Audio-visual Machine Perception for Socially Interacting Robots

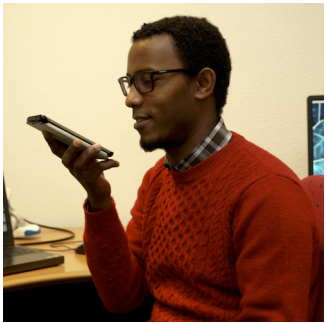
Radu Horaud
Perception team, Inria Grenoble
Radu.Horaud@inria.fr

29 November 2019

Robotics Science (and Technology)

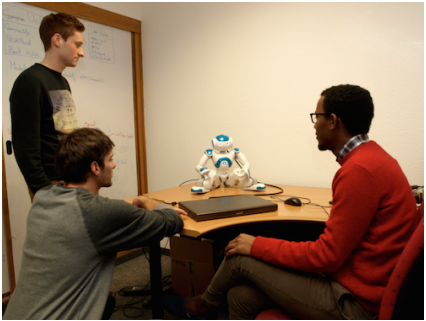
- The art of building a robot able to plan and execute physical tasks.
- Vision (color, range, infrared, etc.) is the primarily sensorial modality in robotics.
- Speech allows humans to communicate.
- Multimodality, e.g. vision and audio, enables social interaction.
- Can we build intelligent robots without hearing/speaking capabilities?

Interaction with Intelligent Devices



- Alexa, GoogleHome, Siri, etc. perform Q&A dialog, e.g. "Who is the general secretary of the People's Republic of China"?
- They cannot answer questions such as "Who is the person seating next to you"?

Robots and People



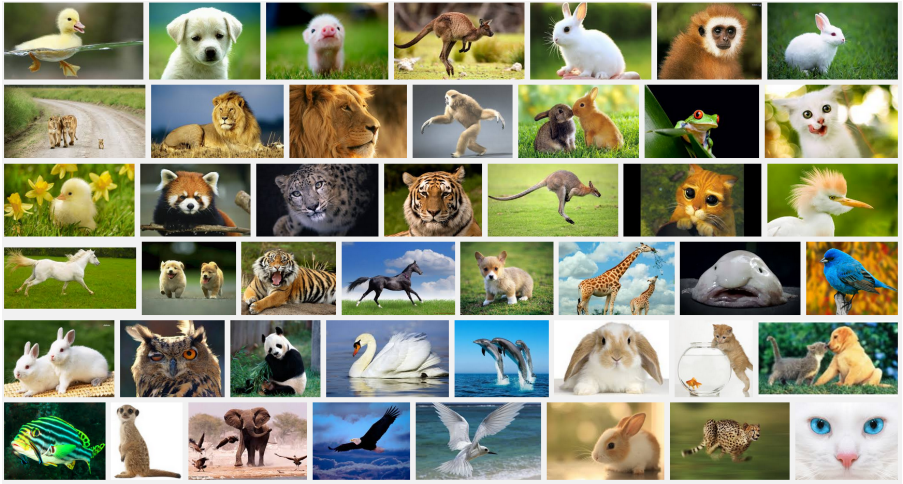
- Robust human-robot interaction needs good audio and visual features, e.g. clean speech and frontal faces;
- Robots must learn actions leading to clean-data collection: this leads to *sensorimotor learning*

