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

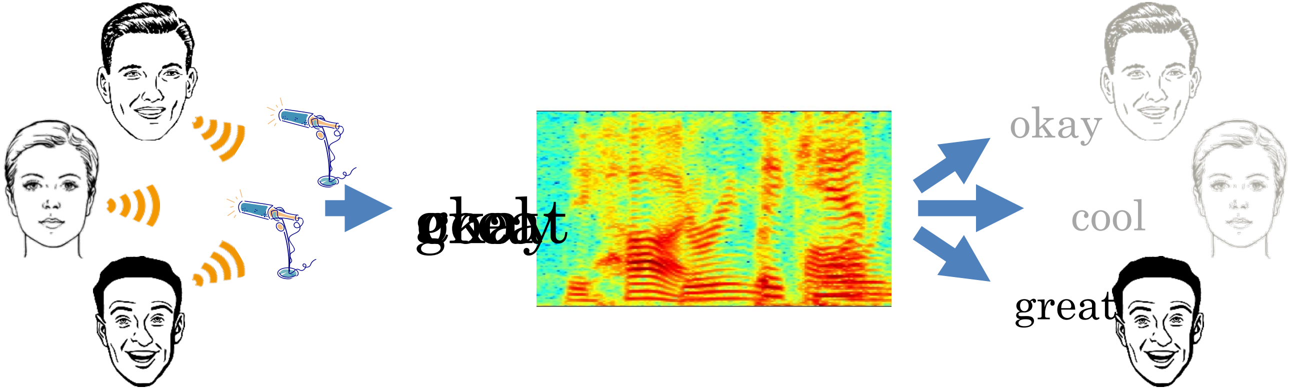
[3]Carnegie Mellon University, USA

http://zqwang7.github.io/

IJCAI JEJU 2024

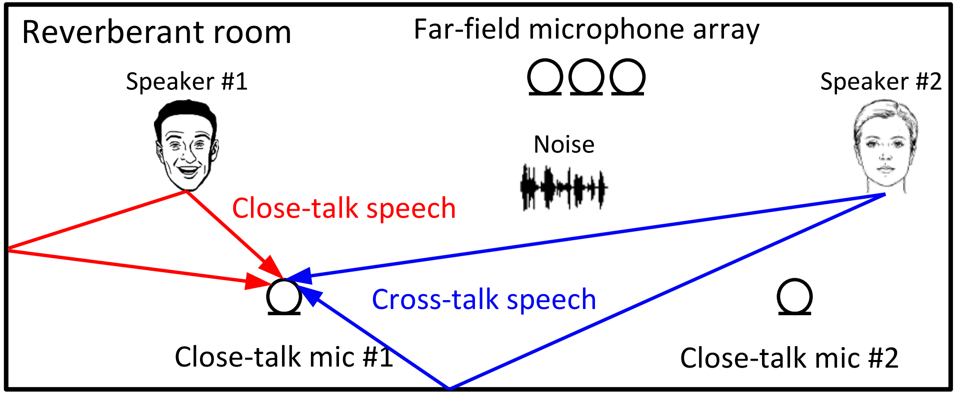
Introduction I

- In many ML / AI applications
  - Sensors usually capture a mixture of target and non-target signals
  - Non-target signals dramatically degrade machine perception
- Multi-speaker audio source separation (*a.k.a.*, the cocktail party problem)
  - Separate mixed speaker signals to individual speaker signals
  - Cross-talk reduction falls into this domain



