INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

TridentSE: Guiding Speech Enhancement with 32 Global Tokens

Dacheng Yin[1*], *Zhiyuan Zhao*[2], *Chuanxin Tang*[2], *Zhiwei Xiong*[1], *Chong Luo*[2]

[1]University of Science and Technology of China, Hefei, China
[2]Microsoft Research Asia, Beijing, China

ydc@mail.ustc.edu.cn, zwxiong@ustc.edu.cn, {zhiyzh, chutan, cluo}@microsoft.com

Abstract

This paper presents TridentSE, a new and innovative architecture for speech enhancement that efficiently combines local details and global information. The architecture uses time-frequency bin level representation for capturing detailed information and a small number of global tokens for processing global information. It employs cross attention modules to transfer information between the local and global representation, and separates the global tokens into two groups to process inter- and intra-frame information. A metric discriminator is utilized to increase perceptual quality and achieve improved performance compared to previous speech enhancement methods. With lower computational cost, TridentSE achieved a PESQ of 3.47 on the VoiceBank+DEMAND dataset and a PESQ of 3.44 on the DNS no-reverb testset, outperforming most previous methods. Visualization shows that the global tokens demonstrate diverse and interpretable global patterns.

Index Terms: Speech enhancement, global representation

1. Introduction

Speech enhancement (SE) aims to improve the quality of speech when it is contaminated with noise. With the advent of the deep learning era, significant progress has been made in speech enhancement techniques. One line of research is the time-domain methods [1, 2, 3], which process speech directly in waveform domain. Another line of research is the frequency-domain methods [4, 5, 6], which process speech in the T-F spectrogram domain. Our method falls into the second category, with the goal of designing an effective frequency-domain approach for single-channel speech enhancement.

For frequency-domain SE methods, the input is a time-frequency (T-F) spectrogram. Research [7] has shown that it is more effective to use T-F masks as the prediction target instead of T-F values. As a result, SE addresses a dense classification or prediction problem, where each T-F bin has its corresponding prediction output. The importance of T-F bin level details, especially the phase structure, has increased with the development of masking methods [8, 9, 10, 11], requiring the SE network to accurately capture local details. However, prior work has indicated that a successful SE network must also understand global (long-range) information on both the frequency [4] and time [12] axis. In short, an SE network must learn both local details and global information.

The simultaneous learning of these two types of information is a non-trivial problem. Existing frequency-domain SE methods either adopt a cylindrical network structure [4, 6, 12] or a U-shaped structure [13, 5]. In the first category, the feature map retains its original T-F resolution as it is transformed by the SE network. While dense local information is naturally processed in each T-F bin, sparse global information is also aggregated by each T-F bin without much coordination, which is computationally inefficient. In the second category, the feature map undergoes gradual down-sampling during its transformation. At its lowest resolution, the global semantic information can be computed efficiently. The transformed feature map is then up-sampled to its original size, with skip connections used to merge the low-level and high-level features. It is important to note that the full-resolution feature is not merged with high-level features until the end of the network. This limitation reduces the network's ability to fuse information compared to the cylindrical architecture, which processes both low- and high-level information at every layer.

In this work, we present a novel third network structure for the speech enhancement task, called TridentSE. It consists of a main network that maintains full-resolution feature maps and two companion branches, each of which only holds 16 global tokens. The main network is responsible for processing dense, low-level details, while the companion branches handle global information. Temporal and frequency tokens are extracted from the original feature map through a cross-attention operation and processed in separate units. The global temporal and frequency information is then injected back into the main network via another cross-attention operation. The three-branch network architecture resembles a trident, hence the name TridentSE.

The use of two dedicated branches in TridentSE provides several advantages. Compared to the cylindrical network structure, it greatly reduces the redundancy of computing global information. And compared to the U-shaped network structure, it has the ability to perform long-range computation from the start. Additionally, the fusion of information through cross-attention is stronger than the simple addition method commonly used in skip connections.

Experimental results demonstrate that TridentSE outperforms previous methods by achieving higher enhancement quality with lower computational complexity. Visualization confirms that the global tokens learned by TridentSE are diverse and interpretable.