Multimodal Interaction with a Robot

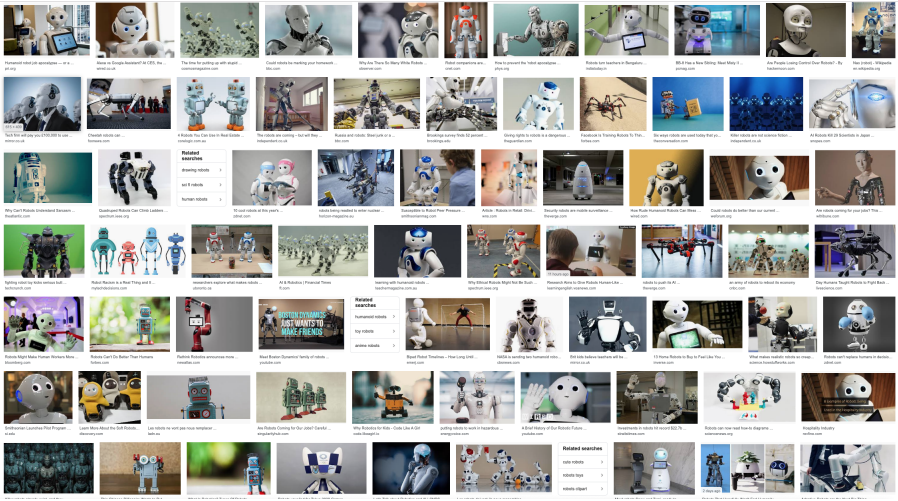
Audio-visual fusion enables interaction beyond hands-free speech recognition and touchscreen interfaces:

- *conversational speech*: speech recognition & speaker diarization,
- *individual- and group-behavior understanding*: social roles, correlation between visual cues (face, head, hands) and verbal cues (speech and non-speech)
- *situated dialog*, e.g. combine speech technologies with visual object recognition.
- *sensorimotor learning*: learn to execute actions leading to robust interaction.

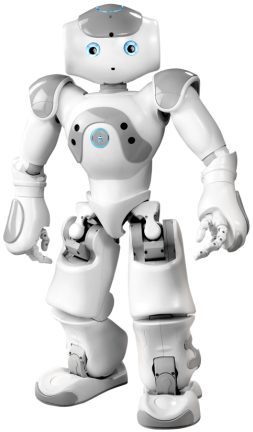
Biodiversity

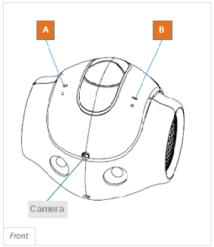


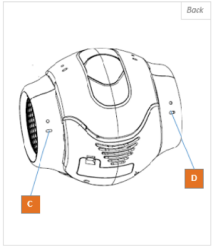
Robots


Related searches
drawing robots
sci fi robots
human robots
Related searches
humanoid robots
toy robots
anime robots
Related searches
cute robots
robots toys
robots clipart
MOTION DYNAMICS JUST WANTS TO MAKE FRIENDS

Nao







Stereo camera pair (EU Humavips)

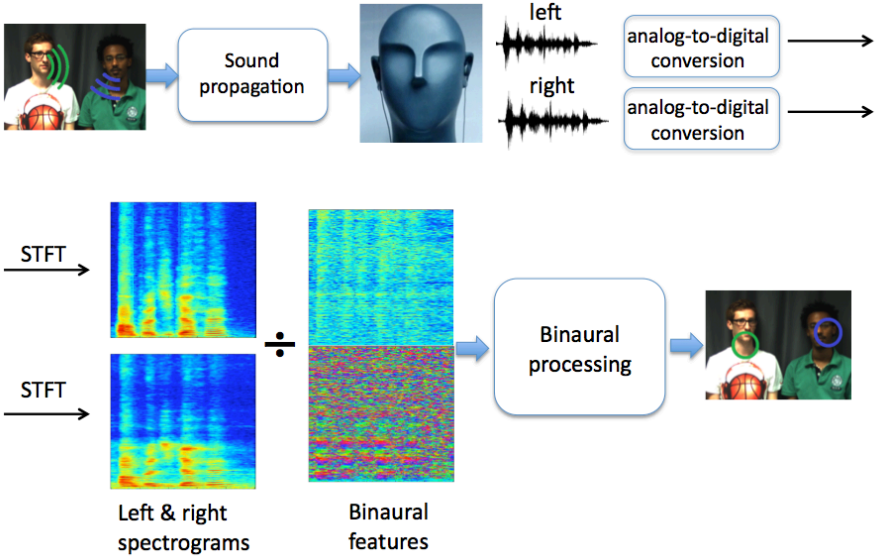
12 microphone array (EU Ears)

http://perception.inrialpes.fr/Free_Access_Data/ICMI2015/ICMI_Demo.mp4

Audio-visual Machine Perception

- Audio-visual alignment:
  - acoustic model and signal processing,
  - mapping sounds on images.
- Audio-visual tracking:
  - intractability,
  - variational inference,
  - online appearance learning.
- Audio-visual gaze control:
  - reward-based robot control,
  - deep reinforcement learning

