30 August – 3 September 2021
INTERSPEECH 2021 BRNO | CZECHIA
Speech everywhere!
www.interspeech2021.org

A Causal U-net based Neural Beamforming Network for Real-Time Multi-Channel Speech Enhancement

—Kuaishou's System for ConferencingSpeech 2021 Challenge

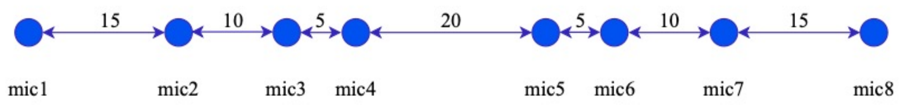
Xinlei Ren, Xu Zhang, Xiguang Zheng, Lianwu Chen,
Chen Zhang, Liang Guo and Bing Yu

Kuaishou Technology
Beijing, China

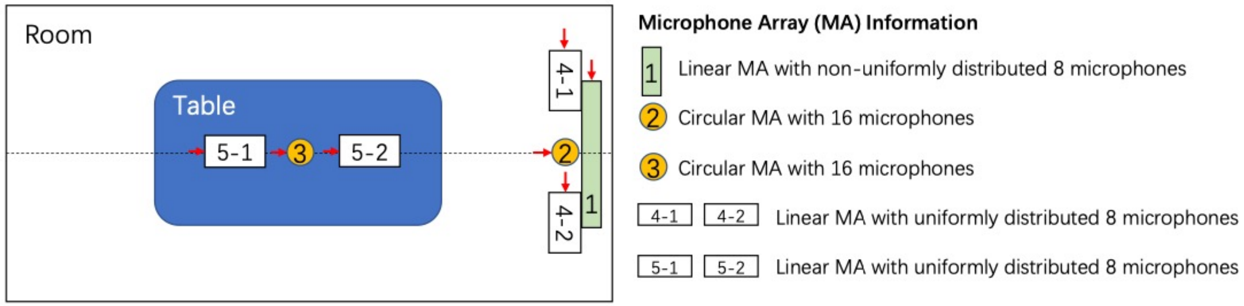
KUAISHOU

Introduction

- Challenge Tasks [1]
  - Task1: single microphone array(real-time)



  - Task2: multiple distributed microphone arrays(nonreal-time)



Introduction

➢ Methods

- Signal processing
  DS[2], MVDR[2], GSC[2]…
- Signal processing combined with deep learning [3] [4]

$$W_{\mathrm{MVDR}} = \frac{R^{-1} a(\theta)}{a^H(\theta) \boxed{R^{-1}} a(\theta)}$$

deep learning

- Deep learning
  MISO [5], SISO + IPD…
- **Deep learning combined with signal processing**
  FaSNet [6], **MIMO + BF**…

Problem formulation

➢ Time-domain signal model

$$y_m(t) = x(t) * h_m(t) + n_m(t)$$

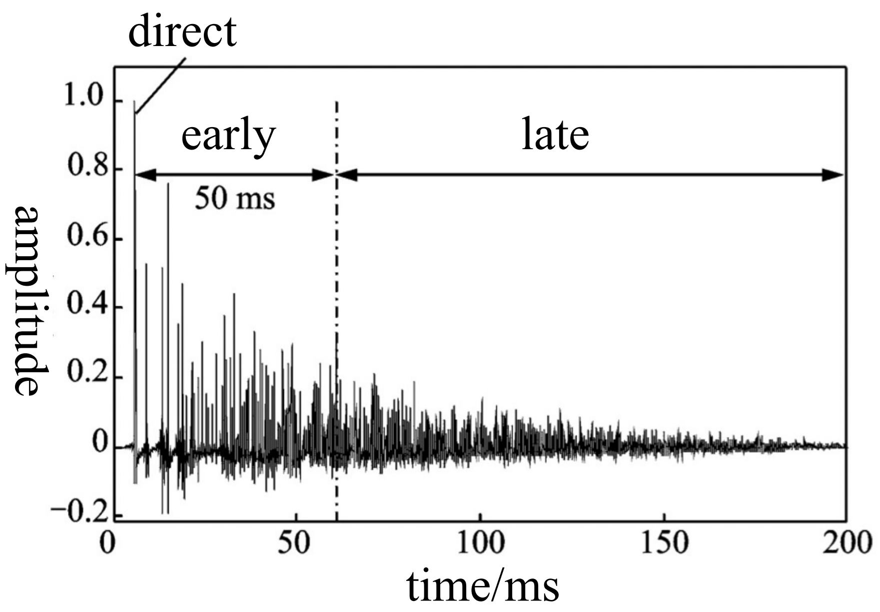
$y_m(t)$	noisy signal recorded by m-th microphone, m=0~(*M*-1)
$x(t)$	clean speech
$h_m(t)$	room impulse response from the clean speech to m-th microphone
$n_m(t)$	noise signal recorded by m-th microphone
$*$	convolution operation

