LOCATION AS SUPERVISION FOR WEAKLY SUPERVISED MULTI-CHANNEL SOURCE SEPARATION OF MACHINE SOUNDS

MITSUBISHI ELECTRIC
Changes for the Better

Aalto University School of Electrical Engineering

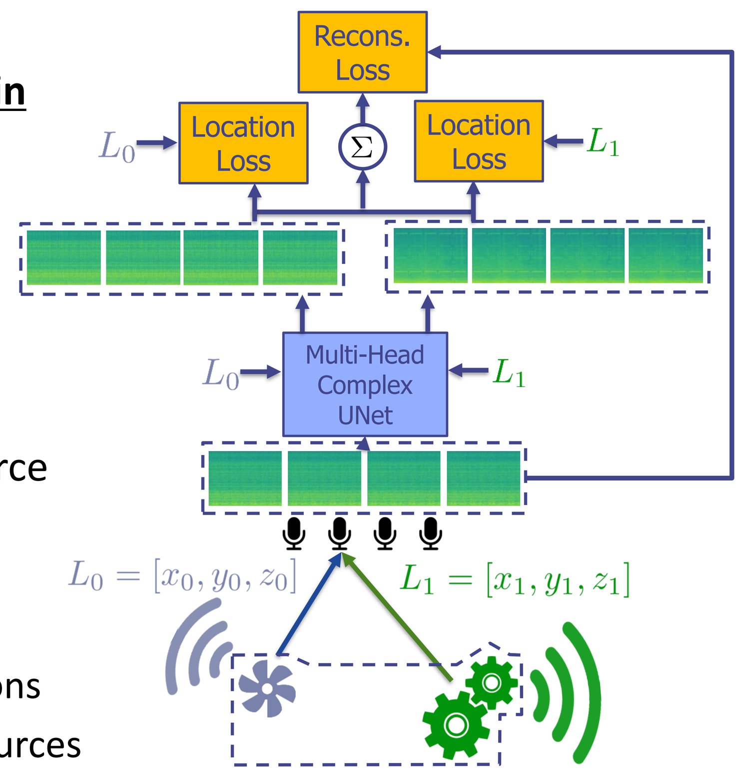
IEEE WASPAA 2023

Ricardo Falcon-Perez[1,2], Gordon Wichern[1], François G. Germain[1], Jonathan Le Roux[1]
[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA,
[2]Acoustics Lab, Aalto University, Espoo, Finland

Overview

Weakly supervised source separation

- We want to separate sounds that **cannot be recorded in isolation**
- Application: Machines with multiple noise-generating parts (e.g., fans, gearboxes, valves) that need to be operated simultaneously
- We propose a loss function based on the *difference between expected and measured time delays* across a microphone array, under the assumption that the source location is known a priori



Experiments & Results

- We simulate a dataset with challenging acoustical conditions
- We use samples from DCASE 2021 Task 2 dataset [1] as sources
- NN architecture: Complex Unet [2] with 1 decoder output per source
- NN input features: complex STFT, IPDs, directional features [2], frequency positional encodings
- Results show **better separation than signal-agnostic beamformers**
- However, performance still lags fully-supervised setting

Method

Feature extraction based on measured interchannel phase differences (IPD), target phase differences (TPD), and directional features for P channels:

$$\text{Real-IPD}^p_{t,f}(\mathbf{Y}) = \angle \mathbf{Y}^{p_0}_{t,f} - \angle \mathbf{Y}^p_{t,f} \in \mathbb{R}$$

$$\text{IPD}^p_{t,f}(\mathbf{Y}) = \cos(\text{Real-IPD}^p_{t,f}(\mathbf{Y})) + j\sin(\text{Real-IPD}^p_{t,f}(\mathbf{Y})) \in \mathbb{C}$$

$$\text{TPD}^p_f(L_i) = \cos(2\pi f \tau(L_i, p)) + j \sin(2\pi f \tau(L_i, p)) \in \mathbb{C}$$

$$d_{t,f}(\mathbf{Y}, L_i) = \sum_{p=1}^{P-1} \text{TPD}^p_f(L_i)\overline{\text{IPD}}^p_{t,f}(\mathbf{Y}) \in \mathbb{C}$$

Measured phase difference of the input mixture

Expected phase difference, based on the time delay of a point source arriving at each mic

