Clarity Workshop

Machine Listening in Dynamic Environments

Dr Christine Evers
University of Southampton, School of Electronics & Computer Science

17 September 2021

Data Challenge

Sound Source LOCAlisation & TrAcking (LOCATA)

University of Southampton

	Static Loudspeakers		Moving Human Talkers	
Array	**Single**	**Multiple**	**Single**	**Multiple**
Fixed	Task 1	Task 2	Task 3	Task 4
Moving	-	-	Task 5	Task 6

Development dataset

- Multichannel recordings + close-talking mouth signals
- Ground-truth positions & orientations for all sources and microphones
- Voice activity labels for each source

Evaluation dataset

- Multichannel recordings for all tasks and arrays
- Ground-truth data for array positions and orientations

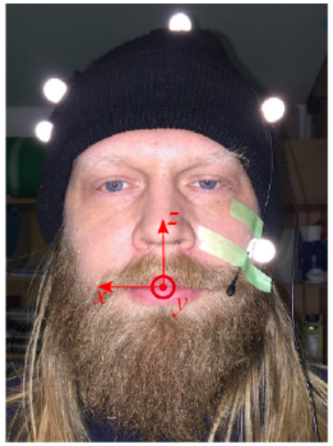
C. Evers *et al.*, "The LOCATA Challenge: Acoustic Source Localization and Tracking," in *IEEE/ACM Trans. Audio, Speech, and Lang. Proc.*, Apr. 2020
C. Evers *et al.*, "Data Corpus for the IEEE-AASP Challenge on Acoustic Source Localization and Tracking (LOCATA)," Data set. Zenodo. http://doi.org/10.5281/zenodo.3630471

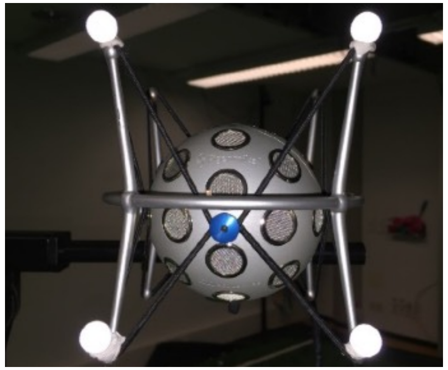
LOCATA Data Challenge

Sound Sources:
- Tasks 1 & 2: Static loudspeakers
- Tasks 2 – 6: Moving human talkers

Microphone Arrays:
- Pseudospherical robot head (12 mics)
- Spherical Eigenmike (32 mics)
- Siemens Signia hearing aids in head-torso simulator (4 mics)
- Planar DICIT array (15 mics)

Close-talking mouth signals:
- DPA d:screet SC4060





C. Evers *et al.*, "The LOCATA Challenge: Acoustic Source Localization and Tracking," in *IEEE/ACM Trans. Audio, Speech, and Lang. Proc.*, Apr. 2020
C. Evers *et al.*, "Data Corpus for the IEEE-AASP Challenge on Acoustic Source Localization and Tracking (LOCATA)," Data set. Zenodo. http://doi.org/10.5281/zenodo.3630471

LOCATA Data Challenge

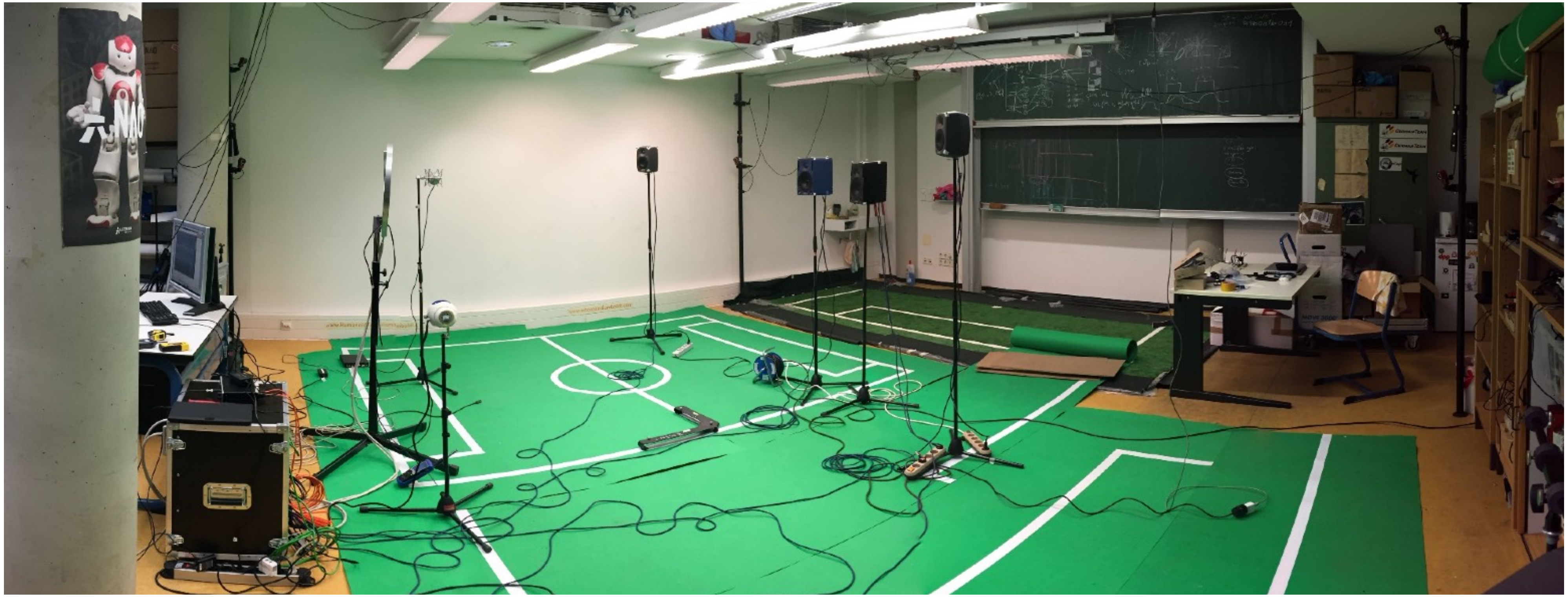
UNIVERSITY OF Southampton

OptiTrack system (10 synchronized IR cameras)

- ➤ Position & orientation of all sources & arrays across time and space

Synchronization:

- ➤ Audio data + tracking data: 120 Hz



C. Evers *et al.*, "The LOCATA Challenge: Acoustic Source Localization and Tracking," in *IEEE/ACM Trans. Audio, Speech, and Lang. Proc.*, Apr. 2020

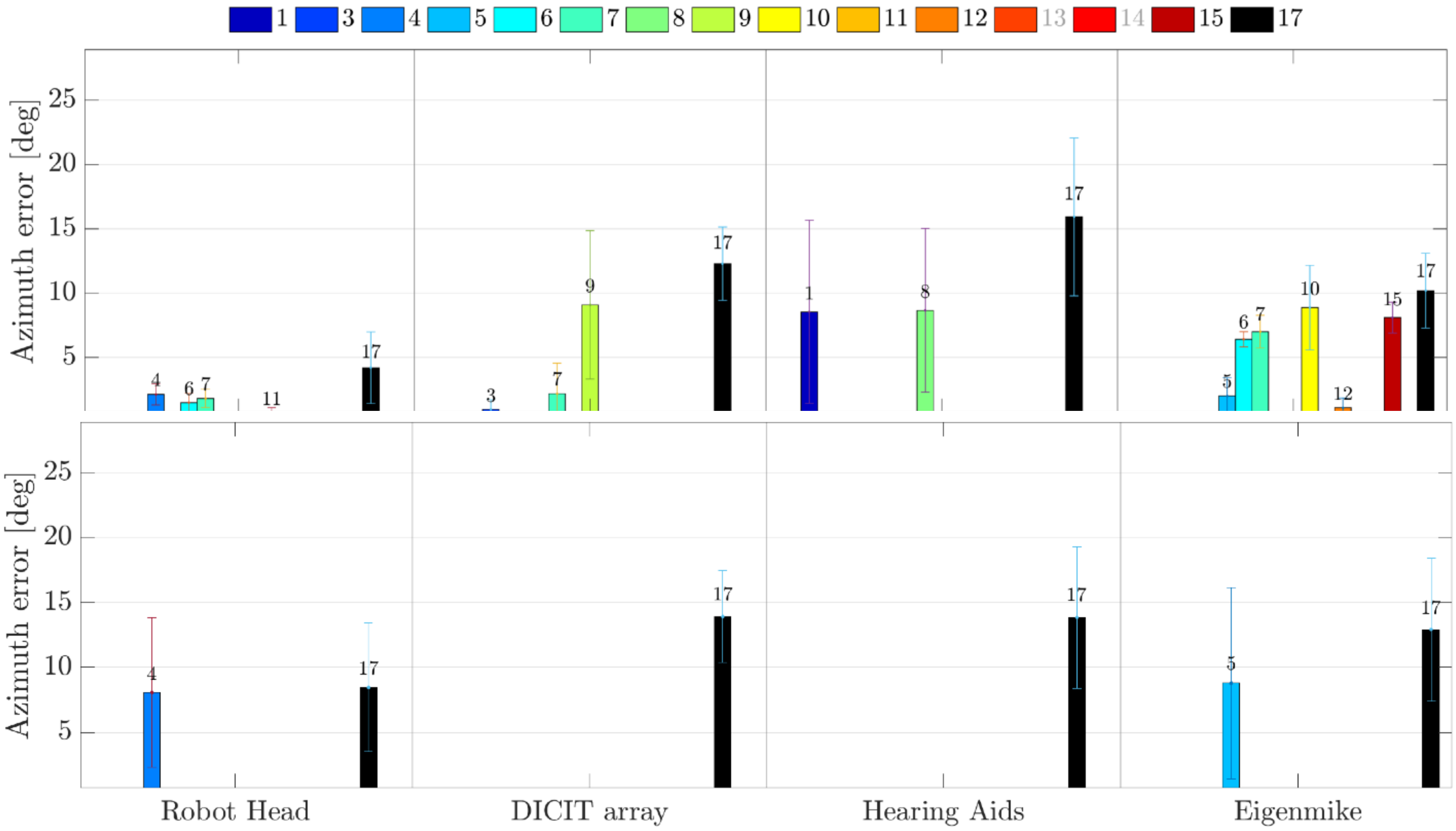
C. Evers *et al.*, "Data Corpus for the IEEE-AASP Challenge on Acoustic Source Localization and Tracking (LOCATA)," Data set. Zenodo. http://doi.org/10.5281/zenodo.3630471

