Speech Enhancement with Neural Homomorphic Synthesis

Wenbin Jiang, Zhijun Liu, Kai Yu, Fei Wen

MoE Key Lab of Artificial Intelligence, AI Institute

X-LANCE Lab, Department of Computer Science and Engineering

Shanghai Jiao Tong University, Shanghai, China

Problem Formular: Monaural Speech Enhancement

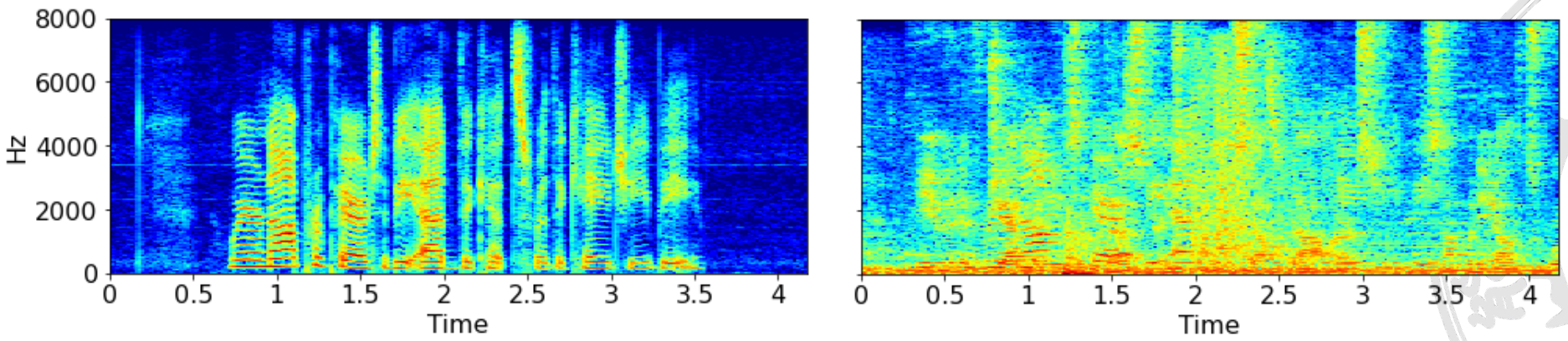
- **Problem formular**
  - Let $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{n}$ denote noisy speech, clean speech, and additive noise
  - The corresponding signal model is

$$\mathbf{x} = \mathbf{y} + \mathbf{n}$$

  - The goal of speech enhancement is to get an estimation $\hat{\mathbf{y}}$ of the clean speech $\mathbf{y}$
  - Given observed noisy speech $\mathbf{x}$, finding a function f such that

