UNSSOR: Unsupervised Neural Speech Separation by Leveraging Over-determined Training Mixtures

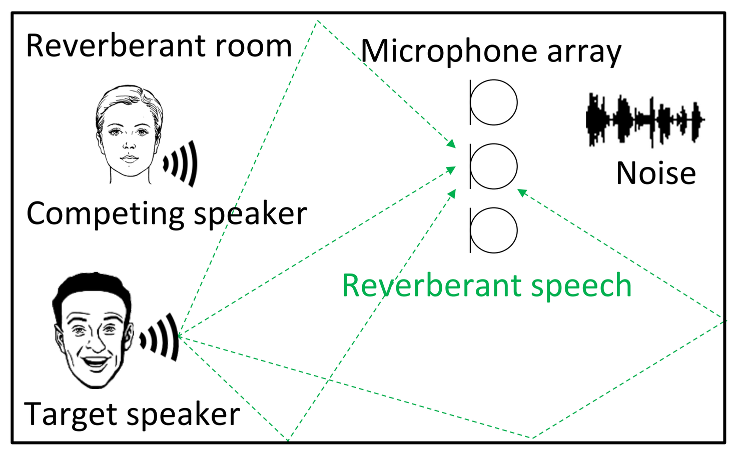
NeurIPS | 2023

Zhong-Qiu Wang and **Shinji Watanabe**

Carnegie Mellon University

Unsupervised separation: motivations

- Speech separation, *a.k.a.* cocktail party problem, aims at separating multi-speaker mixture to individual speaker signals



- Supervised separation
  - Use synthetic training data, exhibiting generalization problems
- Unsupervised separation
  - Leverage unlabeled data for training, alleviating generalization issues
  - Earlier studies still require synthesized mixtures (e.g., MixIT), or rely on conventional spatial processing (e.g., unsupervised deep clustering)
- Our solution: training DNNs directly on mixtures for separation

Problem formulation

- Unsupervised monaural separation is ill-posed
  - Assuming C speakers, 1 microphone
  - Physical model: $Y_1(t,f) = \sum_{c=1}^{C} X_1(c,t,f) + \varepsilon_1(t,f)$
  - A possible solution

Mixture-constraint loss:

$$\mathcal{L}_{\text{MC}} = \sum_{t,f} \left| Y_1(t,f) - \sum_{c=1}^{C} \hat{X}_1(c,t,f) \right|^2$$

$\hat{X}_1(1)$ ... $\hat{X}_1(c)$ ... $\hat{X}_1(C)$

TF-GridNet [Wang+2023] (single-channel complex spectral mapping)

Y_1

#Equations		#Unknowns
$T \times F$	<	$C \times T \times F$
One for each $Y_1(t,f)$		One for each $X_1(c,t,f)$

ill-posed problem with infinite solutions

- Turning into a well-posed problem
  - Assuming C speakers, P microphones
  - Each mixture can add a **constraint** to narrow down the solutions

$\begin{bmatrix} X_1(c,t-A,f), \\ \dots \\ X_1(c,t,f) \\ \dots \\ X_1(c,t+(K-A-1),f) \end{bmatrix} \in \mathbb{C}^K$ stacks nearby T-F bins

Mic 1: $Y_1(t,f) = \sum_{c=1}^{C} X_1(c,t,f) + \varepsilon_1(t,f)$

...

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

...

Mic P: $Y_P(t,f) = \sum_{c=1}^{C} X_P(c,t,f) + \varepsilon_P(t,f)$

➡

$Y_1(t,f) = \sum_{c=1}^{C} X_1(c,t,f) + \varepsilon_1(t,f)$

...

$Y_p(t,f) = \sum_{c=1}^{C} g_p(c,f)^{\mathsf{H}} \widetilde{X}_1(c,t,f) + \varepsilon_p(t,f)$

...

