Machine Listening in Dynamic Environments

Dr Christine Evers
University of Southampton, School of Electronics & Computer Science
17 September 2021
Data Challenge
Sound Source LOCALisation & TrAking (LOCATA)

<table>
<thead>
<tr>
<th>Static Loudspeakers</th>
<th>Moving Human Talkers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Array</td>
<td>Single</td>
</tr>
<tr>
<td>Fixed</td>
<td>Task 1</td>
</tr>
<tr>
<td>Moving</td>
<td>-</td>
</tr>
</tbody>
</table>

**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

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

LOCATA Data Challenge

OptiTrack system (10 synchronized IR cameras)
- Position & orientation of all sources & arrays across time and space

Synchronization:
- Audio data + tracking data: 120 Hz

LOCATA Evaluation

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


Clarity Workshop, Sep’ 2021
Christine Evers: “Machine Listening in Dynamic Environments”
Spatio-Temporal Variations

Avg human walking speed: 1.4 m/s
Spatio-Temporal Variations

- Frequency [kHz]
- Time [s]

\[ \phi_t \]
Spatio-Temporal Variations

Frequency [kHz]

Time [s]

$\phi_{t+1}$
Spatio-Temporal Variations

\[ \phi_{t+2} \]
LOCATA: Impact of source inactivity

Spatio-temporal correlations

\[ p(\phi_t | x_{1:t-1} ; r_t) = \int p(\phi_t | \phi_{t-1} ; r_{t-1:t}) p(\phi_{t-1} | x_{1:t-1} ; r_{t-1}) d\phi_{t-1} \]

Current posterior pdf \[ p(\phi_t | x_{1:t} ; r_t) \propto p(\phi_t | \phi_{t-1} ; r_{t-1:t}) p(\phi_t | x_{1:t-1} ; r_{t-1}) \]

Predictive pdf

\[ t : \text{Time frame index}; \ \phi_t : \text{Source state}; \ r_t : \text{Robot state}; \ x_t : \text{Data at time frame } t \]


Christine Evers: "Machine Listening in Dynamic Environments"
Challenges

01 Source motion

02 Competing sources

03 Ego-motion
Christine Evers: "Machine Listening in Dynamic Environments"
LOCATA - Competing sources

Task 1
Static Source

Task 4
Multiple moving sources


Clarity Workshop, Sep’ 2021
Christine Evers: “Machine Listening in Dynamic Environments”
LOCATA: Impact of source ambiguity

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

Personally identifying features

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

Uncertainty in source cardinality

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

$$p(\Phi_t | x_{1:t}, r_t) = p(\emptyset | x_{1:t}, r_t) + \sum_{p \in P} p\left(\{\phi_t^{(p(1))}, \ldots, \phi_t^{(p(|p|))}\} | x_{1:t}, r_t\right)$$

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

Challenges

01. Source motion
02. Competing sources
03. Ego-motion
Stabilising perception in ego-motion

LOCATA - Competing sources

**Task 1**
Static Source

**Task 6**
Moving microphones

Unknown Source-Sensor Distance

\[
p(\phi_t | x_{1:t-1}; r_t) = \int p(\phi_t | \phi_{t-1}; r_{t-1}) p(\phi_{t-1} | x_{1:t-1}; r_{t-1}) \, d\phi_{t-1}
\]

\[
p(\phi_t | x_{1:t}; r_t) \propto p(x_t | \phi_t; r_t) p(\phi_t | x_{1:t-1}; r_t)
\]

\(t\) : Time frame index; \(\phi_t\) : Source state; \(r_t\) : Robot state; \(x_t\) : Data at time frame \(t\)


Clarity Workshop, Sep’ 2021

Christine Evers: “Machine Listening in Dynamic Environments”
\[ p(S_t, r_t | x_{1:t}, y_{1:t}) \approx p(r_t | y_{1:t}) \prod_{m=1}^{M_t} p(\omega_{t,m} | r_t) \lambda(s_t | r_t, \Omega_{1:t}) \]

- \( t \): Time frame; \( r_t \): Robot state; \( x_t \): Data at frame \( t \); \( \omega_t \): Source detections; \( y_t \): IMU data;
- \( S_t \): Set of Cartesian source states at frame \( t \); \( s_t^{(n)} \): Cartesian position of source \( n \); \( P \): Set of partitions of \( S_t \)

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