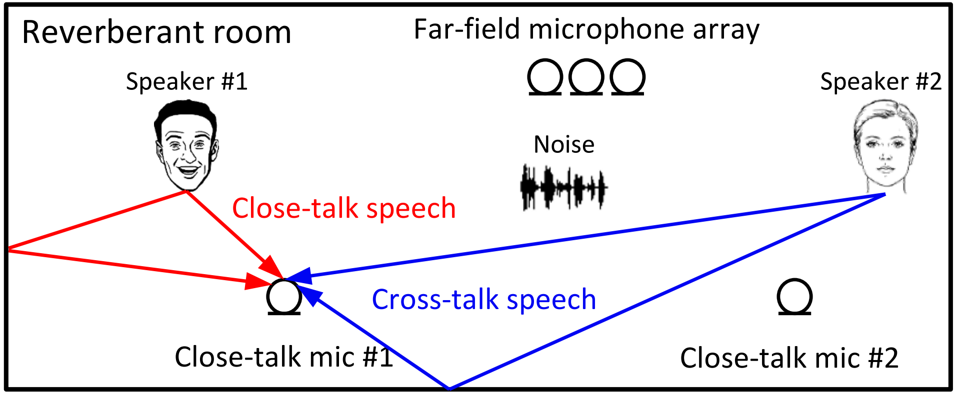
Introduction II

- During data collection, close-talk mixtures are often recorded along with far-field mixtures using close-talk microphones
  - e.g., binaural / lapel microphones
- Close-talk mixture = close-talk speech + cross-talk speech + non-speech signals (e.g., noises)
  - Close-talk speech is often very strong
  - Cross-talk speech by other speakers could also be strong



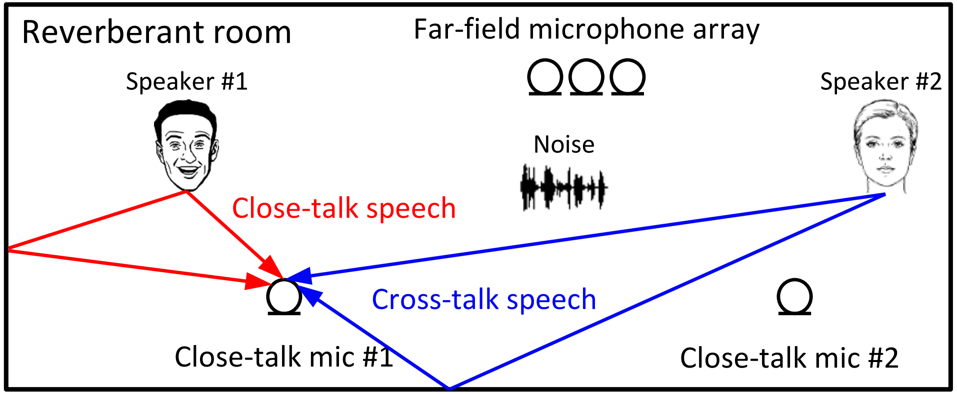
Introduction III

- We propose a novel task: cross-talk reduction (CTR)
  - Reduce cross-talk speech and enhance close-talk speech in each close-talk mixture
- CTR could enable many applications
  - Generate pseudo-labels for real-recorded far-field mixtures
  - Generate pseudo-reference signals for metric computation
  - Reduce labeling efforts of annotators



Introduction IV

- Supervised CTRnet on simulated data ?
  - Leverage room simulators
  - Train supervised DNNs on simulated pairs of close-talk mixtures and clean speech
  - Usually have limited generalizability to real-recorded mixtures
- We propose unsupervised / weakly-supervised CTRnet
  - Can be trained directly on real data, potentially realizing better generalizability



Formulating CTR as blind deconvolution

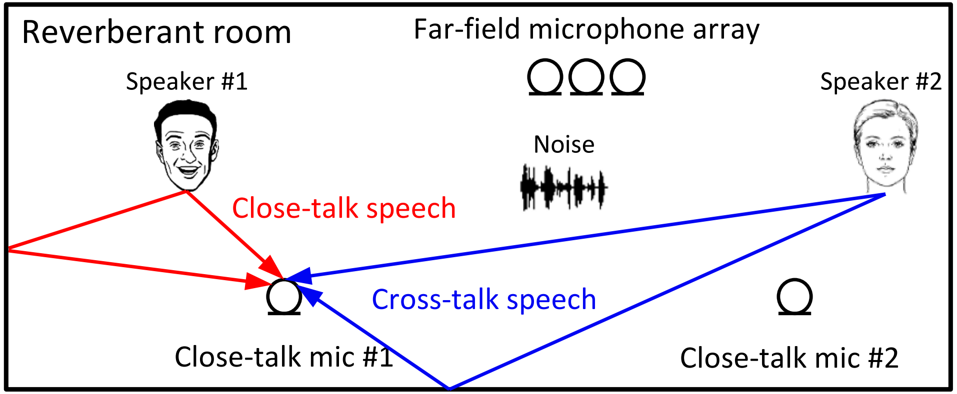
- Physical model
  - Assuming P far-field mics, and C speakers (each wearing a close-talk mic)