$$\hat{\mathbf{y}} = f(\mathbf{x}) \approx \mathbf{y}$$

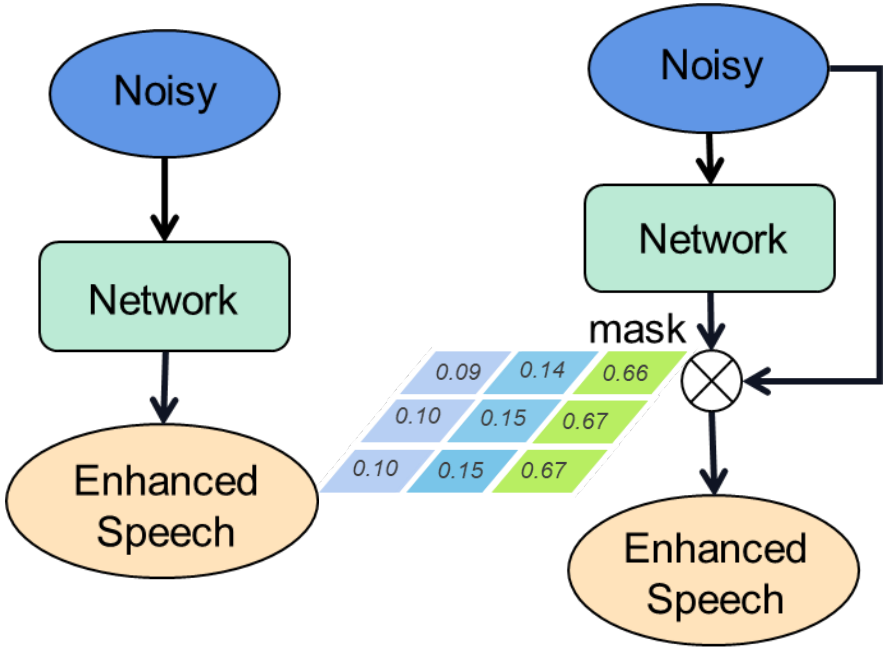
- **Example**



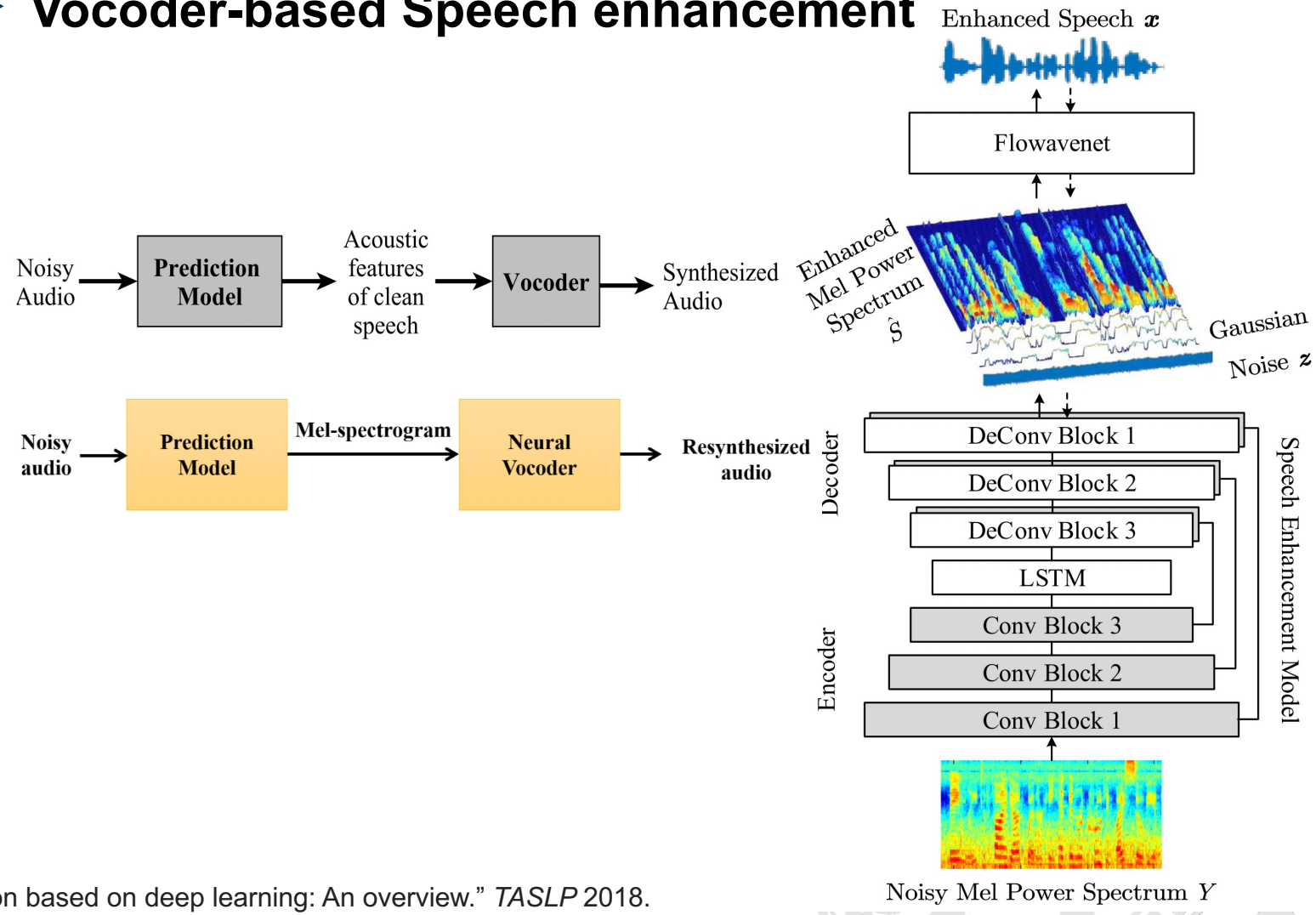
Left: clean speech spectrogram. Right: noisy speech spectrogram

