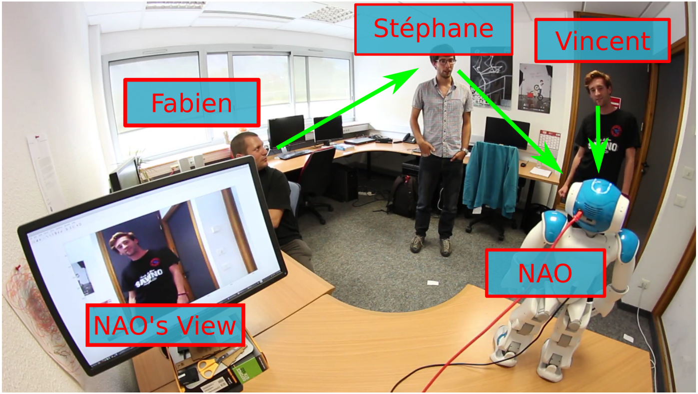
# Audio-Visual Fusion for Human-Robot Interaction

Radu Horaud
PERCEPTION team
INRIA Grenoble Rhône-Alpes
Radu.Horaud@inria.fr

# Human-Computer Interaction

Single user that can control the device.

# Human-Robot Interaction



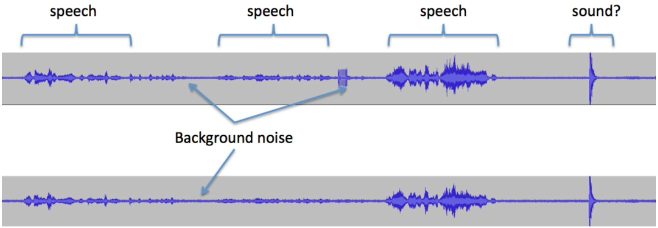

The robot must take decisions!

# Outline

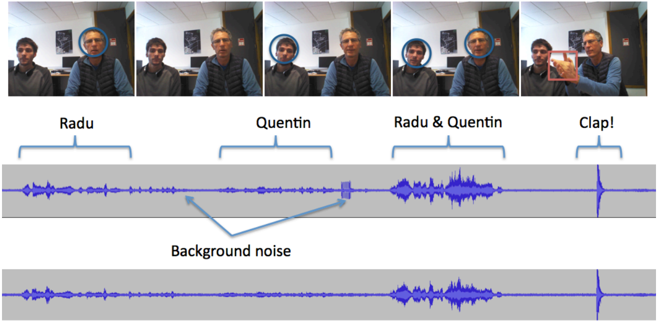
- Audio, vision, and audio-visual challenges
- Datasets
- Supervised sound-source localization
- Mapping sounds onto images
- Clustering audio and visual features
- Multiple-person visual tracking
- Tracking a single speaker
- Tracking multiple speakers
- Conclusions and future work

