The 33rd International Joint Conference on Artificial Intelligence, 3 – 9 August 2024, Jeju Island, South Korea

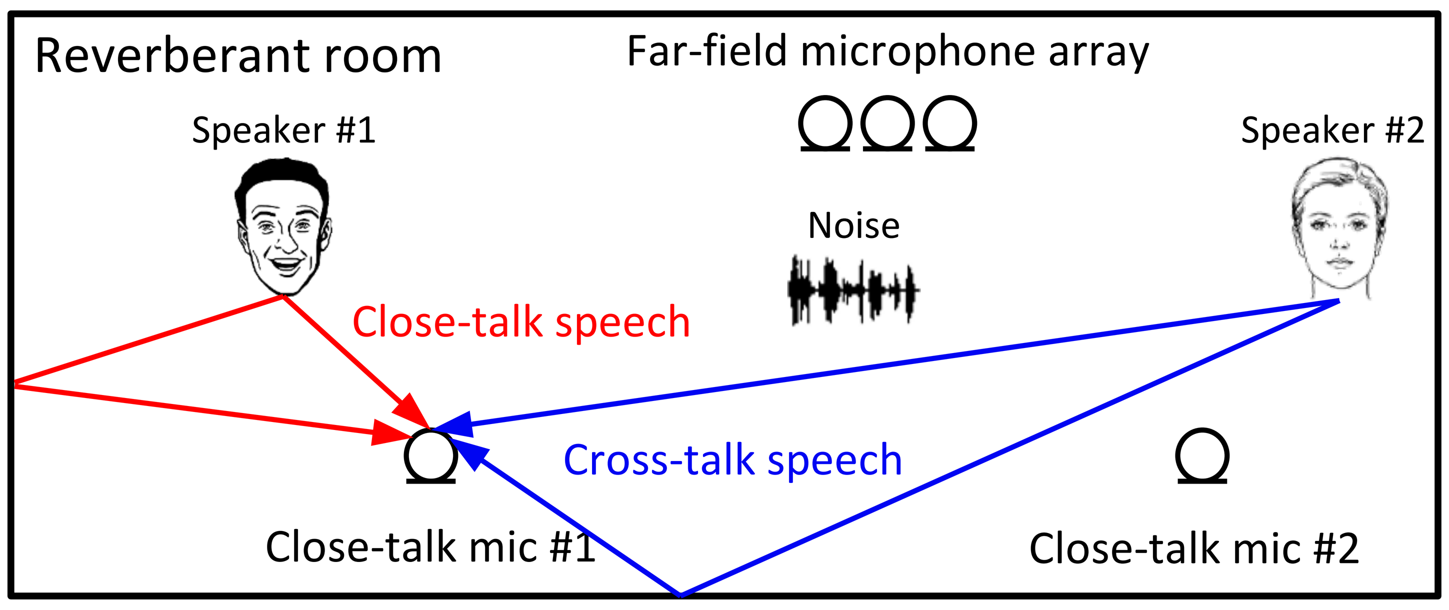
# Cross-Talk Reduction

**Zhong-Qiu Wang**[1], Anurag Kumar[2], and Shinji Watanabe[3]

[1]Southern University of Science and Technology, China
[2]Meta Reality Labs Research, USA [3]Carnegie Mellon University, USA

## 1. Motivation

❑Close-talk mixture has a high input SNR of target speaker, but often contains significant cross-talk speech




❑Cross-talk reduction (CTR) aims at reducing cross-talk speech and enhancing close-talk speech

- Could enable many applications, e.g.,
  - Generate pseudo-labels for real-recorded far-field mixtures
  - Generate pseudo-reference signals for metric computation
  - Reduce labeling efforts of annotators

❑ Supervised CTRnet on simu. data?

