End-to-End Dereverberation, Beamforming, and Speech Recognition in a Cocktail Party

讲者：张王优

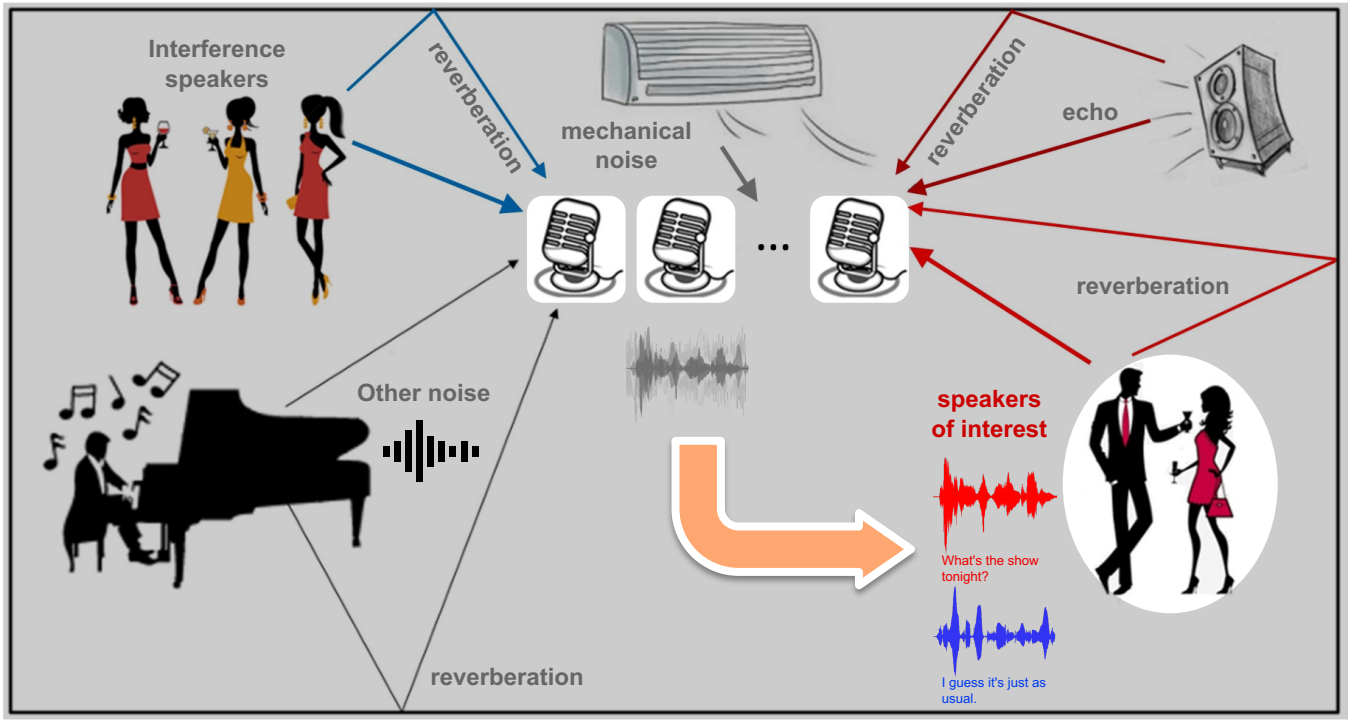
导师：钱彦旻

2022.10.22

SJTU Cross Media
Language Intelligence Lab
上海交通大學跨媒體智能實驗室

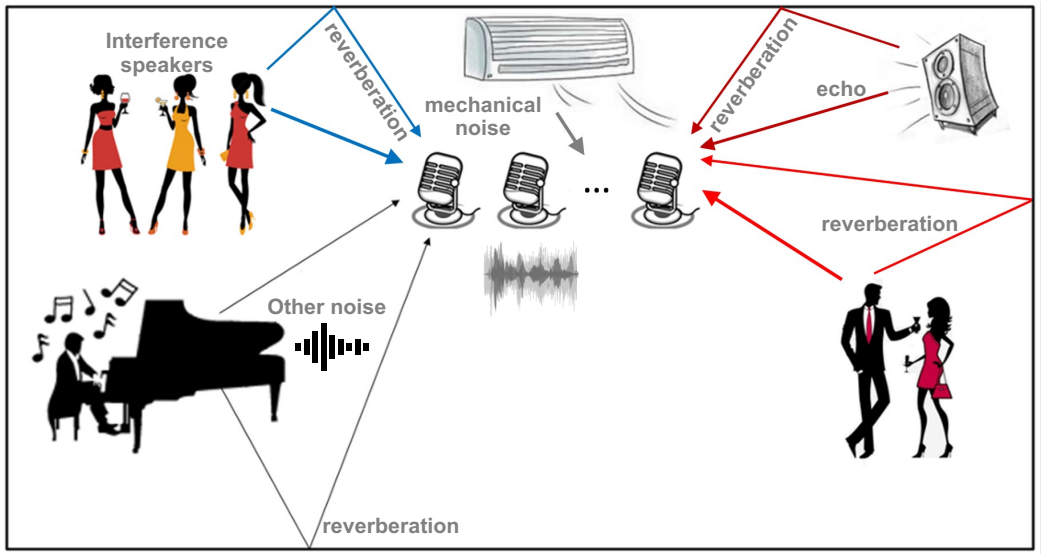
Background

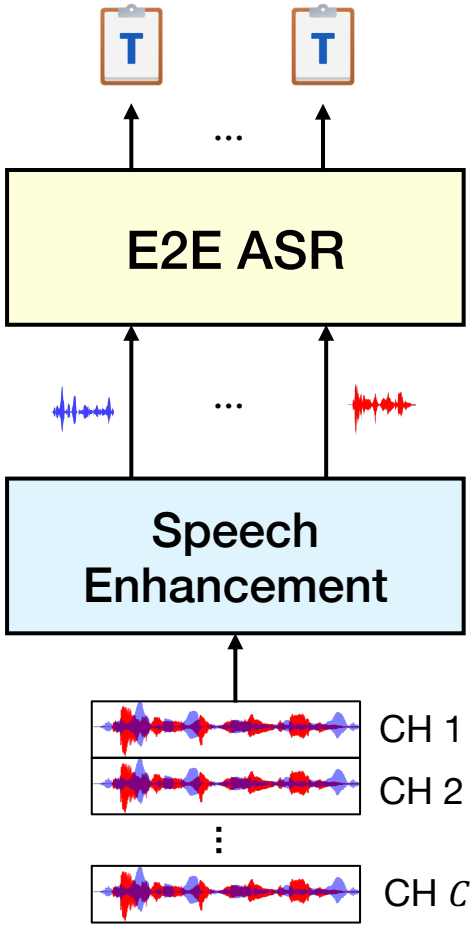
- **The cocktail party problem [Cherry 1953]**
  - Multiple talkers
  - background noise and reverberation
  - **Goal:** to separate speakers of interest from the mixed signal and recognize their speech.



Background

- **Speech processing in the cocktail party scenario**
  - Frontend processing:
    - Speech enhancement
  - Backend processing:
    - Speech recognition
    - ...





Background

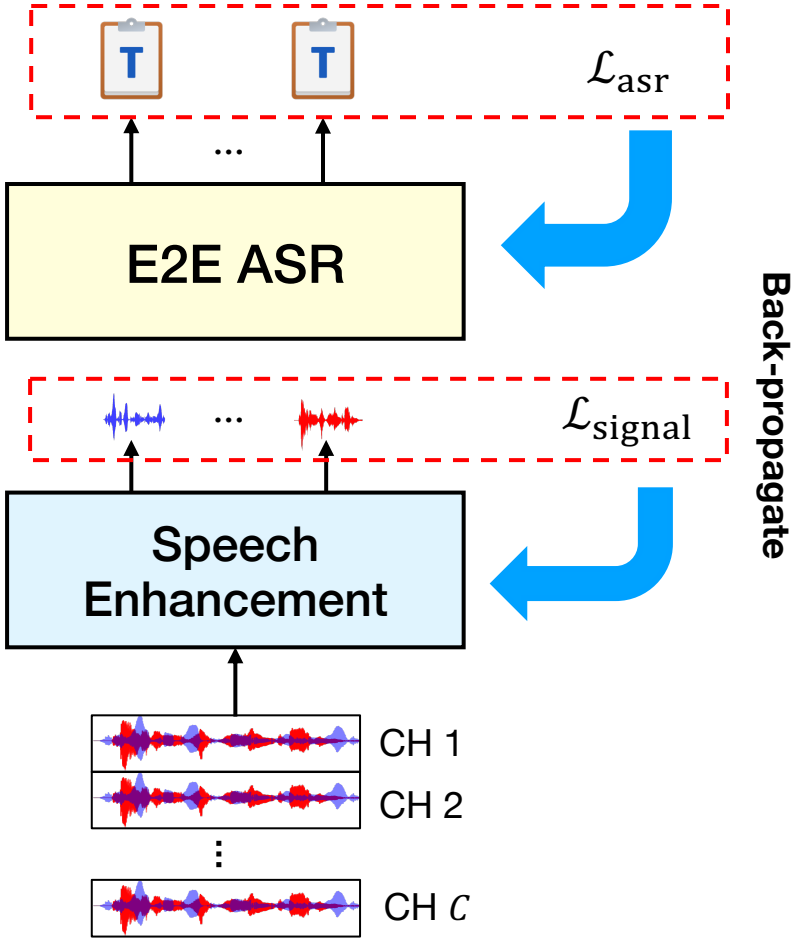
- **Speech processing in the cocktail party scenario**
  - Existing training schemes
    1. Independent training

Pros

- Each module can be designed and updated individually.
- It allows reuse and fast development of different modules.
- Training data of different modules can be prepared separately.

