A Cascaded Speech Enhancement for Hearing Aids in noisy-reverberant Conditions

Clarity 2021

Xi Chen, Yupeng Shi, Wei Xiao, Tingzhao Wu, Meng Wang,
Shidong Shang, Nengheng Zheng, Qinglin Meng

September 17th 2021

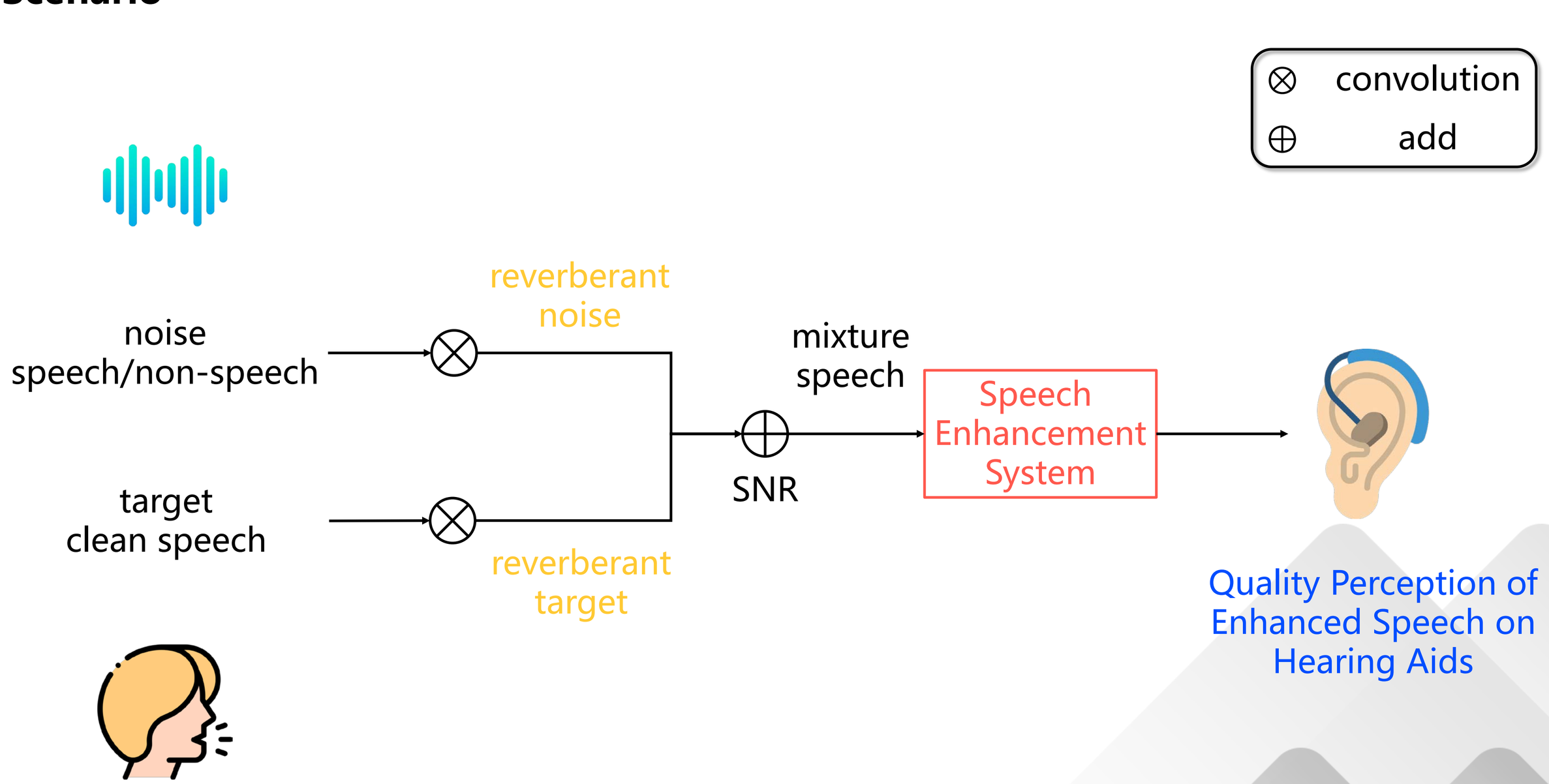
01 Scenario

02 Proposed System

03 Evaluations

04 Conclusions

Scenario


convolution
add
noise
speech/non-speech
reverberant
noise
target
clean speech
reverberant
target
SNR
mixture
speech
Speech
Enhancement
System
Quality Perception of
Enhanced Speech on
Hearing Aids

