Tencent AI Lab

ENHANCING END-TO-END MULTI-CHANNEL SPEECH SEPARATION VIA SPATIAL FEATURE LEARNING

Rongzhi Gu[1], Shi-Xiong Zhang[2], Lianwu Chen[3], Yong Xu[2], Meng Yu[2], Dan Su[3], Yuexian Zou[1,4], Dong Yu[2]

[1] ADSPLAB, School of ECE, Peking University, Shenzhen, China
[2] Tencent AI Lab, Bellevue, WA, USA
[3] Tencent AI Lab, Shenzhen, China
[4] Peng Cheng Laboratory, Shenzhen, China

ICASSP2020

Outline

- Introduction
- Proposed method
- Experiments and results
- Conclusion

Speech separation

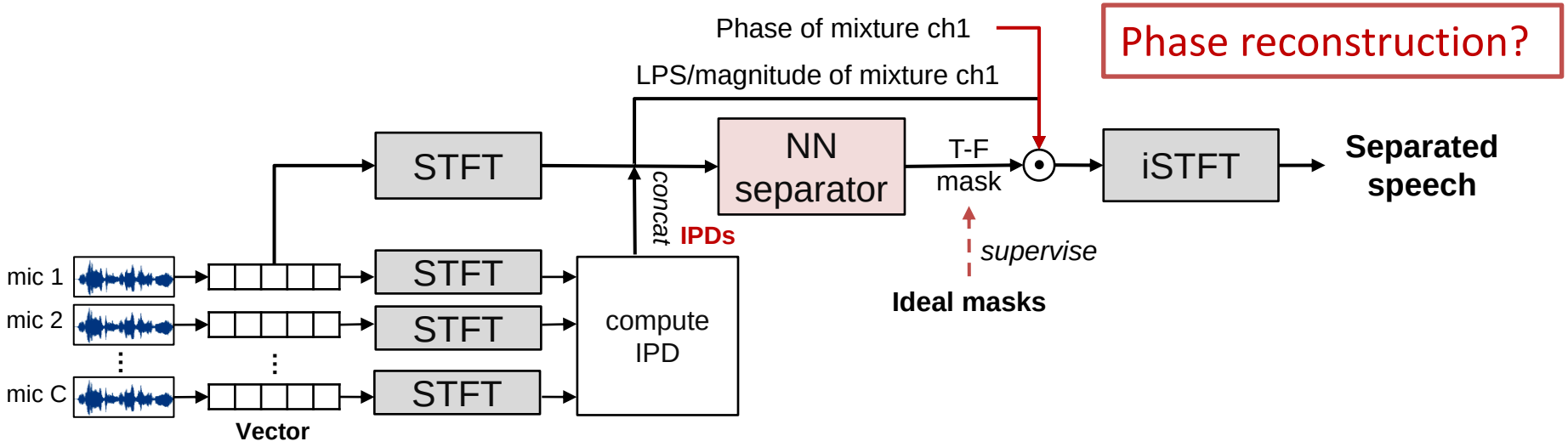
- **Cocktail Party problem** [Cherry 1953]
  - recover the speech of each speaker from overlapped speech mixture

$$\mathbf{y}[n] = \sum_{s=1}^{S} \mathbf{x}_s[n]$$

- $\mathbf{y}[n] = [y^1[n], \ldots, y^C[n]]^T$: observed multi-channel mixture signal
- $\mathbf{x}_s[n] = [x_s^1[n], \ldots, x_s^C[n]]^T$: reverberant image for source s