Related Works

Mapping or Masking



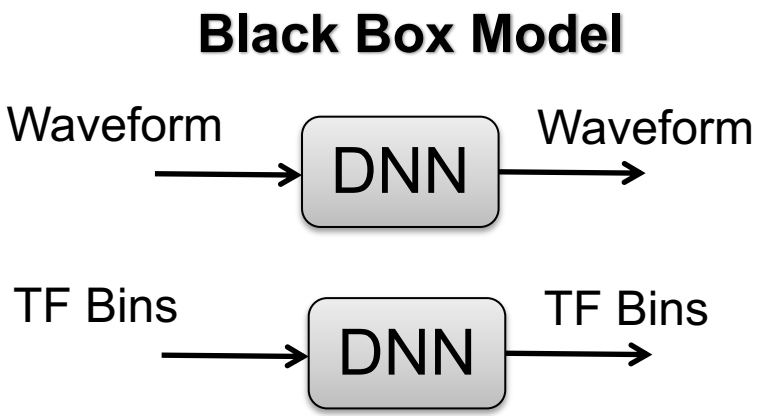
Vocoder-based Speech enhancement



- Wang, DeLiang, and Jitong Chen. “Supervised speech separation based on deep learning: An overview.” *TASLP* 2018.
- Maiti, Soumi, and Michael I. Mandel. "Speech denoising by parametric resynthesis." *ICASSP* 2019.
- Du, Zhihao, Xueliang Zhang, and Jiqing Han. “A joint framework of denoising autoencoder and generative vocoder for monaural speech enhancement.” *TASLP* 2020.

Challenges of the existing methods

- Most deep learning-based speech enhancement methods operate directly on time-frequency representations or learned features without making use of the model of speech production.



- The computational complexity of the vocoder-based methods is too high to be applied in practice.

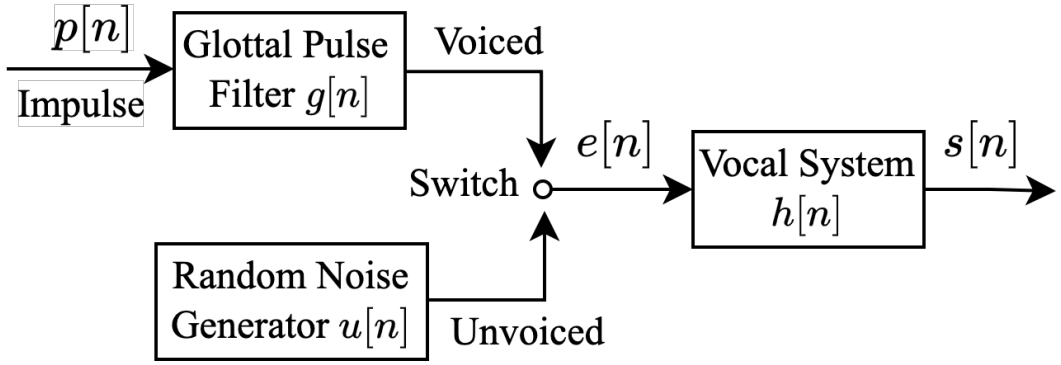
Number of model parameters of Typical vocoders

Vocoder	#Params
WaveNet	21.56 million
WaveGlow	87.88 million
HiFi-GAN	13.92 million

- We need a speech enhancement method with clearer physical meaning and lower computational complexity

