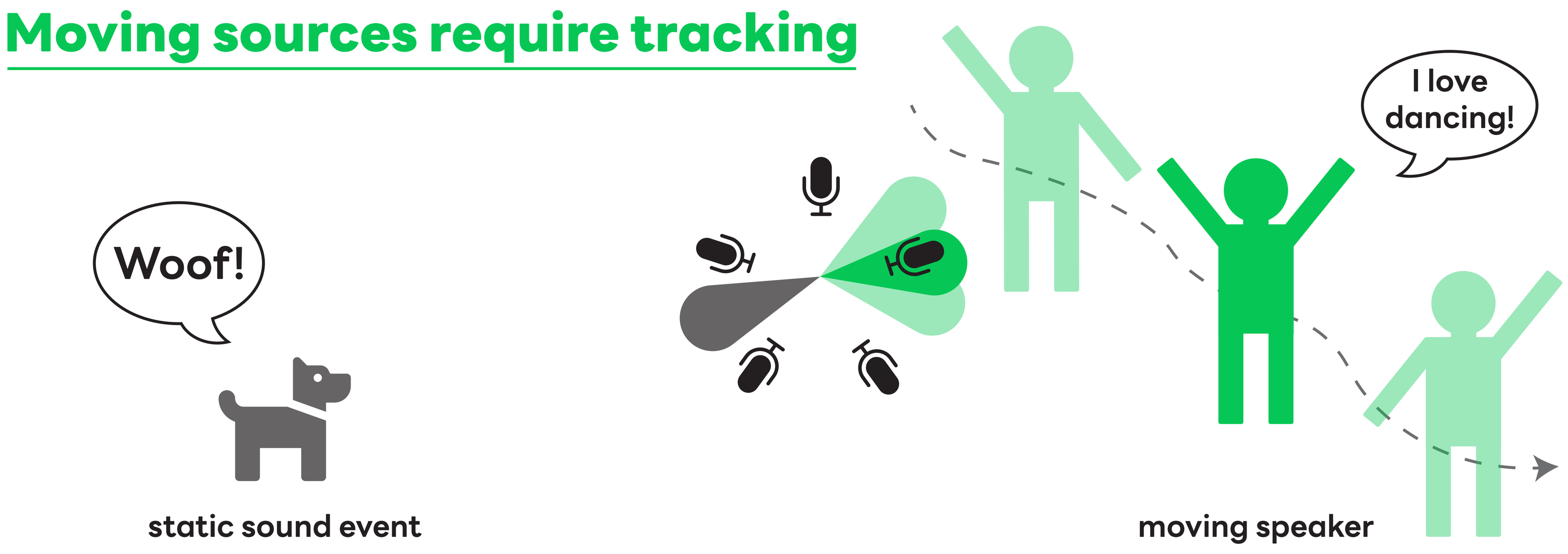
MULTI-CHANNEL SEPARATION OF DYNAMIC SPEECH AND SOUND EVENTS

Takuya Fujimura[1,2] and Robin Scheibler[1] ([1]LINE Corp., [2]Nagoya University)

LINE

Separation of Dynamic Sources

Moving sources require tracking



Contributions of this Work

1. Multi-channel source separation with attention-based tracking
2. Investigate MVDR and Independent Vector Analysis (IVA)
3. Evaluation for speech and sound event detection

Time-Invariant Multi-channel Separation

We investigate two methods

MVDR

$$\mathbf{w} = \frac{(\boldsymbol{\Phi}^N)^{-1}\boldsymbol{\Phi}^S\mathbf{e}_1}{\mathrm{tr}\left((\boldsymbol{\Phi}^N)^{-1}\boldsymbol{\Phi}^S\right)}$$

where S = signal and N = noise.