- Suffers from generalization issues, as simu. data often mismatches real data

❑ We propose un-/weakly-supervised CTRnet

- Can be trained directly on real data, realizing better generalizability

## 2. Formulating CTR as blind deconvolution

❑Physical model

- Assuming $P$ far-field mics, and $C$ speakers, each wearing a close-talk mic

Close-talk mixture $c$: $Y_c(t,f) = \sum_{c'=1}^{C} X_c(c',t,f) + \varepsilon_c(t,f)$

Far-field mixture $p$: $Y_p(t,f) = \sum_{c=1}^{C} X_p(c,t,f) + \varepsilon_p(t,f)$

Note: subscript indexes mics

❑Let $Z(c) = X_c(c)$ denotes close-talk speech of speaker $c$

$$Y_c(t,f) = Z(c,t,f) + \sum_{c'=1,c'\neq c}^{C} X_c(c',t,f) + \varepsilon_c(t,f)$$

$$= Z(c,t,f) + \sum_{c'=1,c'\neq c}^{C} \boldsymbol{g}_c(c',f)^{\mathsf{H}}\, \widetilde{\mathbf{Z}}(c',t,f) + \varepsilon'_c(t,f)$$

$$Y_p(t,f) = \sum_{c=1}^{C} \boldsymbol{g}_p(c,f)^{\mathsf{H}}\, \widetilde{\mathbf{Z}}(c,t,f) + \varepsilon'_p(t,f)$$