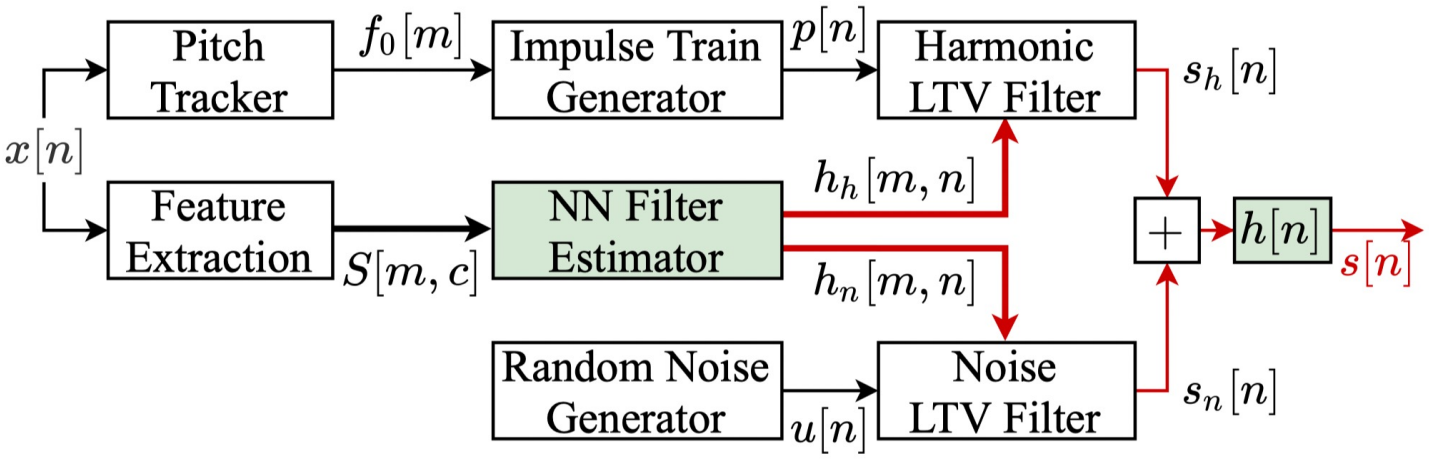
Motivation

- **Source-filter Model of Speech Production**



The discrete-time speech signal $s[n]$ is generated by a convolution of the excitation signal $e[n]$ and the vocal system $h[n]$

- **Neural Homomorphic Vocoder**



Contributions

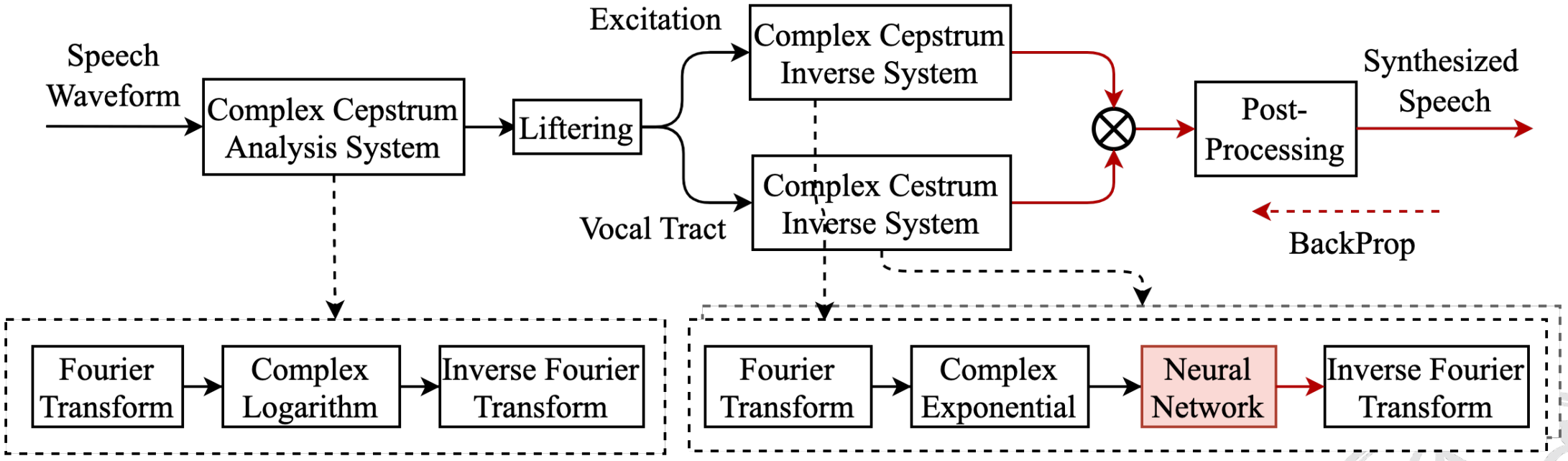
Integrating speech enhancement and the DSP-based vocoder into an ensemble has not been studied yet. Our contributions are as follows:

- We propose a new speech enhancement method that combines the advantages of DSP-based vocoder and complex-valued neural network based spectrum denoiser

- We investigate numerous loss functions and found that multi-resolution STFT loss benefits speech enhancement

Proposed Method

Block Diagram



1. Segment speech waveform into frames
2. Transform time-domain signal into cepstral-domain via complex cepstrum analysis
3. Apply liftering to get the excitation and vocal tract

1. Transform the excitation and vocal tract cepstrum into time-domain via complex cepstrum inverse pipe-line
2. Apply circular convolution and post-processing to get the synthesized speech

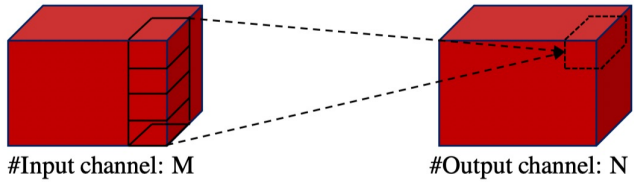
Excitation and vocal tract estimation with complex neural network

Complex neural network

- **Use complex neural networks to process the complex-valued spectrum**
  - Complex neural network is an extension of the real-valued neural network
  - Let $\mathbf{x} = \mathbf{x}_r + j\mathbf{x}_i$ be the input complex vector, and $\mathbf{W} = \mathbf{W}_r + j\mathbf{W}_i$ be the weight matrix of the complex-valued Conv2d
  - The complex-valued convolution is defined as

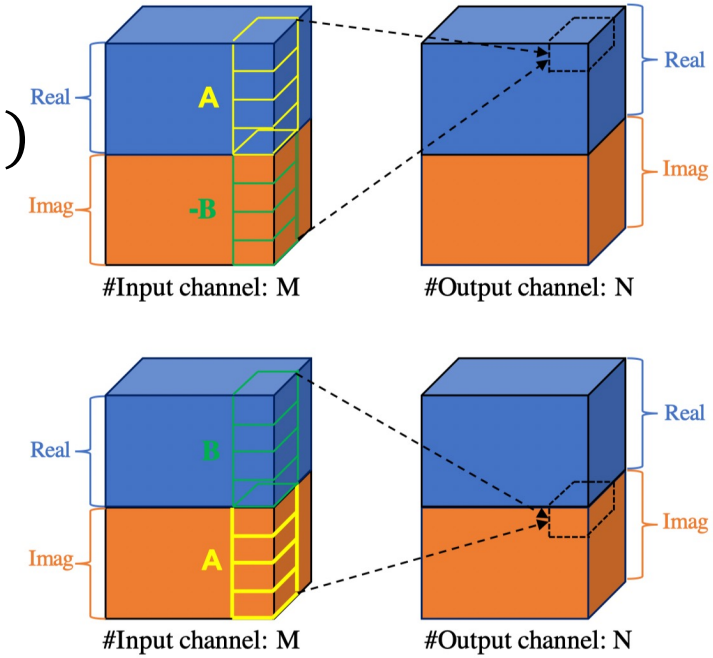
$$\mathbf{W} * \mathbf{x} = (\mathbf{W}_r * \mathbf{x}_r - \mathbf{W}_i * \mathbf{x}_i) + j(\mathbf{W}_i * \mathbf{x}_r + \mathbf{W}_r * \mathbf{x}_i)$$