Motivation

Challenges

- Largely varying **noise** types and SNRs
- **Reverberation** with varying room conditions
- Enhancement for **hearing aid** users

Proposed Solutions

- Deep learning based **denoiser**
- Signal processing based **dereverberation**
- **Equalization** post processing for hearing impairment

01

Signal Model

02

Proposed System

03

Evaluations

04

Conclusions

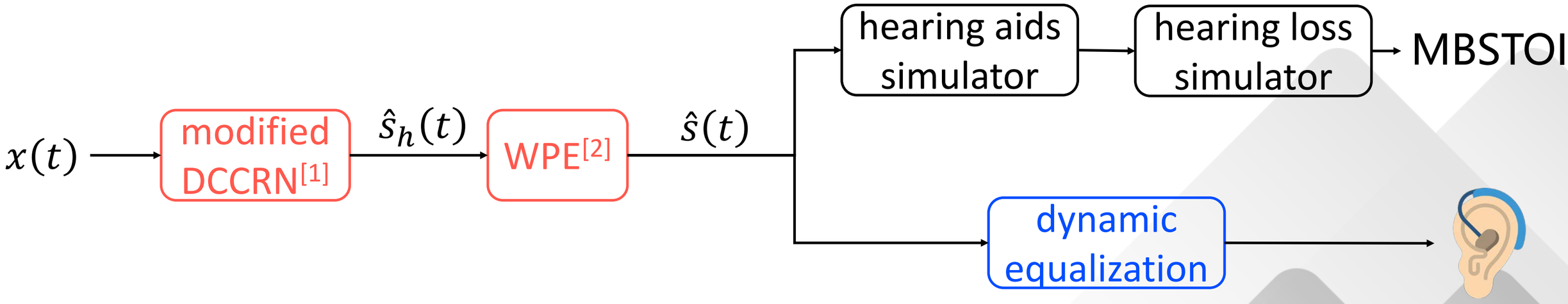
System Overview

$$s_h(t) = s(t) * h_s(t)$$

$$n_h(t) = n(t) * h_n(t)$$

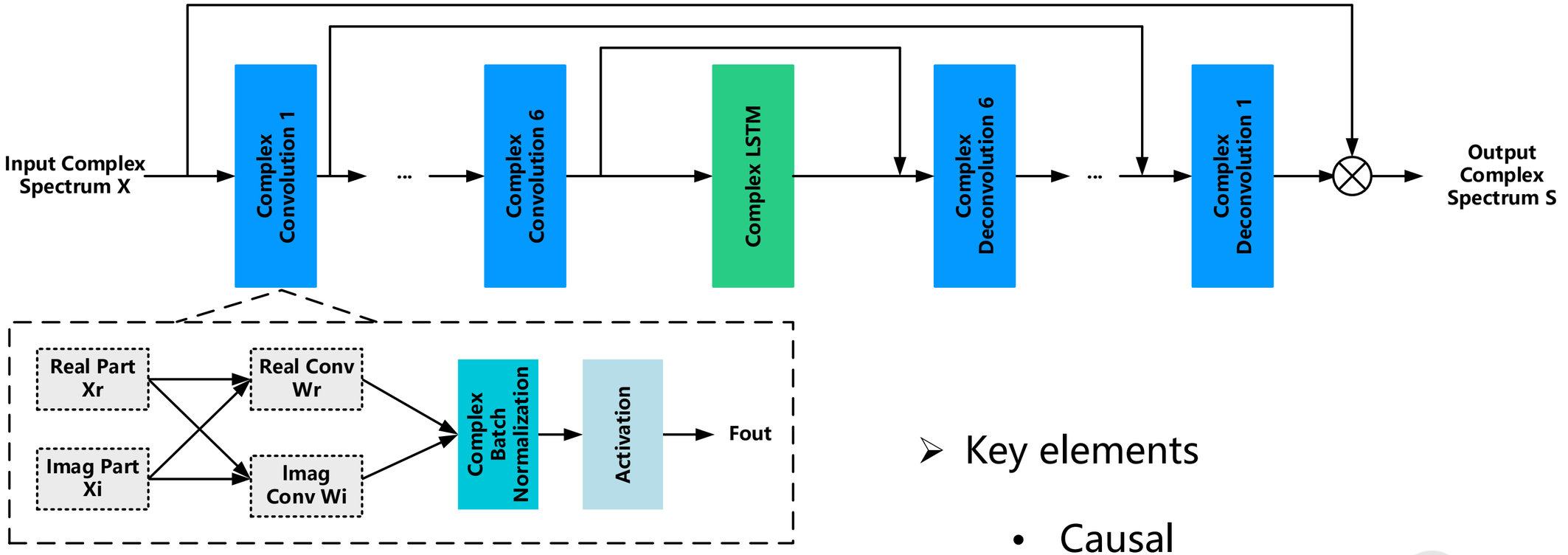
$$x(t) = s_h(t) + n_h(t)$$

$x(t)$	:	mixture speech signal
$s(t)$	:	target speech signal
$*$	:	convolution operation
$h_s(t)$	:	room impulse response for target
$n(t)$	:	noise signal