# Visual Processing

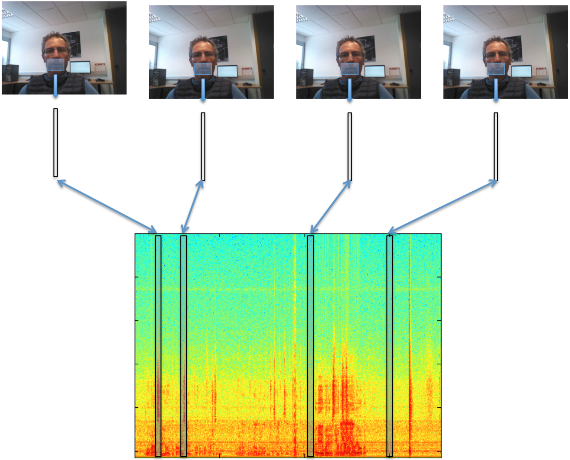
# Audio Processing



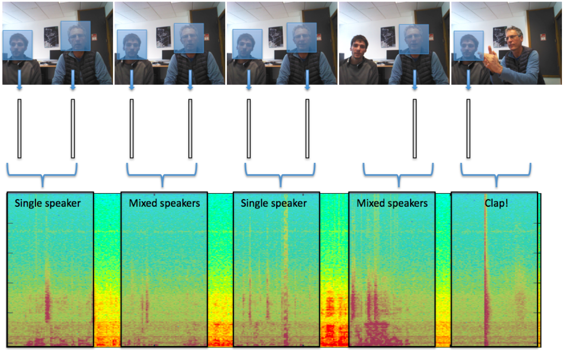
# Audio-visual Scene Analysis



# 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 ($n$):

$$m_1(t) = \underbrace{h_1 \star s(t)}_{\text{convolution}} + n_1(t) \in \mathbb{R}$$

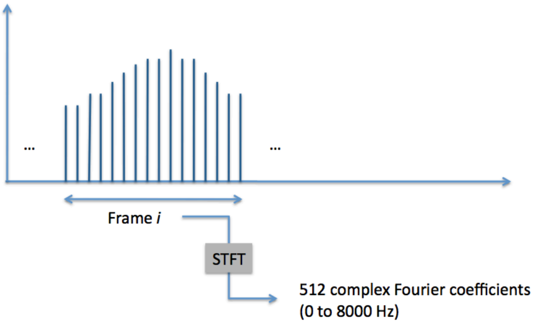
$$m_2(t) = h_2 \star s(t) + n_2(t) \in \mathbb{R}$$

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

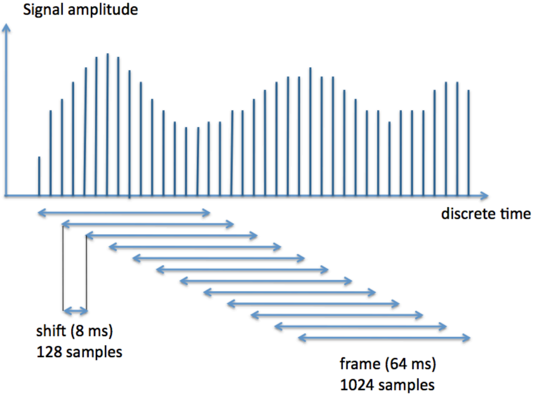
$$M_{1,fl} = H_{1,f} S_{fl} + N_{1,f} \in \mathbb{C}$$
$$M_{2,fl} = H_{2,f} S_{fl} + N_{2,f} \in \mathbb{C}$$

- $f$ (frequency bin) and $l$ (time) are the indexes of a spectrogram point.

# The Short-time Fourier Transform



# Spectrogram: A Sequence of Overlapping Frames



# Binaural Features for a Single Source

- Noise-free binaural signals:

$$m_1(t+\tau_1) = h_1 \star s(t)$$
$$m_2(t+\tau_2) = h_2 \star s(t)$$

- In the STFT domain:

$$M_{1,fl} e^{-2\pi j f \tau_1} = H_{1,f} S_{1,fl}$$
$$M_{2,fl} e^{-2\pi j f \tau_2} = H_{2,f} S_{1,fl}$$

- 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)

$$\begin{aligned} \text{cross-PSD}: \quad & \Phi_{1,2,fl} = M_{1,fl} M_{2,fl}^{*} \\ \text{auto-PSD}: \quad & \Phi_{2,2,fl} = M_{2,fl} M_{2,fl}^{*} \end{aligned}$$

- 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:

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

[X. Li et al 2015] IEEE ICASSP 2015

[X. Li et al 2016] IEEE ICASSP 2016

# Binaural Vector over $L$ Frames

$$
\boldsymbol{H}^{\text{bin}} = \begin{pmatrix} |H_1^{\text{bin}}| \\ \vdots \\ |H_F^{\text{bin}}| \\ \text{Re}(H_1^{\text{bin}}) \\ \text{Im}(H_1^{\text{bin}}) \\ \vdots \\ \text{Re}(H_F^{\text{bin}}) \\ \text{Im}(H_F^{\text{bin}}) \end{pmatrix} \in \mathbb{R}^{1536}
\qquad
\begin{array}{cc}
\text{white noise} & \text{speech} \\
\begin{pmatrix} \bullet \\ \vdots \\ \bullet \\ \vdots \\ \bullet \\ \bullet \\ \bullet \\ \vdots \\ \bullet \\ \bullet \end{pmatrix} &
\begin{pmatrix} \bullet \\ \vdots \\ \times \\ \vdots \\ \bullet \\ \times \\ \times \\ \vdots \\ \bullet \\ \bullet \end{pmatrix}
\end{array}
$$