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

Image of speaker c' at close-talk mic of speaker c

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

Image of speaker c at far-field mic p



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

Each speaker's image at each mic can be reproduced by linearly filtering its close-talk speech

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

$$Y_c(t,f) = Z(c,t,f) + \sum_{c'=1,c'\neq c}^{C} X_c(c',t,f) + \varepsilon_c(t,f)$$
$$= Z(c,t,f) + \sum_{c'=1,c'\neq c}^{C} \mathbf{g}_c(c',f)^{\mathrm{H}}\, \tilde{\mathbf{Z}}(c',t,f) + \varepsilon'_c(t,f)$$

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

Each speaker's image at each mic can be reproduced by linearly filtering its close-talk speech

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

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

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

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

Formulating CTR as blind deconvolution

$$\underset{Z(\cdot,\cdot,\cdot),\, \boldsymbol{g}_\cdot(\cdot,\cdot)}{\operatorname{argmin}} \sum_{c=1}^{C} \sum_{t,f} \left| Y_c(t,f) - Z(c,t,f) - \sum_{c'=1, c' \neq c}^{C} \boldsymbol{g}_c(c',f)^{\mathrm{H}} \tilde{\mathbf{Z}}(c',t,f) \right|^2$$

Find source and filter most consistent with physical model

$$+ \sum_{p=1}^{P} \sum_{t,f} \left| Y_p(t,f) - \sum_{c=1}^{C} \boldsymbol{g}_p(c,f)^{\mathrm{H}} \tilde{\mathbf{Z}}(c,t,f) \right|^2$$

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

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

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

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

Formulating CTR as blind deconvolution

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

A blind deconvolution problem [Levin+2011]

(not solvable if not assuming prior knowledge about the filter or source)

Our solution:

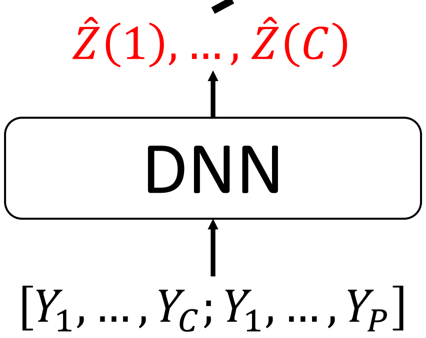
model speech pattern via unsupervised deep learning

Unsupervised CTRnet

$$\mathcal{L}_{\mathrm{MC}} = \sum_{c=1}^{C}\sum_{t,f}\left|Y_c(t,f) - \hat{Z}(c,t,f) - \sum_{c'=1,c'\neq c}^{C} \hat{\boldsymbol{g}}_c(c',f)^{\mathsf{H}}\, \widetilde{\hat{\mathbf{Z}}}(c',t,f)\right|^2$$

$$+\sum_{p=1}^{P}\sum_{t,f}\left|Y_p(t,f) - \sum_{c=1}^{C} \hat{\boldsymbol{g}}_p(c,f)^{\mathsf{H}}\, \widetilde{\hat{\mathbf{Z}}}(c,t,f)\right|^2$$

Optimizing mixture-constraint (MC) loss



- How to compute each $\hat{\boldsymbol{g}}_p(c,f)$?

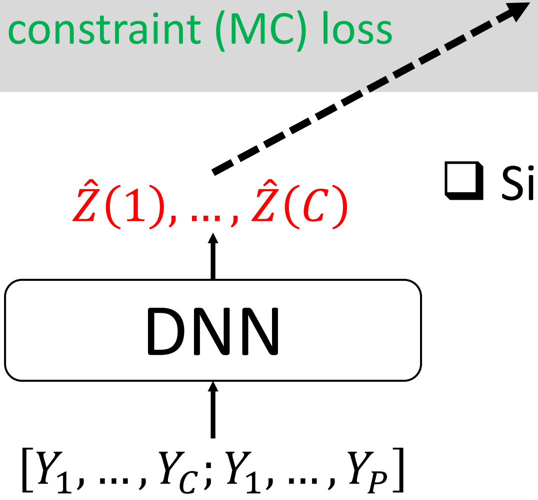
$$\hat{\boldsymbol{g}}_p(c,f) = \underset{\boldsymbol{g}_p(c,f)}{\arg\min} \sum_t \frac{\left|Y_p(t,f) - \boldsymbol{g}_p(c,f)^{\mathsf{H}}\, \widetilde{\hat{\mathbf{Z}}}(c,t,f)\right|^2}{\left|Y_p(t,f)\right|^2}$$