$Y_P(t,f) = \sum_{c=1}^{C} g_P(c,f)^{\mathsf{H}} \widetilde{X}_1(c,t,f) + \varepsilon_P(t,f)$

$X_1(c,t,f)$ $\quad g_p(c,f) \in \mathbb{C}^K$

#Equations		#Unknowns
$P \times T \times F$	>	$C \times T \times F + (P-1) \times C \times F \times K$

When $P > C$ (over-determined) and T is reasonably large (enough observations)

Well-posed problem where a solution for separation exists

Can ***minimize*** mixture constraints at all mics

$$\sum_{t,f} \left(\left| Y_1(t,f) - \sum_{c=1}^{C} X_1(c,t,f) \right|^2 + \sum_{p=2}^{P} \left| Y_p(t,f) - \sum_{c=1}^{C} g_p(c,f)^{\mathsf{H}} \widetilde{X}_1(c,t,f) \right|^2 \right)$$

Proposed algorithm: UNSSOR

- Solve a blind deconvolution problem [Levin+2021]

$$\underset{X_1(\cdot,\cdot,\cdot),\, g_\cdot(\cdot,\cdot)}{\text{argmin}} \sum_{t,f} \left(\left| Y_1(t,f) - \sum_{c=1}^{C} X_1(c,t,f) \right|^2 + \sum_{p=2}^{P} \left| Y_p(t,f) - \sum_{c=1}^{C} g_p(c,f)^{\mathsf{H}} \widetilde{X}_1(c,t,f) \right|^2 \right)$$

  - Not solvable if not assuming prior knowledge about the filter or the source
  - Our solution: model speech patterns via unsupervised deep learning
- UNSSOR

Mixture constraint at all microphones

$$\mathcal{L}_{\text{MC}} = \sum_{t,f} \left(\left| Y_1(t,f) - \sum_{c=1}^{C} \hat{X}_1(c,t,f) \right|^2 + \sum_{p=2}^{P} \left| Y_p(t,f) - \sum_{c=1}^{C} \hat{g}_p(c,f)^{\mathsf{H}} \widetilde{\hat{X}}_1(c,t,f) \right|^2 \right)$$

$\hat{X}_1(1)$ ... $\hat{X}_1(c)$ ... $\hat{X}_1(C)$

TF-GridNet [Wang+2023] (multi-channel complex spectral mapping)

$[Y_1, \dots, Y_p, \dots, Y_P]$

- $\mathcal{L}_{\text{MC}}$ determines whether $\hat{X}_1$ is good, providing sample-level supervision
- How to compute $\hat{g}_p(c,f)$?

$$\hat{g}_p(c,f) = \underset{g_p(c,f)}{\text{arg min}} \sum_t \frac{\left| Y_p(t,f) - g_p(c,f)^{\mathsf{H}} \widetilde{\hat{X}}_1(c,t,f) \right|^2}{|Y_p(t,f)|^2}$$

Forward convolutive prediction (FCP) [Wang+2021]

- FCP filters $\hat{X}_1$ to approximate speaker images at other mics
  - When $\hat{X}_1$ is reasonably accurate

Let $Y_p = X_p(c) + V_p(c)$

Ideally, $V_p(c)$ is independent from $\hat{X}_1(c)$ and $X_p(c)$

$$\hat{g}_p(c,f) = \underset{g_p(c,f)}{\text{arg min}} \sum_t \frac{\left| Y_p(t,f) - g_p(c,f)^{\mathsf{H}} \widetilde{\hat{X}}_1(c,t,f) \right|^2}{|Y_p(t,f)|^2}$$

$$= \underset{g_p(c,f)}{\text{arg min}} \sum_t \frac{\left| X_p(c,t,f) + V_p(c,t,f) - g_p(c,f)^{\mathsf{H}} \widetilde{\hat{X}}_1(c,t,f) \right|^2}{|Y_p(t,f)|^2}$$

$$= \underset{g_p(c,f)}{\text{arg min}} \sum_t \frac{\left| X_p(c,t,f) - g_p(c,f)^{\mathsf{H}} \widetilde{\hat{X}}_1(c,t,f) \right|^2 + |V_p(c,t,f)|^2}{|Y_p(t,f)|^2}$$

$$= \underset{g_p(c,f)}{\text{arg min}} \sum_t \frac{\left| X_p(c,t,f) - g_p(c,f)^{\mathsf{H}} \widetilde{\hat{X}}_1(c,t,f) \right|^2}{|Y_p(t,f)|^2}$$

