S-DCCRN

SUPER WIDE BAND DCCRN WITH LEARNABLE COMPLEX FEATURE FOR SPEECH ENHANCEMENT

https://arxiv.org/pdf/2111.08387.pdf

Chen Chen

2021/11/24

Research on different samplerate

- 8K Hz Narrow Band
- **16K Hz Wide Band**
- 32K Hz Super Wide Band
- 44K Hz Full Band

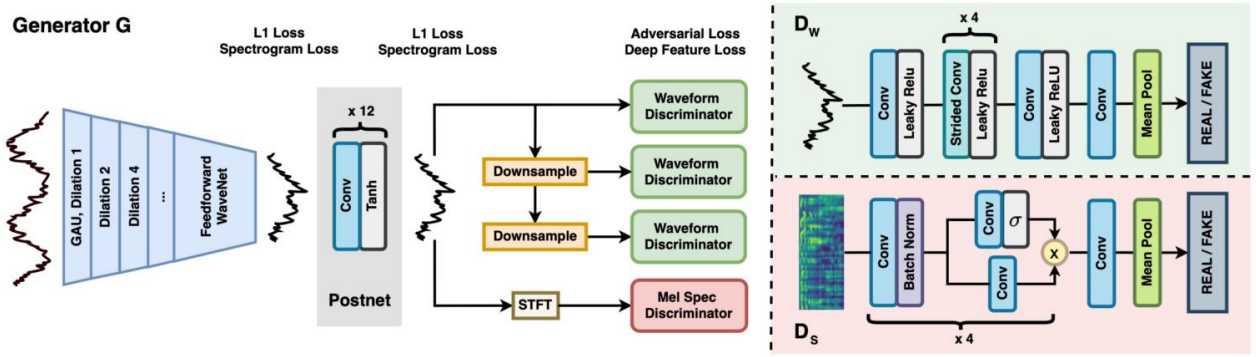
- Most of the recent speech enhancement approaches mainly focus on wide-band signal with a sampling rate of 16K Hz.
- Research on super-wide-band or even full-band denoising is still lacked.

Challenges

- The challenges exist in modeling more frequency bands and particularly high frequency components.
  - How should we model more frequency bands ?
  - How should we use the information in high frequency components ?
- Modeling with larger dimensional features will cause higher complexity of the modeling.
  - How should we compress the dimension of feature ?

Answers of other approaches

- HiFi-GAN
  - Use same network on different samplerate
- RNNoise
  - Bark scale: based on human perception
- PercepNet
  - Represent envelope from 0 to 20 kHz using 34 bands, spaced according to the human hearing equivalent rectangular bandwidth (ERB)



Bark scale

- The **Bark scale** is a psychoacoustical scale proposed by Eberhard Zwicker in 1961. It is named after Heinrich Barkhausen who proposed the first subjective measurements of loudness.