Problem formulation

- Time-Frequency-domain signal model

$$\mathbf{Y}_m(l,f) = \mathbf{X}_m^{direct}(l,f) + \mathbf{X}_m^{early}(l,f) + \mathbf{X}_m^{late}(l,f) + \mathbf{N}_m(l,f)$$

$\mathbf{X}_m^{direct}$	the direct sound
$\mathbf{X}_m^{early}$	the early reflections of the speech
$\mathbf{X}_m^{late}$	the late reverberations of the speech
l	the frame index
f	the frequency index



Problem formulation

- Proposed solution

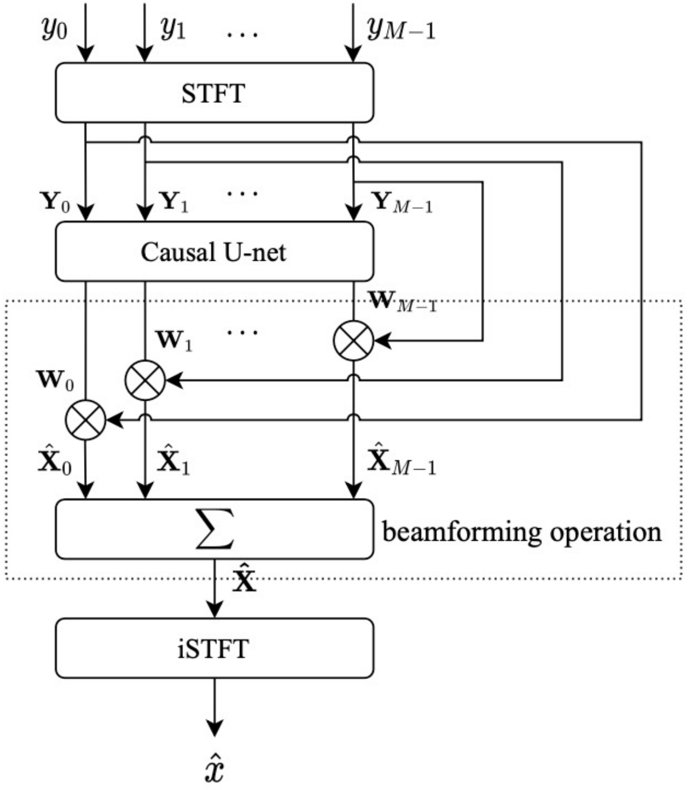
$$\hat{\mathbf{X}}^{direct\_early}(l,f) = \sum_{m=0}^{M-1} \{\mathbf{Y}_m(l,f) \cdot \mathbf{W}_m(l,f)\}$$

- Step1: estimate the complex filters $\mathbf{w}_m$ with U-net
- Step2: enhance the speech using beamforming

Proposed system

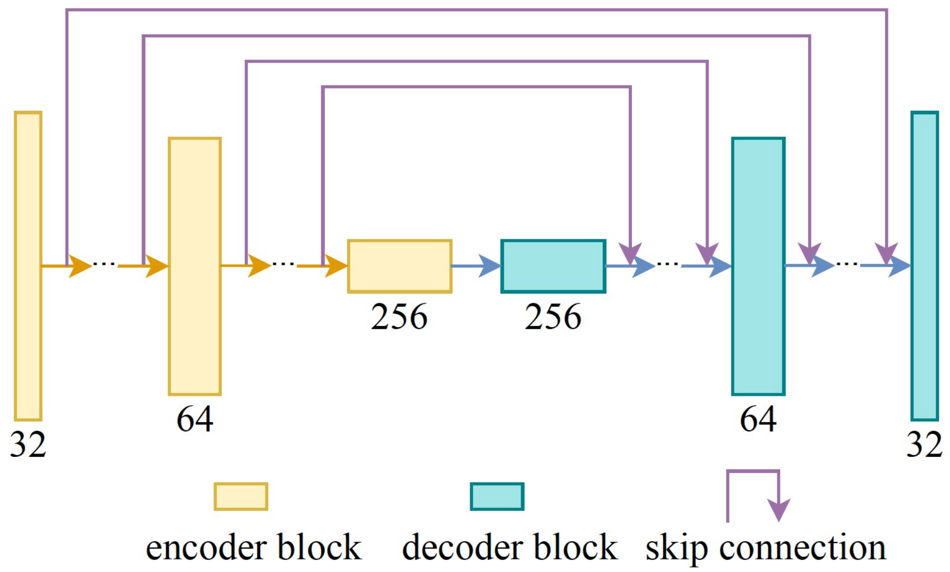
- System architecture
  - causal U-net multi-channel
  - beamforming structure