- $\bullet$ - observed, $\times$ - missing (absent)

# Supervised Sound-source Localization

**Training**

white noise $\leftrightarrow$ direction

$$\begin{pmatrix}\bullet\\ \bullet\\ \bullet\\ \bullet\\ \bullet\\ \bullet\\ \bullet\\ \bullet\\ \bullet\end{pmatrix} \underset{?}{\Longrightarrow} \begin{pmatrix}\bullet\\ \bullet\end{pmatrix}$$

**Localization**

speech $\rightarrow$ direction

$$\begin{pmatrix}\bullet\\ \times\\ \bullet\\ \times\\ \bullet\\ \times\\ \times\\ \bullet\\ \times\end{pmatrix} \Longrightarrow \begin{pmatrix}?\\ ?\end{pmatrix}$$

$\bullet$ - observed, $\times$ - missing

# Gaussian Locally-Linear Mapping (GLLiM)

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

$$\boldsymbol{Y} = \sum_{k=1}^{K} \mathbb{I}(Z = k)(\mathbf{A}_k \boldsymbol{X} + \boldsymbol{b}_k + \boldsymbol{e}_k),$$

# Mixture of Piecewise-linear Inverse Regressions

- Low-dimensional to high-dimensional model:

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

- with:

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{x}, Z = k; \boldsymbol{\theta}) &= \mathcal{N}(\boldsymbol{y}; \mathbf{A}_k \boldsymbol{x} + \boldsymbol{b}_k, \boldsymbol{\Sigma}_k) \\ p(\boldsymbol{x}|Z = k; \boldsymbol{\theta}) &= \mathcal{N}(\boldsymbol{x}; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k) \\ p(Z = k; \boldsymbol{\theta}) &= \pi_k \end{aligned}$$

$\boldsymbol{\Sigma}_k = \mathrm{Diag}(\sigma_{k1}, \ldots, \sigma_{kD}) \in \mathbb{R}^{D \times D}$

# Expectation-Maximization Algorithm

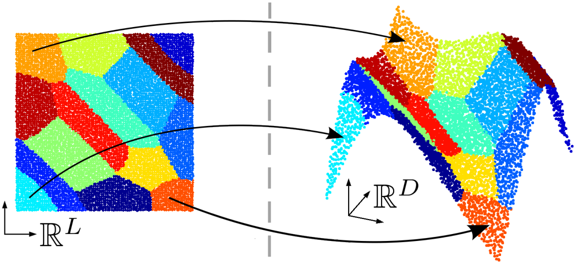
E-step:

$$r_Z^{(i)} = p(\mathbf{Z}_{1:N} | (\boldsymbol{y}, \boldsymbol{x})_{1:N}; \boldsymbol{\theta}^{(i-1)})$$

M-step:

$$\boldsymbol{\theta}^{(i)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left( \mathbb{E}_Z [\log p((\boldsymbol{x}, \boldsymbol{y}, Z)_{1:N}; \boldsymbol{\theta} | (\boldsymbol{x}, \boldsymbol{y})_{1:N}; \boldsymbol{\theta}^{(i-1)})] \right).$$

# Generative Manifold Learning




[Deleforge, Forbes, Horaud 2015] Statistics and Computing

# Inverse and Forward Predictive Distributions

*Inverse predictive distribution (low-to-high):*

$$p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \nu_k \mathcal{N}(\boldsymbol{y}; \mathbf{A}_k \boldsymbol{x} + \boldsymbol{b}_k, \boldsymbol{\Sigma}_k),$$