(a) Real-valued Convolution



#Parameters = Filter Size × M × N

(b) Complex-valued Convolution



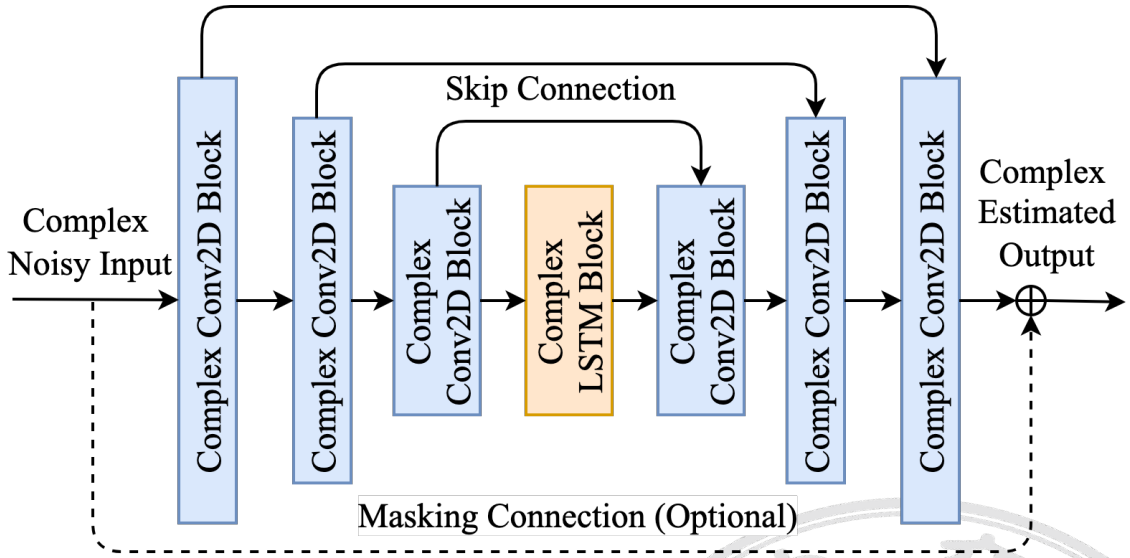
#Parameters = 2 × Filter Size × M × N

- Choi, Hyeong-Seok, et al. "Phase-aware speech enhancement with deep complex u-net." *ICLR* 2018.
- Hu, Yanxin, et al. “DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement.” *InterSpeech* 2020.

Excitation and vocal tract estimation with complex neural network

Complex neural network

- The network follows an U-Net architecture and applies complex recurrent network blocks for temporal modeling
- When the connection is applied, the output of the neural network is considered to be a mask , an additional bounding process is applied



$$\hat{\mathbf{M}}_{t,f} = \left|\hat{\mathbf{M}}_{t,f}\right| \cdot e^{j\theta_{\hat{\mathbf{M}}_{t,f}}} = \hat{\mathbf{M}}^{mag}_{t,f} \cdot \hat{\mathbf{M}}^{phase}_{t,f}$$

$$\hat{\mathbf{M}}^{mag}_{t,f} = tanh(|\mathbf{O}_{t,f}|),\ \hat{\mathbf{M}}^{phase}_{t,f} = \mathbf{O}_{t,f} / |\mathbf{O}_{t,f}|$$

- Choi, Hyeong-Seok, et al. "Phase-aware speech enhancement with deep complex u-net." *ICLR* 2018.

Excitation and vocal tract estimation with complex neural network

Loss functions

- **We consider the following loss functions for neural network training**
  - Scale Invariant SNR loss

$$L_{SI-SNR}(\mathbf{y}, \hat{\mathbf{y}}) = -10 \log 10 \left(\frac{\|\mathbf{y}_{\text{target}}\|_2^2}{\|\mathbf{e}_{\text{noise}}\|_2^2} \right)$$

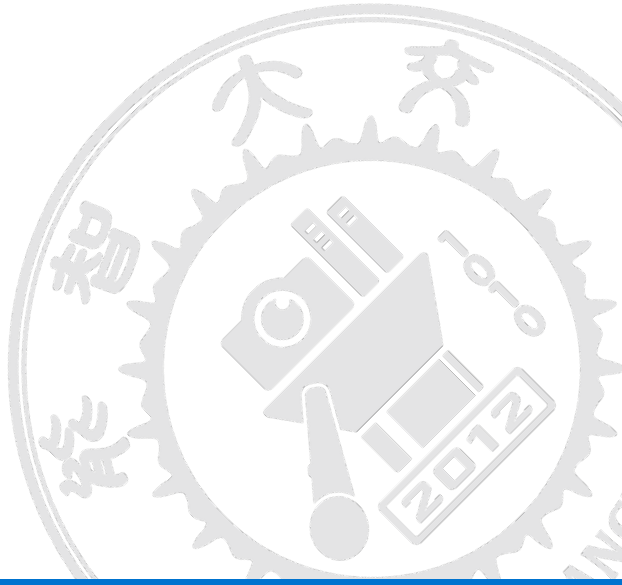
  - Weighted SDR loss

$$L_{wSDR}(\mathbf{x}, \mathbf{y}, \hat{\mathbf{y}}) = \alpha L_{SDR}(\mathbf{y}, \hat{\mathbf{y}}) + (1-\alpha) L_{SDR}(\mathbf{n}, \hat{\mathbf{n}})$$