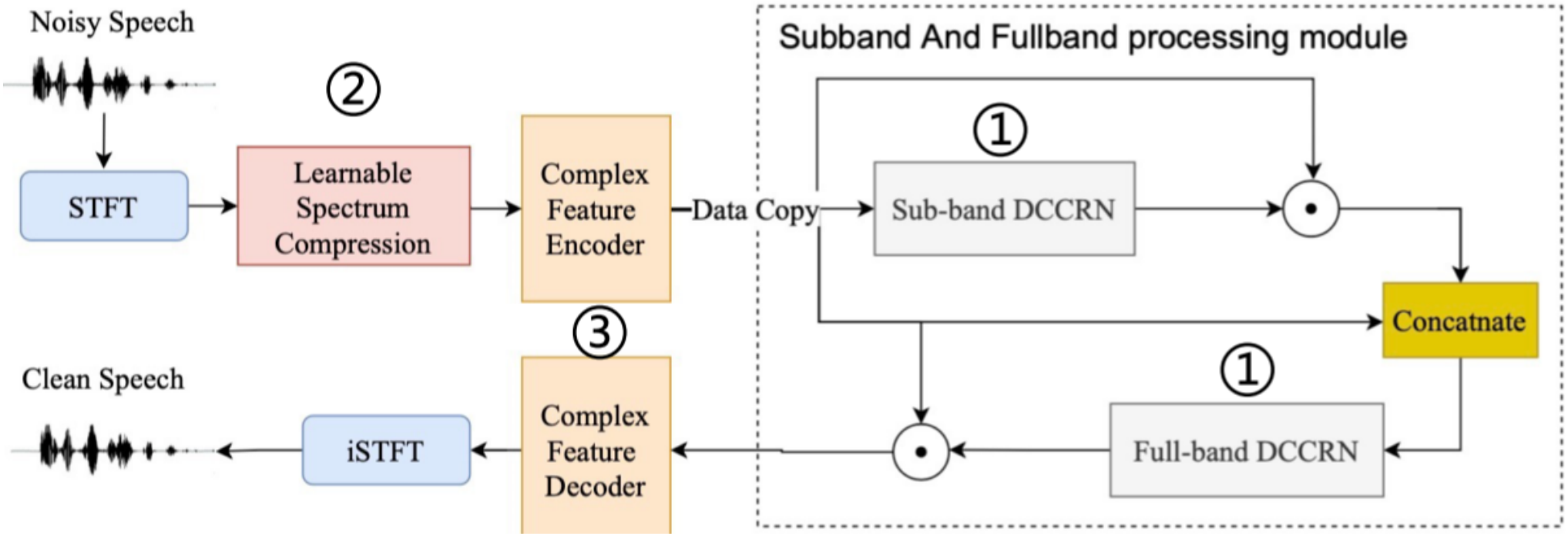
$$B = 13\tan^{-1}\left(\frac{0.76f}{1000}\right) + 3.5\tan^{-1}\left(\frac{f}{7500}\right)^2$$

Critical Band (Bark)	Center Frequency (Hz)	Bandwidth (Hz)
1	50	100
2	150	100
3	250	100
4	350	100
5	450	110
6	570	120
7	700	140
8	840	150
9	1000	160
10	1170	190
11	1370	210
12	1600	240
13	1850	280
14	2150	320
15	2500	380
16	2900	450
17	3400	550
18	4000	700
19	4800	900
20	5800	1100
21	7000	1300
22	8500	1800
23	10500	2500
24	13500	3500

https://www.prosoundtraining.com/2019/07/26/why-equalize-in-1-3-octave-bands/

HiFi-GAN https://arxiv.org/pdf/2006.05694.pdf
RNNoise https://arxiv.org/pdf/1709.08243.pdf
PercepNet https://arxiv.org/pdf/2008.04259.pdf

Answer of S-DCCRN



#	*S*-DCCRN 贡献	动机
①	Sub-band & Full-band DCCRN **(SAF)**	首先子带DCCRN精细化学习高低频信息，然后全带DCCRN结合高低频信息，起到平滑衔接作用
②	Learnable Spectrum Compression (**LSC**)	通过网络学习，动态调整不同频带能量
③	Complex Feature Encoder/Decoder **(CFE/CFD)**	在同16K降噪模型保持相同的较低频率分辨率的同时，通过复数特征编码从谱上获取跟多信息

Complex Feature Encoder/Decoder

① extract high dimensional information

② capture long-term contextual features from time scale

③ extract complex local features

④ process the estimated real/imag features

⑤ pixel convolution is considered as a better alternative for transposed convolution to avoid checkerboard artifacts

