A fully convolutional neural network for complex spectrogram processing in speech enhancement

Author: Zhiheng Ouyang, Hongjiang Yu, Wei-Ping Zhu, Benoit Champagne

Presenter: Hongjiang Yu

Ouline

- Overview
- Proposed method
- Experiments

Overview: DNN-based methods in speech enhancement

- DNN-based methods are widely and are employed as mapping function
  - It maps noisy speech features to certain target
  - Then the target is used to estimate the clean magnitude of speech

Input features

A combination of acoustic features:
Log-power spectrum, MFCC, GFCC, AMS,
RASTA-PLP, etc.

DNN

Output (Target) of DNN

- Mask
  - Ideal Binary Mask (IBM)
  - Ideal Ratio Mask (IRM)
  - Complex Ideal Ratio Mask (CIRM)
- Log-power spectrum

- Fully-connected DNN is most often adopted, 3 to 4 hidden layers, Units >= 1024 per layer

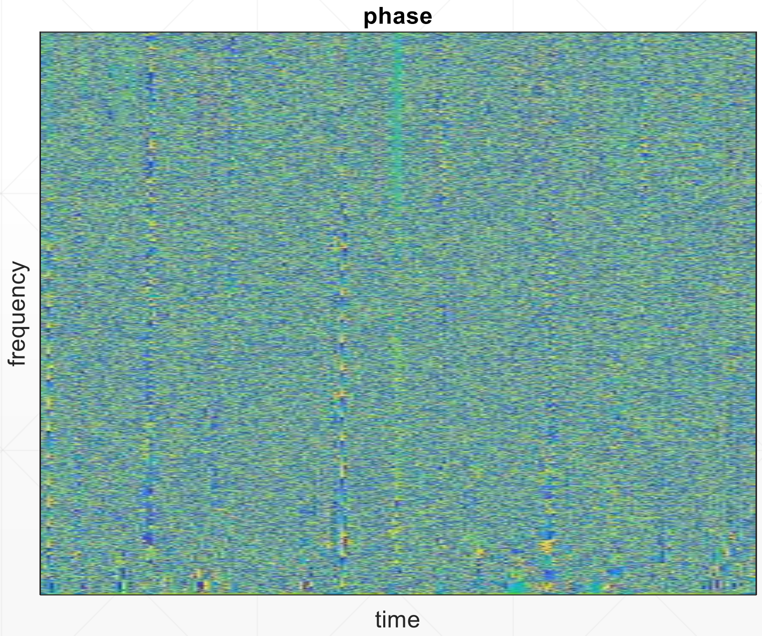
Overview: DNN-based methods in speech enhancement - limitations

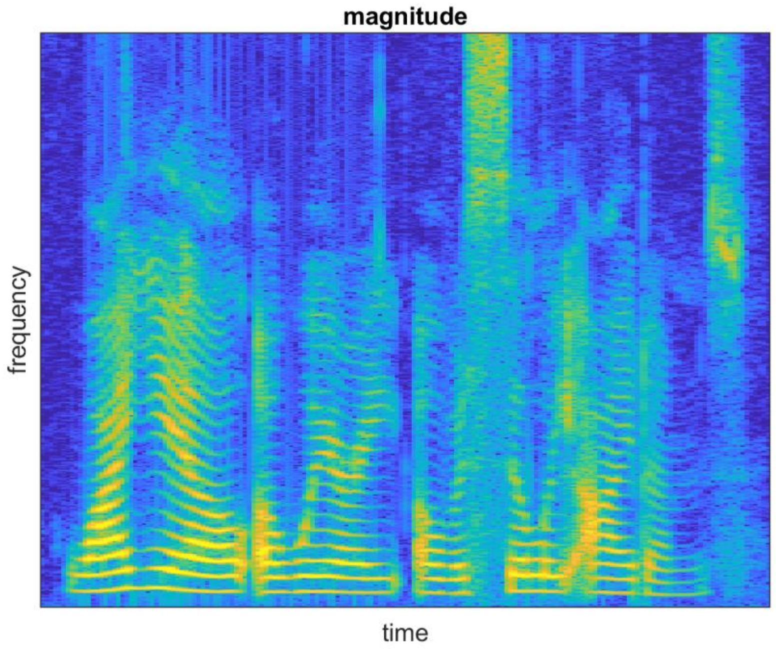
- On one hand, fully-connected DNN is classic but often comes with a high complexity
  - 4 hidden layers, 1024 units per layer → number of parameter >= 3.15 million
- On the other hand, usually noisy spectral phase is directly used to reconstruct the speech
  - Noise like babble could bring much distortion to phase
  - Phase estimation is hard due to the characteristic of phase
- To address the first issue, fully-connected DNN has been replaced by CNN or RNN in some work.
- For the second problem, complex spectrogram estimation has been brought up as a walk-around of phase estimation.