Methods for multi-channel speech separation

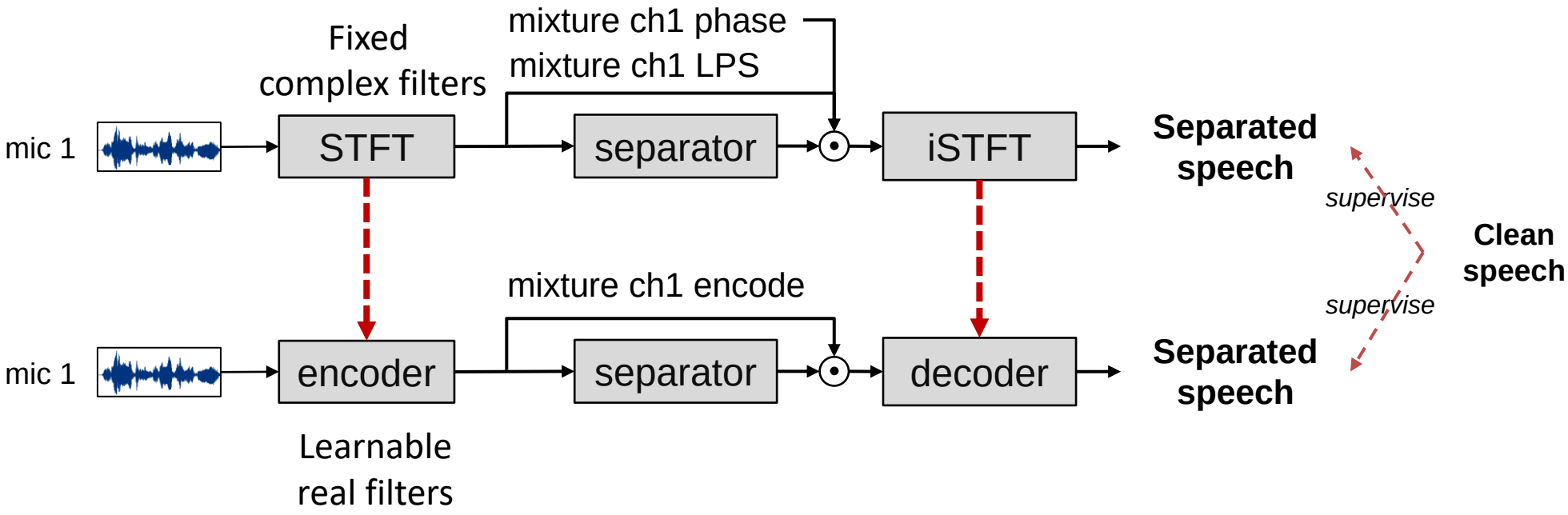
- **T-F masking**
  - formulate speech separation as a supervised learning task in frequency domain



