GTCRN: A SPEECH ENHANCEMENT MODEL REQUIRING ULTRALOW COMPUTATIONAL RESOURCES

Xiaobin Rong[1,2], Tianchi Sun[1,2], Xu Zhang[3], Yuxiang Hu[2], Changbao Zhu[2], Jing Lu[1,2]

[1]Key Laboratory of Modern Acoustics, Nanjing University, Nanjing 210093, China
[2]NJU-Horizon Intelligent Audio Lab, Horizon Robotics, Beijing 100094, China
[3]Jiangsu Thingstar Information Technology Co., Ltd., Nanjing 210046, China

INTRODUCTION

We introduce an ultra-lightweight speech enhancement model

- Only **23.7 K** parameters and **39.6 MMACs** per second
- Achieves a PESQ of **2.87** on the VCTK-DEMAND dataset and a DNSMOS of **3.44** on the DNS3 blind test set

Speech Enhancement (SE)

- Aims at recovering clean speech from its noise-contaminated mixture
- Has been advanced by deep learning models, enabling the high performance of noise suppression

Challenges

- Current SOTA SE models call for substantial computational resources, making them undeployable on edge devices
- Existing lightweight SE models are still too large

Thus, we propose to design an ultra-light SE model that can achieve competitive performance with recent baseline models

EXPERIMENTS

Datasets

- VCTK-DEMAND
  - Resampled to 16 kHz
  - 10,000 for training, 1,572 for validation, 824 for test
- DNS3
  - SNR: -5 – 15 dB
  - Capacity: 2000 hours
  - Argumentation: Include Mandarin corpus from DiDiSpeech[1]
  - 40,000 pairs of 8-second data for training per epoch, 840 for validation, 800 for test
  - DNS3 blind test set is also used for test