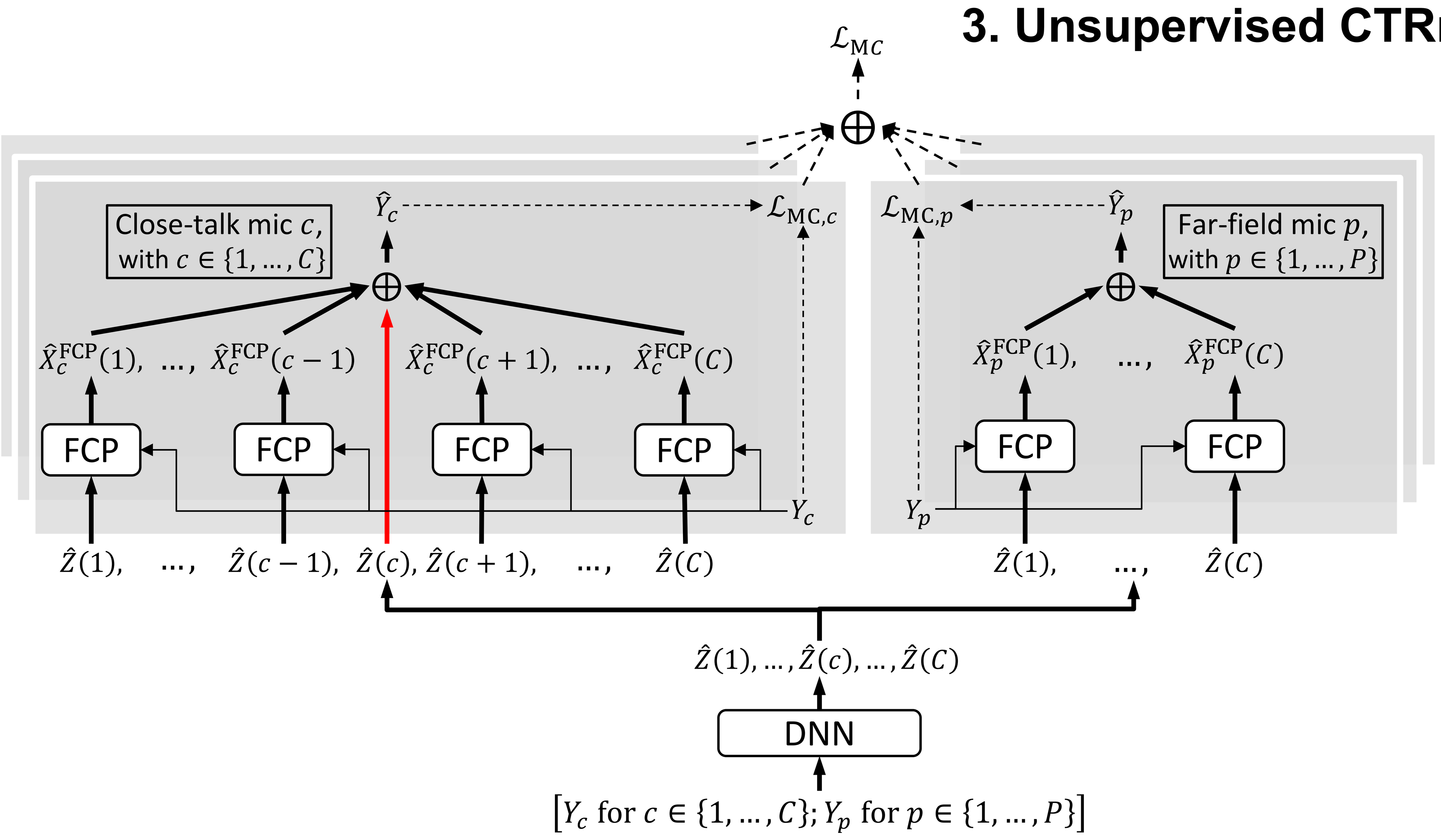
❑CTR via blind deconvolution

$$\underset{Z(\cdot,\cdot,\cdot),\, \boldsymbol{g}.(\cdot,\cdot)}{\operatorname{argmin}} \sum_{c=1}^{C} \sum_{t,f} \left| Y_c(t,f) - Z(c,t,f) - \sum_{c'=1,c'\neq c}^{C} \boldsymbol{g}_c(c',f)^{\mathsf{H}}\, \widetilde{\mathbf{Z}}(c',t,f) \right|^2 + \sum_{p=1}^{P} \sum_{t,f} \left| Y_p(t,f) - \sum_{c=1}^{C} \boldsymbol{g}_p(c,f)^{\mathsf{H}}\, \widetilde{\mathbf{Z}}(c,t,f) \right|^2$$

- Not solvable, if not assuming prior knowledge about filter or source

❑Our solution: model speech patterns via unsupervised deep learning

## 3. Unsupervised CTRnet




❑Input: real & imag. of all close-talk and far-field mixtures

❑Output: real & imag. of close-talk speech of each speaker

❑Loss: mixture-constraint loss

$$\mathcal{L}_{\text{MC}} = \sum_{c=1}^{C} \mathcal{L}_{\text{MC},c} + \alpha \times \sum_{p=1}^{P} \mathcal{L}_{\text{MC},p}$$

$$\mathcal{L}_{\text{MC},c} = \sum_{t,f} \mathcal{F}\left( Y_c(t,f), \hat{Z}(c,t,f) + \sum_{c'=1,c'\neq c}^{C} \hat{\boldsymbol{g}}_c(c',f)^{\mathsf{H}}\, \widetilde{\hat{\mathbf{Z}}}(c',t,f) \right)$$

$$\mathcal{L}_{\text{MC},p} = \sum_{t,f} \mathcal{F}\left( Y_p(t,f), \sum_{c=1}^{C} \hat{\boldsymbol{g}}_p(c,f)^{\mathsf{H}}\, \widetilde{\hat{\mathbf{Z}}}(c,t,f) \right)$$

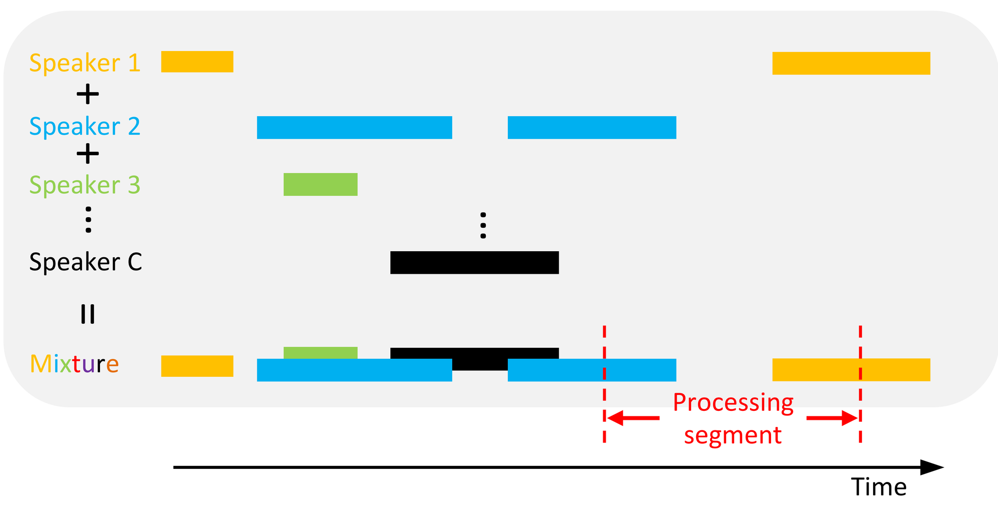
❑Linear filter $\hat{\boldsymbol{g}}.(\cdot,\cdot)$ is estimated via FCP [Wang+2021]