Overview: Spectrogram

- Most methods process magnitude in STFT domain
  - $Y(f,t) = STFT[y(t)] = Y_r(t,f) + jY_i(t,f)$

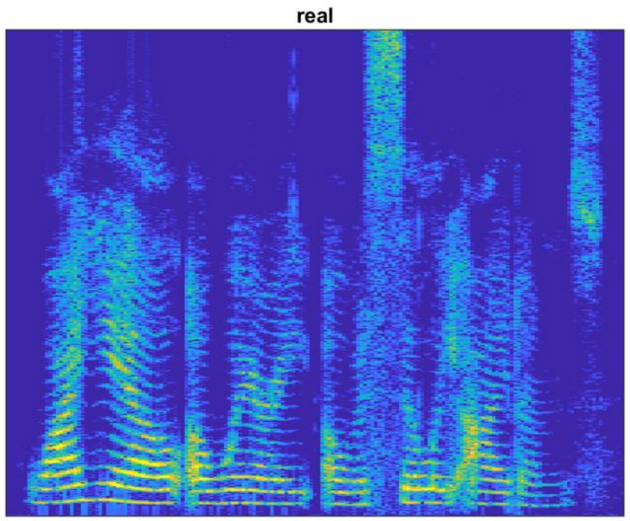
    $Y_r(t,f)$: spectrogram of real part $Y_i(t,f)$: spectrogram of imaginary part
  - Magnitude: $\sqrt[2]{Y_r^2 + Y_i^2}$ Phase: $\arctan(\frac{Y_i}{Y_r})$