Independent Vector Analysis (IVA) [1]

$$\mathbf{W}^{(n+1)} \leftarrow \arg\min_{\mathbf{W}} \sum_k \mathbf{w}_k^H \boldsymbol{\Phi}^{(k,n)} \mathbf{w}_k - 2\log|\det \mathbf{W}|$$

where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]^H$ is the **separation matrix**, and n is the iteration.

The matrix $\boldsymbol{\Phi}^\nu$ is the **time-invariant** spatial covariance matrix,

$$\boldsymbol{\Phi}^\nu = \frac{1}{T}\sum_t \boldsymbol{\Psi}_t^\nu, \qquad \text{with} \qquad \boldsymbol{\Psi}_t^\nu = \gamma_t^\nu \mathbf{x}_t \mathbf{x}_t^H,$$

where γ_t is a **time-frequency mask** produced by a DNN [2, 3].

From Time-invariant to Moving Sources

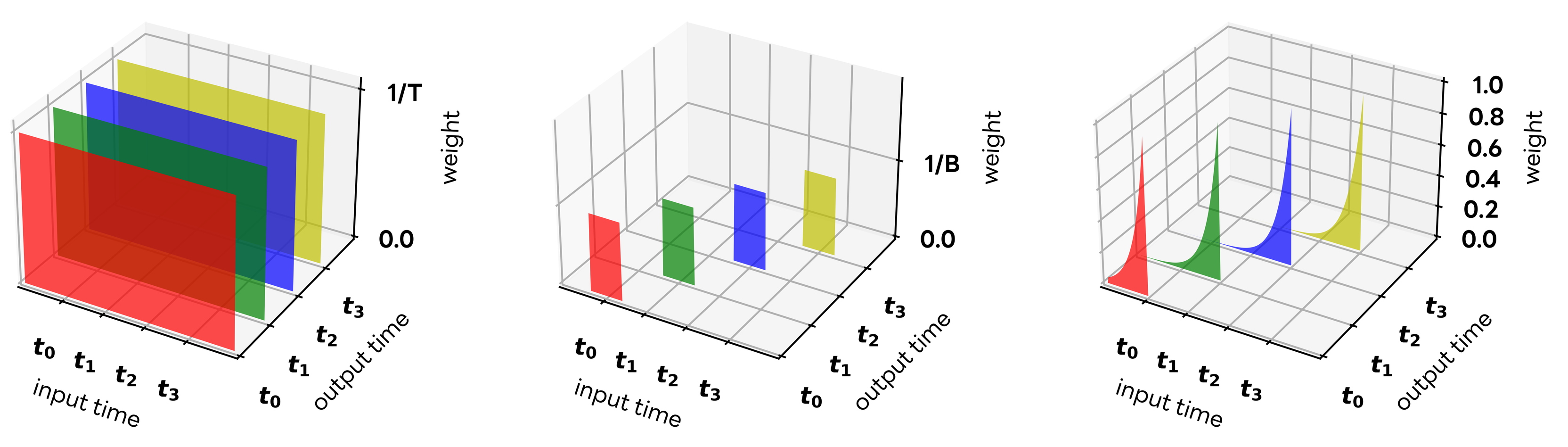
1. Use a **time-varying** spatial covariance matrix

$$\boldsymbol{\Phi}_t^\nu = \frac{1}{T}\sum_{t'} c_{tt'}^\nu \boldsymbol{\Psi}_{t'}^\nu.$$