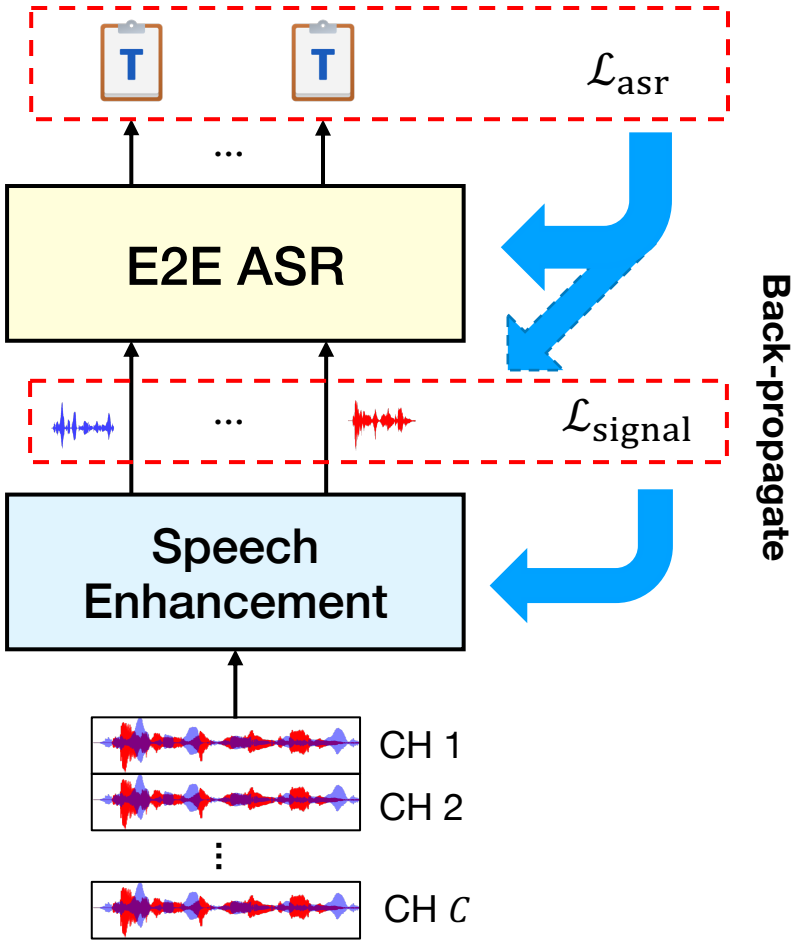
Cons

- There could be a large mismatch between different modules.



Background

- **Speech processing in the cocktail party scenario**
  - Existing training schemes
    1. Independent training
    2. Joint training via multi-task learning



Pros

- The inter-module mismatch could be reduced.
- Different modules are regularized by multiple objectives at the same time.

Cons

- Training data must include reference labels for both modules.
- Performance of an individual module may deteriorate.

Background

❑ Speech processing in the cocktail party scenario

❑ Proposed training scheme

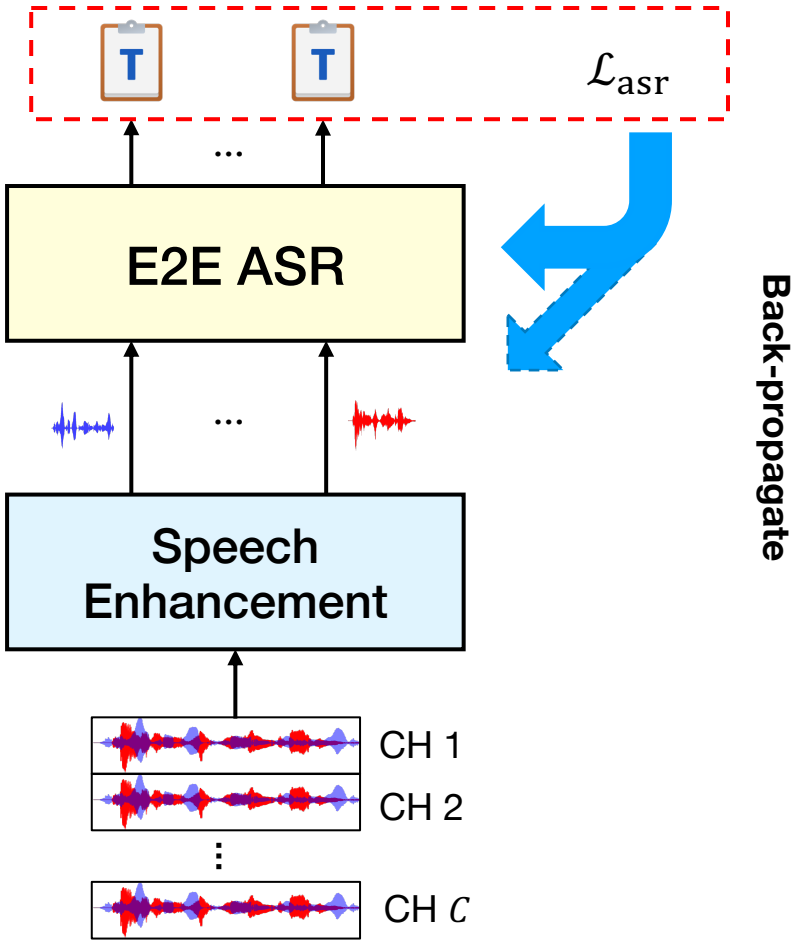
- Fully end-to-end training

Pros

- The inter-module mismatch could be reduced.
- All modules are optimized towards the final objective.
- Only ASR labels are required, making it possible to directly use real recordings for training.

Cons

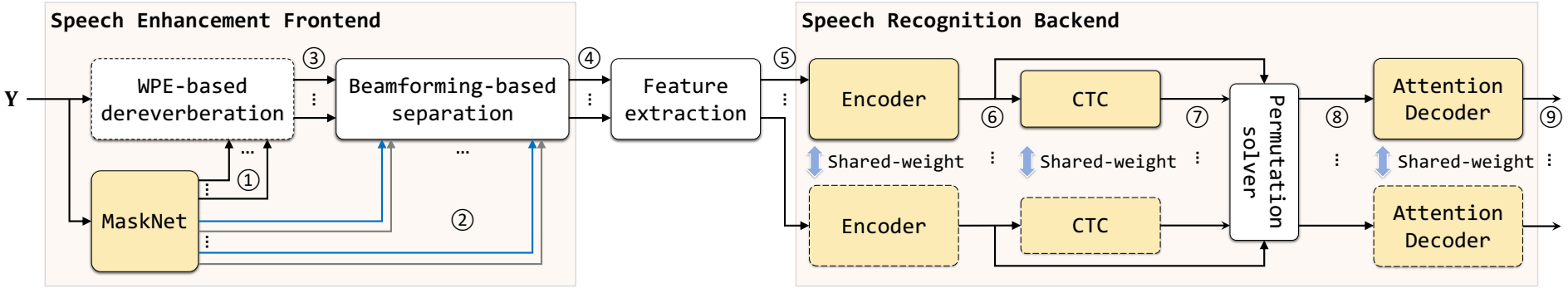
- The frontend module is not explicitly regularized, thus its individual performance is not guaranteed.