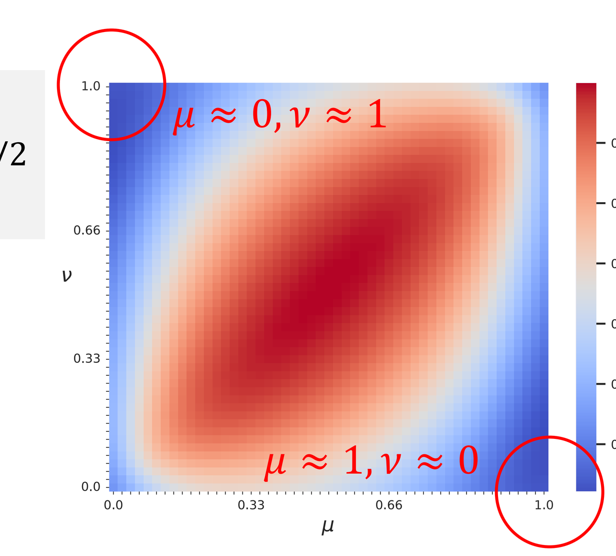
- Minimizing $\mathcal{L}_{\text{MC}}$ promotes separation
  - Hypothesized separation results

$\hat{X}_1(1) = \mu \times X_1(1) + \nu \times X_1(2) + \varepsilon_1/2$

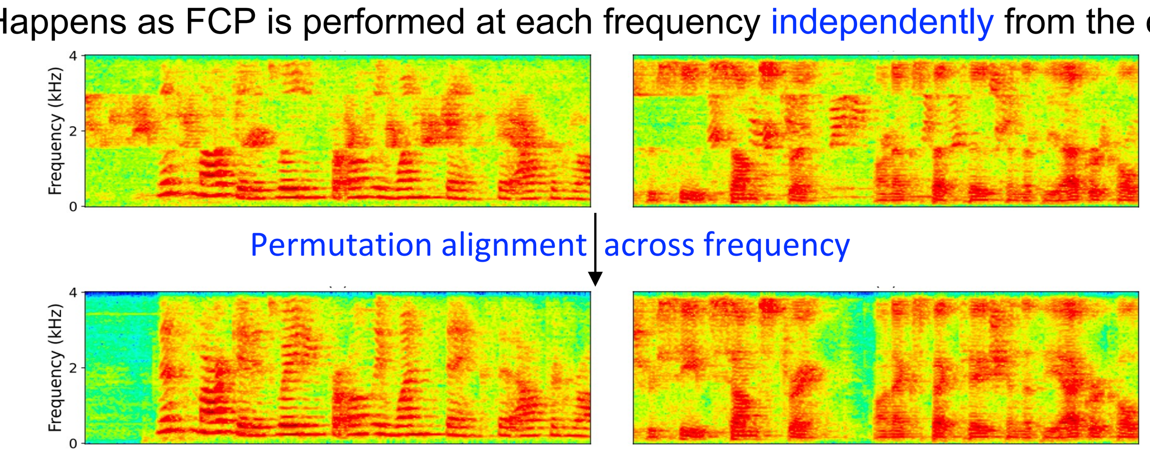
$\hat{X}_1(2) = (1-\mu) \times X_1(1) + (1-\nu) \times X_1(2) + \varepsilon_1/2$

where $0 \le \mu \le 1$, $0 \le \nu \le 1$, and 2 speakers

  - Good separation
    - $\mu \approx 0, \nu \approx 1$ and $\mu \approx 1, \nu \approx 0$
  - Bad separation
    - $\mu \approx 0, \nu \approx 0$ and $\mu \approx 1, \nu \approx 1$
    - μ, ν both away from 0 and 1



- Frequency permutation problem
  - Happens as FCP is performed at each frequency independently from the others



