Audio-Visual Fusion for Human-Robot Interaction

Radu Horaud
PERCEPTION team
INRIA Grenoble Rhône-Alpes
Radu.Horaud@inria.fr
Single user that can control the device.
The robot must take decisions!
Audio, vision, and audio-visual challenges
Datasets
Supervised sound-source localization
Mapping sounds onto images
Clustering audio and visual features
Multiple-person visual tracking
Tracking a single speaker
Tracking multiple speakers
Conclusions and future work
Visual Processing
Audio Processing

Background noise

speech

speech

speech

sound?
Audio-visual Scene Analysis
A Single Audio-Visual Object
Several Audio-Visual Objects
Audio-visual Recordings

- Audio: two omnidirectional microphones, 44100 Hz, acoustic dummy head.
- Vision: two 2MP cameras, 25 FPS, $97^\circ \times 80^\circ$ field of view.
- Room: “natural” acoustic and lighting conditions.
Audio-visual Dataset

- People wander around, turn their faces towards the speaker (and not facing the camera!).
- Casual dialogue, speech turns with overlap, background noise.
Problems to be addressed

- Person detection and pose estimation (head, body, posture, etc.)
- Head orientation, gaze (who looks at whom/what?)
- Tracking persons over long periods of time
- Audio-source localization and separation
- Speech activity detection and speaker diarization (who speaks when?)
- Audio-visual association
- Human-robot dialogue
- etc.
Outline

- Binaural audition
- Supervised sound-source localization
- Mapping sounds onto images
- Audio-visual clustering
- Multiple person tracking
- Speaker diarization
Sound Propagation Model

- Two microphones \( (m) \), a single sound source \( (s) \), room impulse response \( (h) \), noise \( (n) \):

\[
m_1(t) = h_1 * s(t) + n_1(t) \in \mathbb{R}
\]

\[
m_2(t) = h_2 * s(t) + n_2(t) \in \mathbb{R}
\]

- Representation in the spectral domain with the short-time Fourier transform (STFT):

\[
M_{1,f,l} = H_{1,f} S_{f,l} + N_{1,f} \in \mathbb{C}
\]

\[
M_{2,f,l} = H_{2,f} S_{f,l} + N_{2,f} \in \mathbb{C}
\]

- \( f \) (frequency bin) and \( l \) (time) are the indexes of a spectrogram point.
The Short-time Fourier Transform

Frame $i$ → STFT → 512 complex Fourier coefficients (0 to 8000 Hz)
Spectrogram: A Sequence of Overlapping Frames

- Signal amplitude
- Discrete time
- Shift (8 ms) 128 samples
- Frame (64 ms) 1024 samples
Binaural Features for a Single Source

- Noise-free binaural signals:

\[
\begin{align*}
m_1(t + \tau_1) &= h_1 \ast s(t) \\
m_2(t + \tau_2) &= h_2 \ast s(t)
\end{align*}
\]

- In the STFT domain:

\[
\begin{align*}
M_{1,fl} e^{-2\pi j f \tau_1} &= H_{1,f} S_{1,fl} \\
M_{2,fl} e^{-2\pi j f \tau_2} &= H_{2,f} S_{1,fl}
\end{align*}
\]

- Relative transfer function (a Fourier coefficient):

\[
H_f^{\text{bin}} = \frac{H_{1,f}}{H_{2,f}} = \frac{|M_{1,fl}|}{|M_{2,fl}|} e^{2\pi j f \tau} \in \mathbb{C}
\]

- Time difference of arrival (TDOA): \( \tau = \tau_2 - \tau_1 \).
Power Spectral Density (PSD)

cross-PSD: \[ \Phi_{1,2,fl} = M_{1,fl} M_{2,fl}^* \]

auto-PSD: \[ \Phi_{2,2,fl} = M_{2,fl} M_{2,fl}^* \]

Average Cross- and auto-PSD:

\[ \Phi_{1,2,f} = \frac{1}{L} \sum_{l=1}^{L} M_{1,fl} M_{2,fl}^* \]

\[ \Phi_{2,2,f} = \frac{1}{L} \sum_{l=1}^{L} M_{2,fl} M_{2,fl}^* \]

By averaging over several frames, the content of a **non-stationary** speech source is cancelled out!
Estimation of the Relative Transfer Function

Power spectral density estimates, $\tilde{\Phi}$, can be computed for non-stationary speech signals in the presence of either stationary or non-stationary noise, then:

$$H_{f}^{\text{bin}} \approx \frac{\tilde{\Phi}_{1,2,f}}{\tilde{\Phi}_{2,2,f}}$$

Binaural Vector over $L$ Frames

$$H_{\text{bin}} = \begin{pmatrix}
|H_{1}^{\text{bin}}| \\
\vdots \\
|H_{F}^{\text{bin}}| \\
\Re(H_{1}^{\text{bin}}) \\
\Im(H_{1}^{\text{bin}}) \\
\vdots \\
\Re(H_{F}^{\text{bin}}) \\
\Im(H_{F}^{\text{bin}})
\end{pmatrix} \in \mathbb{R}^{1536}$$