Reconstruction loss ensures consistency:

$$\mathcal{L}_{\text{spec}} = \left\|\mathbf{Y} - \hat{\mathbf{Y}}\right\|_2^2 + \left\||\hat{\mathbf{Y}}| - |\mathbf{Y}|\right\|_2^2$$

$$\mathcal{L}_{\text{spat}} = \left\|\mathbf{y}\mathbf{y}^{\mathsf{T}} - \hat{\mathbf{y}}\hat{\mathbf{y}}^{\mathsf{T}}\right\|_2^2$$

Location loss ensures separation:

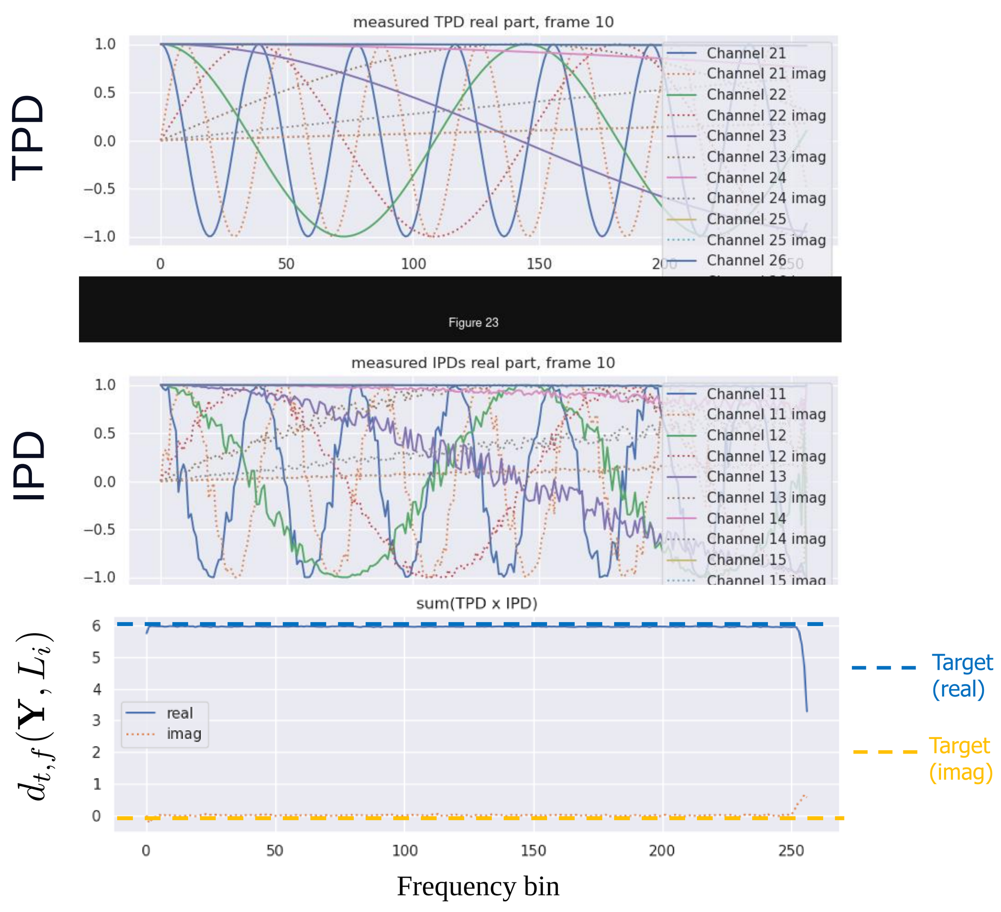
$$\mathcal{L}_{\text{loc}} = \sum_{i=1}^{2}\sum_{f}\sum_{t}\left(\left\|\Re(d_{t,f}(\hat{\mathbf{S}}_i, L_i)) - P\right\|_2^2 + \left\|\Im(d_{t,f}(\hat{\mathbf{S}}_i, L_i)) - 0\right\|_2^2\right)$$

Total loss, combines all of them:

$$\mathcal{L} = \beta_{\text{spec}}\mathcal{L}_{\text{spec}} + \beta_{\text{spat}}\mathcal{L}_{\text{spat}} + \beta_{\text{loc}}\mathcal{L}_{\text{loc}}$$