  - Propose to addressing frequency permutation during training
  - Intra-source magnitude scattering (ISMS) loss
    - Source magnitudes are more scattered when frequency permutation happens

$$\mathcal{L}_{\text{ISMS}} = \sum_{p=1}^{P} \frac{\sum_t \frac{1}{C} \sum_{c=1}^{C} \text{var}\left(\log\left|\hat{X}_p^{\text{FCP}}(c,t,\cdot)\right|\right)}{\sum_t \text{var}\left(\log|Y_p(t,\cdot)|\right)}$$

  - Combine $\mathcal{L}_{\text{MC}}$ and $\mathcal{L}_{\text{ISMS}}$ for training

UNSSOR for under-determined separation

- Monaural input, but loss on multiple microphones

Mixture constraint at all microphones

$$\mathcal{L}_{\text{MC}} = \sum_{t,f} \left(\left| Y_1(t,f) - \sum_{c=1}^{C} \hat{X}_1(c,t,f) \right|^2 + \sum_{p=2}^{P} \left| Y_p(t,f) - \sum_{c=1}^{C} \hat{g}_p(c,f)^{\mathsf{H}} \widetilde{\hat{X}}_1(c,t,f) \right|^2 \right)$$

$\hat{X}_1(1)$ ... $\hat{X}_1(c)$ ... $\hat{X}_1(C)$

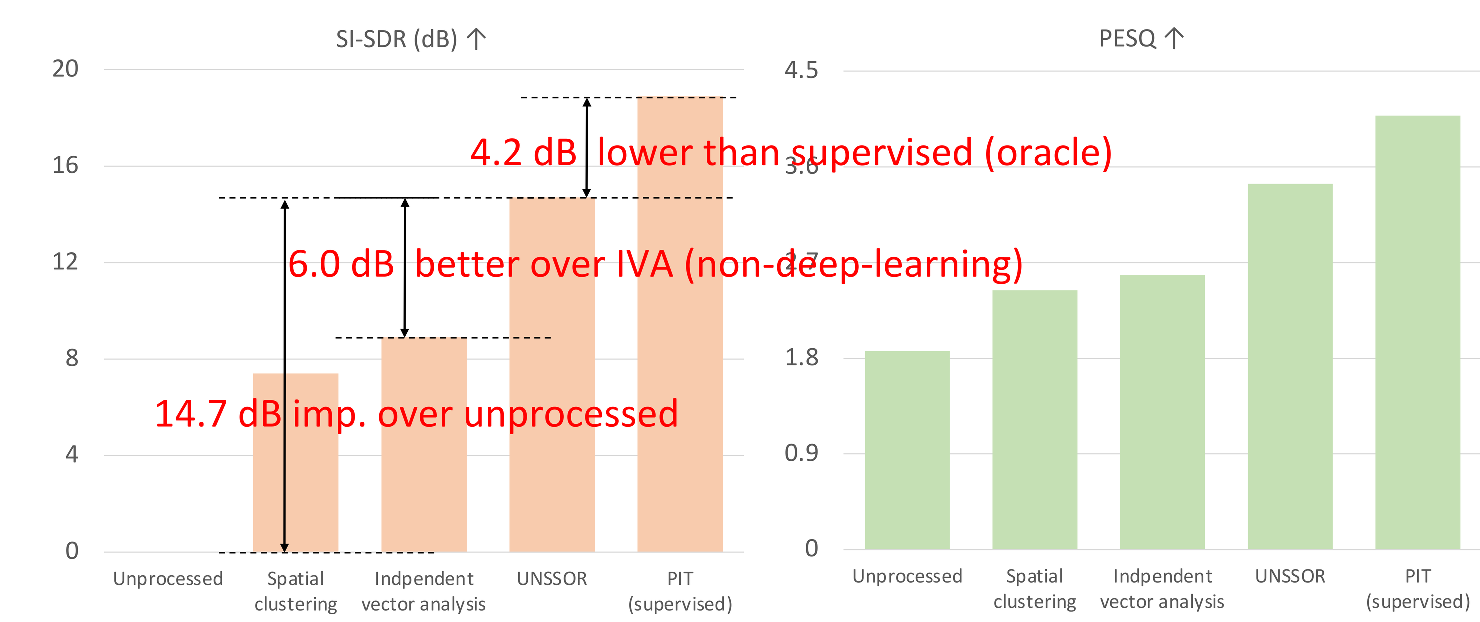
TF-GridNet [Wang+2023] (multi-channel complex spectral mapping)