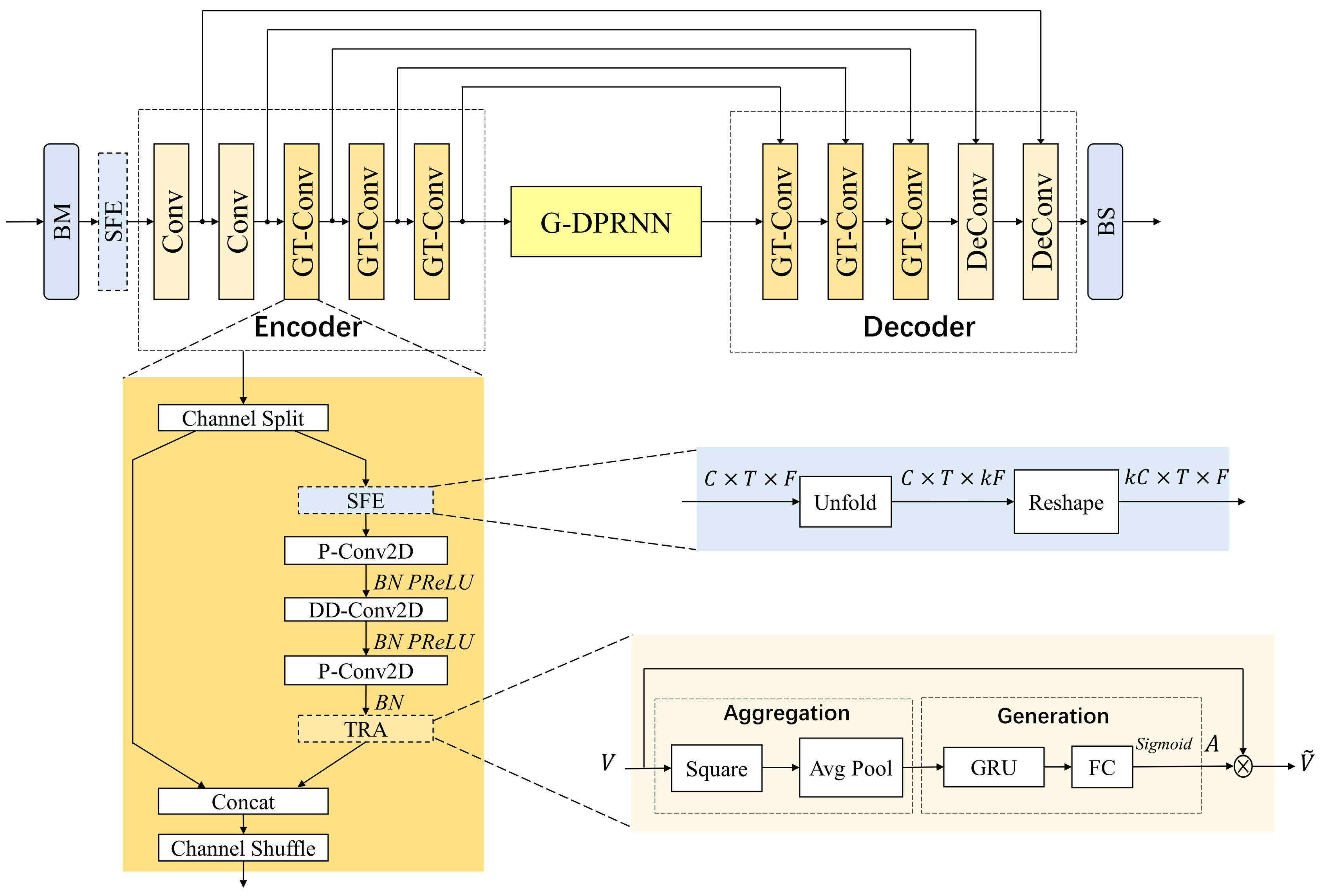
Metrics

- Intrusive: SISNR, PESQ, STOI
- Non-intrusive: DNSMOS

Baseline Models:

- RNNoise[2], PercepNet[3], DeepFilterNet[4], S-DCCRN[5]

METHODS



Grouped Temporal Convolutional Recurrent Network (GTCRN)

- BM/BS: band merging / band splitting
- Conv: convolution block
- DeConv: deconvolution block
- GT-Conv: grouped temporal convolutional block
- SFE: subband feature extraction
- TRA: temporal recurrent attention
- G-DPRNN: grouped dual-path RNN

Loss Function

$$\mathcal{L} = \alpha \mathcal{L}_{SISNR}(\tilde{s}, s) + (1-\beta)\mathcal{L}_{mag}(\tilde{S}, S) + \beta\left(\mathcal{L}_{real}(\tilde{S}, S) + \mathcal{L}_{imag}(\tilde{S}, S)\right)$$

$$\mathcal{L}_{SISNR} = -\log_{10}\left(\frac{\|s_t\|^2}{\|\tilde{s} - s_t\|^2}\right); s_t = \frac{\langle \tilde{s}, s \rangle s}{\|s\|^2}$$

$$\mathcal{L}_{mag}(\tilde{S}, S) = \mathrm{MSE}\left(|\tilde{S}|^{0.3}, |S|^{0.3}\right)$$

$$\mathcal{L}_{real}(\tilde{S}, S) = \mathrm{MSE}\left(\tilde{S}_r / |\tilde{S}|^{0.7}, S_r / |S|^{0.7}\right)$$

$$\mathcal{L}_{imag}(\tilde{S}, S) = \mathrm{MSE}\left(\tilde{S}_i / |\tilde{S}|^{0.7}, S_i / |S|^{0.7}\right)$$