Forward convolutive prediction [Wang+2021]

Unsupervised CTRnet

$$\mathcal{L}_{\text{MC}} = \sum_{c=1}^{C}\sum_{t,f}\left|Y_c(t,f) - \hat{Z}(c,t,f) - \sum_{c'=1,c'\neq c}^{C} \hat{\mathbf{g}}_c(c',f)^{\mathsf{H}}\, \widetilde{\hat{\mathbf{Z}}}(c',t,f)\right|^2 + \sum_{p=1}^{P}\sum_{t,f}\left|Y_p(t,f) - \sum_{c=1}^{C}\hat{\mathbf{g}}_p(c,f)^{\mathsf{H}}\,\widetilde{\hat{\mathbf{Z}}}(c,t,f)\right|^2$$

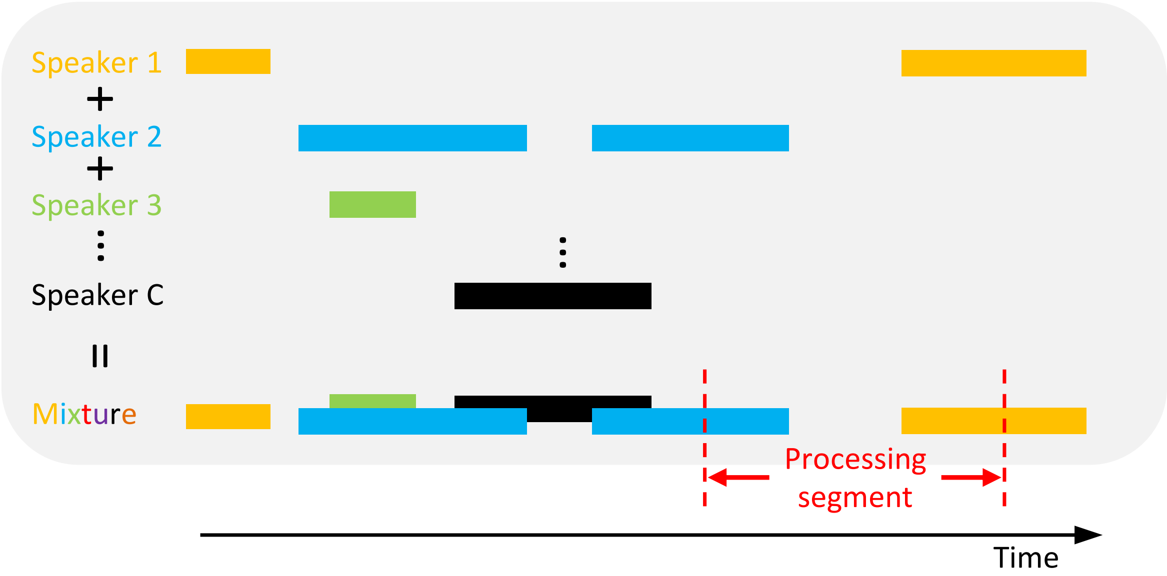
Optimizing mixture-constraint (MC) loss



- Similar to unsupervised clustering
  - Use C source estimates to **explain** $C + P$ mixture signals