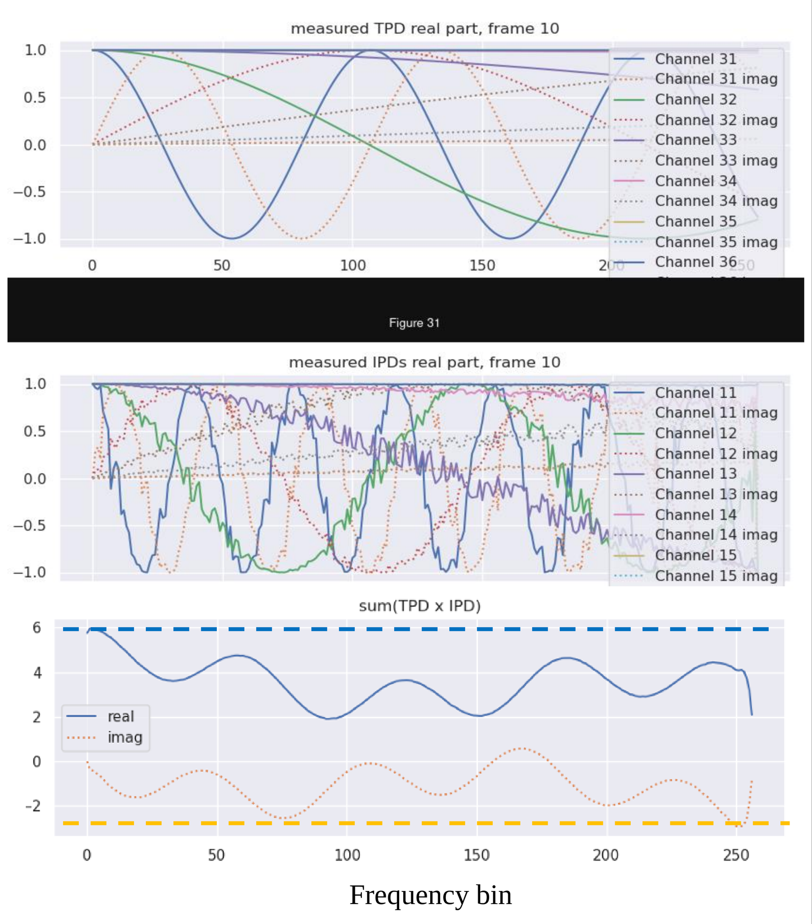
Location Loss

TPD(L_0), IPD(Y)



When the sound source is located where we expect it to be (L_0), the **TPD and the IPD match**.
The real part is equal to the number of microphones (P = 6 in this example).
The imaginary part is equal to zero.

TPD(L_1), IPD(Y)



When we expect the sound source to be somewhere else (L_1), the **TPD and the IPD do not match**.

Dataset

- We simulate challenging reverberant conditions
- 2 sources, 11 mics linear array, harmonically spaced
- 2 machines from the DCASE2021 Task 2 Dataset as sources
- Simulated using PyRoomAcoustics:
  - Shoebox rooms, with randomized multiband materials
  - Image source for early reflections
  - Ray tracing for late part
- In total:
  - 24,000 mixtures of 10 seconds
  - Split into *15,000 / 6,000 / 3,000*