LOCATA Evaluation

UNIVERSITY OF Southampton

Task 1
Static Source
Static Mic Arrays

Task 6
Moving Sources
Moving Mic Arrays



Challenges

01

Source motion

02

Competing sources

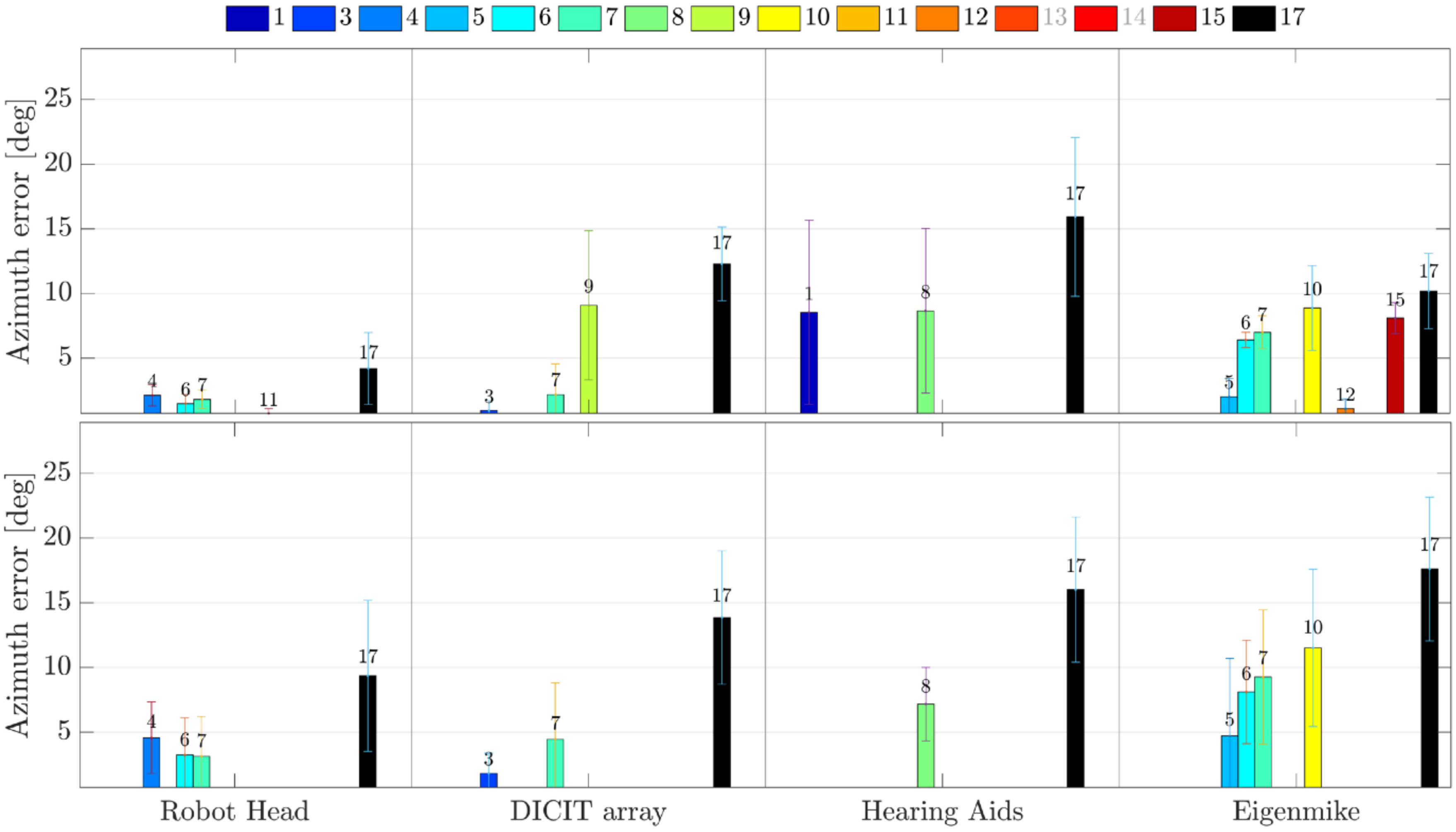
03

Ego-motion

LOCATA - Moving Source

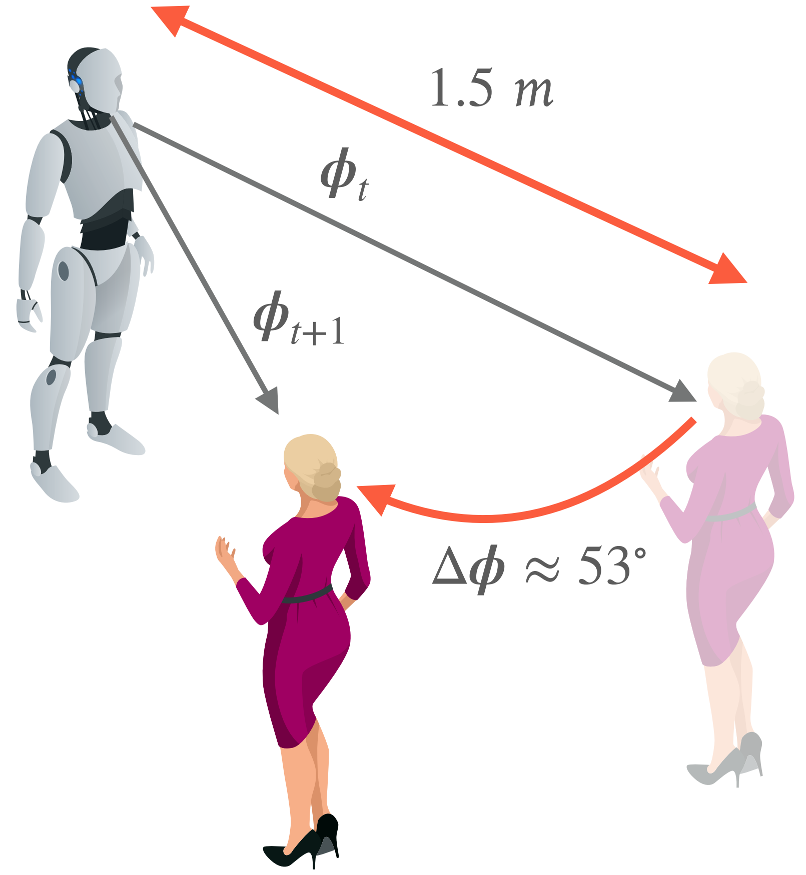
Task 1
Static Source

Task 3
Moving Source



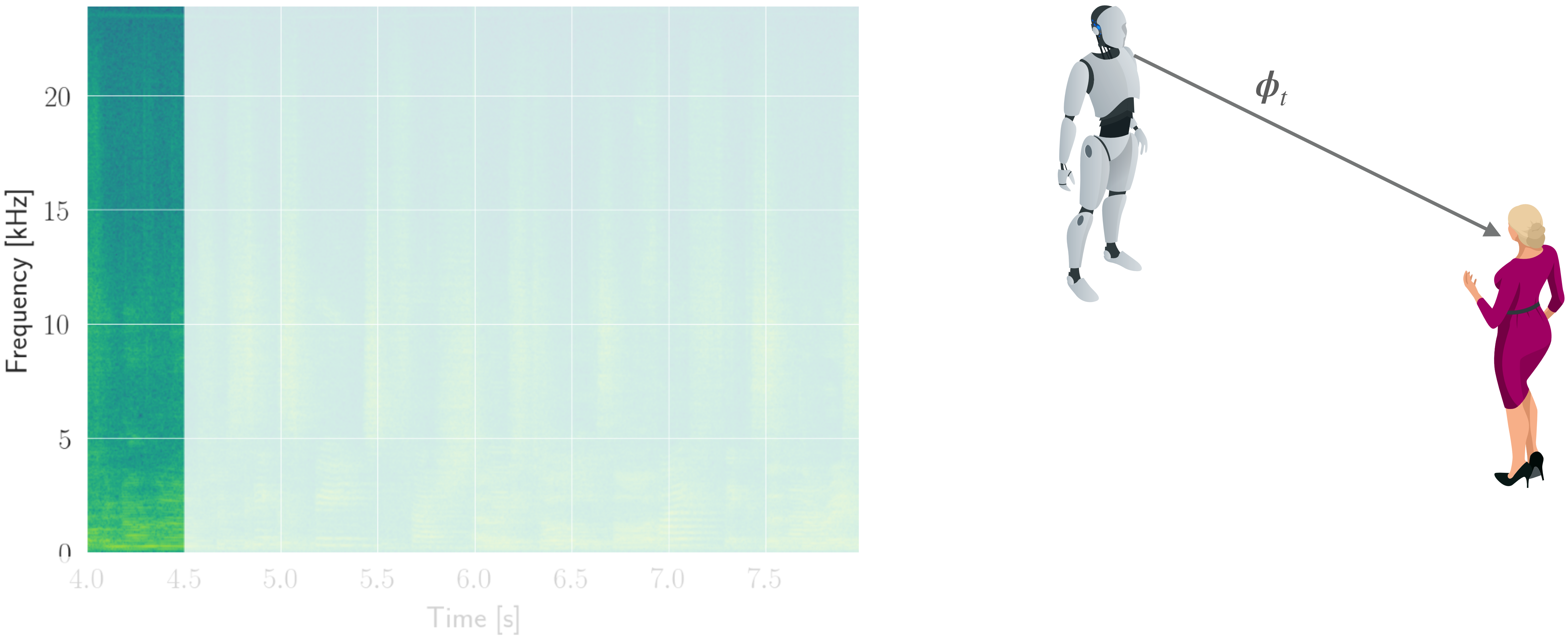
C. Evers *et al.*, "The LOCATA Challenge: Acoustic Source Localization and Tracking," in *IEEE/ACM Trans. Audio, Speech, and Lang. Proc.*, Apr. 2020