2. Method

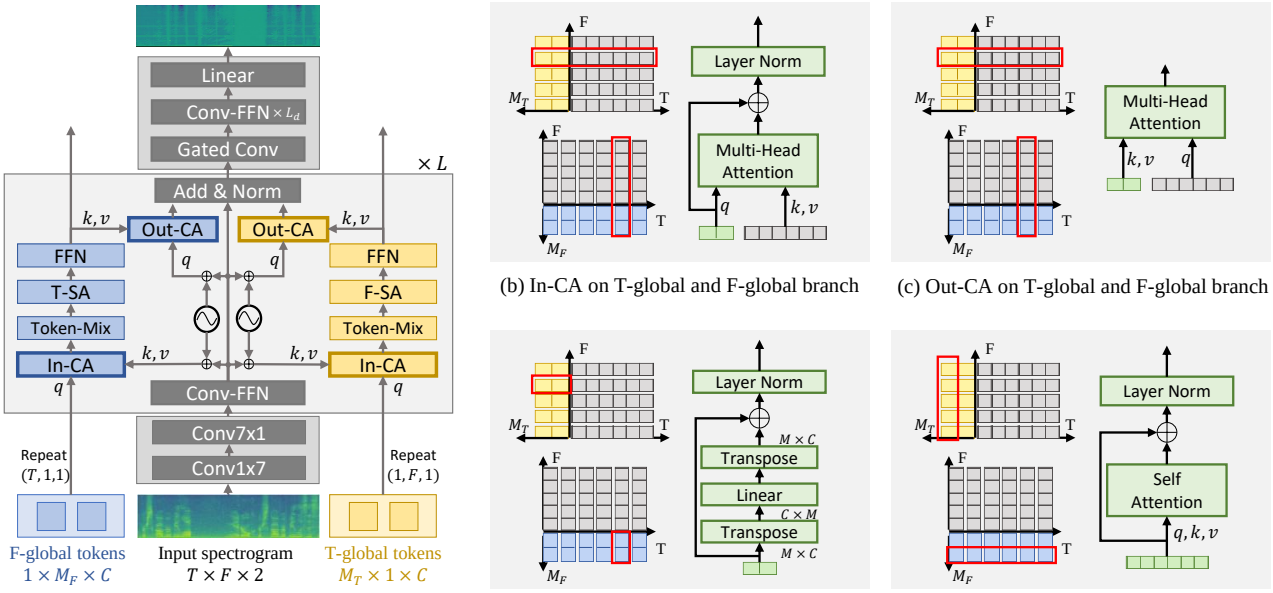
2.1. Overview of TridentSE

The T-F masking framework (as discussed in [12]) is adopted to perform the speech enhancement task, which is to estimate the clean waveform from the input noisy signal. The overall architecture of the proposed T-F mask prediction network, TridentSE, is shown in Fig.1.

TridentSE is composed of three main components: the en-

*Work done during internship at Microsoft Research Asia.

10.21437/Interspeech.2023-565



(a) TridentSE architecture

(b) In-CA on T-global and F-global branch

(c) Out-CA on T-global and F-global branch

(d) Token-Mix on T-global and F-global branch

(e) T-SA and F-SA

Figure 1: *Architecture of the proposed TridentSE. The query, key, and value inputs of the multi-head attention operations are denoted by q, k, and v, respectively.*

coder, the backbone, and the decoder. The encoder extracts local time-frequency features from the input spectrogram, while the decoder decodes the time-frequency representation into a complex ratio mask ($\boldsymbol{M_c}$) with a shape of ($T \times F \times 2$).

The encoder contains two convolutional blocks, each with C output channels. The kernel sizes of these two blocks are 1×7 and 7×1, respectively. Both blocks include batch normalization (BN) and ReLU activation after the convolution operation.