Example of locations for mic array and sources

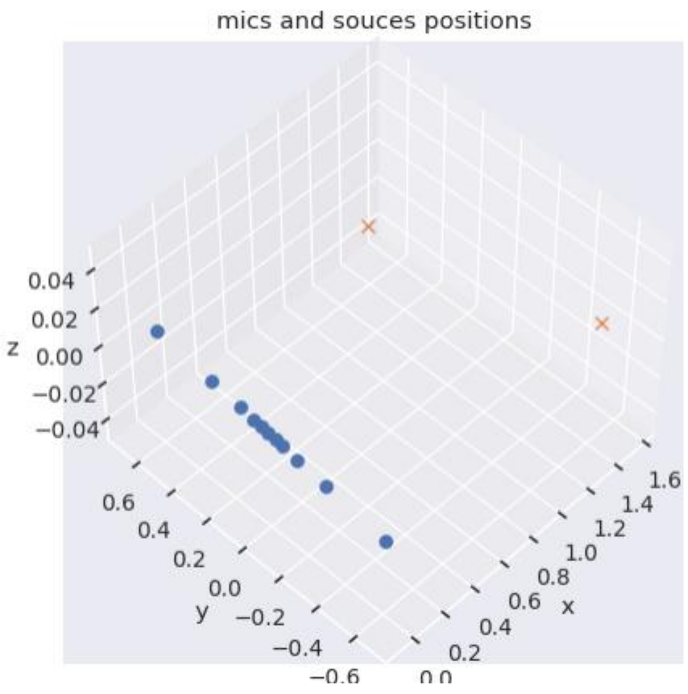


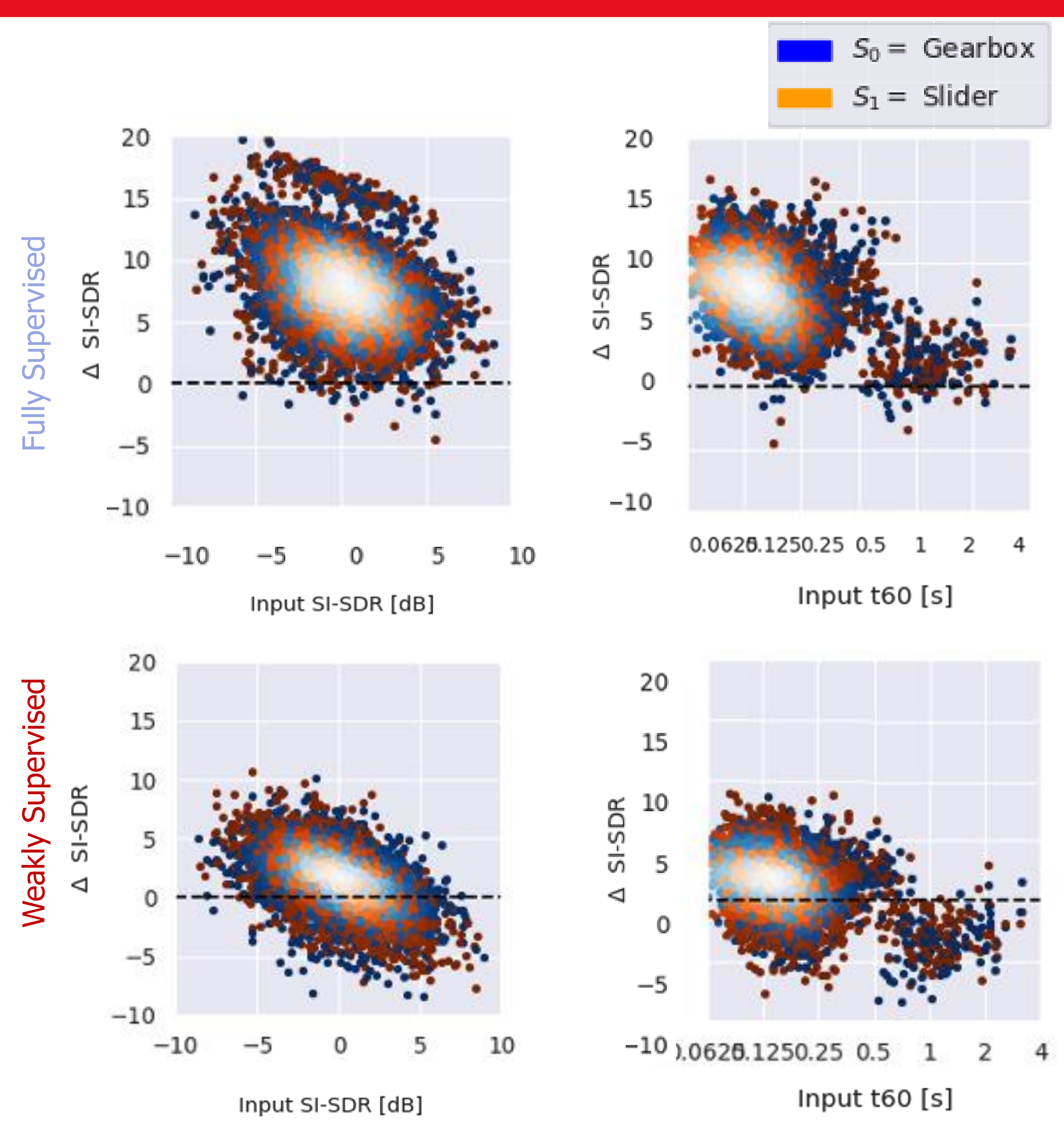
Table 1: Simulation constraints for array and source placement.