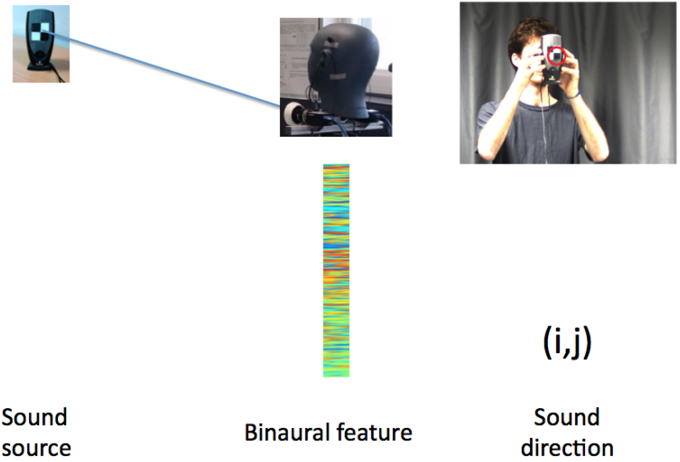
$$\text{with } \nu_k = \frac{\pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}; \boldsymbol{c}_j, \boldsymbol{\Gamma}_j)}$$

*Forward predictive distribution (high-to-low):*

$$p(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}^*) = \sum_{k=1}^{K} \nu_k^* \mathcal{N}(\boldsymbol{x}; \mathbf{A}_k^* \boldsymbol{y} + \boldsymbol{b}_k^*, \boldsymbol{\Sigma}_k^*)$$

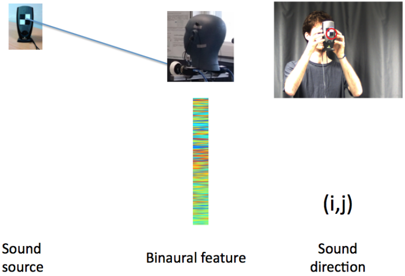
$$\text{with } \nu_k^* = \frac{\pi_k^* \mathcal{N}(\boldsymbol{y}; \boldsymbol{c}_k^*, \boldsymbol{\Gamma}_k^*)}{\displaystyle\sum_{j=1}^{K} \pi_j^* \mathcal{N}(\boldsymbol{y}; \boldsymbol{c}_j^*, \boldsymbol{\Gamma}_j^*)}$$

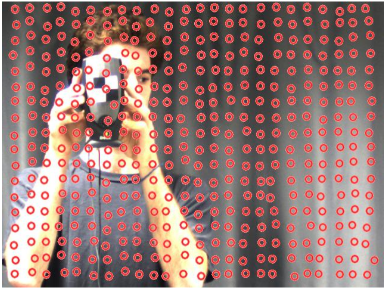
# Audio-visual Localization



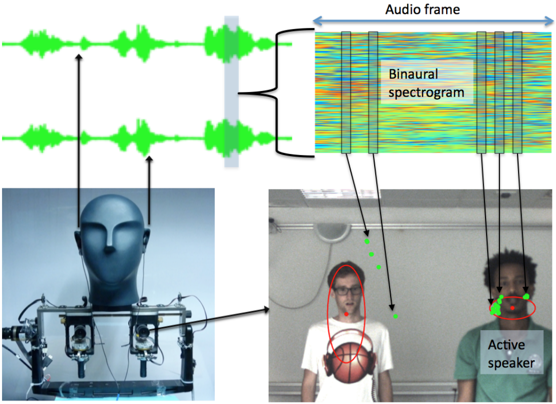

[Deleforge et al. 2015] IEEE Trans. Audio, Speech & Lang. Proc.

# Audio-visual Training Dataset

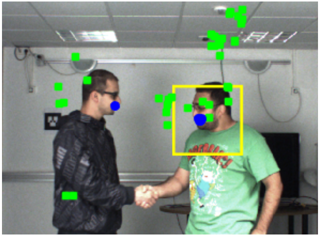


# Audio-visual Alignment Pipeline



# Spatial Alignment



audio (green) & visual (blue)
ground truth (yellow square)

clustering result
active speaker (blue disk)

# Weighted-Data Gaussian Mixture Model

- A weight $w_i > 0$ is associated with each observation $\boldsymbol{x}_i$ (audio or visual) and plugged into a GMM:

$$P(\boldsymbol{x}_i | w_i; \boldsymbol{\Theta}_k) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \frac{1}{w_i}\boldsymbol{\Sigma}_k\right)$$

- A high weight corresponds to a reliable data point.

