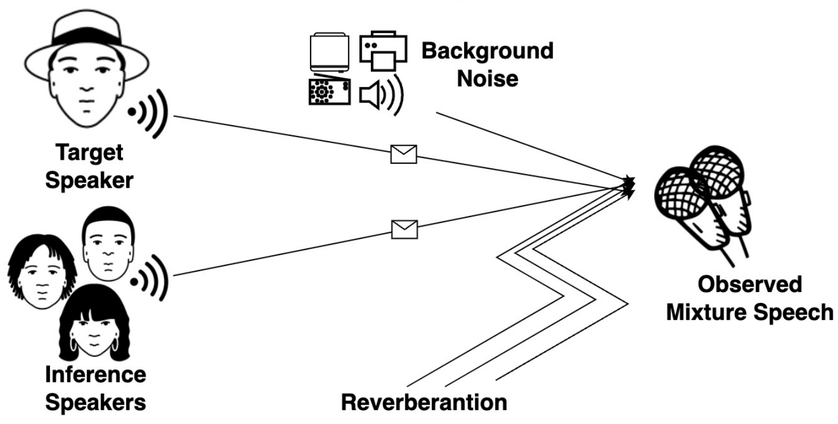
MIIS 2022
International Workshop On Mathematical Issues In Information Sciences
数据科学国际研讨会
17th-18th, Dec. 2022

L-SpEx: Localized Target Speaker Extraction

Meng Ge, Chenglin Xu, Longbiao Wang, Eng Siong Chng, Jianwu Dang, Haizhou Li

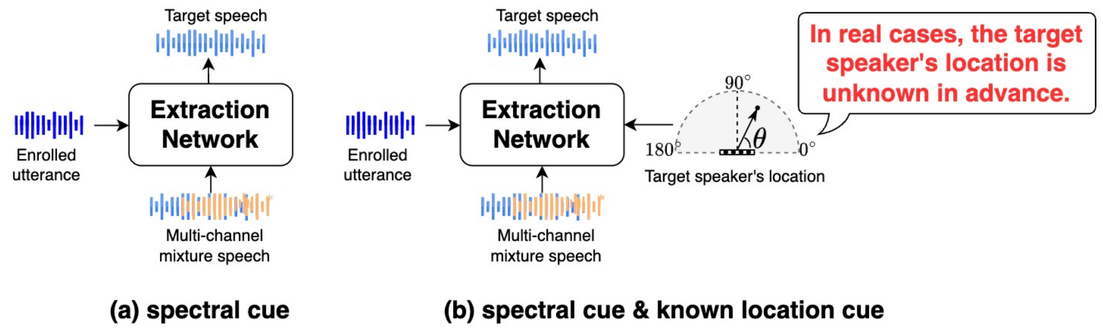
Introduction

- In real-world speech communication, target speech is always mixed with **background interference.**
- **Speaker extraction** aims to extract the target speaker's voice from the mixture speech given a reference utterance of target speaker.



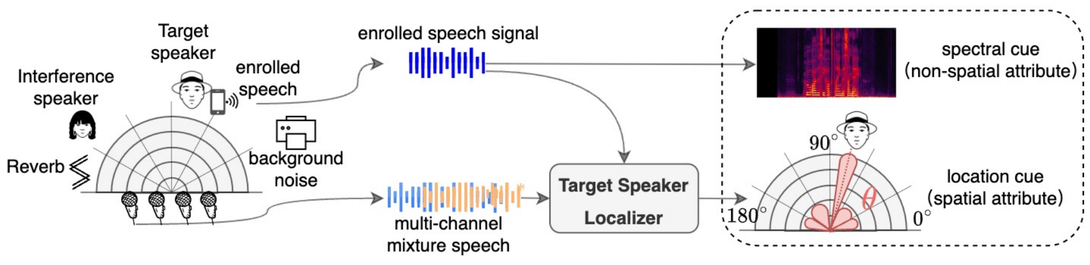
Related Work & Motivation

- **Related Work:** Existing speaker extraction methods extract target speech driven by spectral or spatial cues of target speaker. However, these studies often require the target speaker location is known in advance or detected using an extra visual cue.