Proposed model

E2E Dereverberation, Beamforming, and Multi-Talker ASR

- **Goal:**
  - Build an E2E multi-channel system to achieve joint dereverberation, speech separation, and multi-talker ASR
- **Model design[1]:**
  - Frontend:
    - WPE-based dereverberation
    - Beamforming-based separation
  - Backend: E2E ASR

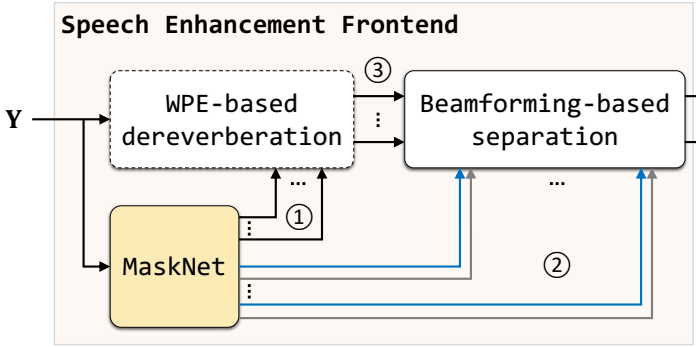


[1] W. Zhang, X. Chang, C. Boeddeker, T. Nakatani, S. Watanabe, and Y. Qian, "End-to-end dereverberation, beamforming, and speech recognition in a cocktail party," IEEE/ACM Trans. ASLP., vol. 30, pp. 3173–3188, 2022.

Proposed model

E2E Dereverberation, Beamforming, and Multi-Talker ASR

- **Model design:**
  - Frontend:
    - WPE-based dereverberation
    - Beamforming-based separation
  - Both WPE and beamforming are linear filtering techniques, and introduce very few distortion[2], thus friendly to ASR.



WPE

1. Estimate the reverberation tail from the past time frames

$$\hat{\mathbf{x}}_{t,f}^{(j),\text{tail}} = \left(\mathbf{G}_f^{(j)}\right)^{\text{H}} \tilde{\mathbf{y}}_{t-\Delta,f}$$

$$\tilde{\mathbf{y}}_{t-\Delta,f} = \left[\mathbf{y}_{t-\Delta,f}^{\text{T}} \quad \cdots \quad \mathbf{y}_{t-\Delta-K+1,f}^{\text{T}}\right]^{\text{T}} \in \mathbb{C}^{CK}$$

2. Subtract it from the observation

$$\hat{\mathbf{x}}^{(j),\text{dereverb}} = \mathbf{y}_{t,f} - \hat{\mathbf{x}}_{t,f}^{(j),\text{tail}} \in \mathbb{C}^{C}$$

[2] Z.-Q. Wang, G. Wichern, and J. Le Roux, “Leveraging low-distortion target estimates for improved speech enhancement,” 2021, arXiv:2110.00570.

Proposed model

E2E Dereverberation, Beamforming, and Multi-Talker ASR

- **Model design:**
  - Frontend:
    - WPE-based dereverberation
    - Beamforming-based separation
  - Both WPE and beamforming are linear filtering techniques, and introduce very few distortion, thus friendly to ASR.

WPE

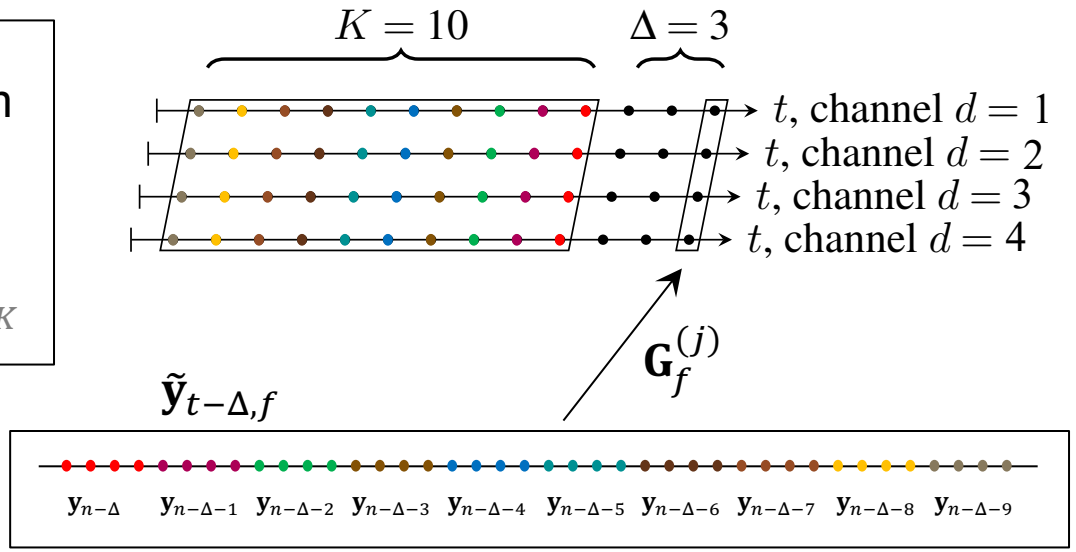
1. Estimate the reverberation tail from the past time frames

$$\hat{\mathbf{x}}_{t,f}^{(j),\text{tail}} = \left(\mathbf{G}_f^{(j)}\right)^{\text{H}} \tilde{\mathbf{y}}_{t-\Delta,f}$$

$$\tilde{\mathbf{y}}_{t-\Delta,f} = \left[\mathbf{y}_{t-\Delta,f}^{\text{T}} \quad \cdots \quad \mathbf{y}_{t-\Delta-K+1,f}^{\text{T}}\right]^{\text{T}} \in \mathbb{C}^{CK}$$

2. Subtract it from the observation

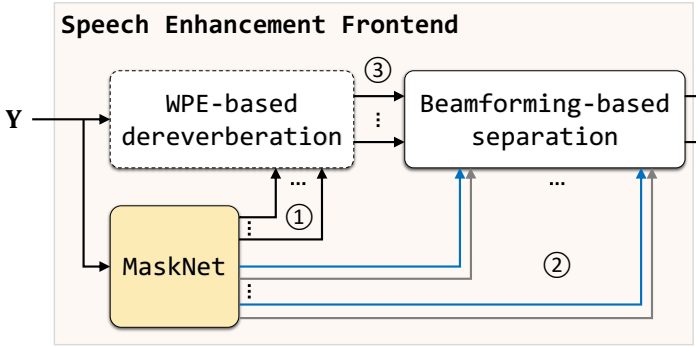
$$\hat{\mathbf{x}}^{(j),\text{dereverb}} = \mathbf{y}_{t,f} - \hat{\mathbf{x}}_{t,f}^{(j),\text{tail}} \in \mathbb{C}^{C}$$



Proposed model

E2E Dereverberation, Beamforming, and Multi-Talker ASR

- **Model design:**
  - Frontend:
    - WPE-based dereverberation
    - Beamforming-based separation
  - Both WPE and beamforming are linear filtering techniques, and introduce very few distortion, thus friendly to ASR.



MVDR Beamforming

1. Find a complex filter $\mathbf{h}_f^{(j)} \in \mathbb{C}^C$ that solves the following optimization problem:

$$\begin{cases} \min\limits_{\mathbf{h}_f} & \left(\mathbf{h}_f^{(j)}\right)^{\mathrm{H}} \mathbf{\Phi}_{\mathrm{n},f}^{(j)} \mathbf{h}_f^{(j)} \\ \text{subject to} & \left(\mathbf{h}_f^{(j)}\right)^{\mathrm{H}} \mathbf{v}_f = 1 \end{cases}$$

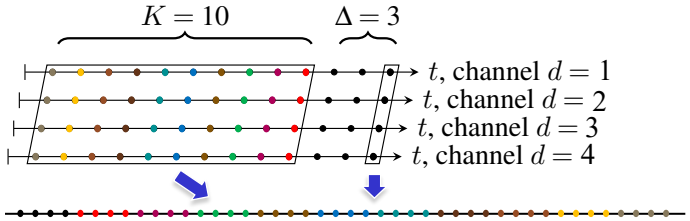
2. Apply beamforming:

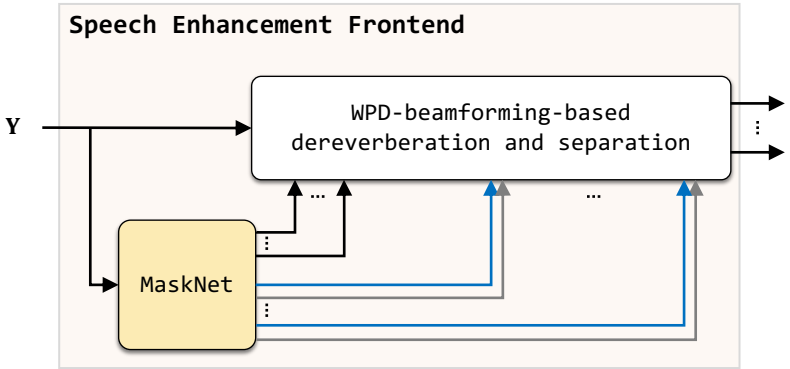
$$\hat{\mathbf{x}}_f^{(j)} = \left(\mathbf{h}_f^{(j)}\right)^{\mathrm{H}} \hat{\mathbf{x}}^{(j),\mathrm{dereverb}}$$