- observed, $\times$ - missing (absent)
Supervised Sound-source Localization

Training:
white noise $\leftrightarrow$ direction

Localization:
speech $\rightarrow$ direction

- observed, $\times$ - missing
Gaussian Locally-Linear Mapping (GLLiM)

- \( Y \in \mathbb{R}^D \) (high-dimensional space)
- \( X \in \mathbb{R}^L \) \((L \ll D)\)
- Piecewise linear mapping:

\[
Y = \sum_{k=1}^{K} \mathbb{I}(Z = k)(A_k X + b_k + e_k),
\]
Mixture of Piecewise-linear Inverse Regressions

- Low-dimensional to high-dimensional model:

\[
p(y, x; \theta) = \sum_{k=1}^{K} p(y|x, Z = k; \theta)p(x|Z = k; \theta)p(Z = k; \theta),
\]

- with:

\[
p(y|x, Z = k; \theta) = \mathcal{N}(y; A_k x + b_k, \Sigma_k)
\]
\[
p(x|Z = k; \theta) = \mathcal{N}(x; c_k, \Gamma_k)
\]
\[
p(Z = k; \theta) = \pi_k
\]

\[
\Sigma_k = \text{Diag}(\sigma_{k1}, \ldots, \sigma_{kD}) \in \mathbb{R}^{D \times D}
\]
Expectation-Maximization Algorithm

E-step:

\[ r_{Z}^{(i)} = p(Z_{1:N} | (y, x)_{1:N}; \theta^{(i-1)}) \]

M-step:

\[ \theta^{(i)} = \arg\max_{\theta} \mathbb{E}_{Z} [\log p((x, y, Z)_{1:N}; \theta | (x, y)_{1:N}; \theta^{(i-1)})] \].
Generative Manifold Learning

Inverse and Forward Predictive Distributions

Inverse predictive distribution (low-to-high):

\[ p(y|x; \theta) = \sum_{k=1}^{K} \nu_k \mathcal{N}(y; A_k x + b_k, \Sigma_k), \]

with \( \nu_k = \frac{\pi_k \mathcal{N}(x; c_k, \Gamma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x; c_j, \Gamma_j)} \)

Forward predictive distribution (high-to-low):

\[ p(x|y; \theta^*) = \sum_{k=1}^{K} \nu_k^* \mathcal{N}(x; A_k^* y + b_k^*, \Sigma_k^*), \]

with \( \nu_k^* = \frac{\pi_k^* \mathcal{N}(y; c_k^*, \Gamma_k^*)}{\sum_{j=1}^{K} \pi_j^* \mathcal{N}(y; c_j^*, \Gamma_j^*)} \)
Audio-visual Localization

Audio-visual Training Dataset

- Sound source
- Binaural feature
- Sound direction

(i,j)
Audio-visual Alignment Pipeline
Spatial Alignment

audio (green) & visual (blue) clustering result
ground truth (yellow square)
active speaker (blue disk)
A weight $w_i > 0$ is associated with each observation $x_i$ (audio or visual) and plugged into a GMM:

$$P(x_i | w_i; \Theta_k) = \sum_{k=1}^{K} \pi_k N \left( x_i; \mu_k, \frac{1}{w_i} \Sigma_k \right)$$

A high weight corresponds to a reliable data point.

http://arxiv.org/abs/1509.01509
Weight Model

The weights are random variables drawn from a Gamma distribution:

\[
P(w; \phi) = \mathcal{G}(w; \alpha, \beta) = \frac{1}{\Gamma(\alpha)} \beta^\alpha w^{\alpha-1} e^{-\beta w},
\]

\[
E[w] = \frac{\alpha}{\beta},
\]

\[
\text{var}[w] = \frac{\alpha}{\beta^2}.
\]
The marginal posteriors have closed-form expressions:

\[ P(z_i = k | x_i; \theta, \phi_i) \propto \pi_k P(x_i; \mu_k, \Sigma_k, \alpha_i, \beta_i) \]

with 

\[ P(x; \mu, \Sigma, \alpha, \beta) = \frac{\Gamma(\alpha + d/2)}{|\Sigma|^{1/2} \Gamma(\alpha) (2\pi\beta)^{d/2}} \left( 1 + \frac{\|x - \mu\|_2^2}{2\beta \Sigma} \right)^{-(\alpha + d/2)} \]

and:

\[ P(w_i | z_i = k, x_i; \theta, \phi_i) = \mathcal{G}(w_i; a_i, b_{ik}) \]

\[ a_i = \alpha_i + \frac{d}{2} \]

\[ b_{ik} = \beta_i + \frac{1}{2} \|x_i - \mu_k\|_2^2 \Sigma_k \]
Maximization

\[ \pi_k = \frac{1}{n} \sum_{i=1}^{n} \eta_{ik} \]

\[ \mu_k = \frac{\sum_{i=1}^{n} \overline{w}_{ik} \eta_{ik} \mathbf{x}_i}{\sum_{i=1}^{n} \overline{w}_{ik} \eta_{ik}} , \]

\[ \Sigma_k = \frac{\sum_{i=1}^{n} \eta_{ik} \overline{w}_{ik} (\mathbf{x}_i - \mu_k) (\mathbf{x}_i - \mu_k)^\top}{\sum_{i=1}^{n} \eta_{ik}} . \]
The weight of an audio observation (green) or of a visual observation (blue):

\[ w_i = \sum_{j \in N(i)} \exp^{-d^2(x_i, x_j)} \]
Audio-visual Clustering Results
## Comparison Table

<table>
<thead>
<tr>
<th>Seq.</th>
<th># Seg.</th>
<th>WD-EM</th>
<th>GMM-U Banfield &amp; Raftery'93</th>
<th>FM-uMST Lee &amp; McLachlan'14</th>
</tr>
</thead>
<tbody>
<tr>
<td>FS</td>
<td>28</td>
<td>100.00%</td>
<td>100.00%</td>
<td>71.43%</td>
</tr>
<tr>
<td>MS</td>
<td>43</td>
<td>83.87%</td>
<td>61.90%</td>
<td>72.22%</td>
</tr>
<tr>
<td>CP</td>
<td>115</td>
<td>65.66%</td>
<td>52.48%</td>
<td>49.57%</td>
</tr>
</tbody>
</table>
Visual Tracking

- A well investigated topic in computer vision
- Multiple-person tracking is still challenging:
  - People detection is challenging (no universal method)
  - Robustness to changes in appearance, occlusions, etc.
  - Most methods use time-consuming sampling, e.g. Monte Carlo, techniques.
  - Most state-of-the-art methods are not causal (use of past, present and future frames).
- We proposed a dynamic Bayesian graphical model and an associated variational EM algorithm.

[Ba, Alameda, Xompero, Horaud 2016] Computer Vision and Image Understanding
Visual-tracking Principle

Latent variables

Clutter
Person 1
Person 2
Person 3
Online Variational Bayesian Model

- Variational approximation of the multi-person filtering distribution
- State space of fixed dimension with an existence variable specifying targets that are visible or not visible
- Exploits observations from multiple detectors, e.g. face, upper body, skin, etc.
- Birth and visibility processes for people coming and out of the field of view.
Temporal Alignment: Single Speaker

Latent variables:
- $S_{t-1} = 1$
- $S_t = 0$
- $S_{t+1} = 3$

Diarization:
- Clutter
- Person 1
- Person 2
- Person 3

Radu Horaud | Audio-Visual Fusion
MAP formulation

\[ \hat{s}_t = \arg\max_{s_t} P(S_t = s_t | X_{1:t}, V_{1:t}, Y_{1:t}, A_{1:t}) \].

- Active-speaker latent variables \( S_{1:t} \). At frame index \( t \):
  \( S_t = n, n \in \{1, \ldots, N\} \) if person \( n \) is both visible and emits speech at \( t \), \( S_t = 0 \) if no visible person speaks at \( t \).
- \( X_{tn} \in \mathbb{R}^2 \) is the location of person \( n \) at frame \( t \); \( V_{tn} = 1 \) if person \( n \) is detected at \( t \), and 0 otherwise;
- \( Y_{tk} \in \mathbb{R}^2 \) is the direction (image location) of sound-source \( k \) at \( t \); \( A_t \in \{0, 1\} \) is the output of a voice activity detection (VAD).
Temporal Alignment: Multiple Speakers

Latent variables

$t-1$  $t$  $t+1$

$S_1 = 1$  $S_1 = 1$  $S_1 = 0$

$S_2 = 0$  $S_2 = 0$  $S_2 = 1$

$S_3 = 1$

Radu Horaud  Audio-Visual Fusion
Multiple Speech Sources

\[ M_{1,fl} = H_{11,fl}S_{1,fl} + \cdots + H_{1K,fl}S_{K,fl} + N_{1,f} \]
\[ M_{2,fl} = H_{21,fl}S_{1,fl} + \cdots + H_{2K,fl}S_{K,fl} + N_{2,f} \]