Unsupervised CTRnet

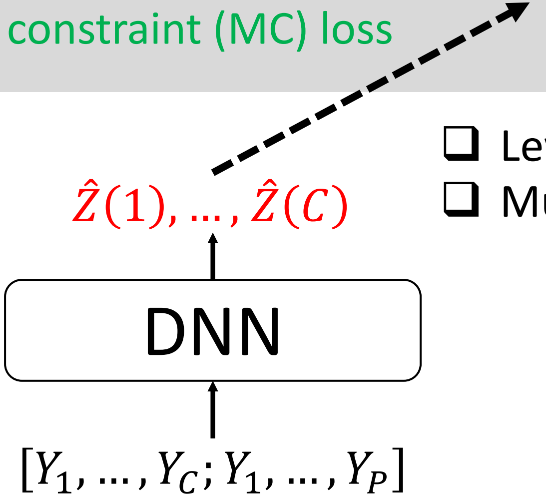
- ❑ Often over-/under-separate mixed speakers, because
  - #active speakers is time-varying
  - Hypothesized #speakers does not match true #speakers



Weakly-supervised CTRnet

$$\mathcal{L}_{\mathrm{MC}} = \sum_{c=1}^{C}\sum_{t,f}\left|Y_c(t,f) - \hat{Z}(c,t,f) - \sum_{c'=1,c'\neq c}^{C}\hat{\boldsymbol{g}}_c(c',f)^{\mathsf{H}}\,\tilde{\hat{\boldsymbol{Z}}}(c',t,f)\right|^2 + \sum_{p=1}^{P}\sum_{t,f}\left|Y_p(t,f) - \sum_{c=1}^{C}\hat{\boldsymbol{g}}_p(c,f)^{\mathsf{H}}\,\tilde{\hat{\boldsymbol{Z}}}(c,t,f)\right|^2$$

Optimizing mixture-constraint (MC) loss



- ❑ Leverage speaker-activity timestamps $d(c) \in \{0,1\}^N$
- ❑ Mute DNN predictions during training

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

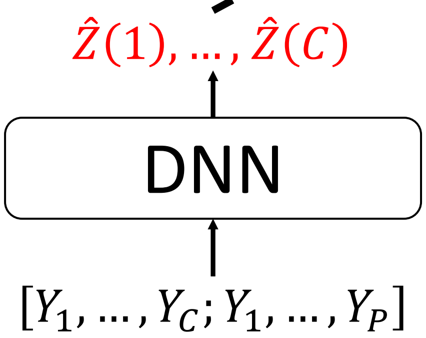
Speaker c active at frame t ?

Weakly-supervised CTRnet

$$\mathcal{L}_{\mathrm{MC}} = \sum_{c=1}^{C} \sum_{t,f} \left| Y_c(t,f) - \hat{Z}(c,t,f) - \sum_{c'=1,c'\neq c}^{C} \hat{\boldsymbol{g}}_c(c',f)^{\mathsf{H}} \, \widetilde{\hat{\boldsymbol{Z}}}(c',t,f) \right|^2$$

$$+ \sum_{p=1}^{P} \sum_{t,f} \left| Y_p(t,f) - \sum_{c=1}^{C} \hat{\boldsymbol{g}}_p(c,f)^{\mathsf{H}} \widetilde{\hat{\boldsymbol{Z}}}(c,t,f) \right|^2$$

Optimizing mixture-constraint (MC) loss



- Leverage speaker-activity timestamps $d(c) \in \{0,1\}^N$
- Mute DNN predictions during training

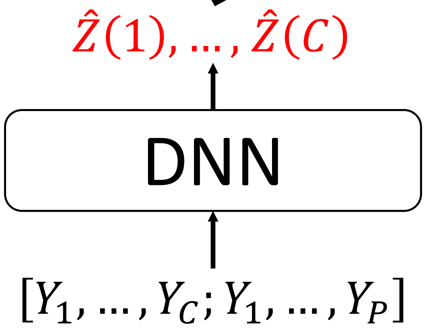
$$\hat{Z}(c,t,f) := \hat{Z}(c,t,f) \quad \times \quad D(c,t)$$

  - $\mathcal{L}_{\mathrm{MC}}$ only penalizes predictions in non-silent ranges

Weakly-supervised CTRnet

$$\mathcal{L}_{\mathrm{MC}} = \sum_{c=1}^{C} \sum_{t,f} \left| Y_c(t,f) - \hat{Z}(c,t,f) - \sum_{c'=1, c' \neq c}^{C} \hat{\boldsymbol{g}}_c(c',f)^{\mathsf{H}} \,\widetilde{\hat{\mathbf{Z}}}(c',t,f) \right|^2 + \sum_{p=1}^{P} \sum_{t,f} \left| Y_p(t,f) - \sum_{c=1}^{C} \hat{\boldsymbol{g}}_p(c,f)^{\mathsf{H}} \widetilde{\hat{\mathbf{Z}}}(c,t,f) \right|^2$$