Spatio-Temporal Variations



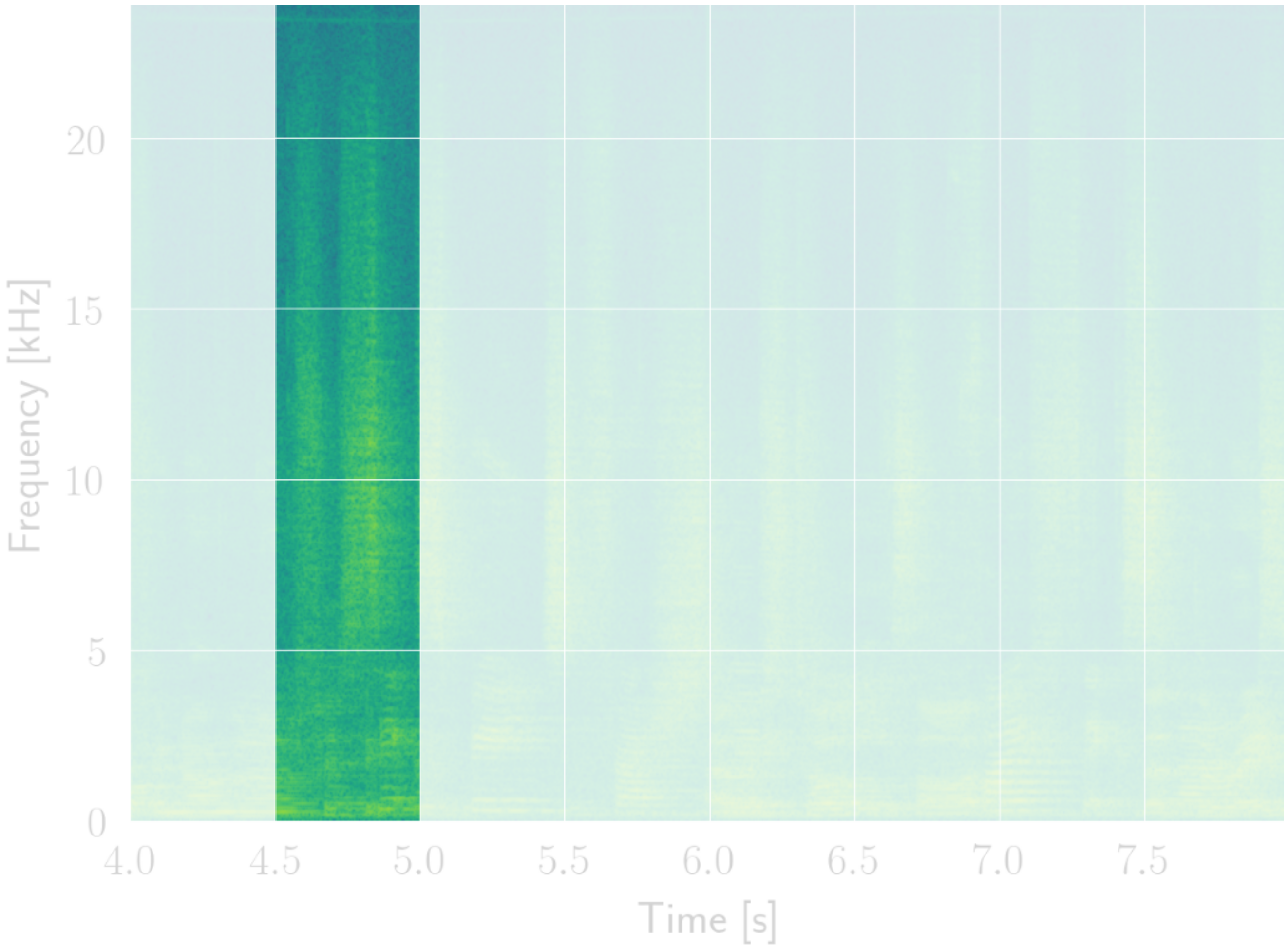
Avg human walking speed: 1.4 m/s

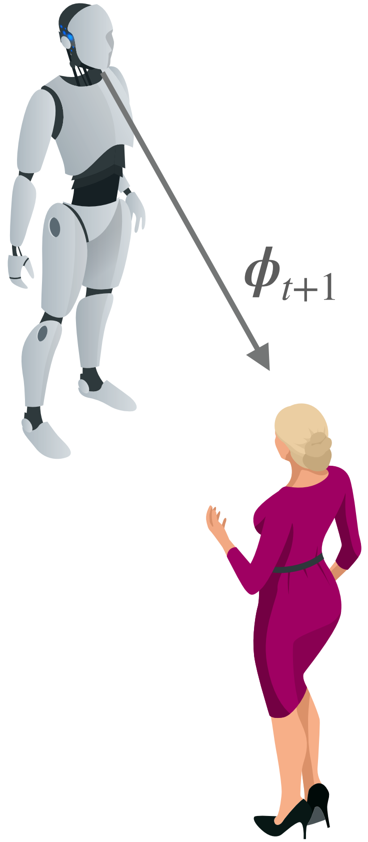
Spatio-Temporal Variations



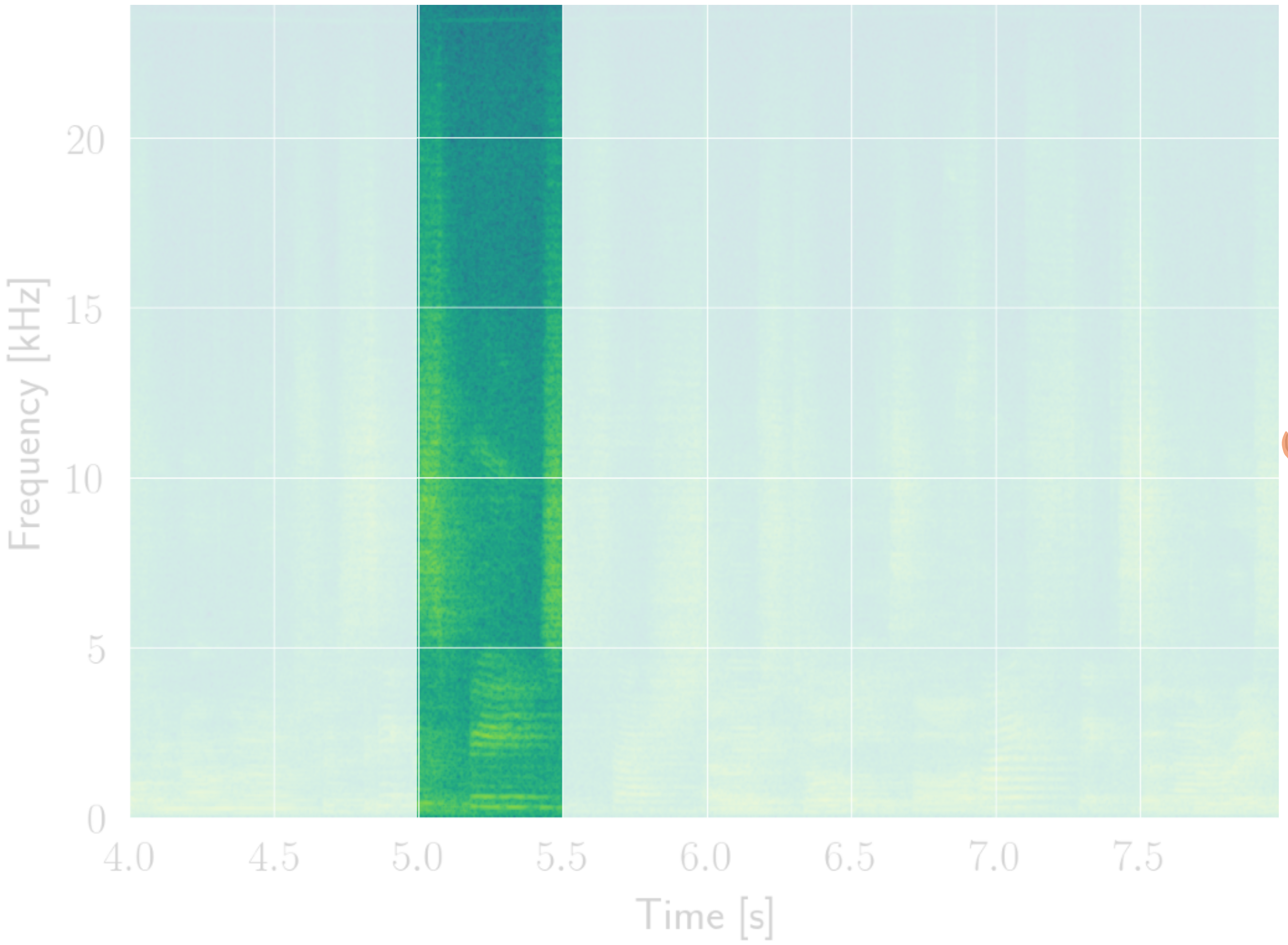
Spatio-Temporal Variations

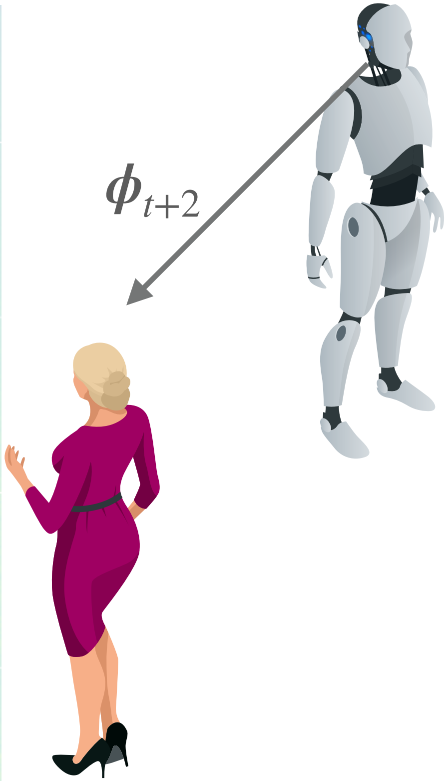
Southampton



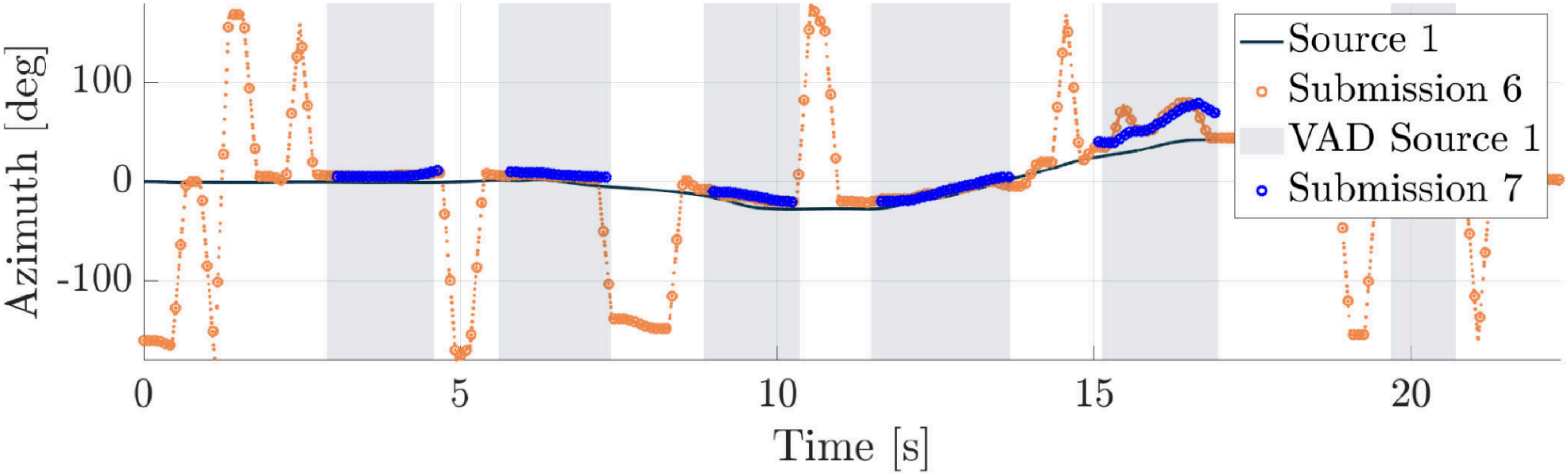


Spatio-Temporal Variations



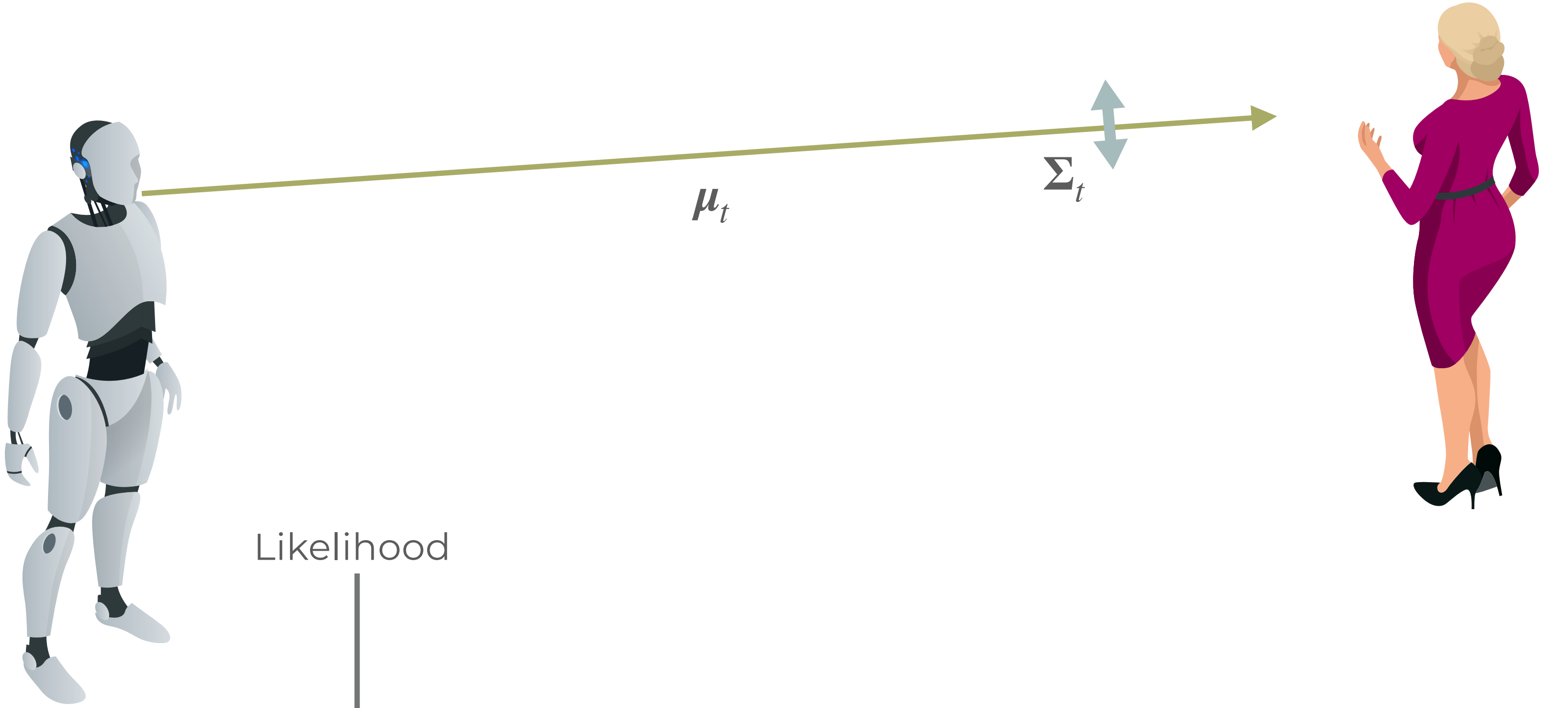


LOCATA: Impact of source inactivity



C. Evers *et al.*, "The LOCATA Challenge: Acoustic Source Localization and Tracking," in *IEEE/ACM Trans. Audio, Speech, and Lang. Proc.*, Apr. 2020

Spatio-temporal correlations



Likelihood

$$p\left(\boldsymbol{\phi}_t \,|\, \boldsymbol{x}_{1:t-1}\,;\, \mathbf{r}_t\right) = \int_{\Phi} p\left(\boldsymbol{\phi}_t \,|\, \boldsymbol{\phi}_{t-1};\, \mathbf{r}_{t-1:t}\right)\, p\left(\boldsymbol{\phi}_{t-1} \,|\, \boldsymbol{x}_{1:t-1}\,;\, \mathbf{r}_{t-1}\right)\, d\boldsymbol{\phi}_{t-1}$$

Current posterior pdf → $\boxed{p\left(\boldsymbol{\phi}_t \,|\, \boldsymbol{x}_{1:t}\,;\, \mathbf{r}_t\right)} \propto \boxed{p\left(\boldsymbol{x}_t \,|\, \boldsymbol{\phi}_t\,;\, \mathbf{r}_t\right)}\,\boxed{p\left(\boldsymbol{\phi}_t \,|\, \boldsymbol{x}_{1:t-1}\,;\, \mathbf{r}_t\right)}$ ← Predictive pdf

t : Time frame index; $\boldsymbol{\phi}_t$: Source state; $\mathbf{r}_t$: Robot state; $\boldsymbol{x}_t$: Data at time frame t

C. Evers, E. Habets, S. Gannot, P. Naylor, "DoA Reliability for Distributed Acoustic Tracking", *in IEEE Signal Proc. Letters*, 2018.