(a) spectral cue (b) spectral cue & known location cue

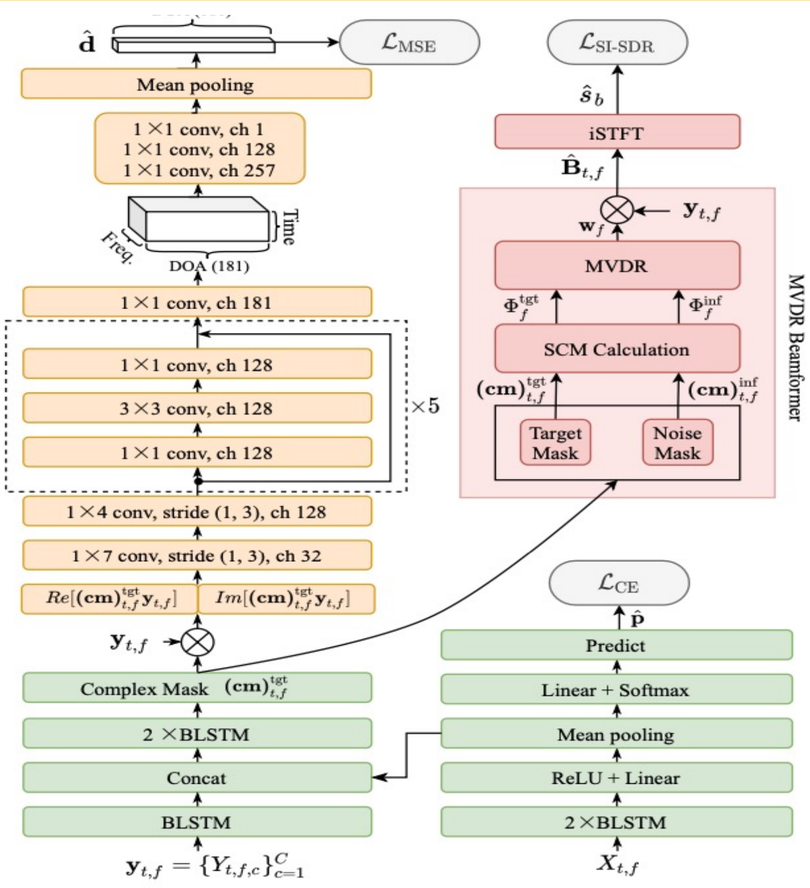
- **Motivation: Taking advantage of the enrolled utterance**, we design a target speaker localizer to estimate target speaker's spatial cues from mixture speech without any assumptions about location.



L-SpEx Architecture

- Target speaker localizer driven by enrolled speech

$$\begin{cases} \hat{\mathbf{d}} = \mathrm{DOAEst}((\mathbf{cm})_{t,f}^{\mathrm{tgt}}\mathbf{y}_{t,f}) \\ \hat{\theta} = \mathrm{argmax}(\hat{\mathbf{d}}) \end{cases} \quad \begin{cases} \mathbf{\Phi}_f^j = \frac{1}{\sum_{t=1}^{T}(\mathbf{cm})_{t,f}^j}\sum_{t=1}^{T}(\mathbf{cm})_{t,f}^j\mathbf{y}_{t,f}\mathbf{y}_{t,f}^H \\ \hat{\mathbf{B}}_{t,f} = \mathbf{w}_f^H\mathbf{y}_{t,f} = \left[\frac{(\mathbf{\Phi}_f^{\mathrm{inf}})^{-1}\mathbf{\Phi}_f^{\mathrm{tgt}}}{\mathrm{tr}((\mathbf{\Phi}_f^{\mathrm{inf}})^{-1}\mathbf{\Phi}_f^{\mathrm{tgt}})}\mathbf{u}\right]^H\mathbf{y}_{t,f} \end{cases}$$



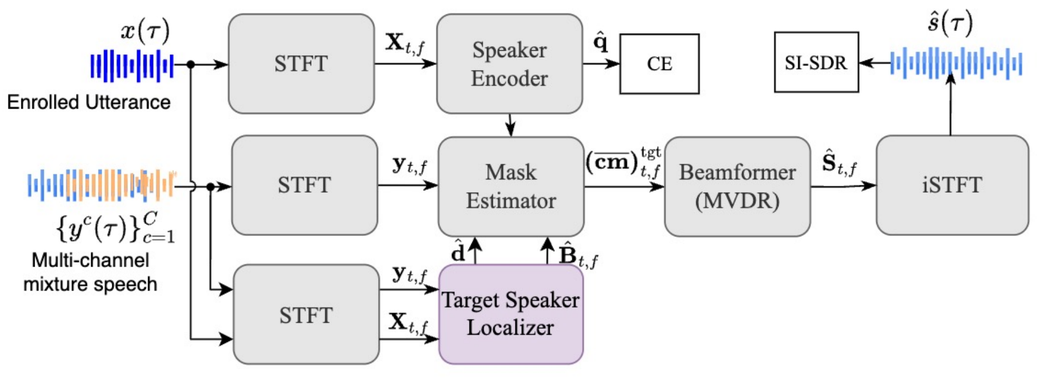
- Target speaker localizer driven by enrolled speech