Proposed model

E2E Dereverberation, Beamforming, and Multi-Talker ASR

- **Model design:**
  - Unified frontend:
    - Convolutional-beamforming-based dereverberation and separation
  - Unified form of convolutional-beamforming: WPD[3]
    - multi-frame beamforming filter





WPD Beamforming

1. Find a complex filter $\mathbf{h}_f^{(j)} \in \mathbb{C}^C$ that solves the following optimization problem:

$$\begin{cases} \min\limits_{\mathbf{h}_f} & \left(\bar{\mathbf{h}}_f^{(j)}\right)^{\mathrm{H}} \mathbf{R}_f^{(j)} \bar{\mathbf{h}}_f^{(j)} \\ \text{subject to} & \left(\mathbf{h}_{0,f}^{(j)}\right)^{\mathrm{H}} \mathbf{v}_f = 1 \end{cases}$$

2. Apply beamforming:

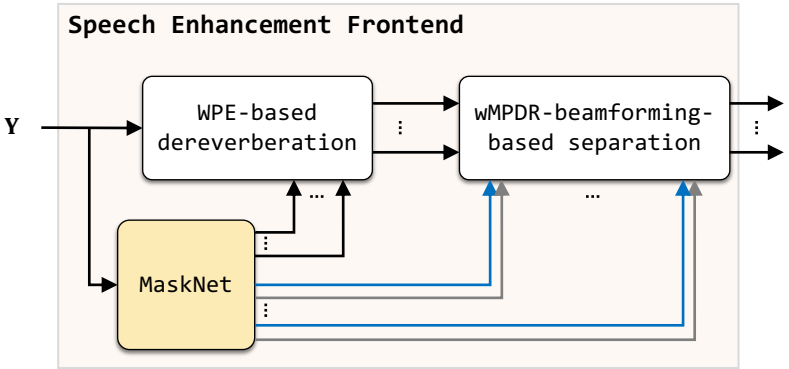
$$\hat{\mathbf{x}}_f^{(j)} = \left(\bar{\mathbf{h}}_f^{(j)}\right)^{\mathrm{H}} \bar{\mathbf{y}}$$

[3] T. Nakatani, and K. Kinoshita, “A unified convolutional beamformer for simultaneous denoising and dereverberation,” IEEE Signal Processing Letters, vol. 26, no. 6, pp. 903–907, 2019.

Proposed model

E2E Dereverberation, Beamforming, and Multi-Talker ASR

- **Model design:**
  - Unified frontend:
    - Convolutional-beamforming-based dereverberation and separation
  - Factorized form of convolutional-beamforming: WPE+wMPDR[4]
    - WPE and beamformer are jointly optimized



wMPDR Beamforming

1. Find a complex filter $\mathbf{h}_f^{(j)} \in \mathbb{C}^C$ that solves the following optimization problem:

$$\begin{cases} \min\limits_{\mathbf{h}_f} \left(\mathbf{h}_f^{(j)}\right)^{\mathrm{H}} \left(\mathbf{\Phi}_{\mathrm{weighted}}^{(j)}\right)_f^{\mathrm{H}} \mathbf{h}_f^{(j)} \\ \text{subject to} \quad \left(\mathbf{h}_f^{(j)}\right)^{\mathrm{H}} \mathbf{v}_f = 1 \end{cases}$$

2. Apply beamforming:

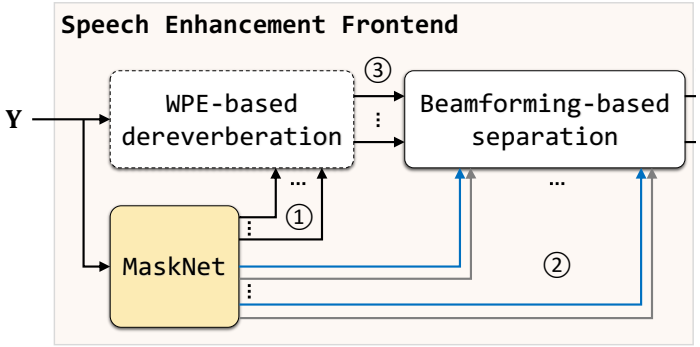
$$\hat{\mathbf{x}}_f^{(j)} = \left(\mathbf{h}_f^{(j)}\right)^{\mathrm{H}} \hat{\mathbf{x}}^{(j),\mathrm{dereverb}}$$

[4] C. Boeddeker, T. Nakatani, K. Kinoshita, and R. Haeb-Umbach, “Jointly optimal dereverberation and beamforming,” in Proc. IEEE ICASSP, 2020, pp. 216–220.

Proposed model

E2E Dereverberation, Beamforming, and Multi-Talker ASR

- **Model design:**
  - Role of NN in Frontend:
    - Providing masks for estimating speech and noise statistics used in WPE and beamforming
    - e.g., $\mathbf{\Phi}_f^{(j)} = \frac{\sum_t \left(\sum_c M_{\text{bf},t,f,c}^{(j)}\right) \mathbf{y}_{t,f}\mathbf{y}_{t,f}^{\mathrm{H}}}{\sum_t \sum_c M_{\text{bf},t,f,c}^{(j)}}$, $\lambda_{t,f}^{(j)} = \frac{1}{C}\sum_c \frac{M_{\text{wpe},t,f,c}^{(j)}}{\sum_t M_{\text{wpe},t,f,c}^{(j)}} \left|\mathbf{y}_{t,f,c}\right|^2$



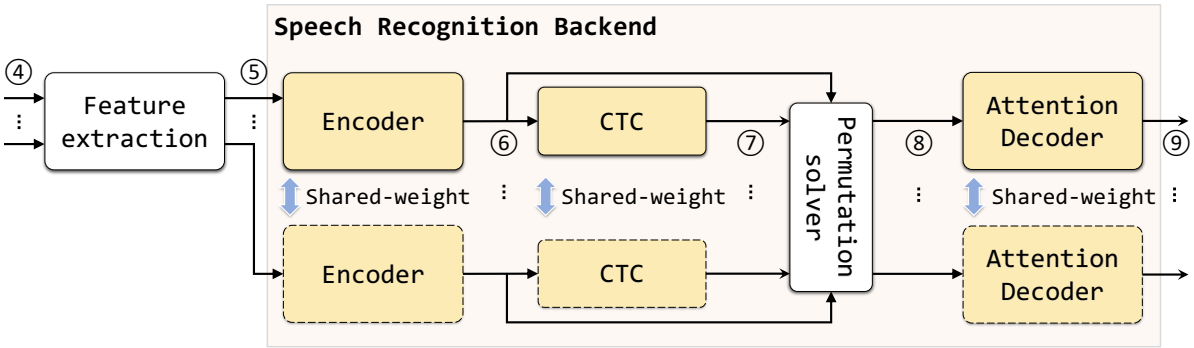
MaskNet

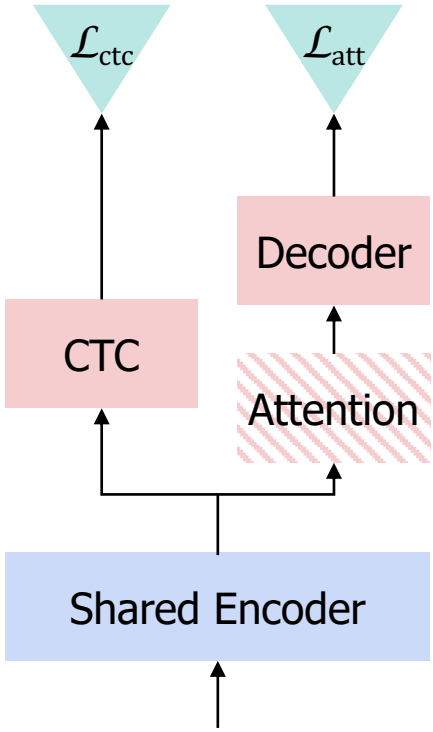
- Each input channel is processed independently.
- The estimated masks for different channels are averaged in the filter part.
- Thus, the MaskNet can be applied to data with different channels and different array geometries.

Proposed model

E2E Dereverberation, Beamforming, and Multi-Talker ASR

- **Model design:**
  - Backend: E2E ASR
    - joint CTC-attention based encoder-decoder[5]
  - The same ASR model is used to process each separated stream.