Challenges

Southampton UNIVERSITY OF

01

Source motion

02

Competing sources

03

Ego-motion

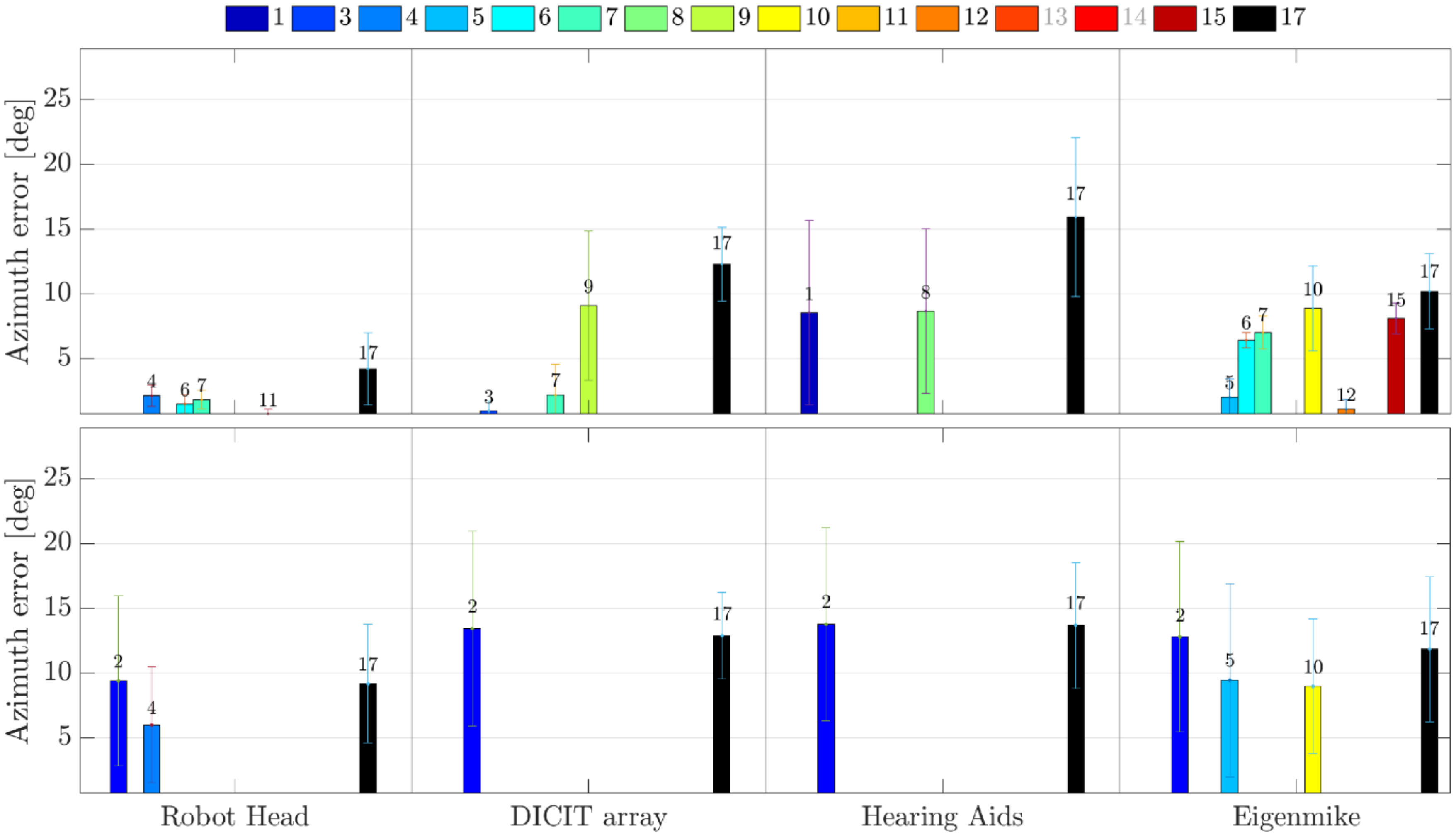

Microphone 1
1.0
0.5
0.0
−0.5
−1.0
Norm. Amplitude

LOCATA - Competing sources

UNIVERSITY OF Southampton

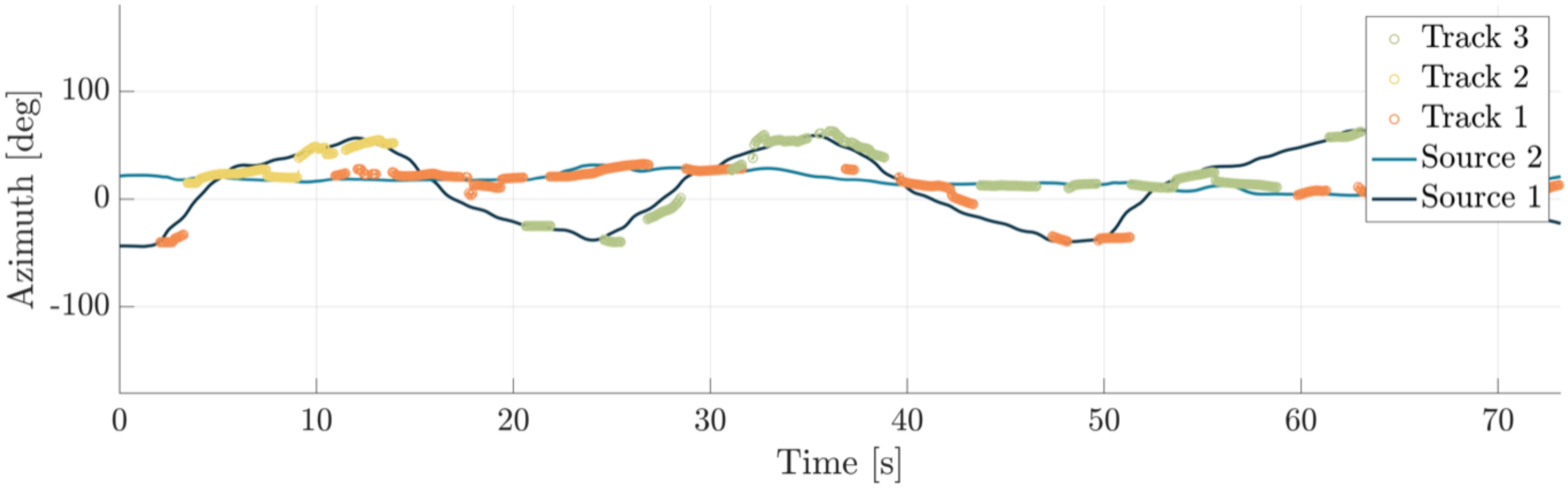
Task 1
Static Source

Task 4
Multiple moving sources



C. Evers *et al.*, "The LOCATA Challenge: Acoustic Source Localization and Tracking," in *IEEE/ACM Trans. Audio, Speech, and Lang. Proc.*, Apr. 2020