$$\hat{\boldsymbol{g}}_a(c,f) = \underset{\boldsymbol{g}_a(c,f)}{\arg\min} \sum_{t} \frac{\left| Y_a(t,f) - \boldsymbol{g}_a(c,f)^{\mathsf{H}}\, \widetilde{\hat{\mathbf{Z}}}(c,t,f) \right|^2}{|Y_a(t,f)|^2}$$

- $a$ indexes all $P$ far-field and $C$ close-talk mics

## 4. Weakly-supervised CTRnet

❑Realistic speaker overlap is sparse and time-varying




❑Unsupervised CTRnet often under-/over-separates mixed speakers

- Like clustering, assuming more clusters → smaller clusters, but some should be merged

❑ Our solution: leverage speaker-activity timestamps

- Let $d(c) \in \{0,1\}^N$ denote timestamps of speaker $c$, with $N$ denoting #samples
- Muting during training: avoid using predictions in silent ranges for FCP

$$\hat{Z}(c,t,f) := \hat{Z}(c,t,f) \quad \times \quad D(c,t) \quad \times \quad E(c)$$

Speaker $c$ active at frame $t$?

Speaker $c$ active in the training segment?

- Speaker-activity loss: predictions in silent ranges should be zero

$$\mathcal{L}_{\text{SA},c} = \frac{\|\hat{z}(c) \odot (1-d(c))\|_1}{\|y_c \odot (1-d(c))\|_1} \times \frac{N - \|d(c)\|_1}{N}$$

## 5. Experiments

❑On a simu. dataset (2-speaker, reverb, weak noise, fully-overlapped)

- Unsupervised CTRnet works, better than spatial clustering (SC) and IVA

| Systems | $I$ | $J$ | $C$ | $P$ | Masking/Mapping | $\alpha$ | $H/L$ | SI-SDR (dB) ↑ | SDR (dB) ↑ | PESQ ↑ | eSTOI ↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Unprocessed mixture | - | - | - | - | - | - | - | 14.7 | 14.7 | 2.92 | 0.875 |
| Unsupervised CTRnet | 30 | 0 | 2 | 6 | Mapping | $1/P$ | $1/-$ | **26.5** | **26.8** | 3.88 | **0.973** |
| SC [Boeddeker, 2019] | - | - | - | 6 | - | - | - | −1.9 | 7.1 | 2.27 | 0.561 |
| IVA [Scheibler and Saijo, 2022] | - | - | - | 6 | - | - | - | 22.6 | 23.7 | 3.66 | 0.948 |

Table 1: Averaged separation results of unsupervised CTRnet on SMS-WSJ-FF-CT.

❑On real-recorded CHiME-7 (4-speaker, reverb, noisy, sparse overlap)

| | | | | | | | DA-WER (%) ↓ | |
|---|---|---|---|---|---|---|---|---|
| Row | Systems | Muting? | $I$ | $J$ | $C$ | $P$ | Val. | Test |
| 0 | Unprocessed mixture | - | - | - | 4 | - | 28.3 | 27.8 |
| 1 | Unsupervised CTRnet | - | 19 | 1 | 4 | 4 | 22.5 | 25.1 |
| 2 | Weakly-supervised CTRnet | ✗ | 19 | 1 | 4 | 4 | 79.1 | 73.0 |
| 3 | Weakly-supervised CTRnet | ✓ | 19 | 1 | 4 | 4 | 20.5 | 22.6 |
| 4 | GSS [Boeddecker *et al.*, 2018] | - | - | - | 4 | 4 | 26.2 | 26.6 |

Table 3: ASR results of CTRnet on CHiME-7 close-talk mixtures.

Weakly-supervised CTRnet better than (1) unsupervised CTRnet; and (2) guided source separation (GSS)

## 6. Conclusion

❑CTRnet can be trained on real data and can effectively reduce cross-talk speech on real data

❑The proposed un-/weakly-supervised learning based methodology for blind deconvolution works on challenging real data such as CHiME-7