-suppression. Furthermore, our proposed lightweight model (0.58M) exhibits performance comparable to other leading models. In summary, the proposed two-stage mask transformation network and information interaction mechanism attain state-of-the-art results with low complexity.

**Acknowledgments.** This work is partly supported by Key Research and Development Program of Science and Technology Department of Sichuan Province under grant No. 2023YFG0145 and the National Key Research and Development Program of China under grant No. 2022YFC3303101.

## References

1. Valentini-Botinhao, C., & Yamagishi, J.: Speech Enhancement of Noisy and Reverberant Speech for Text-to-speech. IEEE Transactions on Audio, Speech, and Language Processing **26**(8), 1420-1433 (2018)
2. Mehrish, A., Majumder, N., Bhardwaj, R., et al: A review of deep learning techniques for speech processing. Information Fusion **99**, 101869 (2023)
3. Lu, Y., Wang, Z., Watanabe, S., et al.: Conditional Diffusion Probabilistic Model for Speech Enhancement. In: 47th International Conference on Acoustics, Speech and Signal Processing, pp. 7402-7406. IEEE, Singapore, Singapore (2022)

4. S. S. Shetu, S. Chakrabarty, O. Thiergart, et al.: Ultra Low Complexity Deep Learning Based Noise Suppression. In: 49th International Conference on Acoustics, Speech and Signal Processing, pp. 466-470. IEEE, Seoul, Korea (2024)
5. S. Abdulatif, R. Cao, B. Yang: CMGAN: Conformer-based Metric GAN for Speech Enhancement. In: 23rd Conference of the International Speech Communication Association, pp. 936-940. IEEE, Incheon, Korea (2022)
6. Pandey, A., Wang, D.: Densely Connected Neural Network with Dilated Convolutions for Real-Time Speech Enhancement in the Time Domain. In: 45th International Conference on Acoustics, Speech and Signal Processing, pp. 6629-6633. IEEE, Barcelona, Spain (2020)
7. Wang, K., He, B., Zhu, W.: TSTNN: Two-Stage Transformer Based Neural Network for Speech Enhancement in the Time Domain. In: 46th International Conference on Acoustics, Speech and Signal Processing, pp. 7098-7102. IEEE, Toronto, Ontario, Canada (2021)
8. Fu, S., Liao, C., Tsao, Y., et al.: MetricGAN: Generative Adversarial Networks Based Black-box Metric Scores Optimization for Speech Enhancement. In: 36th International Conference on Machine Learning, pp. 2031-2041. PMLR, Long Beach, California, USA (2019)
9. Braun, S., Tashev, I.J.: A Consolidated View of Loss Functions for Supervised Deep Learning-based Speech Enhancement. In: 44th International Conference on Telecommunications and Signal Processing, pp. 72-76. IEEE, Virtual Conference (2021)
10. Dang, F., Chen, H., Hu, Q., et al.: First Coarse, Fine afterward: A Lightweight Two-stage Complex Approach for Monaural Speech Enhancement. Speech Communication **146**, 32-44 (2022)
11. Rix, A.W., Beerends, J.G., Hollier, M., et al.: Perceptual Evaluation of Speech Quality (PESQ)-a New Method for Speech Quality Assessment of Telephone Networks and Codecs. In: 26th International Conference on Acoustics, Speech, and Signal Processing, pp.749-752. IEEE, Salt Lake City, UT, USA (2001)
12. Hansen, J.H., Pellom, B.L.: An Effective Quality Evaluation Protocol for Speech Enhancement Algorithms. In: 5th International Conference on Spoken Language Processing, pp. 2819-2822. ISCA, Sydney, Australia (1998)
13. Hu, Y., Loizou, P.C.: Evaluation of Objective Quality Measures for Speech Enhancement. IEEE Transactions on Audio, Speech, and Language Processing **16**(1), 229-238 (2008)
14. Kim, E., Seo, H.: SE-Conformer: Time-Domain Speech Enhancement Using Conformer. In: 22nd Conference of the International Speech Communication Association, pp. 2736-2740. ISCA, Brno, Czechia (2021)
15. Yu, G., Li, A., Wang, Y., et al.: Dual-Branch Attention-In-Attention Transformer for Single-Channel Speech Enhancement. In: 47th International Conference on Acoustics, Speech and Signal Processing, pp. 7847-7851. IEEE, Singapore, Singapore (2022)
16. Schröter, H., Maier, A., Escalante-B, A. N., et al.: Deepfilternet2: Towards Real-time Speech Enhancement on Embedded Devices for Full-band Audio. In: 17th International Workshop on Acoustic Signal Enhancement, pp. 1-5. IEEE, Bamberg, Germany (2022)
17. Chao, R., Yu, C., Fu, S., et al.: Perceptual Contrast Stretching on Target Feature for Speech Enhancement. arXiv preprint arXiv:2203.17152 (2022)
18. Shin, W., Lee, B.H., Kim, J.S., et al.: MetricGAN-OKD: Multi-Metric Optimization of MetricGAN via Online Knowledge Distillation for Speech Enhancement. In: 40th International Conference on Machine Learning, pp. 31521-31538. PMLR, Hawaii, USA (2023)
19. Jiang, W., Yu, K.: Speech Enhancement with Integration of Neural Homomorphic Synthesis and Spectral Masking. IEEE Transactions on Audio, Speech, and Language Processing **27**, 1758-1770 (2021)
20. Wahab, F.E., Ye, Z., Saleem, N., et al.: Compact deep neural networks for real-time speech enhancement on resource-limited devices. Speech Commun **156**, 103008 (2023)