$$\mathrm{DF}_{\mathrm{angle}}(t,f) = \frac{1}{P}\sum_{l,r\in\Omega} cos(\mathbf{o}_{l,r} - \frac{\pi f_s f \Delta_{l,r} cos\hat{\theta}}{(N_{\mathrm{FFT}}-1)v}),$$

$$\mathrm{DF}_{\mathrm{beam}}(t,f) = \sqrt{Re[\hat{\mathbf{B}}_{t,f}]^2 + Im[\hat{\mathbf{B}}_{t,f}]^2}$$

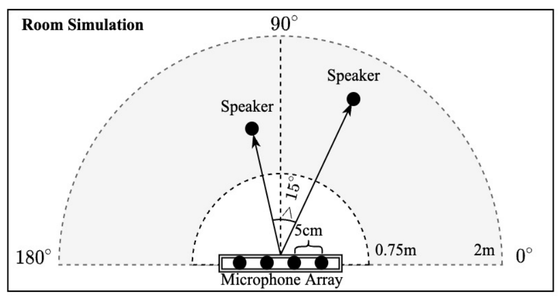
$$\mathbf{y}_{t,f}^{\mathrm{new}} = \mathrm{Concat}[\mathbf{y}_{t,f}, \mathrm{DF}_{\mathrm{beam}}, \mathrm{DF}_{\mathrm{angle}}],$$

$$(\overline{\mathbf{cm}})_{t,f}^{\mathrm{tgt}} = \overline{\mathrm{CMaskEst}}\{\mathbf{y}_{t,f}^{\mathrm{new}}, \overline{\mathrm{Enc}_{\mathrm{speaker}}}(\mathbf{X}_{t,f})\}$$



Dataset, Experimental Results and Discussion

- MC-Libri2Mix Dataset
  - 4-channel reverberated version of Libri2Mix
  - Train/Dev/Test: 127,056 / 2,344 / 6,000



- Results on MC-Libri2Mix

ID	Methods	Mask Type	Spatial Cues		E2E Train	SDR	SI-SDR
			$\mathrm{DF}_{\mathrm{beam}}$	$\mathrm{DF}_{\mathrm{angle}}$			
1	Unprocessed	-	-	-	-	0.46	0.07
2	Mask MVDR (**m**)	**m**	✗	✗	-	8.03	6.36
3	Mask MVDR (**cm**)	**cm**	✗	✗	-	8.02	6.26
4	Pretrained Speaker Localizer	**cm**	✗	✗	-	7.44	5.80
5	L-SpEx	**cm**	✓	✗	✗	8.96	7.17
6		**cm**	✓	✓	✗	9.41	7.29
7		**cm**	✓	✓	✓	**9.68**	**7.45**

- A comparative study of different angle distance

ID	Methods	< 45° (34.6%)		45°-90° (36.5%)		> 90° (28.9%)	
		SDR	SI-SDR	SDR	SI-SDR	SDR	SI-SDR
1	Unprocessed	0.46	0.06	0.43	0.06	0.48	0.08
2	Mask MVDR (**m**)	7.52	5.92	8.35	6.66	8.23	6.51
3	Mask MVDR (**cm**)	7.49	5.82	8.37	6.57	8.22	6.42
4	Pretrained Speaker Localizer	6.95	5.36	7.72	6.09	7.65	6.00
5	L-SpEx	8.27	6.55	9.37	7.56	9.26	7.45
6		8.64	6.63	**9.88**	**7.68**	9.75	7.59
7		**8.97**	**7.06**	9.78	7.66	**9.75**	**7.66**

- A comparative study of different gender mixture

ID	Methods	Diff. Gender (25.2%)		Same Gender (74.8%)	
		SDR	SI-SDR	SDR	SI-SDR
1	Unprocessed	0.35	-0.05	0.49	0.11
2	Mask MVDR (**m**)	9.91	8.22	7.39	5.74
3	Mask MVDR (**cm**)	9.82	8.02	7.42	5.67
4	Pretrained Speaker Localizer	9.02	7.39	6.90	5.28
5	L-SpEx	10.45	8.76	8.46	6.64
6		11.10	9.00	8.85	6.71
7		**11.11**	**9.20**	**8.94**	**6.86**