$$\hat{\mathbf{X}} = \sum_{m=0}^{M-1} \{\mathbf{Y}_m \cdot \mathbf{W}_m\}$$



Proposed system

- Multi-channel causal **U-net**
  - encoder
  - decoder
  - skip connection



Proposed system

➢ Multi-channel causal **U-net**

- encoder (8 blocks)

  conv2d + batch normalization + dropout + LeakyReLu
- decoder (8 blocks)

  replace conv2d with conv2dTranspose
- format

  [BatchSize, Frequency, Frame, Channel]

Layer	**Filter number**	**Kernel**	**Stride**
Conv2d^{1st}	32	$(6, 2)$	$(2, 1)$
Conv2d^{2nd}	32	$(6, 2)$	$(2, 1)$
Conv2d^{3rd}	64	$(7, 2)$	$(2, 1)$
Conv2d^{4th}	64	$(6, 2)$	$(2, 1)$
Conv2d^{5th}	96	$(6, 2)$	$(2, 1)$
Conv2d^{6th}	96	$(6, 2)$	$(2, 1)$
Conv2d^{7th}	128	$(2, 2)$	$(2, 1)$
Conv2d^{8th}	256	$(2, 2)$	$(1, 1)$

Proposed system

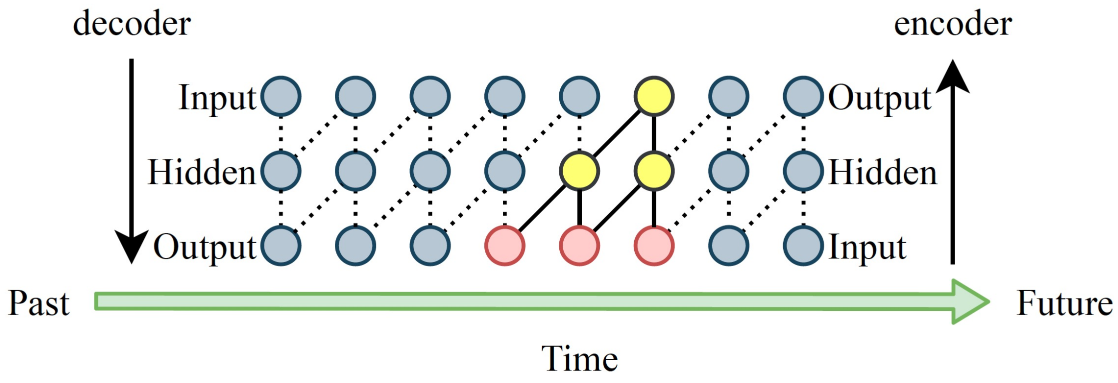
➢ Multi-channel **causal** U-net

- input

  pad K zeros frames

- output

  discard last K frames



Proposed system

- Loss function

  mean absolute error(MAE):

$$loss_{mae} = |x - \hat{x}| + |n - \hat{n}|$$

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

- Post-filter

  wiener filter with the noise estimation algorithm based on minimum tracking.

Experiments and Results

➢ Datasets

speech	aishell-1, aishell-3, vctk and librispeech(train-clean-360)
noise	musan and audioset
rirs	20000 with the image method
total	1000 hours, 70% for training and 30% for validating

➢ Augmentations

reverberation	preserve early 50ms
snr	[-3, 25]dB
scale	[-50, 0.87]dBFS
eq	low-pass filter, de-emphasis filter …

Experiments and Results

- Model Input

feature	complex STFT spectrogram
sample rate	16kHZ
audio length	4 seconds
FFT/hop size	512/256 points
frame number	249 frames
frequency number	256 frequency bins are used

Experiments and Results

- Objective scores of different network structures

	PESQ	**STOI**	**E-STOI**	**Si-SNR**
Noisy	1.278	0.728	0.587	1.893
SISO-U-net	1.841	0.844	0.740	7.475
SISO+IPD-U-net	1.855	0.847	0.746	7.501
MISO-U-net	1.890	0.852	0.758	7.959
MIMO-U-net+BF	1.950	0.861	0.764	8.008
MIMO-U-net+BF+PF (proposed)	**1.919**	**0.857**	**0.759**	**7.935**