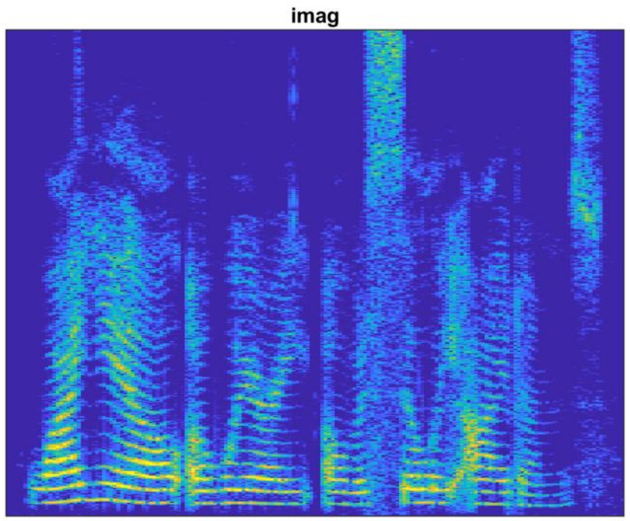


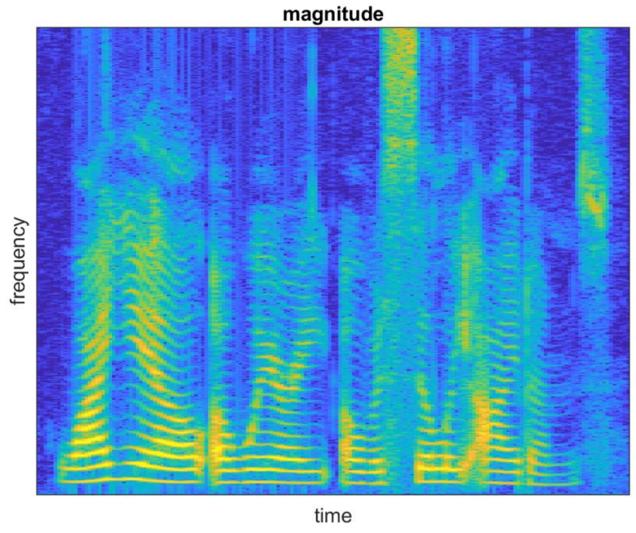


Overview: Complex spectrogram estimation

- Real and imaginary spectrogram of clean speech are similar to magnitude spectrogram
- Through complex spectrogram estimation, we are processing spectral magnitude and phase at the same time.







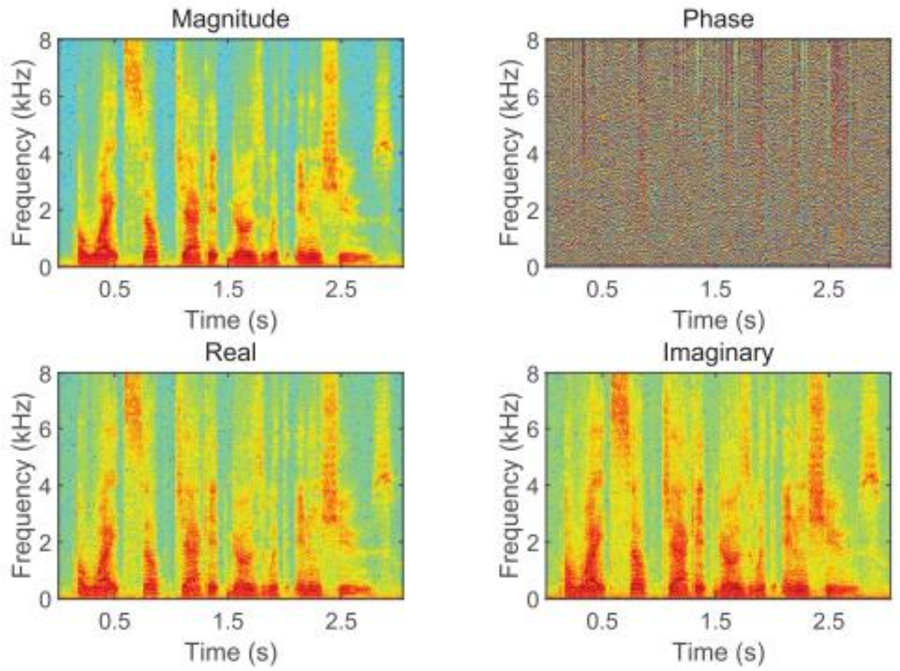
Overview: DNN-based complex spectrogram estimation

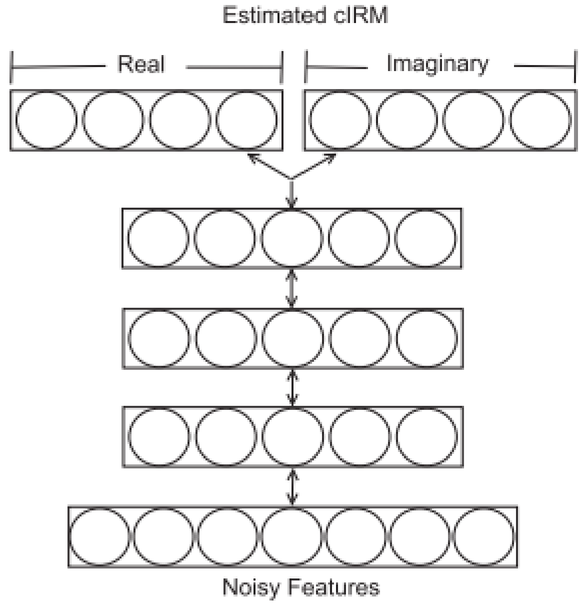
- CIRM is an alternative to complex spectrogram estimation
  - DNN-based CIRM [4]
    - DNN has a fully-connected structure.
  - Outperforms IRM, where only the spectral magnitude is considered

- CNN has been employed for complex spectrogram estimation [3].
  - Input & Output: complex spectrogram
  - Fully-connected Layer is still used
  - Outperformed by fully connected DNN.