[Gebru, Alameda, Forbes, Horaud 2016] IEEE TPAMI
`http://arxiv.org/abs/1509.01509`

# Weight Model

The weights are random variables drawn from a Gamma distribution:

$$
\begin{aligned}
P(w;\boldsymbol{\phi}) &= \mathcal{G}\left(w;\alpha,\beta\right) \\
&= \Gamma\left(\alpha\right)^{-1}\beta^{\alpha}w^{\alpha-1}e^{-\beta w}, \\
\mathrm{E}[w] &= \alpha/\beta, \\
\mathrm{var}[w] &= \alpha/\beta^{2}.
\end{aligned}
$$

# Expectation

The marginal posteriors have closed-form expressions:

$$P(z_i = k | \boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\phi}_i) \propto \pi_k \, \mathcal{P}(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \alpha_i, \beta_i)$$

$$\text{with} \quad \mathcal{P}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \beta) =$$

$$\frac{\Gamma(\alpha + d/2)}{|\boldsymbol{\Sigma}|^{1/2} \, \Gamma(\alpha) \, (2\pi\beta)^{d/2}} \left( 1 + \frac{\|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}}}{2\beta} \right)^{-(\alpha + \frac{d}{2})}.$$

and:

$$\begin{aligned} P(w_i | z_i = k, \boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\phi}_i) = \mathcal{G}(w_i; a_i, b_{ik}) \\ a_i = \alpha_i + \frac{d}{2} \\ b_{ik} = \beta_i + \frac{1}{2} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2_{\boldsymbol{\Sigma}_k} \end{aligned}$$

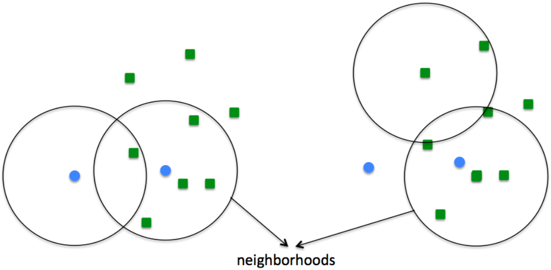
# Maximization

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

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{n} \overline{w}_{ik} \eta_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{n} \overline{w}_{ik} \eta_{ik}},$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{i=1}^{n} \eta_{ik} \overline{w}_{ik} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_k\right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}_k\right)^{\top}}{\sum_{i=1}^{n} \eta_{ik}}.$$

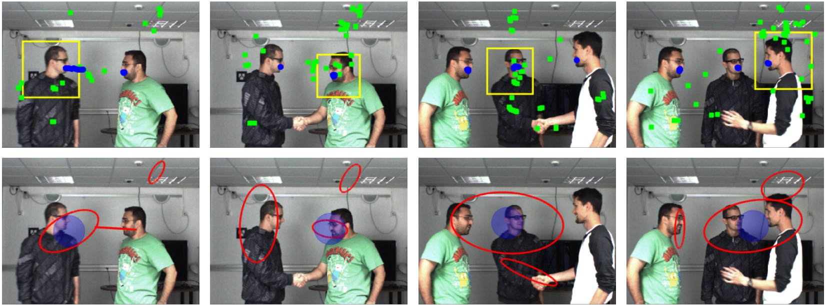
# Weight Initialisation




The weight of an audio observation (green) or of a visual observation (blue):

$$w_i = \sum_{j \in \mathsf{N}(i)} \exp^{-d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)}$$