$h_n(t)$	:	room impulse response for noise
$s_h(t)$	:	reverberant target signal
$n_h(t)$	:	reverberant noise signal



[1] Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., ... & Xie, L. (2020). DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. *arXiv preprint arXiv:2008.00264*.
[2] Drude, L., Heymann, J., Boeddeker, C., & Haeb-Umbach, R. (2018, October). NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing. In Speech Communication; 13th ITG-Symposium (pp. 1-5). VDE.

Modified DCCRN Overview

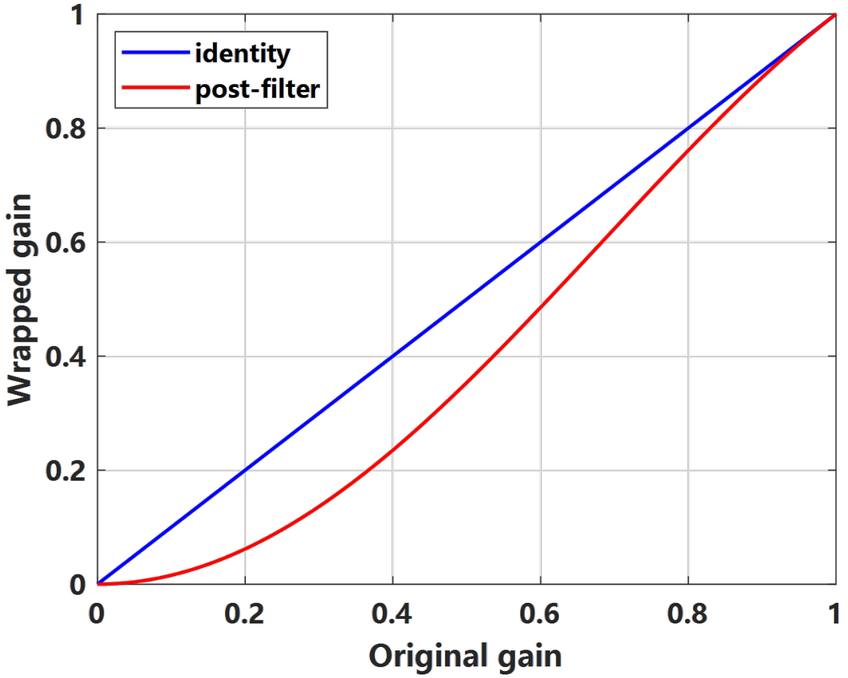


- Short-time Fourier transform (STFT)
  - frame length = 32 ms
  - frame shift = 20 ms

- Key elements
  - Causal
  - Relative lightweight
  - Complex spectrum masks
  - Joint loss function