RESULTS

Table 1: Ablation study results on DNS3 test set

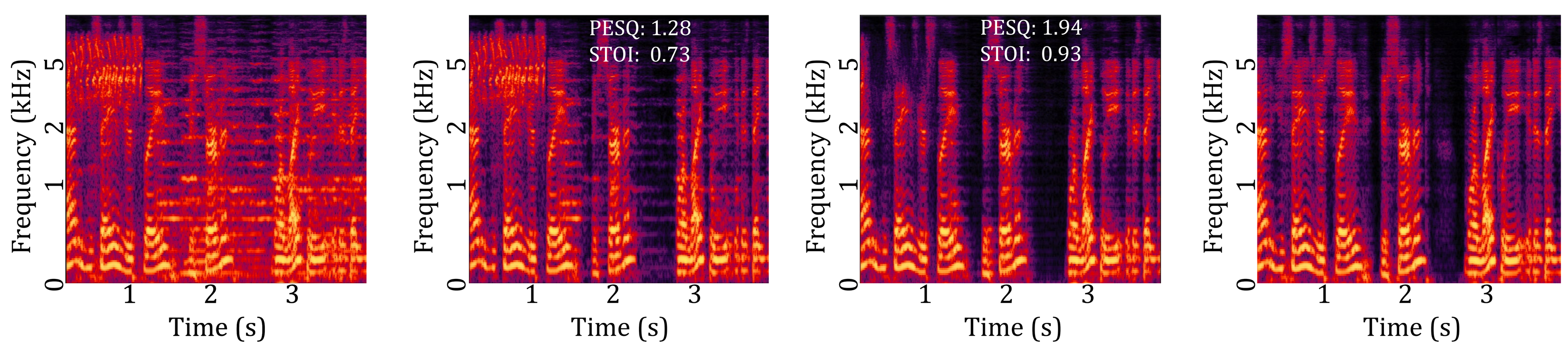
SFE	TA[6]	TRA	Para. (K)	MACs (M/s)	SISNR	PESQ	PESQ
-	-	-	-	-	3.92	1.30	0.789
×	×	×	**13.35**	**33.91**	9.87	1.87	0.834
×	✓	×	14.84	34.00	10.00	1.89	0.838
×	×	✓	21.65	34.47	10.25	1.91	0.840
✓	×	×	15.37	39.07	10.10	1.90	0.838
✓	✓	×	16.86	39.16	10.29	1.92	0.841
✓	×	✓	23.67	39.63	**10.39**	**1.94**	**0.844**

Table 2: Performance on VCTK-DEMAND test set

	Para. (M)	MACs (G/s)	SISNR	PESQ	PESQ
Noisy	-	-	8.45	1.97	0.921
RNNoise[2] (2018)	0.06	0.04	-	2.29	-
PercepNet[3] (2020)	8.00	0.80	-	2.73	-
DeepFilterNet[4] (2022)	1.80	0.35	16.63	2.81	**0.942**
S-DCCRN[5] (2022)	2.34	-	-	2.84	0.940
GTCRN (proposed)	**0.02**	**0.04**	**18.83**	**2.87**	0.940

Table 3: Performance on DNS3 blind test set

	Para. (M)	MACs (G/s)	DNSMOS-P.808	DNSMOS-P.835		
				BAK	SIG	OVRL
Noisy	-	-	2.96	2.65	**3.20**	2.33
RNNoise[2] (2018)	0.06	0.04	3.15	3.45	3.00	2.53
S-DCCRN[5] (2022)	2.34	-	3.43	-	-	-
GTCRN (proposed)	**0.02**	**0.04**	**3.44**	**3.90**	3.00	**2.70**



(a) Noisy (b) Enhanced by RNNoise (c) Enhanced by GTCRN (d) Clean

Source code and audio examples are available at my GitHub:
https://github.com/Xiaobin-Rong/gtcrn

DISCUSSIONS

Current Limitations

- Performance degrades in low-SNR condition
- Generalization ability is constrained due to limited complexity

Future Work

- Transition from model design to framework development for further improvement
- Strengthen generalizability of the model

REFERENCES

[1] Tingwei Guo, Cheng Wen, Dongwei Jiang, Ne Luo, et al., “Didispeech: A Large Scale Mandarin SpeechCorpus,” in ICASSP, 2021, pp. 6968–6972.
[2] Jean-Marc Valin, “A hybrid DSP/deep learning approach to real-time full-band speech enhancement,” in2018 IEEE 20th international workshop on multimediasignal processing (MMSP). IEEE, 2018, pp. 1–5.
[3] Jean-Marc Valin, Umut Isik, Neerad Phansalkar, RitwikGiri, et al., “A Perceptually-Motivated Approach forLow-Complexity, Real-Time Enhancement of FullbandSpeech,” in Proc. Interspeech 2020, 2020, pp. 2482–2486.
[4] Hendrik Schroter, Alberto N Escalante-B, TobiasRosenkranz, and Andreas Maier, “DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering,” in ICASSP, 2022, pp. 7407–7411.
[5] Shubo Lv, Yihui Fu, Mengtao Xing, et al.,“S-DCCRN: Super Wide Band DCCRN with Learnable Complex Feature for Speech Enhancement,” in ICASSP, 2022, pp. 7767–7771.
[6] Qiquan Zhang, Qi Song, et al., “Time-Frequency Atten-tion for Monaural Speech Enhancement,” in ICASSP,2022, pp. 7852–7856.