Y_1

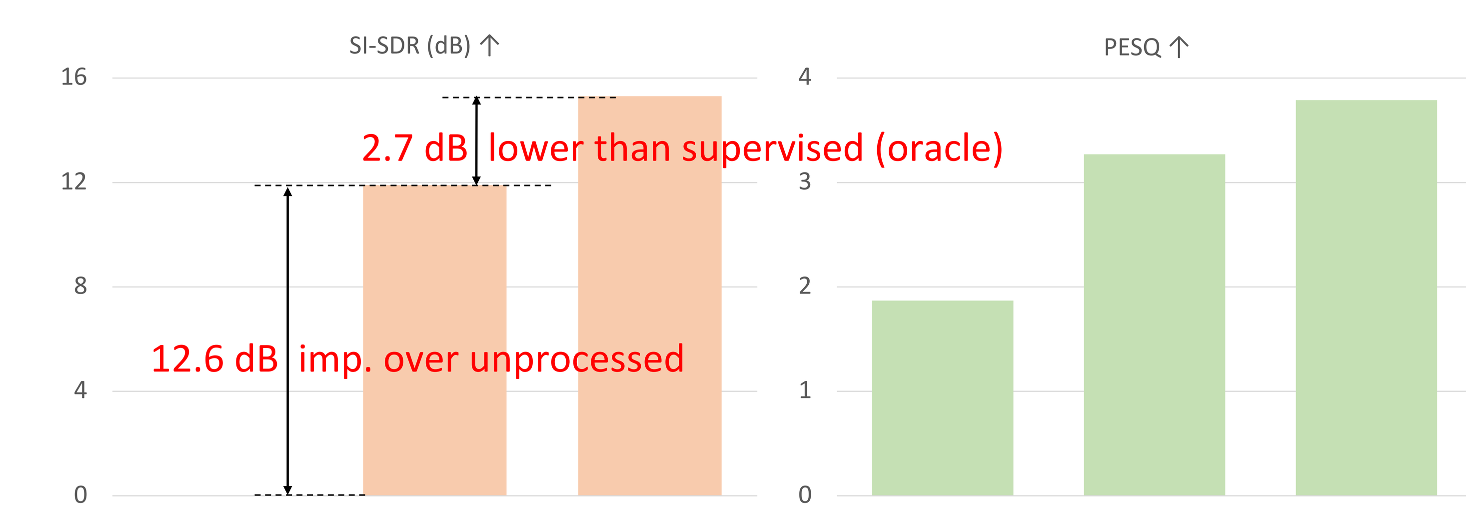
- Require over-determined training mixtures
- At run time, perform under-determined separation (i.e., $P < C$)
- Was considered ill-posed if training mixtures are monaural

Evaluation results

- SMS-WSJ dataset [Drude+19]: reverb 2-speaker mixture with weak noise
- Results on 2-speaker separation (6-channel input and loss)



- Results on 2-speaker separation (1-channel input, 6-channel loss)



References

Levin et al. (2011), "Understanding Blind Deconvolution Algorithms," In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 12, pp. 2354–2367.

Wang et al. (2021), "Convolutive Prediction for Monaural Speech Dereverberation and Noisy-Reverberant Speaker Separation," In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3476–3490.

Wang et al. (2023), "TF-GridNet: Integrating Full- and Sub-Band Modeling for Speech Separation," In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 3221–3236.

Drude et al. (2019), "SMS-WSJ: Database, Performance Measures, and Baseline Recipe for Multi-Channel Source Separation and Recognition," In: arXiv preprint arXiv:1910.13934.

Sound demo