Overview: CIRM

- A DNN is used for learning the mapping from the noisy features to the complex ideal ratio mask



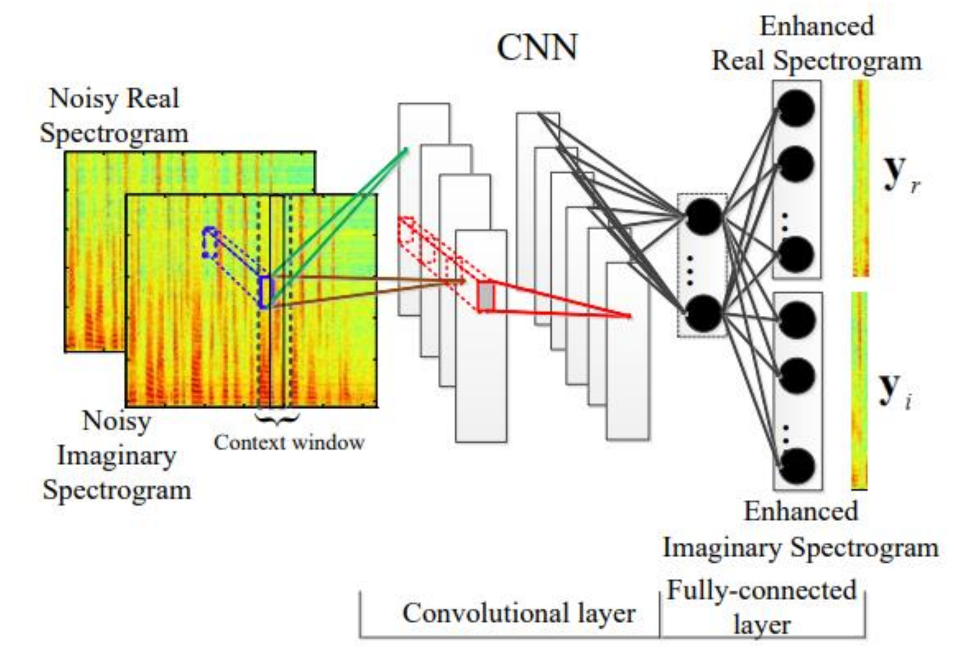


Overview: DNN-based complex spectrogram estimation

- CIRM is an alternative to complex spectrogram estimation
  - DNN-based CIRM [4]
    - DNN has a fully-connected structure.
  - Outperforms IRM, where only the spectral magnitude is considered

- CNN has been employed for complex spectrogram estimation [3].
  - Input & Output: complex spectrogram
  - Fully-connected Layer is still used
  - Outperformed by fully connected DNN.