- **Integration of T-F masking and beamforming**
- **End-to-end approaches**

Single-channel TasNet

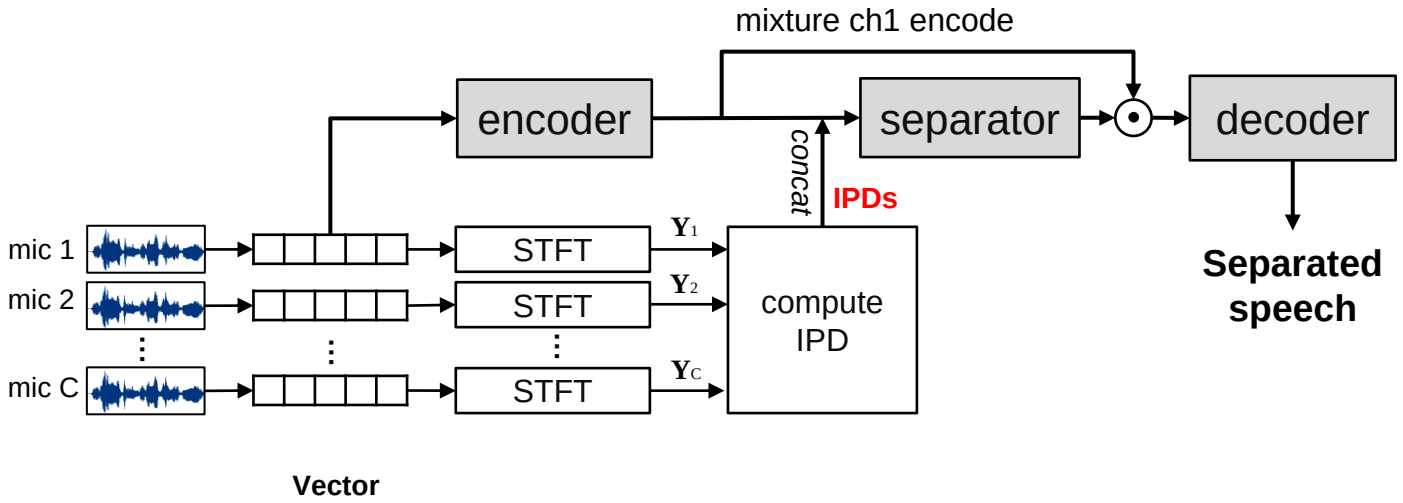
- Encoder-decoder structure

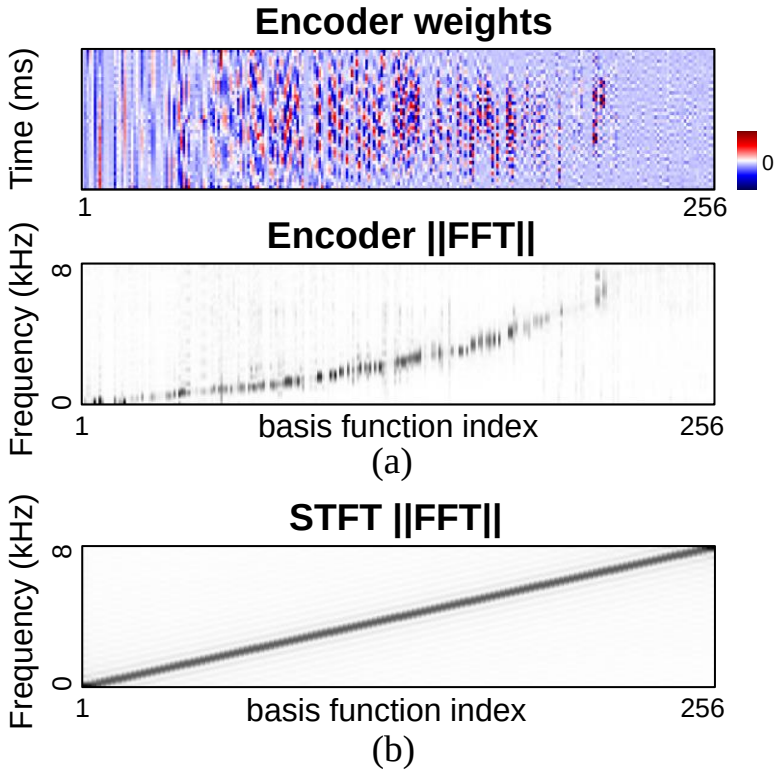


- Unsatisfactory performance under far-field scenario

Our previous try – Data mismatch

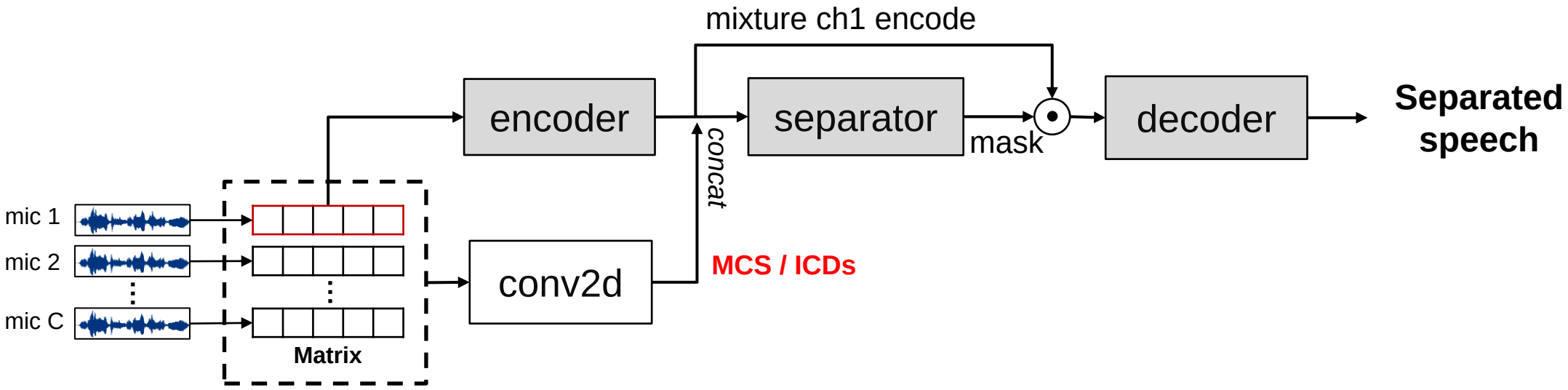
- **A multi-channel speech separation approach:**
  - encode + IPDs
  - a multi-channel speech separation approach