# Audio-visual Clustering Results



# Comparison Table

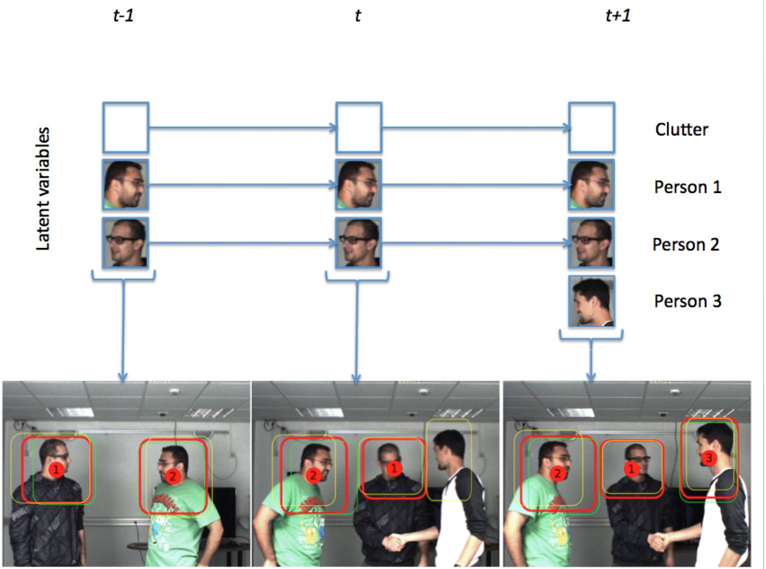
| Seq. | # Seg. | WD-EM | GMM-U Banfield&Raftery'93 | FM-uMST Lee&McLachlan'14 |
|---|---|---|---|---|
| FS | 28 | 100.00% | 100.00% | 71.43% |
| MS | 43 | 83.87% | 61.90% | 72.22% |
| CP | 115 | 65.66% | 52.48% | 49.57% |