⑥ revert the high-dimensional feature to the time-frequency domain

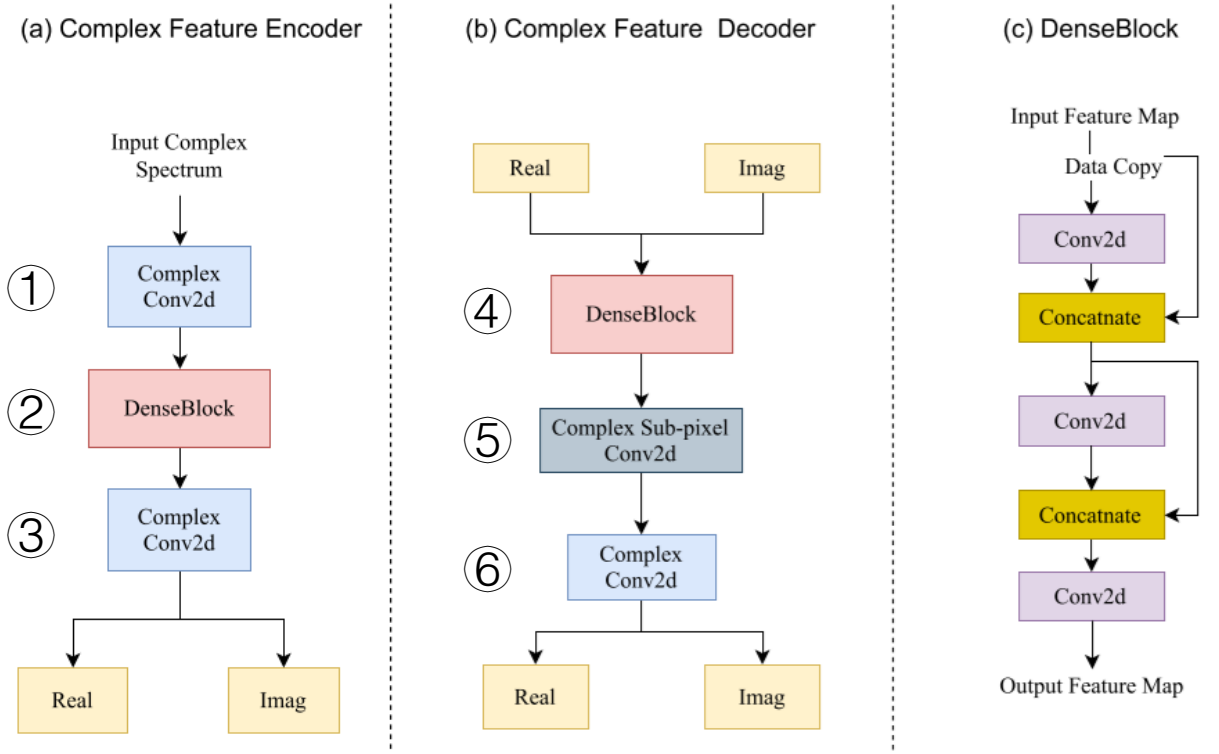


Fig. 2. Complex feature encoder/decoder (CFE/CFD) module

The STFT length is 512.
The hidden channels of the complex feature encoder/decoder module are 32.

Ashutosh Pandey and DeLiang Wang, "Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6629–6633.

Sub-band and Full-band Processing Module

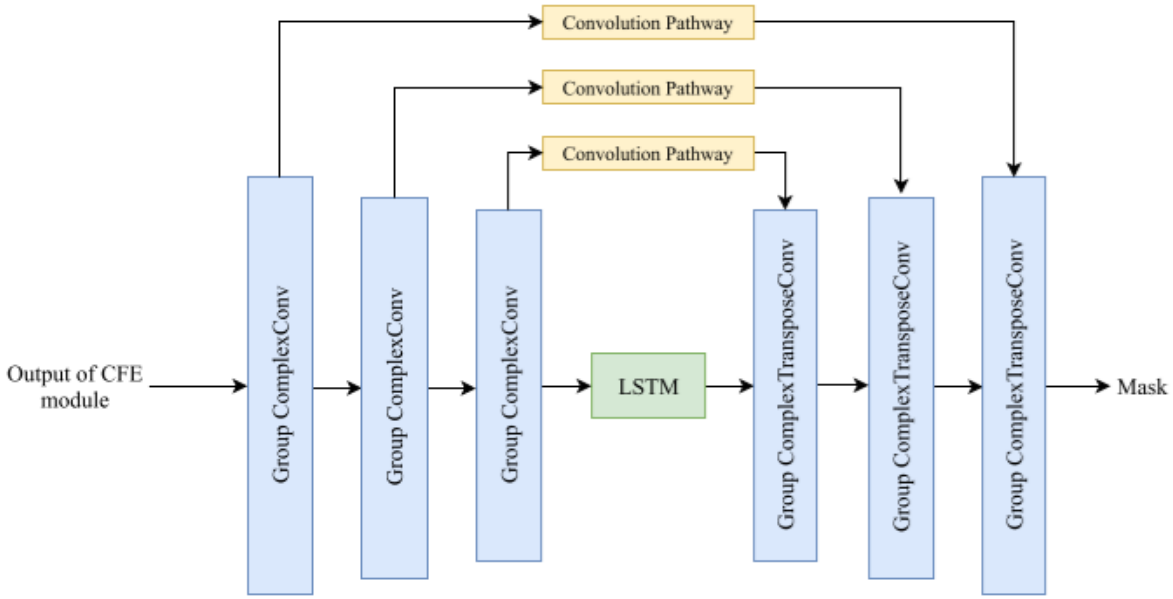


Fig. 3. The sub-band DCCRN module

The skip pathway between encoder and decoder consists of a complex convolution block and batch normalization.