Complex Masks & Joint Loss Function

- Input spectrum: $X = X_{mag}e^{j\varphi_X}$
- Output spectrum: $S = S_{mag}e^{j\varphi_S}$
- Complex mask: $M = M_{mag}e^{j\varphi_M}$
- Envelope postfilter [3]
  - Apply a nonlinear function to gains
  - Normalize to boost cleaner bands



$$S_{mag} = X_{mag} \times M_{mag} \times \sin\left(\frac{\pi}{2} M_{mag}\right) \times G$$

$$\varphi_S = \varphi_X + \varphi_M$$

- Loss function: Torch-stoi[4] + SI-SNR loss

[3] Valin, J. M., Isik, U., Phansalkar, N., Giri, R., Helwani, K., & Krishnaswamy, A. (2020). A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech. arXiv preprint arXiv:2008.04259.

[4] Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2010, March). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In 2010 IEEE international conference on acoustics, speech and signal processing (pp. 4214-4217). IEEE.

Dynamic Equalization

FIG6 Fitting Strategy[5]	
SPL ≤ 40 dB	0~20 dB HL: $G = 0$ 20~60 dB HL: $G = TH - 20$ TH ≥ 60 dB HL: $G = TH - 20 - 0.5 \times (TH - 60)$
40 dB < SPL ≤ 65 dB	0~20 dB HL: $G = 0$ 20~60 dB HL: $G = 0.6 \times (TH - 20)$ TH ≥ 60 dB HL: $G = 0.8 \times (TH - 23)$
65 dB < SPL ≤ 90 dB	0~40 dB HL: $G = 0$ TH ≥ 40 dB HL: $G = 0.1 \times (TH - 40)^{1.4}$

SPL : sound pressure level
HL : hearing loss
G : gain at each frequency band

[5] Killion, M. C. (1993). The 3 types of sensorineural hearing loss: Loudness and intelligibility considerations. Hear. J., 46(11), 31-36.

01 Signal Model

02 Proposed System

03 Evaluations

04 Conclusions

Dataset

- Clean speech
  - the British National Corpus recorded by 40 speakers
- Interferer data
  - Speech: SLR83 database
  - Noise: Freesound database
- Room impulse response
  - RAVEN software
  - RT60: 0.2 ~ 0.4s
- Listeners
  - bilateral pure-tone audiograms
  - [250, 500, 1000, 2000, 3000, 4000, 6000, 8000] Hz

[6] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. Viveros Muñoz, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2021, Brno, Czech Republic, 2021

Evaluations with Challenge Data

- Training Procedure
  - Training: 6000 scenes (train dataset) × 4 channels × bilateral
  - Evaluation: 2000 scenes (80% of development dataset) × 4 channels × bilateral
  - Down-sampled to 16kHz
- For instrumental objective test (MBSTOI)
  - 500 scenes (20% of development dataset) × 3 listeners
- Blind Data for final objective and subjective test
  - Performed by organizers
  - 1500 scenes × 1 listeners
  - Non-overlap on scenes and listeners

Results

Objective evaluations on development dataset

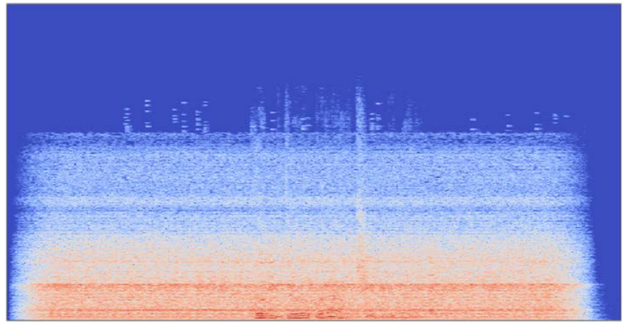
- Complexity: 35 MFLOPs
- Model size: 1.572 MB
- one frame processing time ≈ 1ms (Intel i7-9750)