The decoder includes a 1×1 gated-convolutional layer, L_d depth-wise separable convolutional blocks (Conv-FFN), and a linear layer that maps C-dimensional feature vectors into complex numbers. The amplitude of the final output is restricted by the Tanh activation. Each conv layer in Conv-FFN is followed by GELU activation, and each Conv-FFN is equipped with a residue connection and post-layer normalization.

The backbone of the network consists of L trident blocks and represents the major part of TridentSE.

2.2. Trident block

Each trident block consists of three branches: the main branch ($\mathbf{B}_m$), the time-domain global branch ($\mathbf{B}_t$), and the frequency-domain global branch ($\mathbf{B}_f$). The main branch ($\mathbf{B}_m$) calculates local information for each T-F bin using a Conv-FFN module. This is done by aggregating the local information within the range of a $K \times K$ kernel size using a 2D depth-wise separable convolution.

For the two companion branches ($\mathbf{B}_t$ and $\mathbf{B}_f$), they form a duality pair and can be easily switched by changing the subscripts between T (or t) and F (or f). For simplicity, we will only describe the network architecture in the time-domain global branch ($\mathbf{B}_t$) in this section. The branch ($\mathbf{B}_t$) reduces the number of tokens along the time axis to M_T, while maintaining full frequency resolution. The initial feature of this branch, $X_t^0 \in \mathbb{R}^{M_T \times F \times C}$, is obtained by repeating a bank of M_T initial global tokens $G_t \in \mathbb{R}^{M_T \times 1 \times C}$ along the frequency axis F times. The G_t is a learnable model parameter.

The calculation of $\mathbf{B}_t$ involves three modules. First, the token-mix and the frequency self-attention (F-SA) modules mix the feature vectors along the M_T and F axes, respectively, as shown in Fig.1 (d) and (e). Then, a feed-forward network (FFN) transforms the feature along the channel dimension.

The main branch ($\mathbf{B}_m$) and the companion branches ($\mathbf{B}_t$ and $\mathbf{B}_f$) communicate information through input cross-attention (In-CA) and output-cross-attention (Out-CA) modules, as indicated by the thick border in Fig. 1 (a). As shown in Fig. 1 (b) and (c), each row of the time-domain global feature interacts with its corresponding row of the full-resolution feature and each column of the frequency-domain global feature interacts with its corresponding column of the full-resolution feature. We chose cross-attention (CA) as the method of information communication because of its flexibility in transforming between features with different numbers of tokens. It also calculates dynamic weights for information aggregation and broadcast, which can adapt to variations in different input mixture signals.

The information communication is made aware of the time-frequency structure by concatenating the main branch feature with a sinusoidal 2D-positional encoding before it is fed into the In-CA and Out-CA modules. The Conv-FFN, FFN, In-CA, Token-Mix, and T-SA sub-modules have residual connections and post-layer normalization. The hidden layers of the Conv-FFN and FFN also include GELU activation.

To summarize, there are six hyperparameters that define the model: the number of channels C, the size of the convolution kernel K, the number of global tokens M_T, M_F, the number of trident blocks L, and the number of decoder Conv-FFNs L_d.

2.3. Loss function

Our loss function is applied to both the waveform domain and spectrogram domain. On the spectrogram domain, we use the power-compressed amplitude MSE loss L_a and phase-aware MSE loss L_p as described in [4]. On the waveform domain, we calculate the MSE loss L_w. To optimize the PESQ score di-

rectly, we adopt the MetricGAN method from [14] which introduces a metric discriminator D that predicts the differentiable PESQ score for training the SE network. This generates an additional loss term L_{GAN} for the predicted spectrogram. The total loss L is a combination of these four losses and is calculated according to the following formula:

$$L_a = MSE(|\hat{\boldsymbol{S}}|^p, |\boldsymbol{S}|^p), \tag{1}$$

$$L_p = MSE(\hat{\boldsymbol{S}}/|\hat{\boldsymbol{S}}|^{1-p}, \boldsymbol{S}/|\boldsymbol{S}|^{1-p}), \tag{2}$$