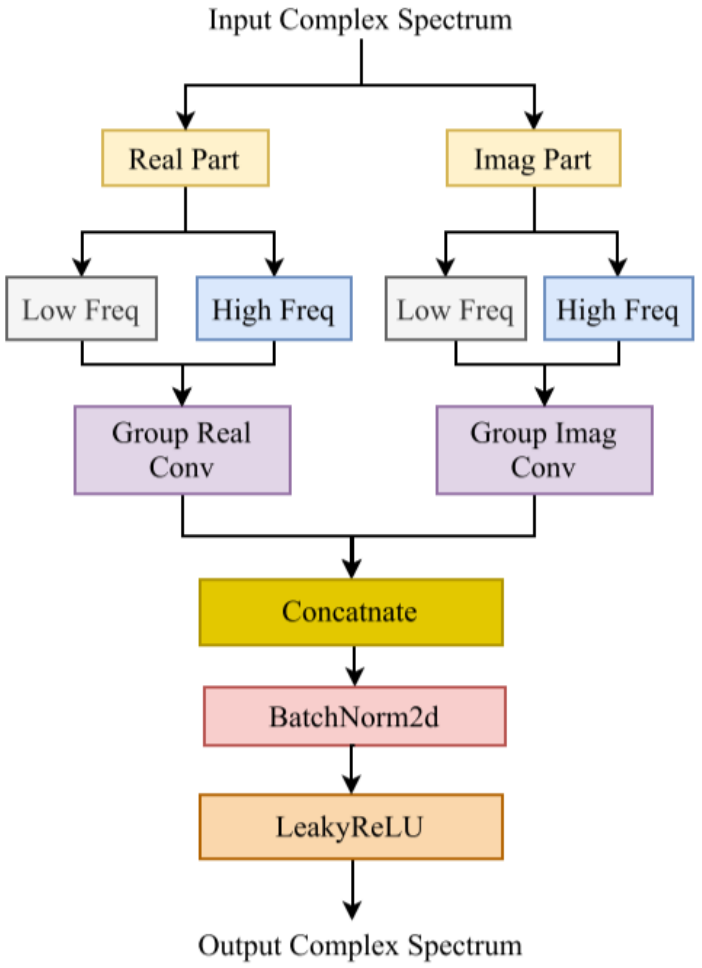


Fig. 4. Complex group convolution

Learnable Spectrum Compression

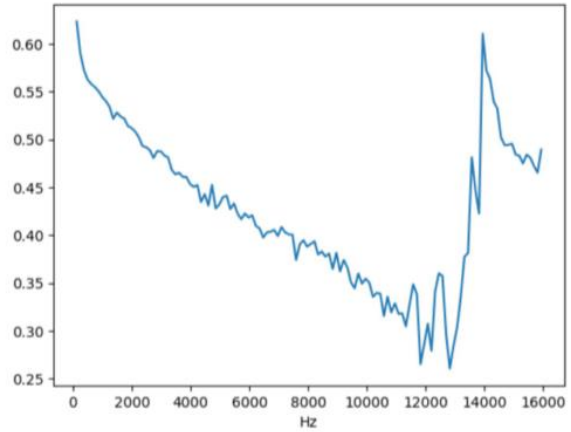


Fig. 5. Compression ratio of different frequency automatically learned by the proposed learnable spectrum compression.