Overview: RI-CNN

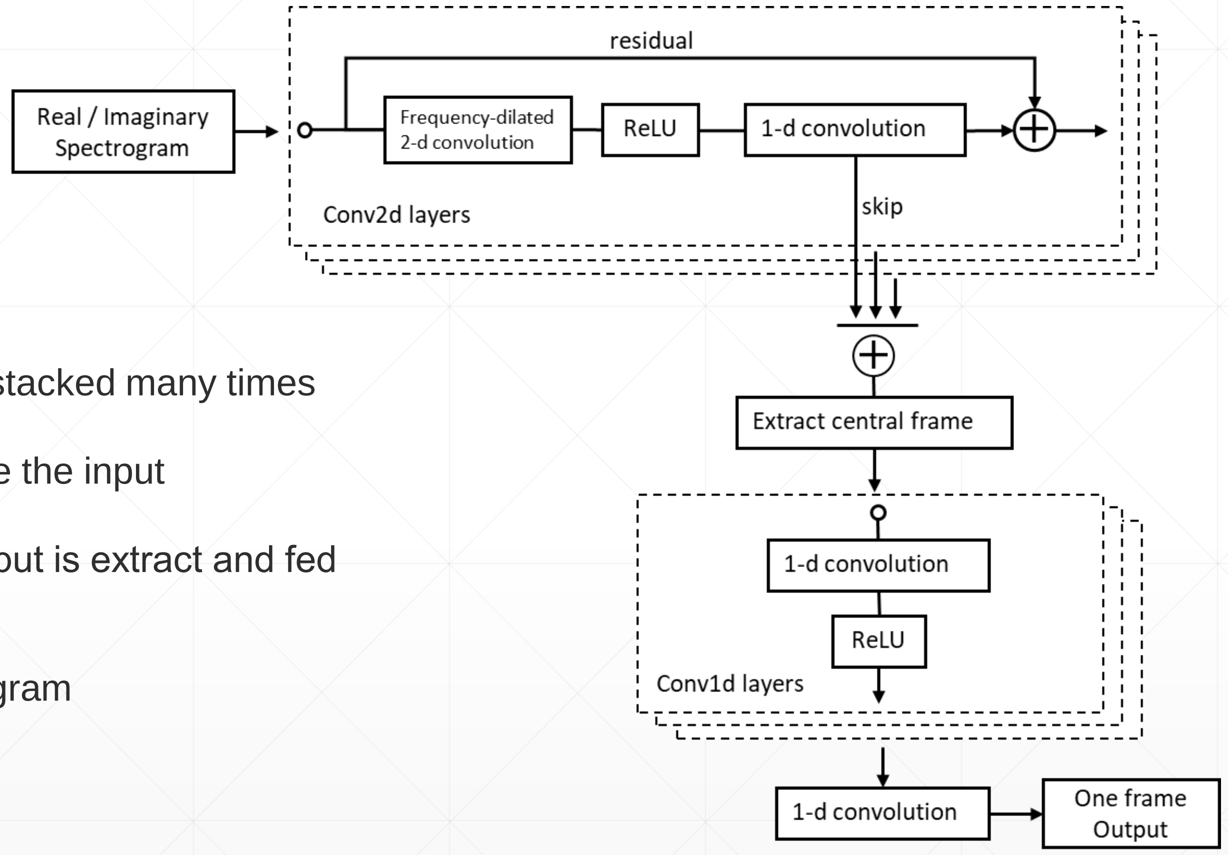


- CNN and DNN are together used for estimating the clean RI spectrograms from the noisy ones
- The estimated clean RI spectrograms are directly used to synthesize enhanced speech

Overview: Proposed method

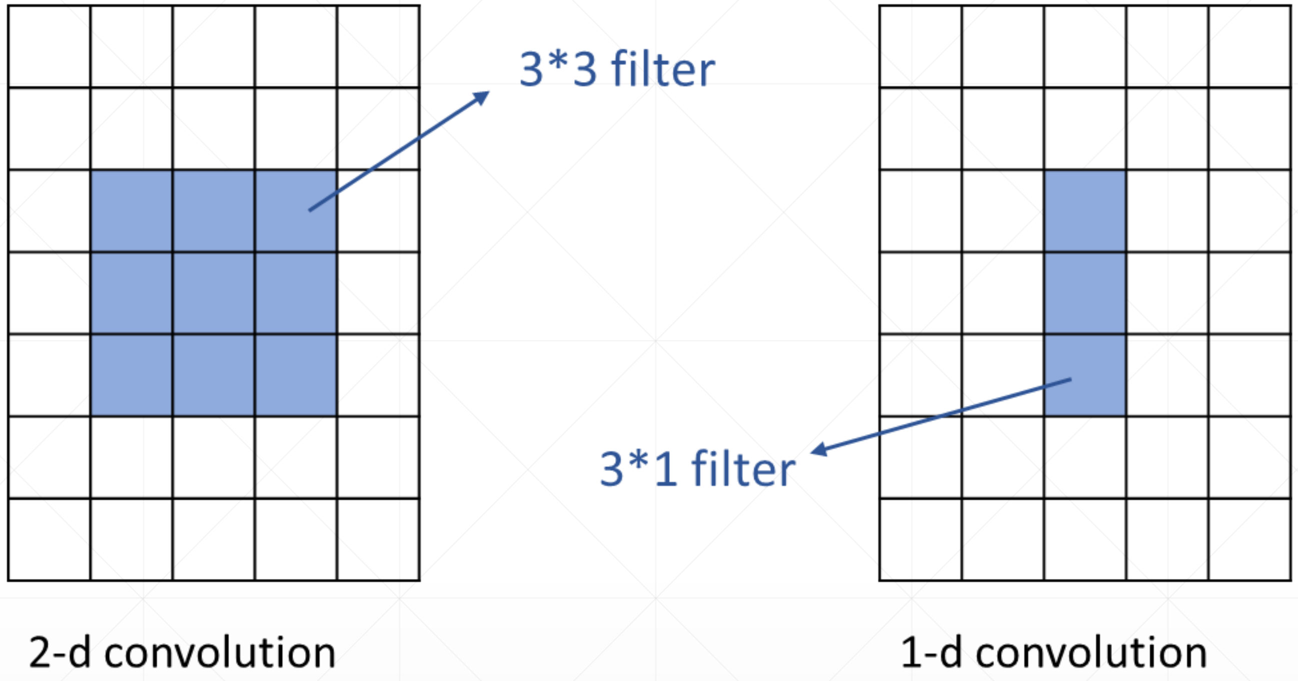
- A new CNN structure is proposed for complex spectrogram estimation.
- Compared with the previous work, the proposed CNN is fully convolutional, which consists of frequency-dilated 2-d convolution and 1-d convolution.