Audio-visual Alignment Pipeline

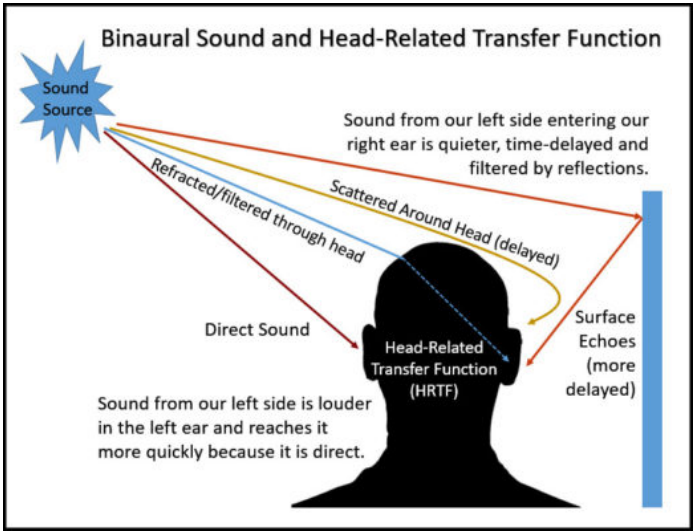


Acoustic Model for a Single Static Source

$$m_1(t) = h_1(t) \star \big(r_1(t) \star s(t)\big) + n_1(t)$$
$$m_2(t) = \underbrace{h_2(t)\star}_{\text{head}} \big(\underbrace{r_2(t)\star}_{\text{room}} s(t)\big) + n_1(t)$$

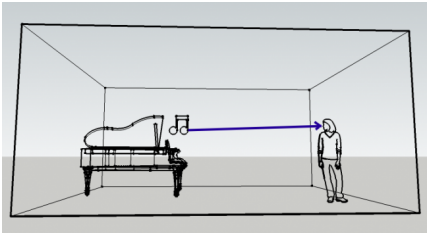
- m: microphone,
- s: sound source,
- r: room impulse response,
- h: head impulse response,
- n: noise,
- $\star$: convolution.