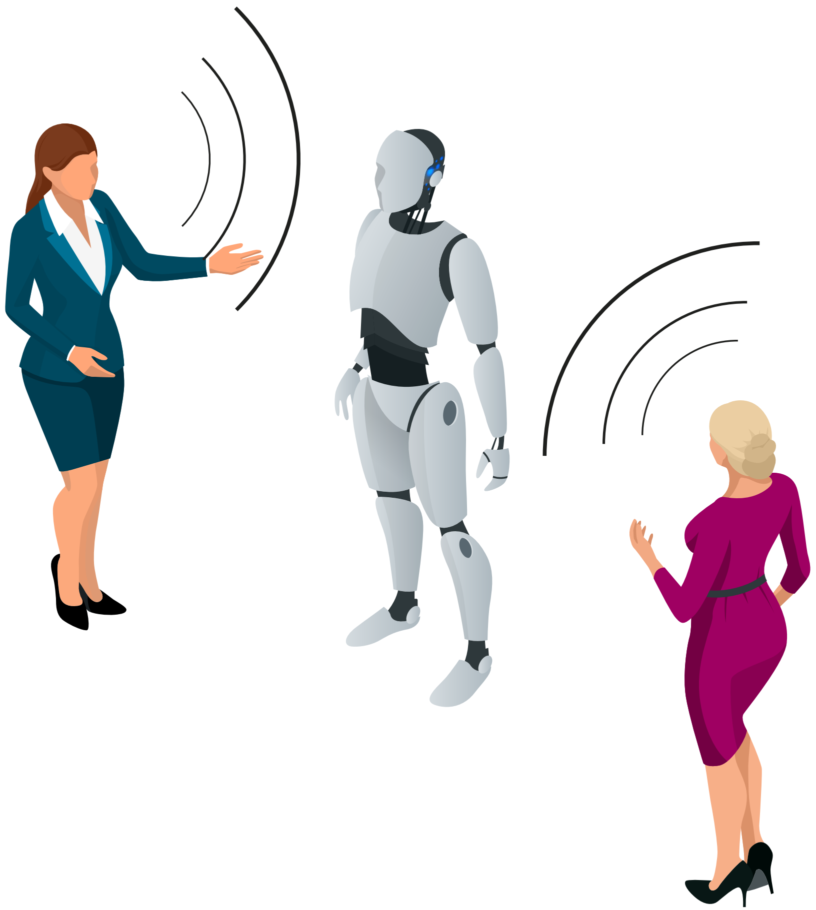
LOCATA: Impact of source ambiguity

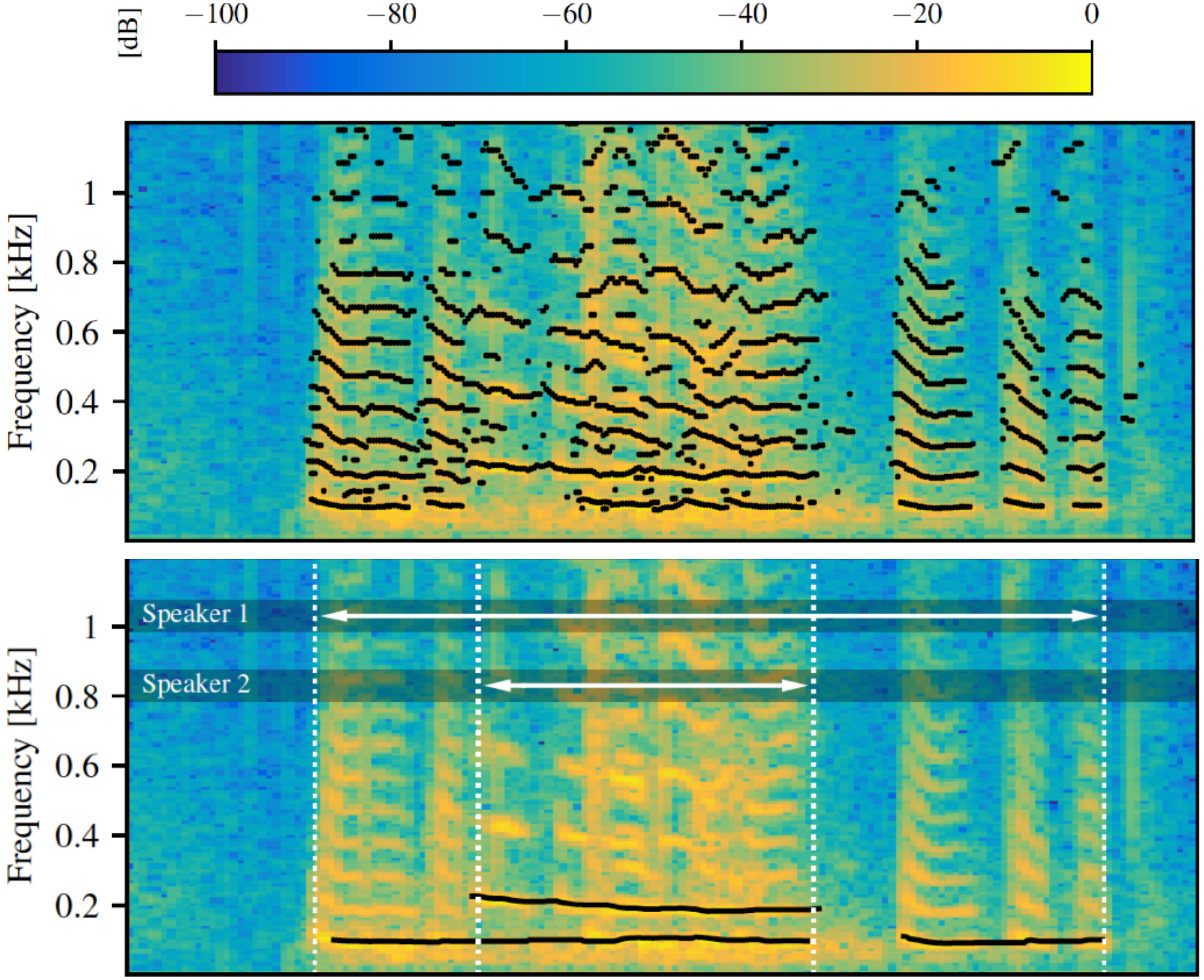


Task 4, Recording 4, Submission ID 4 - Multiple moving sources

C. Evers *et al.*, "The LOCATA Challenge: Acoustic Source Localization and Tracking," in *IEEE/ACM Trans. Audio, Speech, and Lang. Proc.*, Apr. 2020

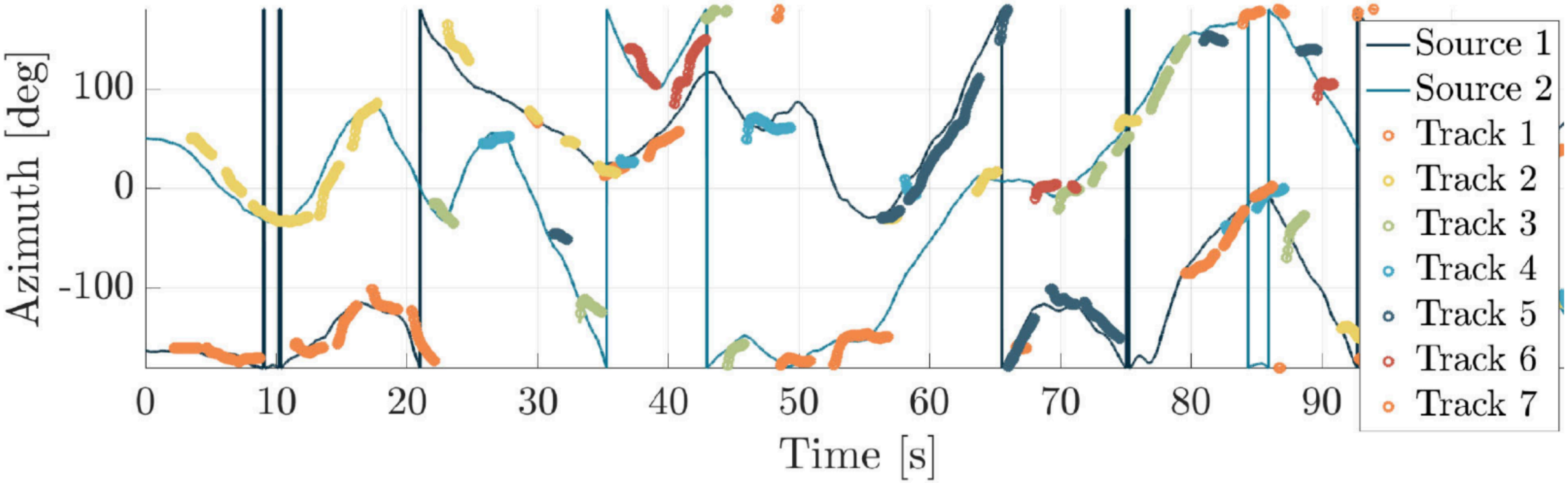
Personally identifying features





A. Hogg *et al.*, "Overlapping speaker segmentation using multiple hypothesis tracking of fundamental frequency", IEEE Trans. Audio, Speech, Lang. Proc., 2021
A. Hogg, C. Evers, P. Naylor, "Multichannel Overlapping Speaker Segmentation Using Multiple Hypothesis Tracking Of Acoustic And Spatial Features," *ICASSP* 2021