# 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

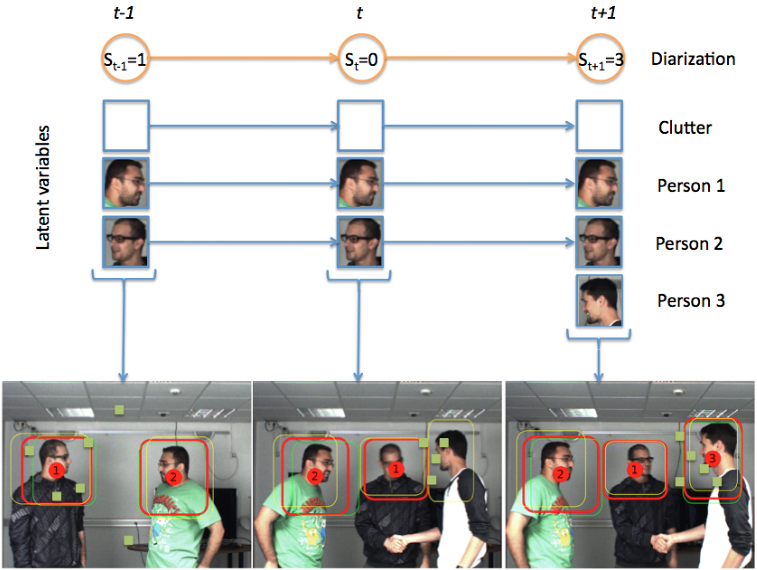
# Visual-tracking Principle



# 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 and out of the field of view.

# Temporal Alignment: Single Speaker

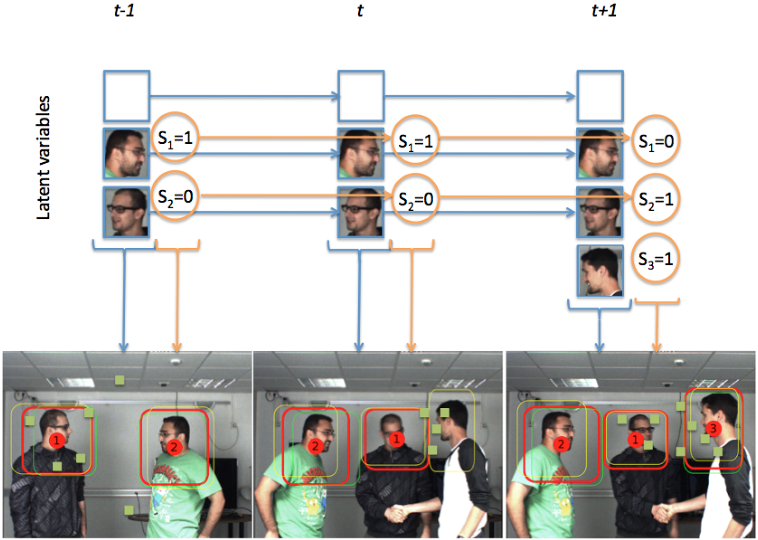


# MAP formulation

$$\hat{s}_t = \operatorname*{argmax}_{s_t} P(S_t = s_t | \boldsymbol{X}_{1:t}, \boldsymbol{V}_{1:t}, \boldsymbol{Y}_{1:t}, \boldsymbol{A}_{1:t}).$$

- Active-speaker latent variables $\mathbf{S}_{1:t}$. At frame index $t$: $S_t = n, n \in \{1, \dots, N\}$ if person $n$ is both visible and emits speech at $t$, $S_t = 0$ if no visible person speaks at $t$.
- $\boldsymbol{X}_{tn} \in \mathbb{R}^2$ is the location of person $n$ at frame $t$; $V_{tn} = 1$ if person $n$ is detected at $t$, and 0 otherwise;
- $\boldsymbol{Y}_{tk} \in \mathbb{R}^2$ 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



# 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

$$\text{cross-PSD}: \quad \Phi_{1,2,fl} = M_{1,fl} M^*_{2,fl}$$
$$\text{auto-PSD}: \quad \Phi_{2,2,fl} = M_{2,fl} M^*_{2,fl}$$

$$H^{\text{bin}}_{fl} = \frac{\Phi_{1,2,fl}}{\Phi_{2,2,fl}} \approx \frac{|H_{1k,fl}|}{|H_{2k,fl}|} e^{2\pi j f \tau} \in \mathbb{C}$$

# Supervised Localization of Multiple Speech Sources

- Complex-valued binaural spectrogram: $\mathbf{H}^{\text{bin}} = \left\{ H^{\text{bin}}_{fl} \right\}_{f=1,l=1}^{f=F,l=L}$
- Training audio-visual dataset of $M$ binaural feature vectors:

$$\widetilde{\mathbf{W}} = \{\widetilde{\boldsymbol{W}}_1, \dots, \widetilde{\boldsymbol{W}}_m, \dots \widetilde{\boldsymbol{W}}_M\} \in \mathbb{C}^{F \times M}$$

- and associated $M$ sound directions (or image locations):

$$\widetilde{\mathbf{X}} = \{\widetilde{\boldsymbol{X}}_1, \dots, \widetilde{\boldsymbol{X}}_m, \dots \widetilde{\boldsymbol{X}}_M\} \in \mathbb{R}^{2 \times M}$$

- Each binaural observation is drawn from a complex-Gaussian distribution centered at $\widetilde{W}_{mf}$:

$$P(H^{\text{bin}}_{fl}; \boldsymbol{\theta}) = \mathcal{N}_c(H^{\text{bin}}_{fl}; \widetilde{W}_{mf}, \sigma_f) \quad \forall\, 1 \leq f \leq F$$
$$\text{with} \quad (\widetilde{W}_{m1}, \dots, \widetilde{W}_{mf}, \dots, \widetilde{W}_{mF}) \leftrightarrow \widetilde{\boldsymbol{X}}_m$$

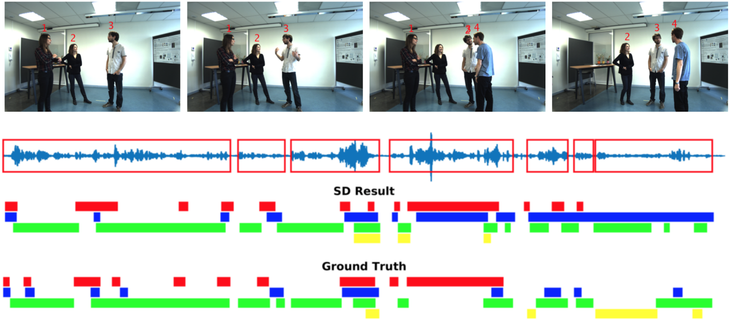
# Spatiotemporal Alignment of Sound-Sources and Persons

- 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.