- We make the assumption that at each frequency-frame point \((f, l)\), only one of the sources is active

  cross-PSD : \( \Phi_{1,2,fl} = M_{1,fl}M_{2,fl}^* \)

  auto-PSD : \( \Phi_{2,2,fl} = M_{2,fl}M_{2,fl}^* \)

\[ H_{fl}^{\text{bin}} = \frac{\Phi_{1,2,fl}}{\Phi_{2,2,fl}} \approx \frac{|H_{1k,fl}|}{|H_{2k,fl}|} e^{2\pi j f \tau} \in \mathbb{C} \]
Supervised Localization of Multiple Speech Sources

- Complex-valued binaural spectrogram:
  \[ H_{\text{bin}} = \left\{ H_{fl}^{\text{bin}} \right\}_{f=1,l=1}^{f=F,l=L} \]

- Training audio-visual dataset of \( M \) binaural feature vectors:
  \[ \tilde{W} = \{ \tilde{W}_1, \ldots, \tilde{W}_m, \ldots, \tilde{W}_M \} \in \mathbb{C}^{F \times M} \]

- and associated \( M \) sound directions (or image locations):
  \[ \tilde{X} = \{ \tilde{X}_1, \ldots, \tilde{X}_m, \ldots, \tilde{X}_M \} \in \mathbb{R}^{2 \times M} \]

- Each binaural observation is drawn from a complex-Gaussian distribution centered at \( \tilde{W}_{mf} \):
  \[ P(H_{fl}^{\text{bin}}; \theta) = \mathcal{N}_c(H_{fl}^{\text{bin}}; \tilde{W}_{mf}, \sigma_f) \quad \forall \ 1 \leq f \leq F \]
  \[ \text{with} \quad (\tilde{W}_{m1}, \ldots, \tilde{W}_{mf}, \ldots, \tilde{W}_{mF}) \leftrightarrow \tilde{X}_m \]
Spatiotemporal Alignment of Sound-Sources and Persons

- Binaural features are clustered using $F$ complex-Gaussian mixture models
- The single-source temporal model has been extended to multiple sources / multiple persons
- Diarization is estimated via a MAP formulation

[Gebru, Ba, Li and Horaud 2017] IEEE TPAMI
http://arxiv.org/abs/1603.09725
Example with Multiple Speakers
Vision vs. Audio

The two modalities are different:

Visual data (rays of light) are dense both in the spatial and temporal domains, are corrupted by photometric effects, occlusions, foreshortening, depth ambiguities, limited field of view, limited range, etc.

→ We seek *illuminant-invariant* features.

Auditory data (acoustic waves) are sparse in the spectral and temporal domains, corrupted by overlapping (mixed) signals, background noise, reverberations, room acoustics, etc.

→ We seek *environment-invariant* features.
Our Robots

NAO with stereo vision  
12 microphones  
PEPPER
HRI Distributed Software Architecture

https://team.inria.fr/perception/research/naolab/
Conclusions

- Auditory and visual data cannot be combined in their original formats.
- We addressed spatial and spatiotemporal alignment, and diarization.
- Unconstrained scenarios, robot-centric sensors, no wearable devices.
Next Steps

- Combine sound separation with sound localization, such at to assign speech-sources to people, not just source locations (on its way).
- Incorporate visual gaze and visual focus of attention, i.e. who is looking at whom/what [Massé, Ba and Horaud 2016] IEEE ICMI.
- Understanding turn-taking, robot pops into the conversation, etc.
- Speech recognition, natural language processing and dialogue (currently not addressed in our group).
Collaborators & Funding

Joint work with:

- Antoine Deleforge, Florence Forbes, Laurent Girin, Sileye Ba, Israel Gebru, Xavier Alameda-Pineda, Xiaofei Li, Sharon Gannot, Stéphane Lathuilière, Benoit Massé, Pablo Mesejo, etc.

Funding:

- EU project HUMAVIPS (2010-2013) “HUManoids with Audio-Visual abilities in Populated Spaces”,
- ERC Proof of Concept VHIALab (2017-2018)
Publications, research pages, datasets, software, etc.

https://team.inria.fr/perception
Data Challenge: Audio-Visual Speaker Diarization
Data Challenge: Audio-Visual Speaker Diarization
Data Challenge: Audio-Visual Speaker Diarization

Observations:
Data Challenge: Audio-Visual Speaker Diarization

Observations:
Data Challenge: Audio-Visual Speaker Diarization

Problems and Difficulties

- Visual tracking of multiple persons
  - How many people?
  - Occlusions
- Associate visual tracks to Sound Source Localisation (SSL)
  - Unknown number of speakers at each time step
  - Noisy SSL

Evaluation

- Tracking metric: Multi-object tracking accuracy (MOTA)
- Speaker detection: diarization error rate (DER)