Proposed CNN Architecture



- The Conv2d layers and Conv1d layers are stacked many times
- 13 frames of real/imaginary spectrogram are the input
- The central frame of the Conv2d layers' output is extract and fed into the Conv1d layer
- The final output is one frame of the spectrogram

2-d and 1-d Convolution



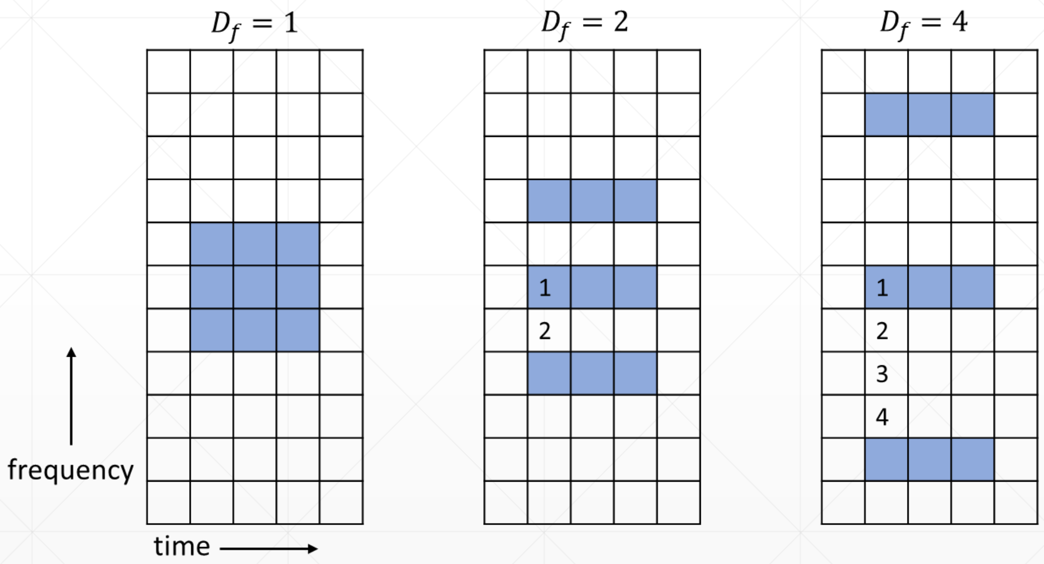
The 1-d convolution is a special case where the filter is only of 1-dim.

1-d convolution is applied along frequency axis.

It is more efficient than 2-d convolution when the goal is to increase the size of receptive field along frequency axis

Frequency-dilated convolution

- The size of input is usually:
  - Time axis: 10 to 20
  - Frequency axis: a few hundreds
- Dilation is applied on frequency axis to accommodate the size of input on frequency as it yields a exponential increment.