Experiments and Results

- Objective scores of the proposed and baseline systems

	PESQ	STOI	E-STOI	Si-SNR
Task1				
Noisy	1.515	0.823	0.690	4.474
baseline	1.999	0.888	0.780	9.159
proposed	**2.125**	**0.908**	**0.817**	**9.287**
Task2				
Noisy	1.506	0.824	0.693	4.504
baseline	1.983	0.887	0.780	9.228
proposed	**2.125**	**0.909**	**0.818**	**9.343**

Experiments and Results

➢ Subjective scores of the some systems

Task1:

Ranking	Team	MOS	S-MOS	N-MOS	dMOS	dS-MOS	dN-MOS	95%CI
1	kuaishou_deep_ns	4.02	3.87	3.87	1.46	0.94	0.84	0.02
2	HKBAT	3.86	3.78	3.80	1.30	0.85	0.77	0.02
3	WavingBrother	3.57	3.62	3.64	1.01	0.69	0.62	0.02
4	CMfeiyi	3.57	3.56	3.58	1.01	0.63	0.55	0.02
5	HNT	3.56	3.59	3.63	1.00	0.66	0.60	0.02
6	SRIB_IISC	3.45	3.38	3.36	0.70	0.32	0.25	0.04
7	GSL	3.44	3.47	3.48	0.69	0.41	0.37	0.04
·	Baseline	3.43	3.55	3.55	0.68	0.49	0.44	0.03
8	Deep Narcissus	3.34	3.47	3.49	0.59	0.41	0.39	0.04
9	Hust3iAsrLab	3.23	3.27	3.26	0.48	0.21	0.15	0.04
10	I2R-ALI	3.22	3.44	3.48	0.47	0.38	0.37	0.04
11	loiu	3.16	3.32	3.33	0.41	0.26	0.23	0.04
12	RoyalFlush	3.14	3.18	3.17	0.39	0.12	0.06	0.04
13	Doreso	2.98	3.12	3.13	0.23	0.06	0.02	0.05
14	SLADKIE	2.90	3.05	3.05	0.15	-0.01	-0.06	0.04
15	NlabAtFiveFloor	2.67	2.65	2.64	-0.08	-0.41	-0.46	0.04
	Noisy	2.56	2.93	3.03	0.00	0.00	0.00	0.02

Experiments and Results

- Subjective scores of the some systems

Task2:

Ranking	Team	MOS	S-MOS	N-MOS	dMOS	dS-MOS	dN-MOS	95%CI
1	kuaishou_deep_ns	4.14	3.93	3.92	1.63	1.05	0.93	0.02
2	HungarianDance	3.60	3.60	3.62	1.09	0.72	0.63	0.02
3	WavingBrother	3.54	3.58	3.60	1.03	0.70	0.61	0.02
·	Baseline	3.32	3.41	3.44	0.92	0.75	0.70	0.03
4	GSL	3.25	3.29	3.31	0.86	0.63	0.57	0.03
5	RoyalFlush	3.24	3.38	3.43	0.85	0.72	0.69	0.03
	Noisy	2.51	2.88	2.99	0.00	0.00	0.00	0.02

Conclusion

- Use the multi-channel causal U-net to estimate multi-channel complex masks
- Combine the U-net with the traditional beamforming structure
- Compare the performances of different network structures
- The proposed system significantly outperforms other systems

Reference

[1] W. Rao, L. Xie, Y. Wang, T. Yu, S. Watanabe, Z.-H. Tan, H. Bu, and S. Shang, "Conferencingspeech 2021 challenge evaluation plan." [Online]. Available: https://arxiv.org/abs/2104.00960.

[2] Benesty J, Sondhi M M, Huang Y. Springer Handbook of Speech Processing[M]. 2008.

[3] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. L. Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Interspeech*, 2016, pp. 1981–1985.

[4] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 196–200.

[5] H. Lee, H. Y. Kim, W. H. Kang, J. Kim, and N. S. Kim, "End-to-End multi-channel speech enhancement using inter-channel time-restricted attention on raw waveform," in *Proc. Interspeech 2019*, 2019, pp. 4285–4289.

[6] Y. Luo, C. Han, N. Mesgarani, E. Ceolini and S. Liu, "FaSNet: Low-Latency Adaptive Beamforming for Multi-Microphone Audio Processing," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 260-267, doi: 10.1109/ASRU46091.2019.9003849.

Thank you!

ICASSP2022
3D Audio Challenge L3DAS22

- Task
  - 3D Speech Enhancement
  - 3D Sound Event Localization and Detection
- Datasets

  Recorded in Real Room with SoundFiled Microphones
- Algorithm

  No Restrictions
- Website

  www.l3das.com