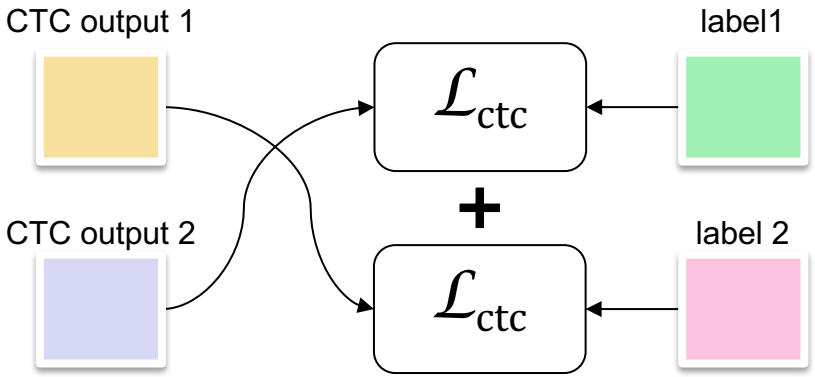


[5] S. Kim, T. Hori, and S. Watanabe, “Joint CTC-attention based end-to-end speech recognition using multi-task learning,” in Proc. IEEE ICASSP, 2017, pp. 4835–4839.

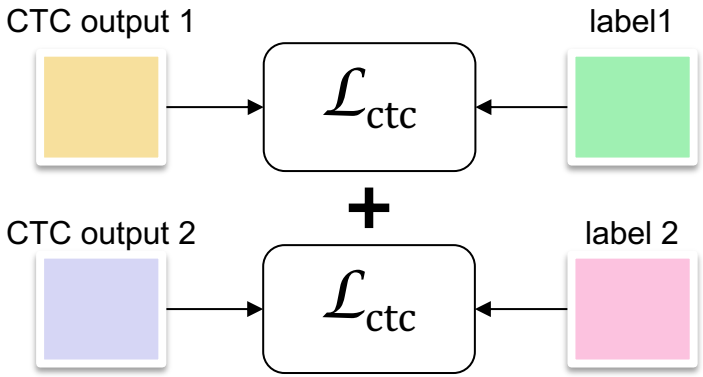
Proposed model

E2E Dereverberation, Beamforming, and Multi-Talker ASR

- **Model design:**
  - Backend: E2E ASR
    - joint CTC-attention based encoder-decoder[5]
  - The same ASR model is used to process each separated stream.
  - The permutation problem is solved using the CTC loss.



Permutation 1

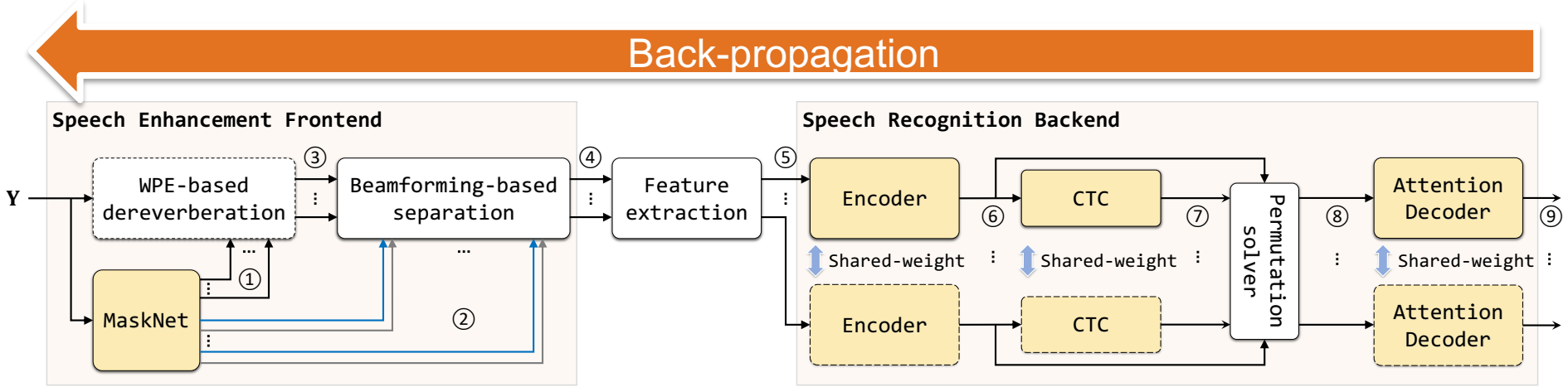


Permutation 2

Proposed approaches

Attacking the numerical instability issue

- **Numerical instability issue in fully end-to-end training**
  - **Observation:** Loss is NaN. / Gradient is NaN.
  - **Consequence:** Hard to converge → Poor performance
  - **Causes:** Unstable complex matrix operations in Frontend
    - e.g., complex matrix inverse, eigenvalue decomposition, ill-conditioned covariance matrix due to poorly-estimated masks



Proposed approaches

Attacking the numerical instability issue

- **Numerical instability issue in fully end-to-end training**
  - **Solutions:**
    1. Diagonal loading

$$\mathbf{\Phi}' = \mathbf{\Phi} + \varepsilon \,\mathrm{Trace}(\mathbf{\Phi})\mathbf{I}$$

    2. Mask flooring

$$\widehat{\mathbf{M}} = \mathrm{Maximum}(\mathbf{M}, \xi)$$

    3. More stable complex matrix operation
       - $\mathbf{\Phi}\mathbf{\Phi}^{-1} = \mathbf{I} \quad \Longrightarrow \quad \mathbf{\Phi}^{-1} = \begin{bmatrix} \mathcal{R}e\{\mathbf{\Phi}^{-1}\} \\ \mathcal{I}m\{\mathbf{\Phi}^{-1}\} \end{bmatrix} = \begin{bmatrix} \mathcal{R}e\{\mathbf{\Phi}\} & \mathcal{I}m\{\mathbf{\Phi}\} \\ -\mathcal{I}m\{\mathbf{\Phi}\} & \mathcal{R}e\{\mathbf{\Phi}\} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix}$
       - $\mathbf{\Phi}^{-1}\mathbf{v} \quad \Longrightarrow \quad \mathrm{Solve}(\mathbf{\Phi}, \mathbf{v})$
    4. Double precision
       - Higher precision mitigates overflow and underflow in some complex operations.

Proposed approaches

Attacking the numerical instability issue

❑ Numerical instability issue in fully end-to-end training

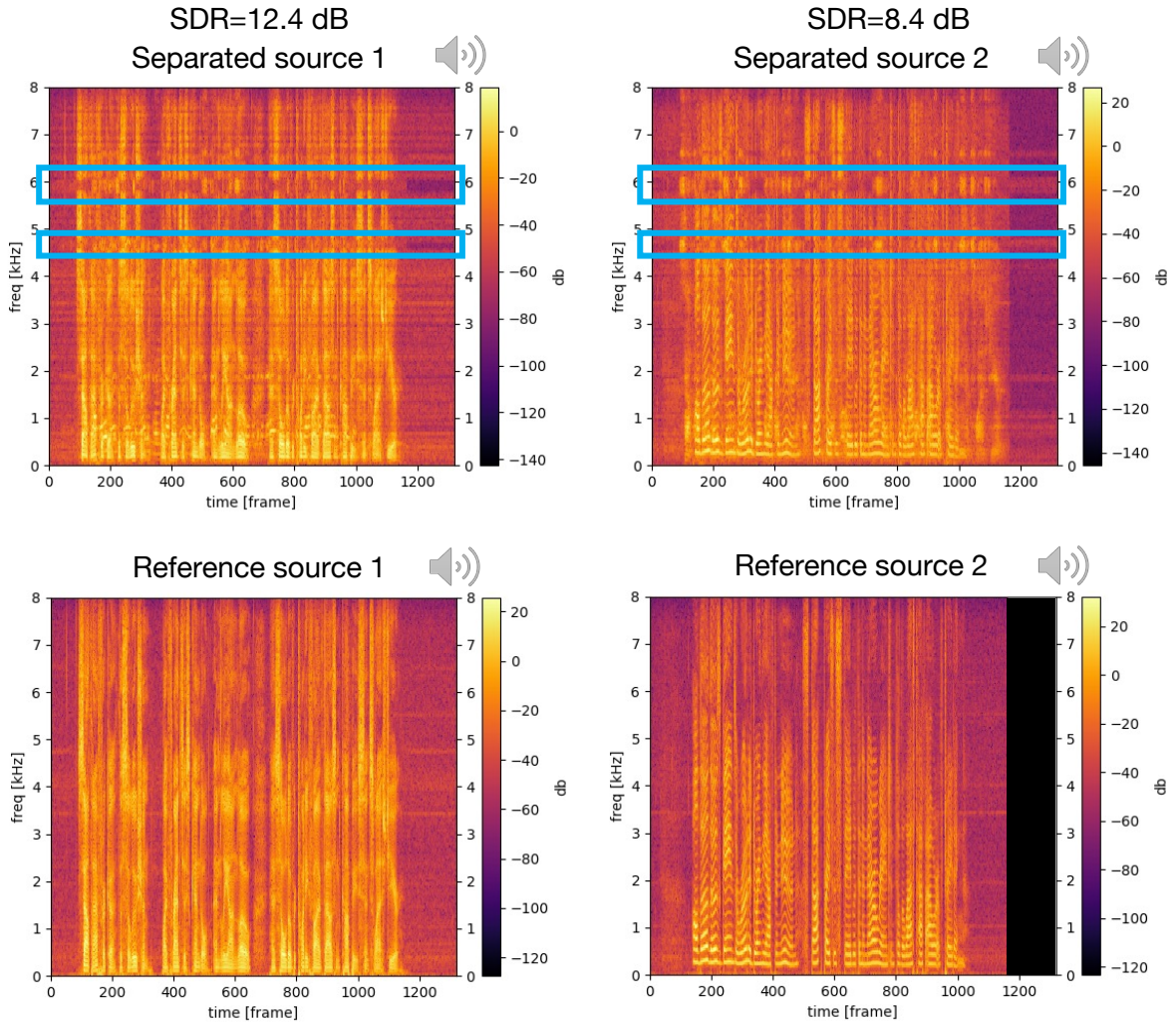
Evaluation of the proposed techniques with the WPE+MVDR+ASR Model of different architectures on the spatialized reverberant WSJ1-2mix evaluation set. The number of filter taps K and channels C are set to 5 and 2 for evaluation (same as training), respectively.