  $\longrightarrow$ *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.

  $\longrightarrow$ *We seek environment-invariant features.*

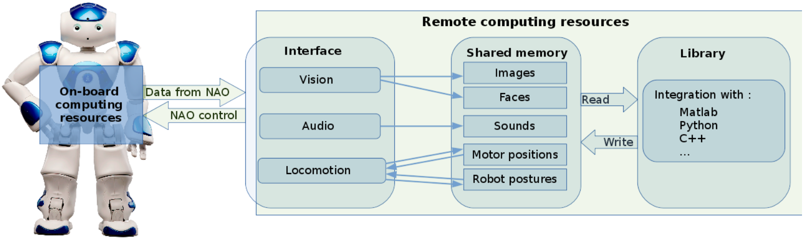
# Our Robots

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.
- Audio-visual attention strategies [Lathuilière, Massé, Mesejo and Horaud 2017] submitted to Pattern Recognition Letters.
- 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.

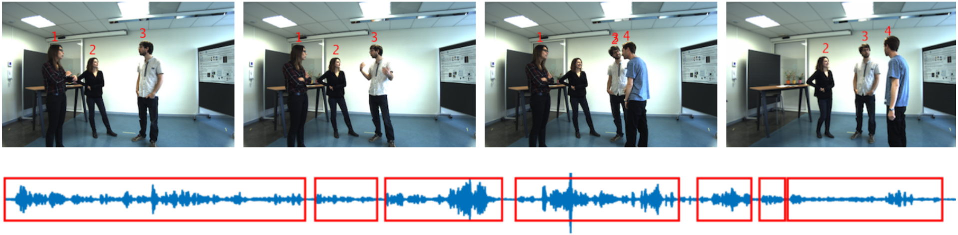
Funding:

- EU project HUMAVIPS (2010-2013) "*HUManoids with Audio-VIsual abilities in Populated Spaces*",
- EU project EARS (2014-2017) "*Embedded Audition for RobotS*"
- ERC VHIA (2014-2019) "*Vision and Hearing In Action*"
- Samsung Electronics (2016-2017) "*Live Together with Robots*"
- ERC Proof of Concept VHIALab (2017-2018)

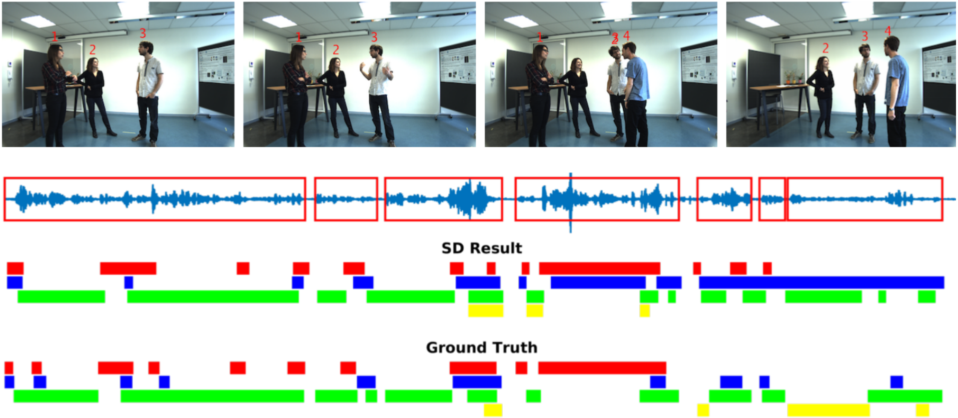
# Publications, research pages, datasets, software, etc.

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)