  - Multi-resolution STFT loss

$$L_{MR-STFT} = -\left(\|\mathbf{y} - \hat{\mathbf{y}}\|_1 + \sum_{i=1}^{I} L_{stft}^{(i)}(\mathbf{y}, \hat{\mathbf{y}}) \right)$$

$$L_{stft}^{(i)}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\| |\mathbf{Y}_i| - |\hat{\mathbf{Y}}_i| \|_F}{\| |\mathbf{Y}_i| \|_F} + \| \log |\mathbf{Y}_i| - \log |\hat{\mathbf{Y}}_i| \|_1$$



Experiments

Datasets

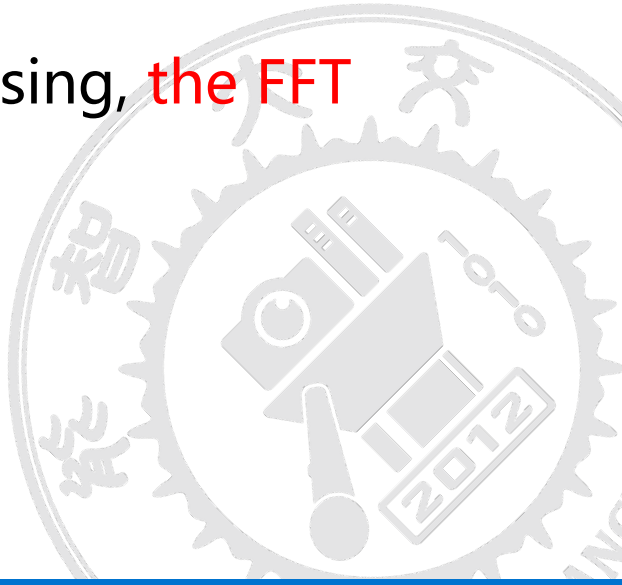
- Speech: Chinese Standard Mandarin Speech Corpus[1]
- Noise: DEMAND[2]
- Four noise categories and one environment of each are selected as the training, validation, and seen-noise test set
- Another two noise categories are selected as unseen test data to evaluate the noise generalization
- The SNR levels for mixing the noisy speech are randomly sampled from a uniform distribution [-5dB, 10dB]

1. https://www.data-baker.com/en/#/data/index/source
2. https://zenodo.org/record/1227121

Experiments

Setups

- All the utterances are framed by a hamming window with a length of 32 ms and a hop size of 8 ms, and the FFT length is 512.
- The quefrency for separating the excitation and vocal tract is 29.
- The network architectures follow the setting of the DCCRN
- The number of channel, kernel size and stride are set to {16, 32, 64, 128, 128, 128}, (5,2) and (2,1), respectively.
- For the proposed method, in order to alleviate cepstrum aliasing, the FFT size is increased to 2048.



Experiments

Results

Table 1. PESQ and eSTOI scores of the compared methods for seen noise types and unseen noise categories.