2. Use them to compute **time-varying** beamforming weights.

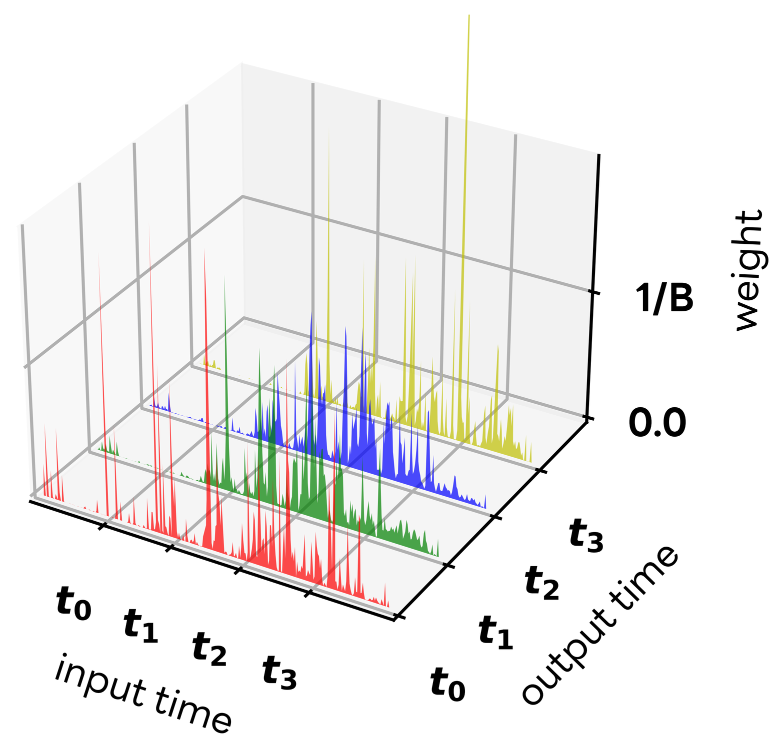
The weights $c_{tt'}$ map input frames to output beamforming weights.

time-invariant (TIV) | block (BLK) | online (ONL)



In this work, we adopt the DNN-predicted attention-based tracking proposed for single source in [4].

attention-based (ATT) [4]



Experiments

Network

Mask 3-layers convnet + GLU activations (similar to [3]).
Attention Mel-spec + spatial features input [5],

$$\begin{aligned} \mathbf{z}_0 &= \mathrm{Mel}(\mathrm{Concat}(|\mathbf{x}|^2, \mathrm{SpaFeat}(\mathbf{x}))) \\ \mathbf{z}_1 &= \mathrm{Conv}(10\log_{10}(\mathbf{z}_0)), \\ c &= \mathrm{SelfAtt}(\mathbf{z}_1). \end{aligned}$$

Speech Separation

Simulated dataset with 0, 1, or 2 moving sources out of 2.
Speech WSJ0+WSJ1 **Noise** CHiME3 **ASR** Whisper

Results

Sources	2 static		moving/static		2 moving	
	SDR↑	WER↓	SDR↑	WER↓	SDR↑	WER↓
Target	—	10.7	—	10.7	—	11.0
Mixture	-0.15	72.2	-0.15	72.3	-0.14	74.0
ATT-MVDR	9.61	13.9	6.65	20.8	4.83	34.3
oracle mask*	11.56	11.6	8.80	13.9	7.18	**19.5**
TIV-IVA	10.65	11.6	4.06	20.8	-0.02	54.9
ONL-IVA	5.14	19.1	1.97	36.1	-0.20	60.1
BLK-IVA	8.84	14.2	4.49	20.5	1.86	44.0
ATT-IVA	**13.54**	**11.3**	**10.78**	**13.1**	**7.65**	**27.7**

Sound Event Detection

Separation network trained on simulated SELD dataset.

SDR (↑) for sound event separation on the synthetic validation set.