- **Data mismatch: encoder & STFT**
  - **Encoder**: learned in purely data-driven fashion
  - **STFT**: fixed complex filters (with evenly distributed frequency responses)

Proposed model



- **Aim: an end-to-end multi-channel speech separation model**
  - **Encoder**: transform mixture waveform into the mixture encode
  - **Conv2d**: spatial feature learning
  - **Separator**: estimate a mask in encoder output domain for each speaker
  - **Decoder**: reconstruct the separated speech waveform

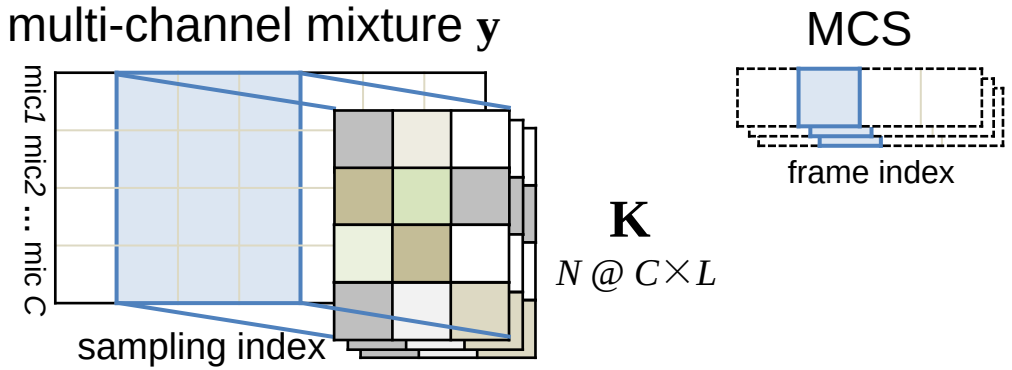
Spatial feature learning – multi-channel convolution sum

- **Main idea:**
  - learn time-domain filters spanning all signal channels $\mathbf{K}=\{\mathbf{k}^{(n)}\}\in\mathbb{R}^{C\times L\times N}$ to perform adaptive spatial filtering

- **Multi-channel convolution sum (MCS):**

$$\mathrm{MCS}^{(n)}=\sum_{c=1}^{C}\mathbf{y}_c\circledast k_c^{(n)}$$

  - a DAS beamformer-like formation
  - Each set of filters $k^{(n)}$ is expected to steer at a different direction



- **Implementation:**
  - Conv2d, kernel: N@C×L

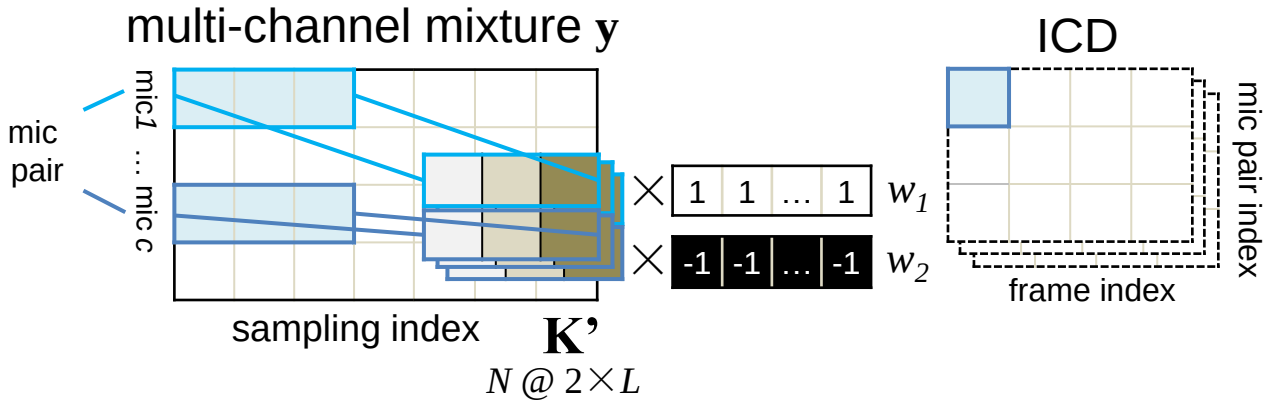
Spatial feature learning – inter-channel convolution difference