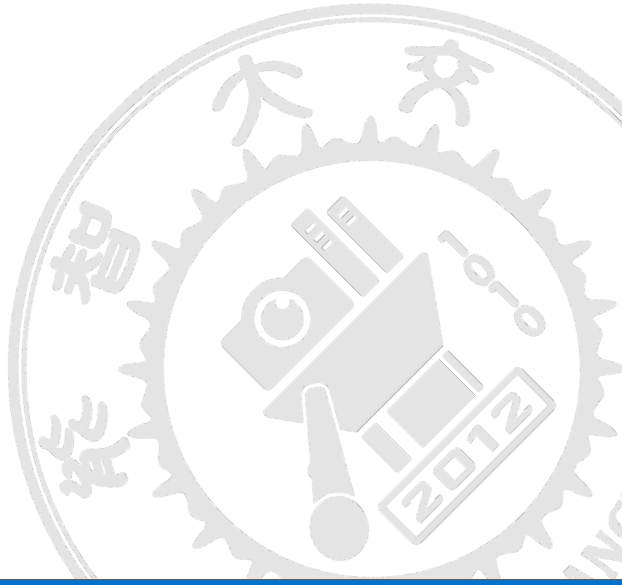
	Metrics	KITC	MEET	CAFE	BUS	Seen	Nature	Street	Unseen	Overall
Noisy	PESQ	1.345	1.117	1.098	1.821	1.345	1.211	1.180	1.196	1.270
	eSTOI	0.925	0.721	0.682	0.930	0.815	0.812	0.793	0.803	0.809
OMLSA	PESQ	2.181	1.195	1.256	2.577	1.802	1.757	1.591	1.674	1.738
	eSTOI	0.936	0.732	0.721	0.947	0.834	0.849	0.835	0.842	0.838
DCCRN-SI-SNR	PESQ	2.993	2.370	2.163	3.199	2.681	2.397	2.430	2.413	2.547
	eSTOI	0.971	0.930	0.906	0.975	0.946	0.924	0.930	0.927	0.936
DCCRN-wSDR	PESQ	2.827	2.325	2.142	2.965	2.565	2.297	2.339	2.318	2.441
	eSTOI	0.966	0.927	0.906	0.969	0.942	0.920	0.926	0.923	0.932
DCCRN-MR-STFT	PESQ	3.270	2.475	2.280	3.229	2.814	2.495	2.518	2.507	2.660
	eSTOI	0.970	0.932	0.914	0.972	0.947	0.928	0.933	0.931	0.939
NHS-SE	PESQ	3.444	2.850	2.598	3.582	**3.119**	2.708	2.859	**2.783**	**2.951**
	eSTOI	0.969	0.943	0.925	0.974	**0.953**	0.933	0.944	**0.939**	**0.946**

- The SI-SNR loss and wSDR loss yield comparable results in term of PESQ and eSTOI
- The MR-STFT loss outperforms the SI-SNR and wSDR losses in terms of PESQ
- The proposed NHS-SE obtains the highest score in all metrics.
- For the unseen noise types, the performance of all methods is degraded to some extent. However, the proposed method still outperforms all the others.

Experiments

Audio samples (online)

https://jiang-wenbin.github.io/NHS-SE/



Experiments

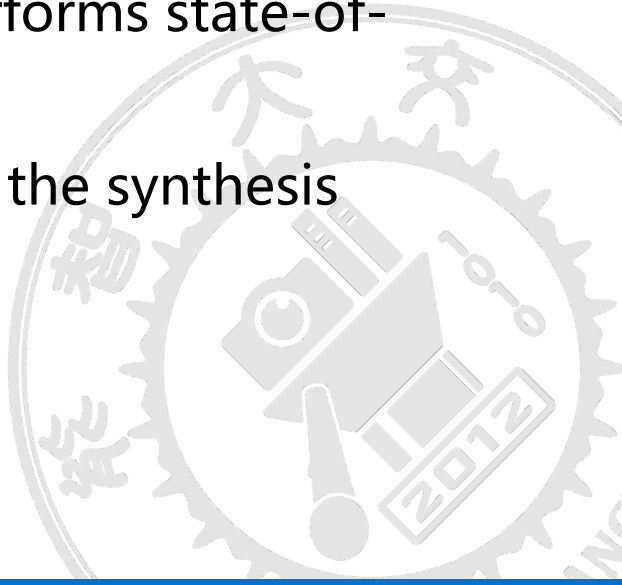
Results: some failures

- We also conducted extra experiments using the proposed NHS-SE model with the SI-SNR and wSDR losses. However, the both models failed to converge.
- Synthesis-based method will cause phase mismatch, the proposed method failed to improve the SI-SNR score.
- Although the proposed method obtained the highest PESQ and eSTOI scores, there is some artificial noise in some case.

Clean Noisy DCCRN NHS-SE

Conclusions and Future Works

- We proposed a novel speech enhancement method based on homomorphic analysis and synthesis
- It makes use of the advantages of the classical vocoder method and the recently popular speech enhancement method based on complex-valued neural networks.
- The results demonstrated that the multi-resolution STFT loss performs better than the others in terms of PESQ and eSTOI.
- Using the multi-resolution STFT loss, the proposed method outperforms state-of-the-art methods on both seen noise and unseen noise.
- Future works include removing some artificial noise introduced by the synthesis procedure and studying speaker generalization of the models.



Thanks!