Optimizing mixture-constraint (MC) loss



- ❑ Leverage speaker-activity timestamps $d(c) \in \{0,1\}^N$
- ❑ Mute DNN predictions during training

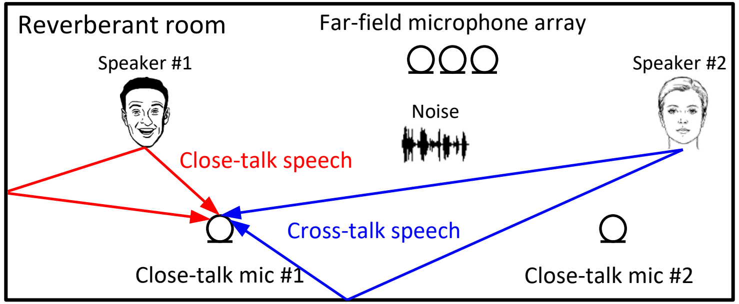
$$\hat{Z}(c,t,f) := \hat{Z}(c,t,f) \quad \times \quad D(c,t)$$

  - ○ $\mathcal{L}_{\mathrm{MC}}$ only penalizes predictions in non-silent ranges
- ❑ Penalizing predictions in silent ranges
  - ○ Predictions in silent ranges should be zero

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

Evaluation Results – Simulated Data

- On a simulated dataset based SMS-WSJ
  - 2-speaker mixtures
  - Reverb + weak noise
  - fully-overlapped speakers

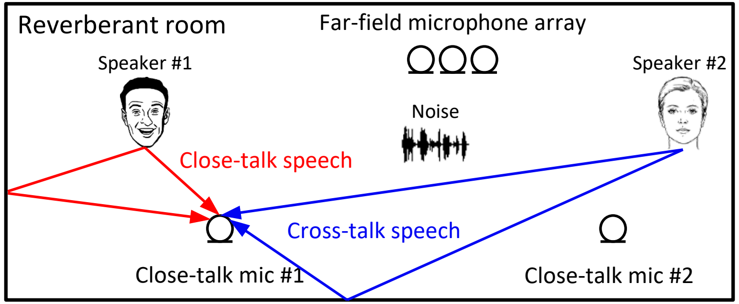


Systems	SI-SDR (dB) ↑	SDR (dB) ↑	PESQ ↑	eSTOI ↑
Unprocessed mixture	14.7	14.7	2.92	0.875
Unsupervised CTRnet	**26.5**	**26.8**	3.88	**0.973**
SC [Boeddeker, 2019]	−1.9	7.1	2.27	0.561
IVA [Scheibler and Saijo, 2022]	22.6	23.7	3.66	0.948

- Unsupervised CTRnet works almost perfectly in simulated cases
- Better than spatial clustering (SC) and independent vector analysis (IVA)

Evaluation Results – Real Data

- CHiME-7 close-talk mixtures
  - 4-speaker mixtures
  - Noisy-reverb
  - Sparse speaker overlap
  - Conversational setup
- Use speech recognition performance for comparison



Row	Systems	Muting?	I	J	C	P	DA-WER (%) ↓ 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

- Weakly-supervised CTRnet better than unsupervised CTRnet
- Better than guided source separation (GSS)

Conclusion

- CTRnet
  - Can be trained directly on real data
  - Can effectively reduce cross-talk speech on real data

- Our learning based methodology for blind deconvolution shows strong potential on challenging real data such as CHiME-7

Thanks!

Definition of $\widetilde{\widetilde{\mathbf{Z}}}(c,t,f)$

$$\widetilde{\widetilde{\mathbf{Z}}}(c,t,f) = \begin{bmatrix} \hat{Z}(c, t-I, f), \\ \dots \\ \hat{Z}(c,t,f), \\ \dots \\ \hat{Z}(c, t+J, f) \end{bmatrix} \in \mathbb{C}^{I+1+J}$$

, stack $I+1+J$ nearby T-F units