Frequency dilation is employed to produce a large receptive field with small filters. Hence the proposed CNN could be configured with fewer parameters while still achieving a competitive performance

Experiments Setup

- Size of input spectrogram: 13*251 (13 frames for time, 251 point for frequency)
- 2-d frequency-dilated convolution is applied to increase the size of receptive field in frequency
  - 500-point DFT $\rightarrow$ length of 251 in frequency
  - Filter size of 5 in frequency, stacked 6 times $\rightarrow$ receptive field size of 253
    - Without frequency dilation, the filter size has to be 43 in order to obtain the same receptive field size
- No need for dilation on time axis
  - 13-frame input
  - Filter size of 3 on time axis, stacked 6 times $\rightarrow$ receptive size of 13, just enough to cover all input frames

Experiments Setup

- Dataset: TIMIT
  - 780 utterances are used for the training and 90 utterances used for testing
- Metric: PESQ, Segmental Signal to Noise Ratio (SSNR)
- Comparison methods: CIRM and RI-CNN
- Noise: babble, street, factory, restaurant
- Window length: 500 (Hamming, 50 percent overlap)
- SNR
  - training: -5, 0, 5, 10
  - Testing: -6, 0, 6, 12

Experiments: Comparison with different models

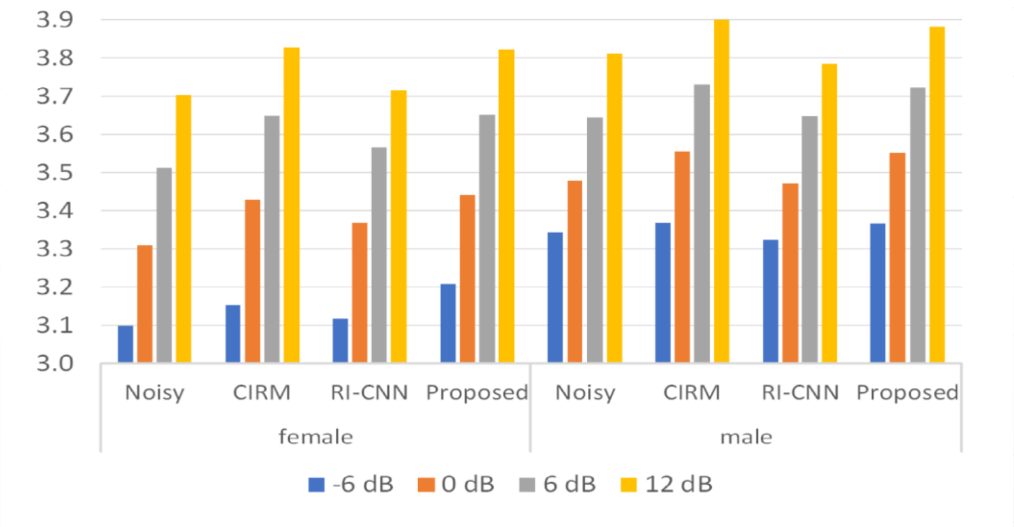
- Comparison with different models (RI-CNN [3], CIRM [4])

metrics	PESQ				SSNR			
SNR	**-6 dB**	**0 dB**	**6 dB**	**12 dB**	**-6 dB**	**0 dB**	**6 dB**	**12 dB**
unprocessed	1.296	1.674	2.124	2.549	-12.454	-8.046	-2.722	2.994
CIRM	1.740	2.267	2.706	3.071	$\mathbf{-0.874}$	**2.242**	5.042	7.504
RI-CNN	1.723	2.018	2.477	2.711	-2.891	0.188	2.710	4.415
proposed	**1.861**	**2.337**	**2.741**	**3.079**	-1.723	2.083	**5.629**	**8.948**

- Proposed model works pretty well considering the number of parameter is kept rather small (243 k), compared with RI-CNN (775k) and CIRM (3.87 M)