Architecture	WER (%)	PESQ	STOI	SDR (dB)
Original mixture	-	1.20	0.65	-1.45
1-ch 2-spkr ASR	24.86	-	-	-
+ Nara-WPE pre-processing	21.29	-	-	-
Proposed model	16.59	1.30	0.74	2.49
+ (1) Diagonal loading	15.12	**1.32**	**0.75**	**3.25**
+ (2) Mask flooring	16.20	1.30	0.74	2.82
+ (3) Stable complex op.	15.77	1.32	0.75	3.13
+ (4) Double precision	16.43	1.31	0.74	2.87
+ Techs (1)–(4)	**15.01**	1.31	0.74	2.81

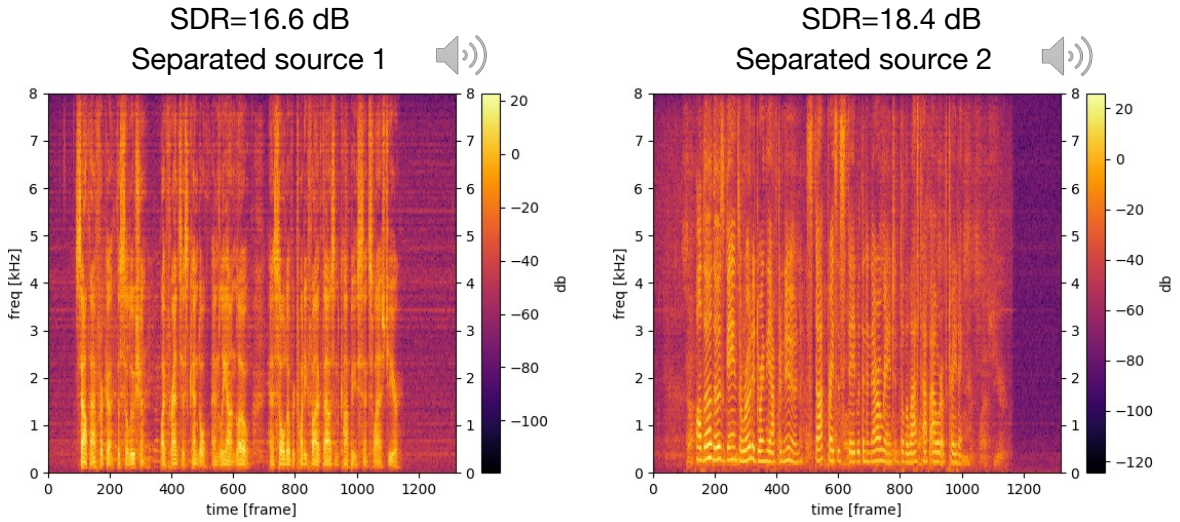
Proposed approaches

Attacking the frequency permutation problem

- **Frequency permutation problem in fully end-to-end training**



- **Frequency permutation problem in fully end-to-end training**
  - **Solutions:**
    1. Voice activity detection (VAD)-like 1-D mask
       - The same mask value is shared among all frequency bins in each frame.



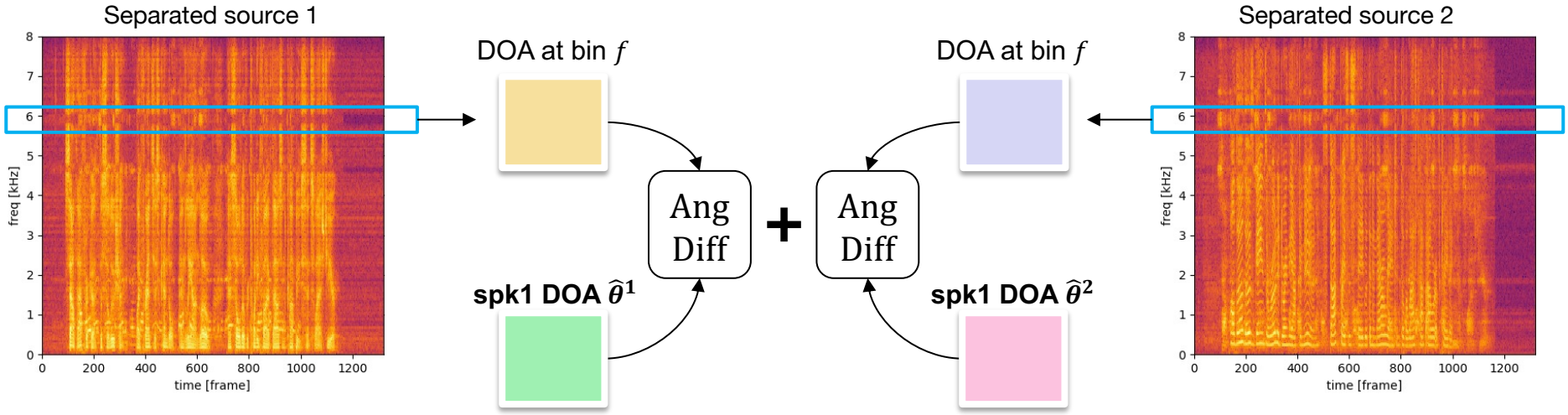
Proposed approaches

Attacking the frequency permutation problem

- **Frequency permutation problem in fully end-to-end training**
  - **Solutions:**
    1. Voice activity detection (VAD)-like 1-D mask
       - The same mask value is shared among all frequency bins in each frame.
    2. Direction-of-arrival (DOA) consistency-based permutation adjustment
       - **Assumption:** The microphone array geometry is known.
       1. Use mask-weighted SRP-PATH[6] to estimate the global DOA of each speaker $\hat{\theta}^{(j)}$
       2. Estimate the DOA at each freq. bin for each separated stream
       3. Adjusting the frequency permutation across the separated stream by matching the estimated freq.-bin-wise DOA with $\hat{\theta}^{(j)}$

[6] Z.-Q. Wang, X. Zhang, and D. Wang, “Robust TDOA estimation based on time-frequency masking and deep neural networks,” in Proc. ISCA Interspeech, 2018, pp. 322–326.

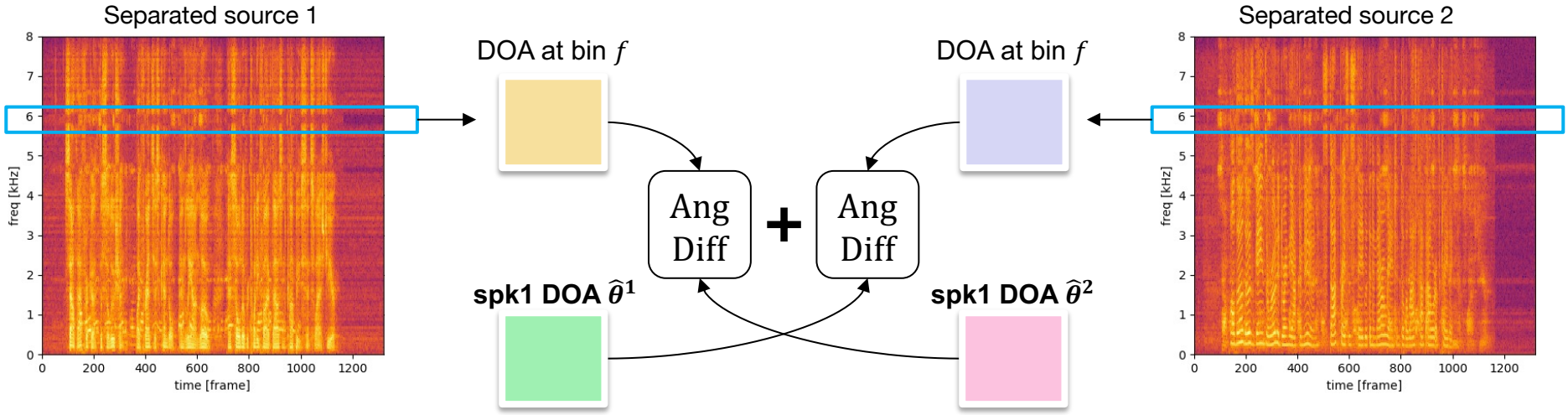
- **Frequency permutation problem in fully end-to-end training**
  - **Solutions:**
    1. Voice activity detection (VAD)-like 1-D mask
       - The same mask value is shared among all frequency bins in each frame.
    2. Direction-of-arrival (DOA) consistency-based permutation adjustment