$$L_w = MSE(\hat{\boldsymbol{s}}, \boldsymbol{s}), \tag{3}$$

$$L_{GAN} = ||1 - D(\boldsymbol{S}, \hat{\boldsymbol{S}})||^2 \tag{4}$$

$$L = (L_a + L_p + L_w)/3 + \lambda L_{GAN}, \tag{5}$$

where $\hat{\boldsymbol{S}}$, $\hat{\boldsymbol{s}}$, $\boldsymbol{S}$ and $\boldsymbol{s}$ represent the enhanced spectrogram and waveform, the ground-truth clean spectrogram, and waveform, respectively. $|\cdot|$ calculates the amplitude of the complex spectrogram and p is the power of the spectrogram compression, set to $p = 0.3$. The weight of the GAN loss term, λ, is set to 0.005 in our experiments.

The loss for training the discriminator D is calculated as follows:

$$L_D = ||1 - D(\boldsymbol{S}, \boldsymbol{S})||^2 + ||Q(\boldsymbol{S}, \hat{\boldsymbol{S}}) - D(\boldsymbol{S}, \hat{\boldsymbol{S}})||^2, \tag{6}$$

where $Q(\boldsymbol{S}, \hat{\boldsymbol{S}})$ is the normalized PESQ score between $\boldsymbol{S}$ and $\hat{\boldsymbol{S}}$ ranged from 0 to 1.

3. Experiments

3.1. Dataset and evaluation metrics

We evaluate our method using two datasets. The first is the widely-used VoiceBank+DEMAND dataset [15] which contains paired clean and pre-mixed noisy speech samples. The clean speech comes from the VoiceBank corpus [16] and consists of 11,572 utterances from 28 speakers in the training set, and 872 utterances from two speakers in the test set. The noisy speech samples in the training set are mixed with 10 types of noise, including eight from the DEMAND database and two artificially generated noises, at signal-to-noise ratios (SNRs) of 0, 5, 10, and 15 dB. In the test set, the speech is mixed with five types of noise from the DEMAND database, at SNRs of 2.5, 7.5, 12.5, and 17.5 dB. None of the noise conditions or speakers in the test set were present in the training set.

The second dataset we use is the large-scale DNS dataset [17], which includes 500 hours of clean speech from 2150 speakers, and over 180 hours of noise waveform from 150 classes. During the training stage, we perform online mixing to obtain noisy-clean speech pairs. To do so, 75% of the clean speech is convolved with randomly selected room impulse responses (RIRs) from the set provided in [18], and the clean or reverberant speech is mixed with randomly selected noise, with a uniformly sampled SNR ranging from -5 to 20 dB. We evaluate our method on two test sets: "no_reverb" and "with_reverb", each of which contains 150 noisy-clean speech pairs.

We evaluate the enhancement quality using a total of five metrics, with higher scores indicating better results. Both datasets are evaluated using wide-band PESQ and short-term objective intelligibility (STOI) to measure perceptual quality and intelligibility, respectively. In VoiceBank+DEMAND dataset, we use three additional MOS-based metrics [19]: MOS prediction of the signal distortion (CSIG), of the intrusiveness of background noise (CBAK), and of the overall effect (COVL). These metrics are scored on a scale of 1 to 5.

Table 1: *The effect of different configurations in TridentSE.*

#	M_T	M_F	MGAN	L	L_D	FLOPS	RTF	PESQ
G1	0	0	w/o	6	6	18.3G	0.23	3.01
G2	1	1	w/o	3	4	23.6G	0.22	3.23
G3	2	2	w/o	3	4	24.0G	0.22	3.24
G4	6	6	w/o	3	4	25.3G	0.23	3.30
G5	16	16	w/o	3	4	28.7G	0.24	3.30
M	16	16	w/	3	4	28.7G	0.24	3.44
A1	axial	axial	w/	3	4	36.4G	0.35	3.41
A2	1-group	1-group	w/	3	4	18.9G	0.20	3.06

In terms of computational complexity, we report the FLOPS for a 3-second input signal and the real-time factor (RTF) on six Intel(R) Xeon(R) E5-2690 v3 CPU cores. Additionally, we report the model size in terms of the number of parameters.