- **inspired by IPD formulation:**
  - T-F bins that dominated by the same source share the same time delay
  - IPDs of these T-F bins naturally form a cluster within each frequency band

- **Inter-channel convolution difference (ICD)**

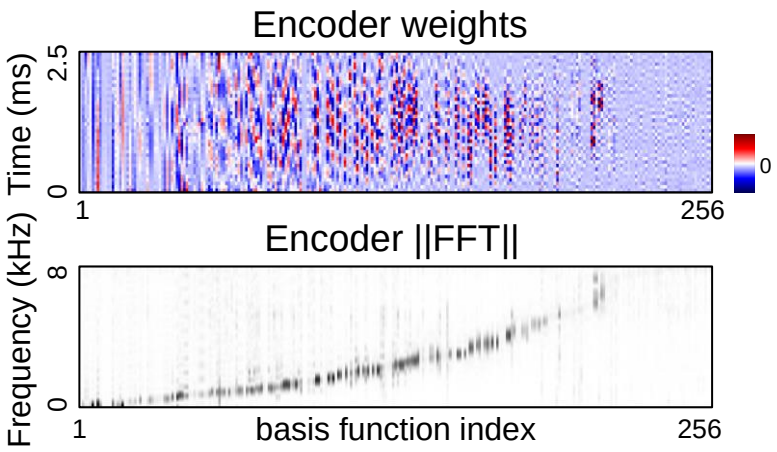
$$\mathrm{ICD}_m^{(n)} = \sum_{c=1}^{2} w_c \cdot \left(\mathbf{y}_{m_c} \circledast k'^{(n)} \right)$$

  - m: microphone pair index
  - $w_c \in \mathbb{R}^{1\times L}$: window function
  - Initialization: w_1=[1, …, 1], w_2 = [-1, …, -1]

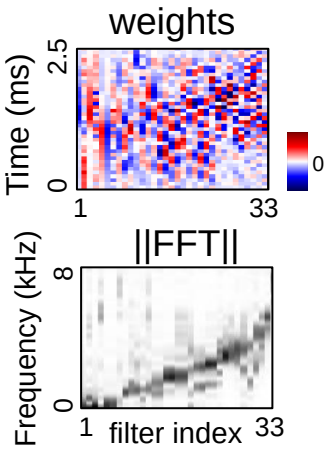


- **Implementation:**
  - A customized conv2d kernel

Visualization of learned filters



Encoder

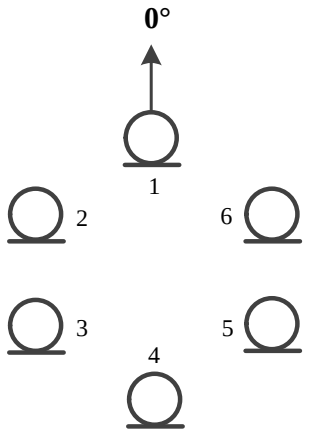


Conv2d kernel **K**

Experiments - Data

- spatialized reverberant WSJ0 2-mix

Train	20000 utterances / 30h / 101 speakers
Validation	8000 utterances / 8h / 101 speakers
Test	5000 utterances / 5h / 18 speakers (unseen)
Sampling rate	16kHz
Mix SNR	[-5, 5] dB
Overlap Ratio	100%
RT60	0.05~0.5s
Microphone array	6-microphone circular array of 7cm diameter
Included angle	0-15: 16%, 15-45: 29%, 45-90: 26%, 90-180: 29%
Selected pairs for IPD/ICD	[1, 4], [2, 5], [3, 6], [1, 2], [3, 4], [5, 6]