Proposed approaches

Attacking the frequency permutation problem

- **Frequency permutation problem in fully end-to-end training**
  - **Solutions:**
    1. Voice activity detection (VAD)-like 1-D mask
       - The same mask value is shared among all frequency bins in each frame.
    2. Direction-of-arrival (DOA) consistency-based permutation adjustment



Proposed approaches

Attacking the frequency permutation problem

- **Frequency permutation problem in fully end-to-end training**

Evaluation of different mask types on the spatialized reverberant WSJ1-2mix evaluation set. We compare the best performance of proposed models based on different beamformer variants, by tuning the configurations of $C \in \{2,4,6\}$ and $K \in \{1,3,5,7,10\}$ for each individual model in the evaluation phase.

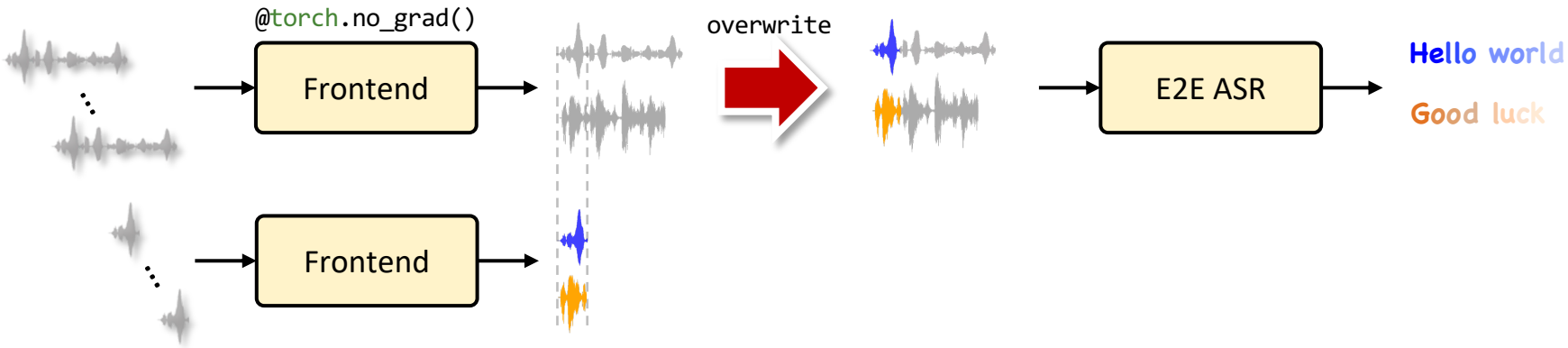
No.	Model (+ASR)	Mask	WER (%)	PESQ	STOI	SDR (dB)
1	WPE+MVDR	T-F	9.50	1.56	0.83	7.73
2	WPE+wMPDR	T-F	9.44	1.63	0.82	8.49
3	WPD	T-F	10.60	1.61	0.82	7.89
4	WPE+MVDR	1-D	9.45	1.95	0.86	12.54
5	WPE+wMPDR	1-D	10.26	1.97	0.86	12.20
6	WPD	1-D	10.48	**2.19**	**0.87**	**14.15**

❑ Frequency permutation problem in fully end-to-end training

Evaluation of the proposed DOA-consistency-based frequency permutation adjustment strategy on the SMS-WSJ evaluation set. The T-F mask-based beamforming is used. We set $K = 5$ and $C = 6$ in the evaluation phase.

Threshold β	MVDR_Souden				MVDR_RTF			
	WER (%)	PESQ	STOI	SDR (dB)	WER (%)	PESQ	STOI	SDR (dB)
$< 0°$	**17.23**	1.69	0.78	3.93	**16.12**	1.68	0.77	3.77
30°	21.67	1.74	0.78	4.67	21.52	1.70	0.76	3.53
60°	25.85	1.74	0.77	4.80	24.86	1.70	0.76	3.63
90°	28.79	1.74	0.77	4.86	28.05	1.70	0.76	3.75
120°	31.95	1.75	0.77	5.00	30.43	1.71	0.76	3.93
150°	34.55	1.75	0.78	5.14	34.00	1.71	0.77	4.10
180°	38.04	1.76	0.78	5.31	37.87	1.73	0.77	4.32
210°	36.83	1.76	0.78	5.35	35.36	1.73	0.77	4.35
240°	34.66	1.76	0.78	5.35	33.03	1.73	0.77	4.36
270°	31.91	1.76	0.78	5.35	29.90	1.73	0.77	4.37
300°	30.48	1.76	0.78	5.36	28.30	1.73	0.78	4.37
330°	29.18	1.76	0.78	5.36	27.08	1.73	0.78	4.38
360°	28.67	**1.76**	**0.78**	**5.37**	26.52	**1.73**	**0.78**	**4.38**

- **Memory consumption issue in fully end-to-end training**
  - **Goal:** to mitigate the large GPU memory consumption issue
  - **Solutions:**
    - Channel sampling
      - Randomly selecting $C' = 2$ channels from training ($C' < C$)
    - Approximated truncated back-propagation through time (TBPTT)[7]



[7] T. von Neumann, K. Kinoshita, L. Drude, C. Boeddeker, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, “End-to-end training of time domain audio separation and recognition,” in Proc. IEEE ICASSP, 2020, pp. 7004–7008.

Proposed approaches

Memory-efficient training strategies

- **Memory consumption issue in fully end-to-end training**

Peak allocated GPU memory when applying different training strategies with batch size 1. The training samples here are all around 24s.

Strategy	Mask	Formula	Mem. (GB)
Plain training (full 6-ch)	T-F	MVDR_Souden	4.484
	1-D	MVDR_Souden	4.057
	T-F	MVDR_RTF	4.889
	1-D	MVDR_RTF	4.484
Channel sampling (2-ch)	T-F	MVDR_Souden	2.538
	1-D	MVDR_Souden	2.383
	T-F	MVDR_RTF	2.366
	1-D	MVDR_RTF	2.393
Approx. TBPTT	T-F	MVDR_Souden	2.019
	1-D	MVDR_Souden	2.000
	T-F	MVDR_RTF	2.013
	1-D	MVDR_RTF	1.996

Proposed approaches

Memory-efficient training strategies

❑ Memory consumption issue in fully end-to-end training

Evaluation of the proposed memory-efficient training strategies on the SMS-WSJ evaluation set. We set $K = 5$ and $C = 6$ in the evaluation phase.

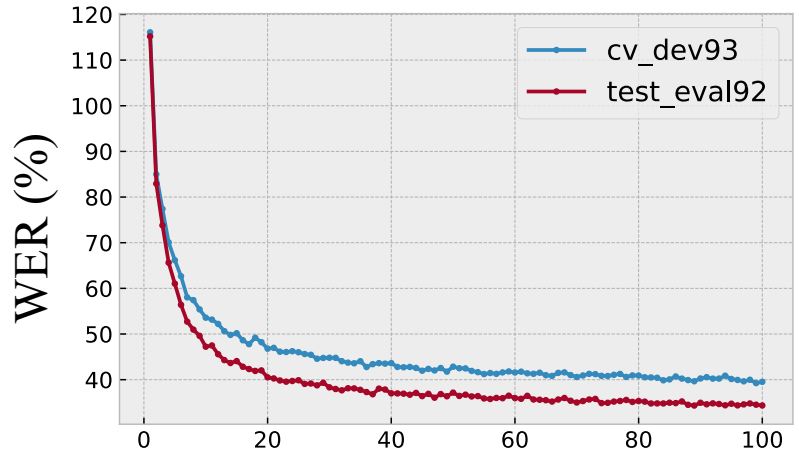
Strategy	Mask	Formula	WER (%)	PESQ	STOI	SDR (dB)
Channel sampling	T-F	MVDR_Souden	17.23	1.69	0.78	3.93
	1-D	MVDR_Souden	17.50	**2.10**	**0.85**	**11.18**
	T-F	MVDR_RTF	16.12	1.68	0.77	3.77
	1-D	MVDR_RTF	17.14	2.05	0.85	10.36
Approx. TBPTT	T-F	MVDR_Souden	18.36	1.71	0.76	4.30
	1-D	MVDR_Souden	16.32	2.09	0.84	10.61
	T-F	MVDR_RTF	**15.94**	1.79	0.78	4.30
	1-D	MVDR_RTF	18.58	2.04	0.83	9.15