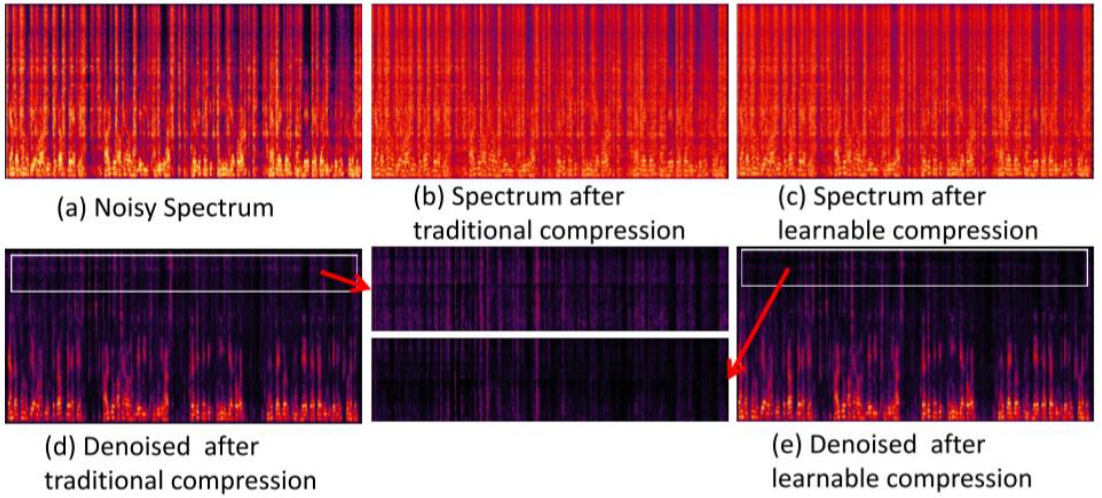


Fig. 6. Comparison on the denoising result on a testing noisy clip for the cases with/without learnable spectrum compression.

In detail, the learnable spectrum compression can be described as

$$Y_{\mathbf{LSC}} = |Y|^{\alpha} e^{j\varphi_Y}$$

where Y and α denote the noisy spectrum and the learnable parameters respectively.

Loss Function

- SI-SNR loss in time-domain
  - scale-invariant source-to-noise ratio

$$\begin{cases} \mathbf{s}_{target} := \frac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle \mathbf{s}}{\|\mathbf{s}\|^2} \\ \mathbf{e}_{noise} := \hat{\mathbf{s}} - \mathbf{s}_{target} \\ \text{SI-SNR} := 10\, log_{10} \frac{\|\mathbf{s}_{target}\|^2}{\|\mathbf{e}_{noise}\|^2} \end{cases}$$

- Complex mean-squared error (MSE) loss in T-F domain

$$\mathcal{L}_{\text{cMSE}} = \frac{1}{T \times F} \sum_{t,f} \left| |X|\, e^{j\varphi_X} - |\hat{X}| e^{j\varphi_{\hat{X}}} \right|$$

- Kullback-Leibler Divergence in T-F domain

$$\mathcal{L}_{\text{KL}} = \frac{1}{T \times F} \sum_{t,f} \hat{X} \cdot log(\frac{\hat{X}}{X})$$

Result

Table 1. Results of various models and ablation experiments on Voice Bank and DEMAND set.

Model	# Para.(M)	PESQ	CSIG	COVL	CBAK	STOI
Noisy	-	1.97	3.35	2.63	2.44	0.921
RNNoise	0.06	2.34	3.40	2.84	2.51	0.922
PercepNet	8	2.73	-	-	-	-
DCCRN	3.7	2.54	3.74	3.13	2.75	0.938
SP	2.76	2.63	3.86	3.23	3.03	0.935
SAF	2.73	2.71	3.94	3.31	3.08	0.937
+ SC	2.73	2.76	3.98	3.36	2.87	0.938
+ LSC	2.73	2.77	3.98	3.35	2.92	0.938
+ CFE/CFD	2.34	2.69	3.90	3.28	**3.08**	0.939
+ SC	2.34	2.77	3.98	3.37	2.87	0.940
+ LSC (*S*-DCCRN)	2.34	**2.84**	**4.03**	**3.43**	2.97	**0.940**

Table 2. MOS and DNSMOS results on DNS-2021 blind test set.

Model	MOS	DNSMOS*
Noisy	1.66	2.94
RNNoise	2.32	3.07
DCCRN	3.30	3.31
SAF	3.20	3.33
S-DCCRN	**3.62**	**3.43**

*: Calculated on downsampled speech (16K Hz)

Demo: https://imybo.github.io/S-DCCRN/