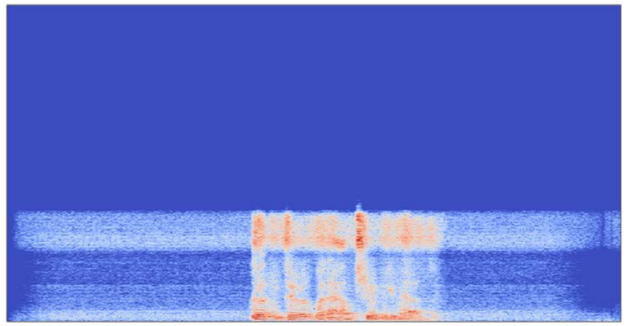
MBSTOI		
Baseline	**Proposed**	Anechoic
0.53	**0.60**	0.71

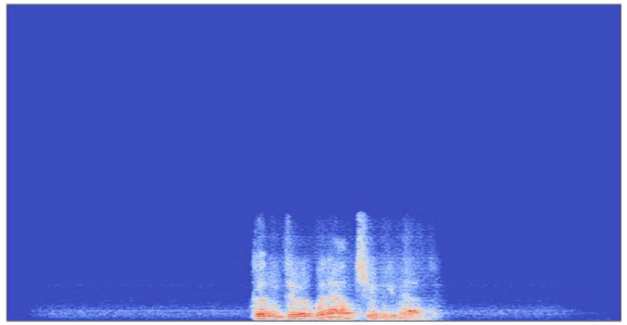
On **blind test data**, baseline MBSTOI score is 0.31, and the proposed achieved **0.56**

Subjective Listening Preference

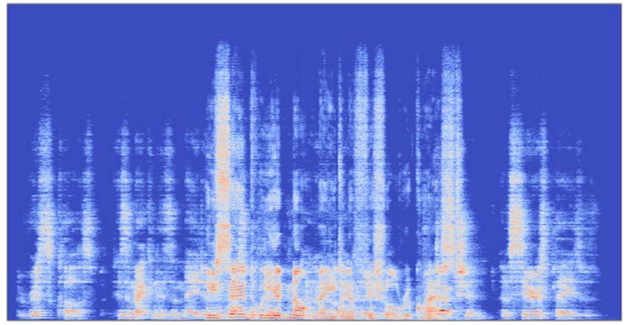
Couldn't do it without you know

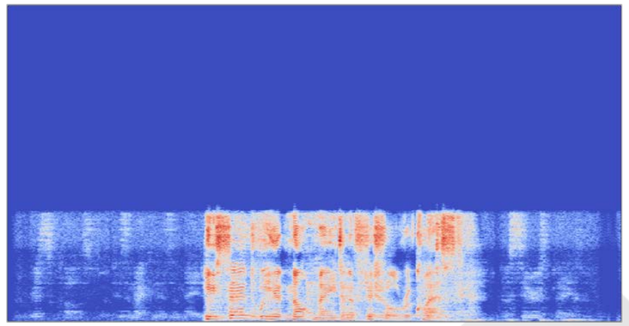


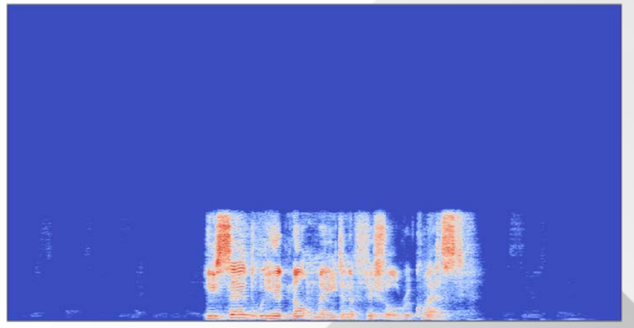




He said we had not made an official request







01

Signal Model

02

Proposed System

03

Evaluations

04

Conclusions

Conclusions

- Cascaded general speech enhancement system
  - Causal
  - Optimizing for perception
  - Low complexity
  - Relative small model size
- Dynamic equalization scheme with personal hearing profiles
  - Improve subjective perception despite hurting MBSTOI

Improve speech perception and intelligibility for hearing impairment

腾讯天籁

Thank you for watching this presentation!

If you have any question, just contact me

1900432053@email.szu.edu.cn