Parameter	Range
Distance between sources	$[0.5, 1.5]$
Distance between sources and mic array center	$[0.75, 2.0]$
Distance between sources or mic, and room surface	$[0.5, \infty]$
Angle between mic array normal and sources	$[0°, 30°]$

Results

Performance in terms of mean ± standard deviation of SI-SDR (dB) for different source separation approaches evaluated on datasets with 2 different sets of machines, and 2 different acoustical conditions. SetA = [s_0 = gearbox, s_1 = slider]; SetB = [s_0 = pump, s_1 = valve].

	Trained on		Anechoic				Reverberant			
			SetA		SetB		SetA		SetB	
Approach	Set	Reverb	SI-SDR$_0$ ↑	SI-SDR$_1$ ↑	SI-SDR$_0$ ↑	SI-SDR$_1$ ↑	SI-SDR$_0$ ↑	SI-SDR$_1$ ↑	SI-SDR$_0$ ↑	SI-SDR$_1$ ↑
Mixture	n/a	n/a	−0.1±4.4	0.1±4.4	0.1±2.5	−0.1±2.5	0.0±2.6	0.0±2.6	0.1±2.6	−0.1±2.6
Delaysum	n/a	n/a	−3.0±7.1	−2.9±6.7	−2.4±5.8	−2.4±5.7	−3.2±5.1	−3.5±5.2	−3.1±5.1	−3.4±5.3
Ideal Binary Masks	n/a	n/a	8.7±5.4	8.8±5.3	8.8±3.4	8.2±3.3	9.0±3.3	8.9±3.2	9.1±3.3	8.7±3.2
Fully Supervised	A	✓	15.2±2.5	15.6±2.6	14.3±2.3	14.4±2.0	7.7±3.3	7.7±3.3	7.4±3.3	7.3±3.2
Fully Supervised	A	✗	21.4±2.8	23.6±3.2	18.9±3.7	21.3±3.3	3.8±5.0	4.2±5.0	3.6±4.8	4.0±4.7
WeakSup	A	✓	4.8±3.4	4.1±2.5	5.7±2.3	4.9±2.0	1.6±2.7	1.2±2.3	1.8±2.8	1.4±2.5
WeakSup	A	✗	7.0±3.4	7.1±3.3	7.7±2.2	7.5±2.3	3.2±2.8	3.2±2.6	3.2±2.9	3.2±2.6
Fully Supervised	B	✓	11.9±2.4	12.3±2.5	11.5±2.0	11.7±1.7	6.5±3.0	6.5±3.1	6.4±3.0	6.3±2.9
Fully Supervised	B	✗	19.0±2.5	19.4±2.7	18.3±2.6	18.8±2.0	4.2±4.4	4.1±4.4	4.1±4.3	4.0±4.3
WeakSup	B	✓	4.0±3.1	3.9±2.8	4.7±2.0	4.4±1.8	1.7±2.4	1.3±2.3	1.7±2.4	1.1±2.2
WeakSup	B	✗	3.9±3.9	3.7±2.6	5.4±2.4	4.8±2.1	1.9±2.7	1.5±2.1	2.3±2.7	1.6±2.2



Main findings:

- 1) Delay-and-sum (Baseline) is quite bad → Challenging scenario
- 2) Weakly supervised (WS) is not as good as Fully Supervised (FS), but there is some separation:
  - FS is best when trained with reverb → Avoids domain shift
  - WS is best when trained with anechoic → Avoids noisy loss function
- 3) Signal content (training set) has little impact
- 4) Performance drops under high reverberation
  - IPDs are noisy and not reliable

Future Work

- Future work includes investigating more complex sound propagation models that are applicable to recorded data
- We will also explore few-shot learning applications, for example, where there is some data available about the acoustics of the environment, such as room impulse responses.

References

[1] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, et al., "Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous detection for machine condition monitoring under domain shifted conditions," in Proc. DCASE, 2021.
[2] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, et al., "Multi-modal multi-channel target speech separation," IEEE J. Sel. Top. Signal Process.,vol. 14, no. 3, pp. 530-541, 2020.
[3] K. Saijo and R. Scheibler, "Spatial loss for unsupervised multi-channel source separation," in Proc. Interspeech, 2022.