3.2. Implementation details

All the utterances are resampled to 16kHz and we use 3-second segments for training. STFT is computed using a Hann window of length 20ms, hop length of 10ms, and FFT size of 324. Four hyper-parameters, M_T, M_F, L, L_d, are tuned in our experiments. The other hyper-parameters are set as follows: $C = 96$, $K = 7$. The head number of T-SA, F-SA, In-CA, and Out-CA are set to 2, 2, 3, and 3, respectively. The hidden size of FFN and Conv-FFN is 96. The sinusoidal 2D-positional encoding has 64 channels. The model is trained using LAMB [20] optimizer with learning rate of 0.0008. The metric discriminator is trained with Adam [21] optimizer with a learning rate of 0.0004. The warm-up steps and batch size are set to 5,000 and 8, respectively. The training epochs are 300 and 120 for VoiceBank+DEMAND dataset and DNS dataset, respectively.

3.3. Global representation and adversarial training

The results of our experiments on the VoiceBank+DEMAND dataset are presented in Table 1. The number of global tokens was gradually increased from G1 to G5 until the PESQ score no longer improved. In experiment G1, companion branches have been removed, resulting in a lower computational cost, which is compensated by adding extra layers. However, the comparison between G1 and G2 shows that companion branches are crucial for a significant improvement in PESQ scores, with an increase of 0.22. As the number of global tokens increases, we observe a relatively large improvement in PESQ scores of 0.07, which eventually reaches saturation at a relatively small number of 16 global tokens, suggesting that processing global information does not require intensive computation. The full model, TridentSE, was achieved in experiment M by adding adversarial training to G5, resulting in a 0.14 improvement in PESQ scores.

3.4. Ablation study

We conducted an ablation study on our global information processing method. In experiment A1 of Table 1, we replace the two companion branches with two axial attention blocks that calculates attention along time and frequency axis respectively. This resulted in a 27% increase in FLOPS, 46% in RTF and a decrease of 0.03 in PESQ, demonstrating the superiority of the Trident architecture over traditional axial attention in speech enhancement tasks. Experiment A2, which uses only one companion branch to directly aggregate information from all T-F bins, resulted in a large drop in PESQ of 0.38, highlighting the ne-

Table 2: *System comparison on VoiceBank+DEMAND dataset. Data with label '*' is our reproduced result.*

	Architecture	PESQ	CSIG	CBAK	COVL	STOI(%)	FLOPS	RTF	#Param.
Noisy	-	1.97	3.35	2.44	2.63	92.1	0	0	0
SEGAN [22]	U-shaped	2.16	3.48	2.94	2.80	-	-	-	-
DEMUCS [2]	U-shaped	3.07	4.31	3.40	3.63	95	77.8G	1.18	60.8M
sudo-rm-rf [23]	U-shaped	3.11*	4.36*	3.58*	3.74*	95	21.9G	0.20	4.85M
SE-Conformer [3]	U-shaped	3.13	4.45	3.55	3.82	95	-	-	-
DCCRN [5]	U-shaped	2.68	3.88	3.18	3.27	94	25.2G	0.26	3.67M
TFT-Net [12]	cylindrical	2.75	3.93	3.44	3.34	-	295G	0.73	5.81M
PHASEN [4]	cylindrical	2.99	4.21	3.55	3.62	-	206G	0.51	20.9M
SN-Net [24]	cylindrical	3.12	4.39	3.60	3.77	-	-	-	-
DB-AIAT [25]	cylindrical	3.31	4.61	3.75	3.96	96	68.0G	3.81	2.81M
DPT-FSNET [6]	cylindrical	3.33	4.58	3.72	4.00	96	55.7G*	1.12*	**0.88**M
CMGAN [26]	cylindrical	3.41	4.63	**3.94**	**4.12**	96	116G	1.02	1.83M
TridentSE-S	Trident	3.36	4.61	3.75	3.99	**96**	**19.8G**	**0.16**	1.00M
TridentSE-M	Trident	**3.44**	**4.65**	3.77	4.06	**96**	28.7G	0.24	1.42M
TridentSE-L	Trident	**3.47**	**4.70**	3.81	4.10	**96**	59.8G	0.49	3.03M

cessity of separate T and F processing. Our results indicate that the proposed Trident architecture is a superior choice compared to evaluated alternatives.