Method	Mixture	TIV-IVA	ATT-IVA
SDR ↑	-6.04	-4.13	**0.71**

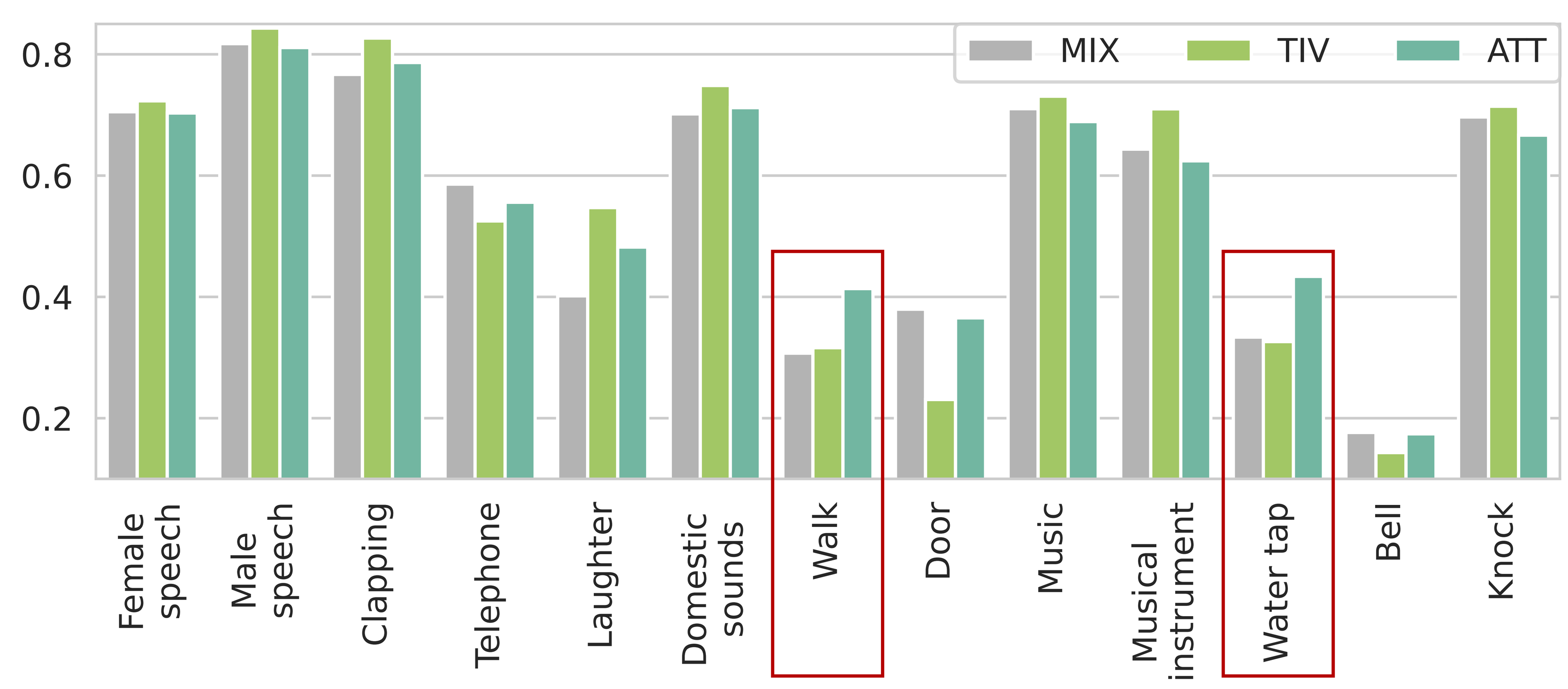
Event detection trained on DCASE 2022 SELD dataset [6].
SED networks:

- MIX: mixture only
- TIV: mixture + time-invariant separation
- ATT: mixture + attention-based separation
- TIV+ATT: ensemble of TIV and ATT

Results macro-F1 score on the STARSS22 dataset [6].

MIX	TIV	ATT	TIV+ATT	Classwise*
0.5559	0.5681	0.5706	**0.5793**	0.6026

Class-wise results



References

[1] Ono, WASPAA, Nov. 2011.
[2] Heymann et al., ICASSP, Mar. 2016.
[3] Scheibler and Togami., ICASSP, Aug. 2021.
[4] Ochiai et al., TASLP, vol. 31, Jan. 2023.
[5] Jarett et al., EUSIPCO, Aug. 2010.
[6] Politis et al., DCASE Workshop, Nov. 2022.

audio samples