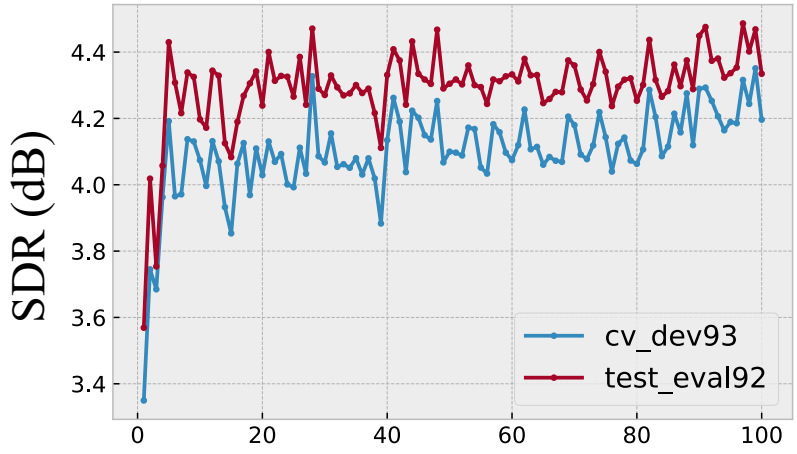
❑ Comparison of different training schemes

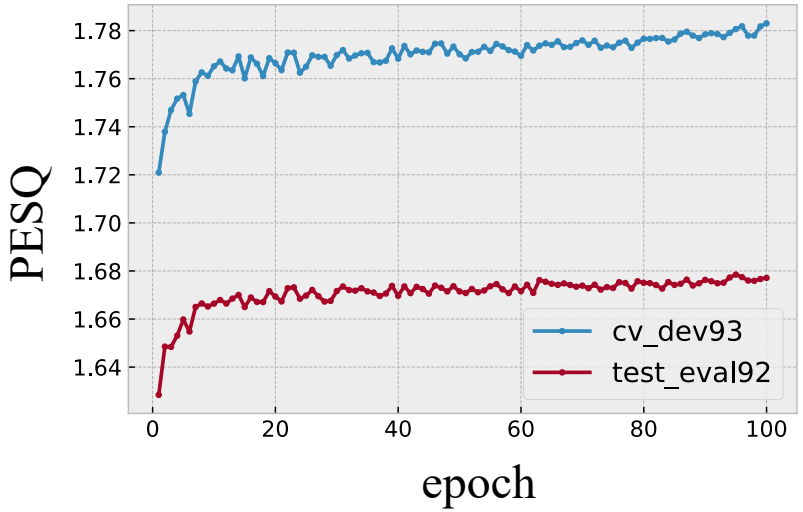
Comparison of different training schemes on the SMS-WSJ evaluation set. We set K = 5 and C = 6 in the evaluation phase.

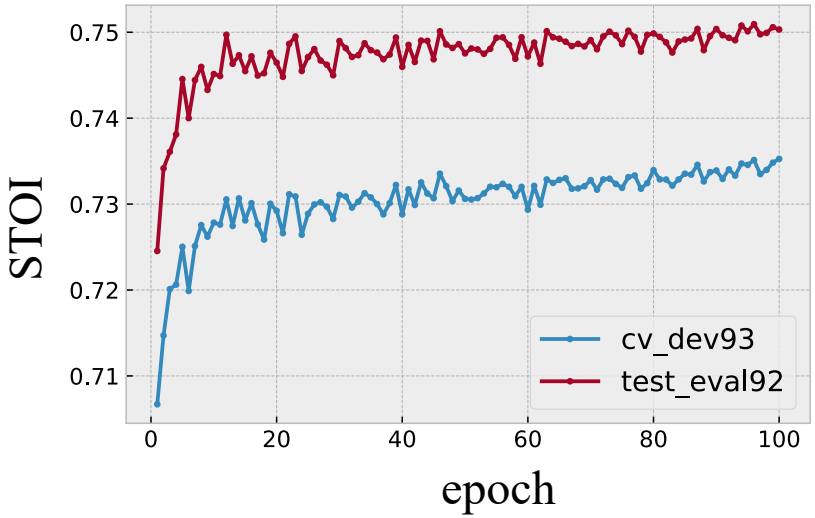
Training scheme	Mask	Formula	WER (%)	PESQ	STOI	SDR (dB)
(1) Independent training	T-F	MVDR_Souden	42.30	2.08	0.83	11.88
	1-D	MVDR_Souden	37.30	**2.13**	0.85	11.95
	T-F	MVDR_RTF	38.20	2.08	0.84	11.38
	1-D	MVDR_RTF	40.30	2.04	0.84	10.90
(2) Fully E2E training	T-F	MVDR_Souden	17.23	1.69	0.78	3.93
	1-D	MVDR_Souden	17.50	2.10	0.85	11.18
	T-F	MVDR_RTF	16.12	1.68	0.77	3.77
	1-D	MVDR_RTF	17.14	2.05	0.85	10.36
(3) Multi-task learning	T-F	MVDR_Souden	15.52	1.98	0.83	10.86
	1-D	MVDR_Souden	**15.36**	2.11	**0.85**	**12.14**
	T-F	MVDR_RTF	17.15	1.82	0.81	7.81
	1-D	MVDR_RTF	15.69	2.07	0.85	11.43

❑ Analysis of full E2E training









Summary & Outlook

Summary

- We proposed a fully E2E method for joint dereverberation, beamforming, and ASR in the cocktail party scenario.
- The proposed model only uses the ASR loss for training, while still achieving decent speech enhancement performance.
- Extensive experiments were conducted to evaluate and analyze the proposed approaches.

Outlook

- Extending E2E training to more tasks
- Multi-channel E2E training for conversational speech (sparsely overlapped)
- Processing of multi-array speech and ad-hoc array speech
- Processing of moving speakers

References

[1] W. Zhang, X. Chang, C. Boeddeker, T. Nakatani, S. Watanabe, and Y. Qian, "End-to-end dereverberation, beamforming, and speech recognition in a cocktail party," IEEE/ACM Trans. ASLP., vol. 30, pp. 3173–3188, 2022. [7]

[2] Z.-Q. Wang, G. Wichern, and J. Le Roux, "Leveraging low-distortion target estimates for improved speech enhancement," 2021, arXiv:2110.00570. [8]

[3] T. Nakatani, and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," IEEE Signal Processing Letters, vol. 26, no. 6, pp. 903–907, 2019. [11]

[4] C. Boeddeker, T. Nakatani, K. Kinoshita, and R. Haeb-Umbach, "Jointly optimal dereverberation and beamforming," in Proc. IEEE ICASSP, 2020, pp. 216–220. [12]

[5] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in Proc. IEEE ICASSP, 2017, pp. 4835–4839. [14]

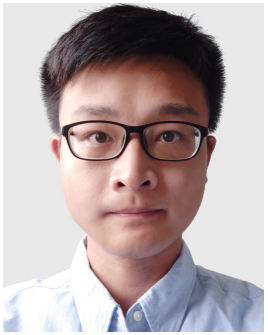
[6] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust TDOA estimation based on time-frequency masking and deep neural networks," in Proc. ISCA Interspeech, 2018, pp. 322–326. [21]

[7] T. von Neumann, K. Kinoshita, L. Drude, C. Boeddeker, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, "End-to-end training of time domain audio separation and recognition," in Proc. IEEE ICASSP, 2020, pp. 7004–7008. [26]

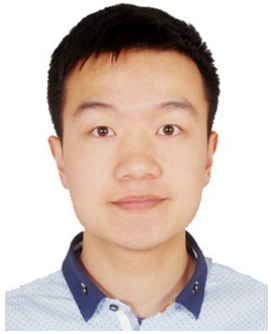
[8] X. Wang, D. Wang, N. Kanda, S. Emre Eskimez, and T. Yoshioka, "Leveraging real conversational data for multi-channel continuous speech separation," in Proc. ISCA Interspeech, 2022, pp. 3814–3818.

THANK YOU!

Q & A


Wangyou Zhang


Yanmin Qian


Xuankai Chang


Christoph Boeddeker


Tomohiro Nakatani


Shinji Watanabe