3.5. System comparison

In Table 2, we compare our method with other time-domain and T-F domain methods with U-shaped and cylindrical architecture on VoiceBank+DEMAND dataset. TridentSE-M is the same model as the experiment M in Table 1. TridentSE-S and TridentSE-L are small and large version of TridentSE, respectively. The only difference is the model depth. In TridentSE-S, $L = L_d = 2$, while in TridentSE-L, $L = 7$ and $L_d = 8$. TridentSE-S has the smallest FLOPS and RTF among all the listed methods, but the enhancement quality outperforms all other methods except CMGAN and the COVL score of DPT-FSNET. Compared with CMGAN, TridentSE-M and -L achieve higher PESQ and CSIG with only one-forth and one-half of computational cost, respectively. In summary, Trident architecture is faster and better than the previous methods. Table 3 shows the results on DNS dataset. With half of the inference time, TridentSE-L achieves a new state-of-the-art on PESQ and outperforms DPT-FSNET by a large margin on no_reverb testset.

Table 3: *Results on DNS no_reverb / with_reverb testset.*

	PESQ	STOI(%)	RTF
Noisy	1.58 / 1.82	91.52 / 86.62	0
PoCoNet[27]	2.75 / 2.83	- / -	-
FullSubNet[28]	2.78 / 2.97	96.11 / 92.62	0.39
DPT-FSNet[6]	3.26 / 3.53	97.68 / 95.23	1.12
TridentSE-L	**3.44** / 3.50	**97.86** / 95.22	0.49

3.6. Attention visualization

In this experiment, we figure out what is learned in the global tokens by visualizing In-CA's attention maps as shown in Figure 2. The attention map is obtained by enhancing a sample with SNR of 1.4dB using TridentSE-L. (b1-b3) and (c1-c3) demonstrate the shallow layer attention of three global tokens in the frequency and time global branch, respectively. They mainly attend to different wide frequency bands and time spans. As the layer goes deeper, the attention map shows more speech-specific patterns, such as harmonics (d2) and formants (d3) which are the important evidence of identifying phonemes. Some other global tokens focus on noise-dominant T-F bins (d1) to capture noise-specific information. All these attention maps are distributed globally in the spectrogram. Therefore we can confirm that the global tokens have learned meaningful global information.

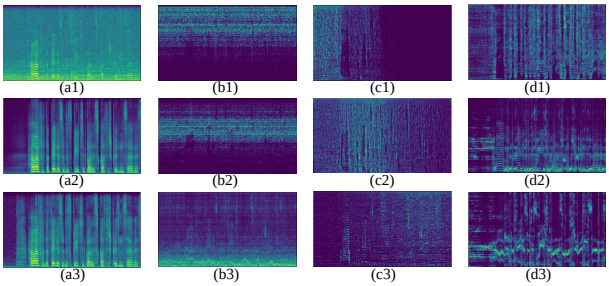


Figure 2: *Visualization of spectrogram and attention maps. (a1-a3): STFT spectrogram of noisy, enhanced, and clean speech. (b1-b3 & c1-c3): Shallow layer attention maps on F-global and T-global branch. (d1-d3): Attention maps that focus on speech or noise patterns*

4. Conclusion

We have introduced a new speech enhancement network called TridentSE. This network features a unique trident structure, with a main network and two accompanying branches. The main network, with its lightweight design, preserves full spectrogram resolution to capture low-level details in each T-F bin. Meanwhile, the two companion branches utilize a total of 32 global tokens to efficiently extract and process concentrated global information. With the added use of adversarial training, TridentSE outperforms previous methods in both the VoiceBank+DEMAND and DNS datasets, while requiring significantly less computational resources. The attention maps demonstrate that the global tokens have acquired a diverse and meaningful range of global information. For future work, we aim to develop a causal version of TridentSE, in order to meet the demands of low-delay, real-time applications.