Experiments

- **Training objective:** SI-SDR

$$\begin{cases} x_{\text{target}} := \dfrac{\langle \hat{x}, x \rangle x}{\|x\|_2^2} \\ e_{\text{noise}} := \hat{x} - x_{\text{target}} \\ \text{SI-SDR} := 10 \log_{10} \dfrac{\|x_{\text{target}}\|_2^2}{\|e_{\text{noise}}\|_2^2} \end{cases}$$

  - $\hat{x}$: estimated reverberant speech
  - x: ground truth reverberant speech
- **Permutation invariant training**
- **Evaluation Metrics**
  - SI-SDR improvement
  - SDR improvement

Results – different conv2d configurations

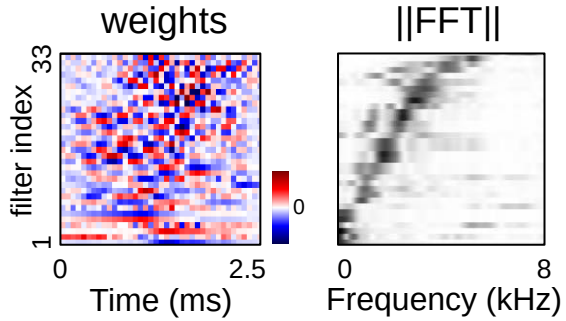
Setup	window w	# filters N	SI-SDR improvement (dB)					SDRi (dB)
			<15	15-45	45-90	>90	Ave.	
Single-channel Conv-TasNet	-	-	8.5	9.0	9.1	9.3	9.1	9.4
+ MCS (conv2d (6×40))	-	256	5.7	10.3	11.9	12.9	10.8	11.2
+ ICD (conv2d (2×40))	fix -1	256	5.5	10.9	12.3	12.9	11.0	11.4
+ ICD (conv2d (2×40))	init. -1	256	6.2	11.2	12.6	13.2	11.4	11.8
+ ICD (conv2d (2×40))	init. randomly	33	8.2	8.1	9.0	9.1	8.9	9.2
+ ICD (conv2d (2×40))	fix -1	33	6.9	11.1	12.3	12.9	11.3	11.7
+ ICD (conv2d (2×40))	init. -1	33	6.7	11.7	13.1	13.9	11.9	12.3

- **The performances with 33 filters are relatively superior to those with 256 filters.**
- **The value of *w* contributes significantly to the separation performance**
  - Init randomly: no subtraction operation
  - Fix -1: exact inter-channel convolution difference
  - Init -1: initialize *w* as -1 and set *w* learnable during training

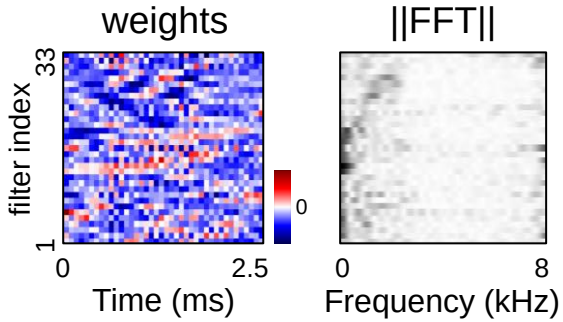
Results – IPD vs ICD

Setup	SI-SDR improvement (dB)		
	<15°	>15°	average
cosIPD, sinIPD	7.7	12.2	11.5
cosIPD, sinIPD (trainable kernel)	7.9	12.3	11.6
ICD	6.7	12.9	11.9
ICD, cosIPD, sinIPD	8.1	13.2	12.4

- IPD provide beneficial spatial information of sources and help the multi-channel speech separation.
- With the trainable kernel, the performance improves slightly.
- ICD based separation model obtains 0.4dB improvement over cosIPD+sinIPD based
- The incorporation of ICDs and IPDs achieves further 0.5dB improvement.



Learned filters K' (ICD)



Learned filters K' (ICD+cosIPD+sinIPD)

Conclusions

- **This work proposes an end-to-end multi-channel speech separation model**
  - learn effective spatial cues directly from the multi-channel speech waveforms
  - end-to-end optimization in a purely data-driven fashion

- **The learned spatial features**
  - can be computed with few parameters and computation cost
  - can be combined with well-designed IPDs and obtain better results

Thank you!

1701111335@pku.edu.cn

Thank you!
Q&A