Experiments: Evaluation of phase processing

- For comparison, clean magnitude is combined with either estimated phase or noisy phase
  - Female speech benefits more than male speech. A maximal improvement of 0.15 is observed on female speech
- For all three methods, the improvement on PESQ is rather limited when use a combination of noisy magnitude and noisy phase



Experiments: Comparison with different model configurations

243k

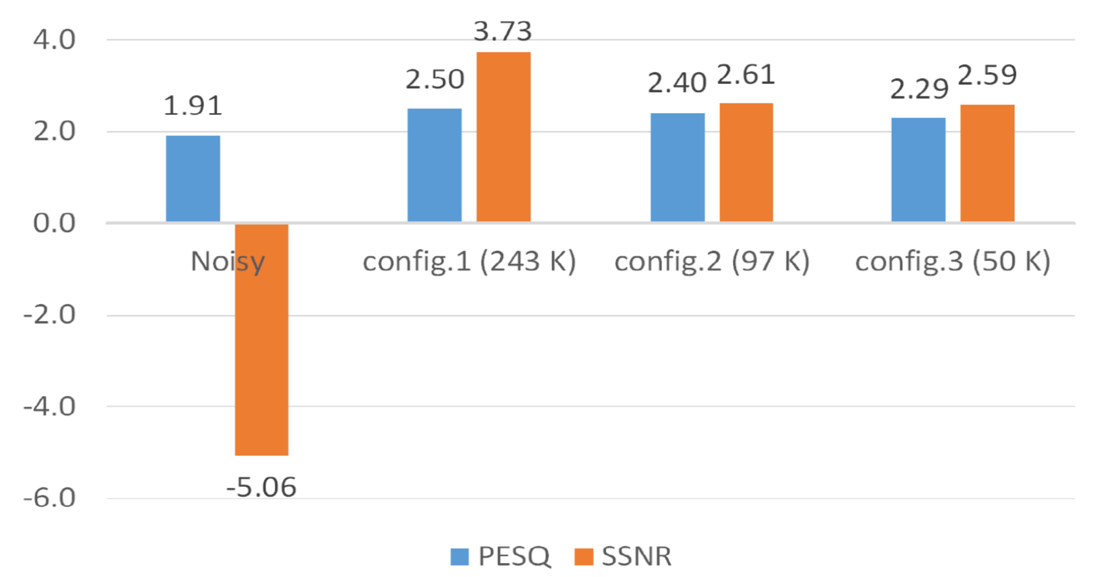
Layer name	Filter name	Height	Width	Channel
Conv2d	dilated 2d	5	3	48
	1d-skip	1	1	48
	1d-residual	1	1	48
Conv1d	1d	3	1	96
Output	1d-real	3	1	1
	1d-imag	3	1	1

97k

Layer name	Filter name	Height	Width	Channel
Conv2d	dilated 2d	5	3	32
	1d-skip	1	1	24
	1d-residual	1	1	24
Conv1d	1d	5	1	64
Output	1d-real	17	1	1
	1d-imag	17	1	1

50k

Layer name	Filter name	Height	Width	Channel
Conv2d	dilated 2d	5	3	32
	1d-skip	1	1	16
	1d-residual	1	1	16
Conv1d	1d	1	1	48
Output	1d-real	17	1	1
	1d-imag	17	1	1



Conclusion

- In this study, we have proposed a fully convolutional neural network with frequency-dilated 2-d convolution for complex spectrogram processing.

- we have demonstrated that the proposed CNN performs very well for complex spectrogram estimation, and results in clean phase estimation.

- We also have paid attention to the memory efficiency of the proposed CNN by considering limited number of parameters and memory footprint, leading to a tradeoff between the model complexity and the achievable performance.

Reference

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” CoRR, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[2] S. R. Park and J. Lee, “A fully convolutional neural network for speech enhancement,” CoRR, vol. abs/1609.07132, 2016. [Online]. Available: http://arxiv.org/abs/1609.07132

[3] S. Fu, T. Hu, Y. Tsao, and X. Lu, “Complex spectrogram enhancement by convolutional neural network with multi-metrics learning,” in 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP), Sept 2017, pp. 1–6.

[4] D. S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for monaural speech separation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 3, pp. 483–492, March 2016.