5. References

[1] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE ACM Trans. Audio Speech Lang. Process.*, 2019.

[2] Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi, "Real time speech enhancement in the waveform domain," in *Interspeech*, 2020.

[3] Eesung Kim and Hyeji Seo, "Se-conformer: Time-domain speech enhancement using conformer," in *Interspeech*, 2021.

[4] Dacheng Yin, Chong Luo, Zhiwei Xiong, and Wenjun Zeng, "PHASEN: A phase-and-harmonics-aware speech enhancement network," in *AAAI*, 2020.

[5] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie, "DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement," in *Interspeech*, 2020.

[6] Feng Dang, Hangting Chen, and Pengyuan Zhang, "Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *ICASSP*, 2022.

[7] Yan Zhao, Zhong-Qiu Wang, and DeLiang Wang, "A two-stage algorithm for noisy and reverberant speech enhancement," in *ICASSP*, 2017.

[8] Guoning Hu and DeLiang Wang, "Speech segregation based on pitch tracking and amplitude modulation," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, 2001.

[9] Soundararajan Srinivasan, Nicoleta Roman, and DeLiang Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, 2006.

[10] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *ICASSP*, 2015.

[11] Donald S Williamson, Yuxuan Wang, and DeLiang Wang, "Complex ratio masking for monaural speech separation," *ITASLP*, 2015.

[12] Chuanxin Tang, Chong Luo, Zhiyuan Zhao, Wenxuan Xie, and Wenjun Zeng, "Joint time-frequency and time domain learning for speech enhancement," in *IJCAI*, 2020.

[13] Yihui Fu, Yun Liu, Jingdong Li, Dawei Luo, Shubo Lv, Yukai Jv, and Lei Xie, "Uformer: A unet based dilated complex & real dual-path conformer network for simultaneous speech enhancement and dereverberation," in *ICASSP*, 2022.

[14] Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, and Shou-De Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *ICML*, 2019.

[15] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech," in *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*, 2016.

[16] Christophe Veaux, Junichi Yamagishi, and Simon 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)*, 2013.

[17] Chandan KA Reddy, Ebrahim Beyrami, Harishchandra Dubey, Vishak Gopal, Roger Cheng, Ross Cutler, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, et al., "The interspeech 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework," *arXiv preprint arXiv:2001.08662*, 2020.

[18] Chandan KA Reddy, Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "Icassp 2021 deep noise suppression challenge," in *ICASSP*, 2021.

[19] Yi Hu and Philipos C Loizou, "Evaluation of objective quality measures for speech enhancement," *ITASLP*, 2007.

[20] Yang You, Jing Li, Sashank J. Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh, "Large batch optimization for deep learning: Training BERT in 76 minutes," in *ICLR*, 2020.

[21] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, "SEGAN: speech enhancement generative adversarial network," in *Interspeech*, 2017.

[23] Efthymios Tzinis, Zhepei Wang, and Paris Smaragdis, "Sudo rm-rf: Efficient networks for universal audio source separation," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2020.

[24] Chengyu Zheng, Xiulian Peng, Yuan Zhang, Sriram Srinivasan, and Yan Lu, "Interactive speech and noise modeling for speech enhancement," in *AAAI*, 2021.

[25] Guochen Yu, Andong Li, Chengshi Zheng, Yinuo Guo, Yutian Wang, and Hui Wang, "Dual-branch attention-in-attention transformer for single-channel speech enhancement," in *ICASSP*, 2022.

[26] Sherif Abdulatif, Ruizhe Cao, and Bin Yang, "Cmgan: Conformer-based metric-gan for monaural speech enhancement," *Interspeech*, 2022.

[27] Umut Isik, Ritwik Giri, Neerad Phansalkar, Jean-Marc Valin, Karim Helwani, and Arvindh Krishnaswamy, "Poconet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss," *arXiv preprint arXiv:2008.04470*, 2020.

[28] Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP*, 2021.