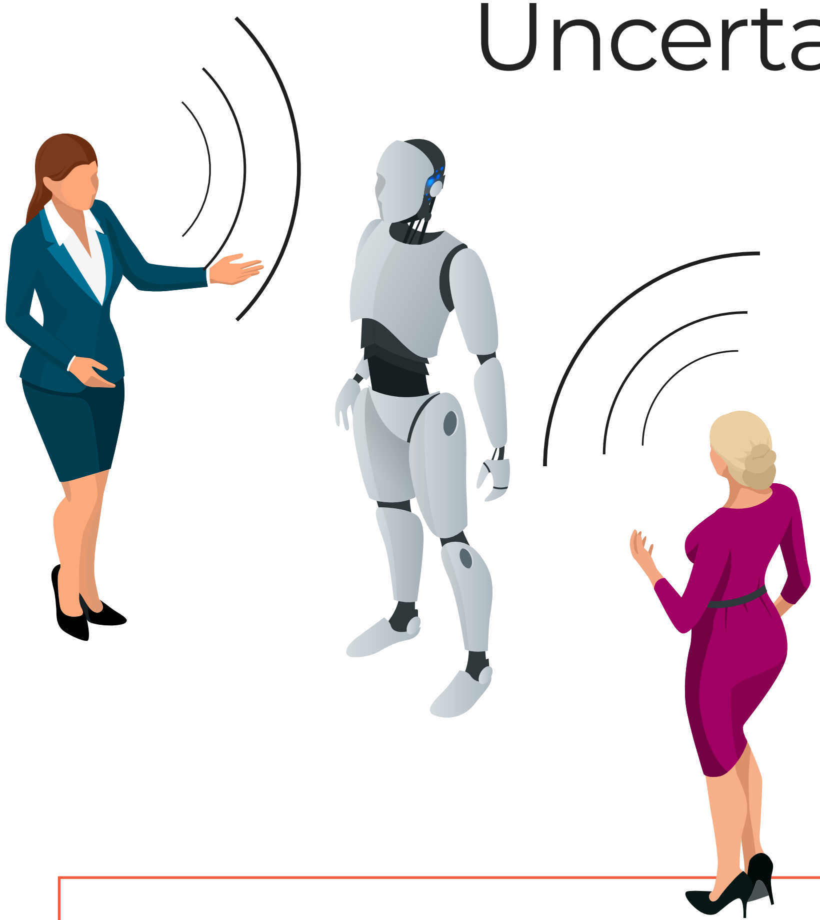
LOCATA: Impact of source cardinality

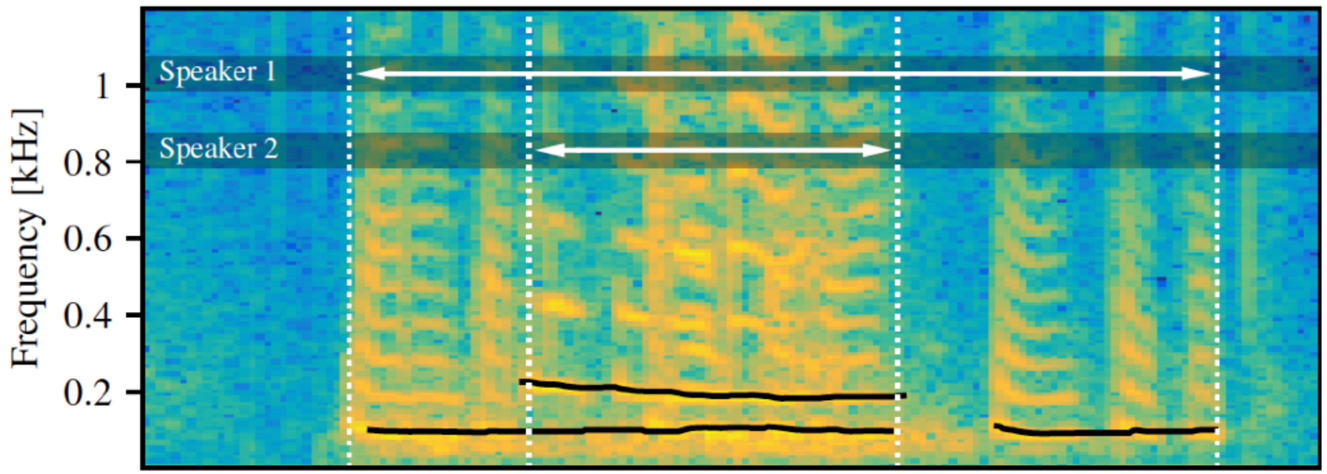


Task 6, Recording 2, Submission ID 4 - Multiple moving sources

C. Evers *et al.*, "The LOCATA Challenge: Acoustic Source Localization and Tracking," in *IEEE/ACM Trans. Audio, Speech, and Lang. Proc.*, Apr. 2020

Uncertainty in source cardinality





Random set of sources: $\Phi_t = \left\{ \phi_t^{(1)}, \ldots, \phi_t^{(N_t)} \right\}$

$$p\left(\Phi_t \,|\, x_{1:t};\, \mathbf{r}_t\right) = p\left(\varnothing \,|\, x_{1:t};\, \mathbf{r}_t\right) + \sum_{\mathbf{p}\in\mathbf{P}} p\left(\left\{ \phi_t^{(p(1))}, \ldots, \phi_t^{(p(|\mathbf{p}|))} \right\} \,|\, x_{1:t};\, \mathbf{r}_t\right)$$

t : Time frame; $\mathbf{r}_t$: Robot state; x_t : Data at frame t;
Φ_t : Set of source states at frame t; $\phi_t^{(n)}$: State of source n; $\mathbf{P}$: Set of partitions of Φ_t

C. Evers & P. Naylor, “Acoustic SLAM,” in IEEE/ACM Trans. Audio, Speech, and Lang. Proc., Sep. 2018.
C. Evers & P. Naylor, "Optimized Self-Localization for SLAM in Dynamic Scenes using Probability Hypothesis Density Filters," in IEEE Trans. Signal Proc., Feb. 2018.

Challenges

01

Source motion

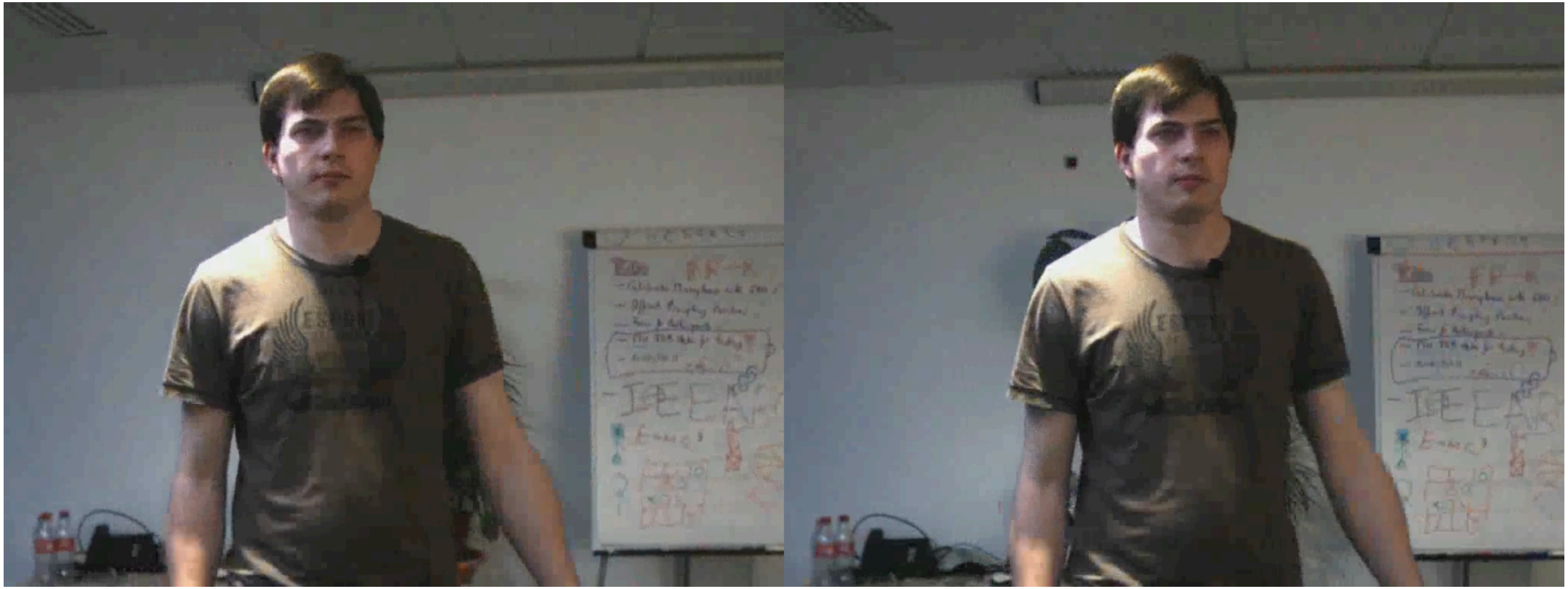
02

Competing sources

03

Ego-motion

Stabilising perception in ego-motion



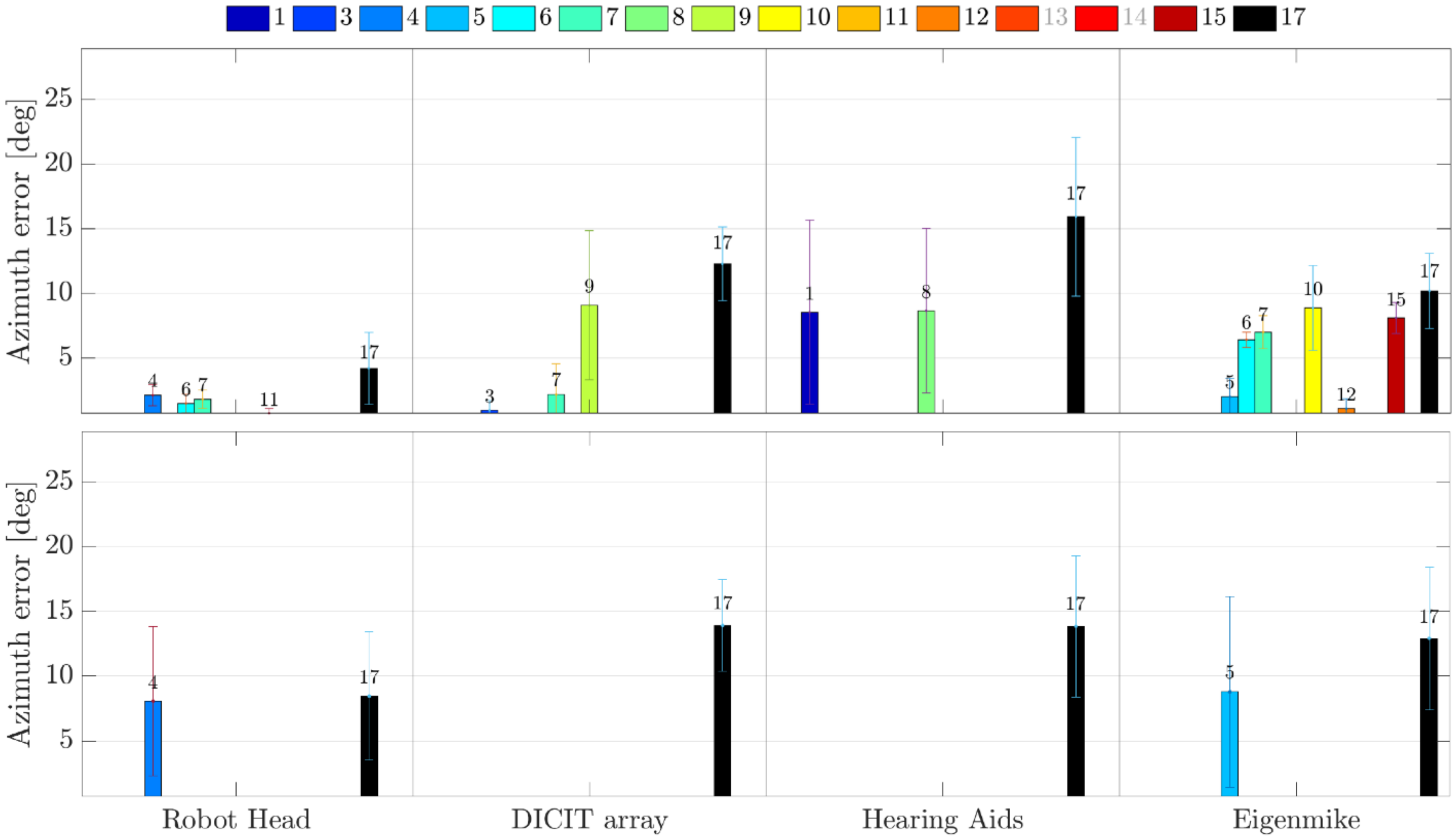
X. Alameda-Pineda *et al.*, “RAVEL: An Annotated Corpus for Training Robots with Audiovisual Abilities,” *Journal of Multimodal User Interfaces*, 2013.

LOCATA - Competing sources

UNIVERSITY OF Southampton

Task 1
Static Source

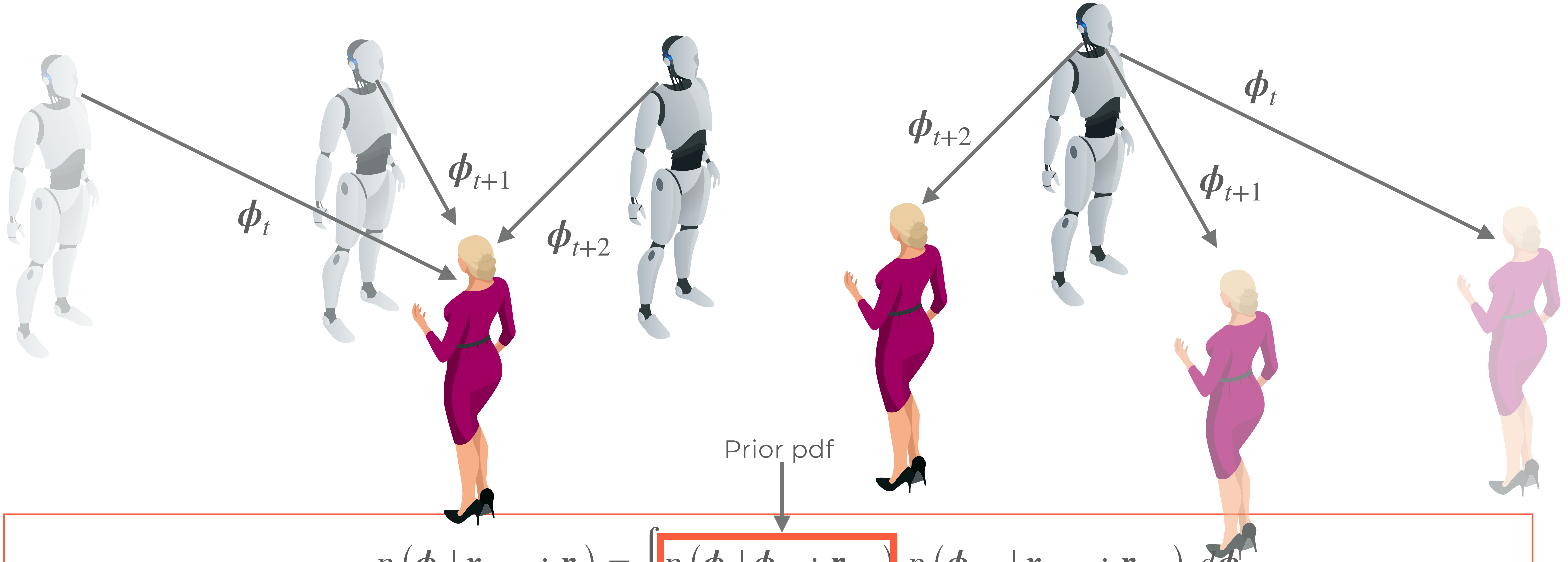
Task 6
Moving microphones



C. Evers *et al.*, "The LOCATA Challenge: Acoustic Source Localization and Tracking," in *IEEE/ACM Trans. Audio, Speech, and Lang. Proc.*, Apr. 2020

Unknown Source-Sensor Distance

UNIVERSITY OF Southampton



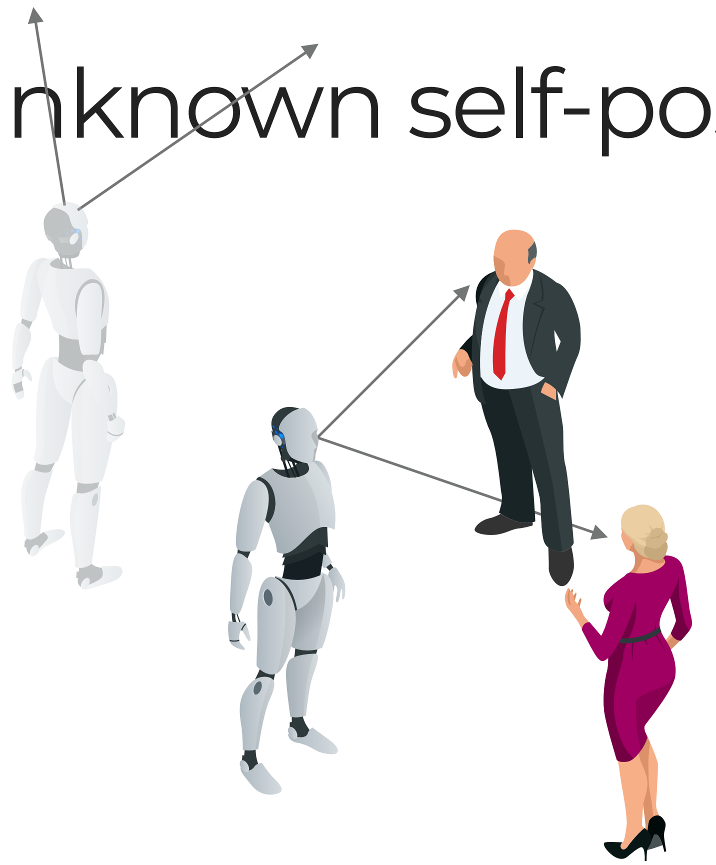
$$p\left(\boldsymbol{\phi}_t \,|\, \boldsymbol{x}_{1:t-1}\,;\, \mathbf{r}_t\right) = \int_{\Phi} \overset{\text{Prior pdf}}{\boxed{p\left(\boldsymbol{\phi}_t \,|\, \boldsymbol{\phi}_{t-1};\, \mathbf{r}_{t-1}\right)}}\; p\left(\boldsymbol{\phi}_{t-1} \,|\, \boldsymbol{x}_{1:t-1}\,;\, \mathbf{r}_{t-1}\right)\, d\boldsymbol{\phi}_{t-1}$$

$$p\left(\boldsymbol{\phi}_t \,|\, \boldsymbol{x}_{1:t}\,;\, \mathbf{r}_t\right) \propto p\left(\boldsymbol{x}_t \,|\, \boldsymbol{\phi}_t\,;\, \mathbf{r}_t\right)\; p\left(\boldsymbol{\phi}_t \,|\, \boldsymbol{x}_{1:t-1}\,;\, \mathbf{r}_t\right)$$

t : Time frame index; $\boldsymbol{\phi}_t$: Source state; $\mathbf{r}_t$: Robot state; $\boldsymbol{x}_t$: Data at time frame t

Y. Ban, X. Alameda-Pineda, C. Evers, R. Horaud, “Tracking Multiple Audio Sources with the von Mises Distribution and Variational EM”, *in IEEE Signal Proc. Letters*, 2019

Unknown self-position



$$p\left(\boldsymbol{S}_t, \mathbf{r}_t \mid \boldsymbol{x}_{1:t}, \mathbf{y}_{1:t}\right) \approx p\left(\mathbf{r}_t \mid \mathbf{y}_{1:t}\right) \prod_{m=1}^{M_t} p\left(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t\right) \lambda\left(\boldsymbol{s}_t \mid \mathbf{r}_t, \boldsymbol{\Omega}_{1:t}\right)$$

t : Time frame; $\mathbf{r}_t$: Robot state; $\boldsymbol{x}_t$: Data at frame t; $\boldsymbol{\omega}_t$: Source detections; $\mathbf{y}_t$: IMU data;
$\boldsymbol{S}_t$: Set of Cartesian source states at frame t; $\boldsymbol{s}_t^{(n)}$: Cartesian position of source n; $\mathbf{P}$: Set of partitions of $\boldsymbol{S}_t$

C. Evers & P. Naylor, "Acoustic SLAM," in IEEE/ACM Trans. Audio, Speech, and Lang. Proc., Sep. 2018.
C. Evers & P. Naylor, "Optimized Self-Localization for SLAM in Dynamic Scenes using Probability Hypothesis Density Filters," in IEEE Trans. Signal Proc., Feb. 2018.

Summary

Challenges impacting on machine listening in dynamic scenarios:

- **Source motion:**
  - Spatio-temporal variations, prohibiting batch-processing of recordings
  - Source inactivity, leading to estimation bias and false estimates
- **Competing sources:**
  - Overlapping speech, resulting in ambiguity in the source identities
  - Uncertainty in the number of sources, resulting in a combinatorial problem
- **Ego-motion:**
  - Unknown source-sensor distance, required for reference frame transformations
  - Unknown self-position, prohibitive for coherent integration of long-term memory


Thank you