Head Impulse Response, $h(t)$

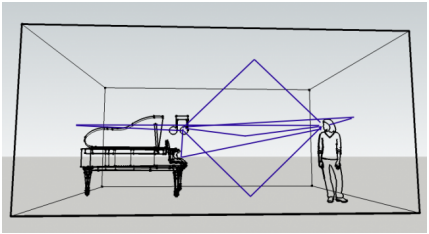


Room Impulse Response, $r(t)$

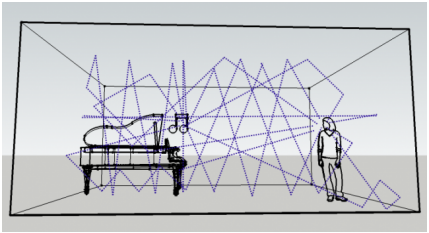
Direct path



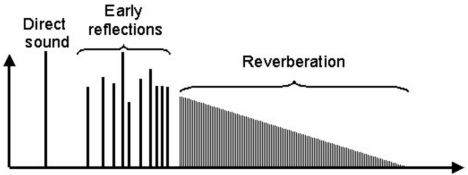
Early reflections



Reverberation



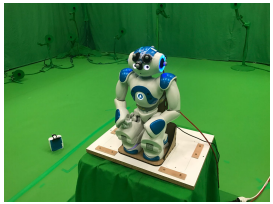
Room impulse response

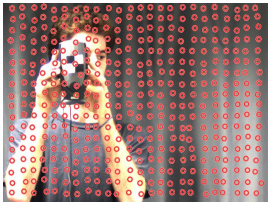


Acoustic Signal Processing

- Noise-free and anechoic room : $m(t) = h(t) \star s(t)$,
- input-output correlation: $m(t) \otimes s(t) = h(t) \star s(t) \otimes s(t)$,
- for a chirp source: $s(t) \otimes s(t) = \delta(t)$ (impulse),
- head impulse response: $h(t) = m(t) \otimes s(t)$,
- discrete Fourier transform : $h(t) \rightarrow \boldsymbol{H} \in \mathbb{C}^F$, with $F = 512$ frequency bins.
- the *head-related transfer function* (HRTF) for two microphones: $\boldsymbol{H}_{21} = \boldsymbol{H}_2 / \boldsymbol{H}_1$

Supervised Audio-Visual Alignment[1]





$$\begin{pmatrix} \boldsymbol{H}_{21} \\ \vdots \\ \boldsymbol{H}_{j1} \\ \vdots \\ \boldsymbol{H}_{J1} \end{pmatrix}_i \in \mathbb{C}^{512(J-1)} \qquad \begin{pmatrix} x_i \\ y_i \end{pmatrix}$$

[1] Deleforge, Horaud, Schechner & Girin, "Co-localization of audio sources in images using binaural features and locally-linear regression," *IEEE/ACM TASLP*, 2015

Room Reverberation Model

- Temporal representation:

$$m(t) = r(t) \star s(t) + n(t)$$

- STFT[2] representation, $f \in [1 \dots F]$, $l \in [1 \dots L]$ frames:
- multiplicative model: $M_{fl} = R_{fl} S_{fl} + N_{fl} \in \mathbb{C}$,
- convolutive model:

$$M_{fl} = R_{fl} \star S_{fl} = \sum_{q=0}^{q=Q-1} R_{lq} S_{l-q}$$

$$= \underbrace{R_{f0} S_{fl}}_{\text{direct path}} + \dots + \underbrace{R_{fq} S_{fl-q} + \dots + R_{fQ-1} S_{fl-Q+1}}_{\text{early reflections + reverberation}}$$

[2] short-time Fourier transform

Direct-Path Relative Transfer Function (DP-RTF)

- Two microphones

$$M_{1,fl} = R_{1,f0} S_{fl} + \cdots + R_{1,fQ-1} S_{fl-Q+1}$$
$$M_{2,fl} = R_{2,f0} S_{fl} + \cdots + R_{2,fQ-1} S_{fl-Q+1}$$

- The ratio $\boxed{R_{21,f0} = \dfrac{R_{2,f0}}{R_{1,f0}}}$ is the *direct-path relative transfer function* at bin f.
- DP-RTF for a single source.[3]
- DP-RTF for multiple sources (up to three).[4]

[3] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the Direct-Path Relative Transfer Function for Supervised Sound-Source Localization," *IEEE/ACM TASLP*, Nov. 2016

[4] X. Li, L. Girin, R. Horaud, and S. Gannot, "Multiple-Speaker Localization Based on Direct-Path Features and Likelihood Maximization with Spatial Sparsity Regularization," *IEEE/ACM TASLP*, Oct. 2017

Supervised Sound-source Localization

Training (HRTF):

$$\begin{pmatrix} \boldsymbol{H}_{21} \\ \vdots \\ \boldsymbol{H}_{j1} \\ \vdots \\ \boldsymbol{H}_{J1} \end{pmatrix}_{i=1}^{i=I} \Longleftrightarrow \begin{pmatrix} x_i \\ y_i \end{pmatrix}$$

Prediction (DP-RTF):

$$\begin{pmatrix} \boldsymbol{R}_{21} \\ \vdots \\ \boldsymbol{R}_{j1} \\ \vdots \\ \boldsymbol{R}_{J1} \end{pmatrix} \Longrightarrow \begin{pmatrix} x \\ y \end{pmatrix}$$

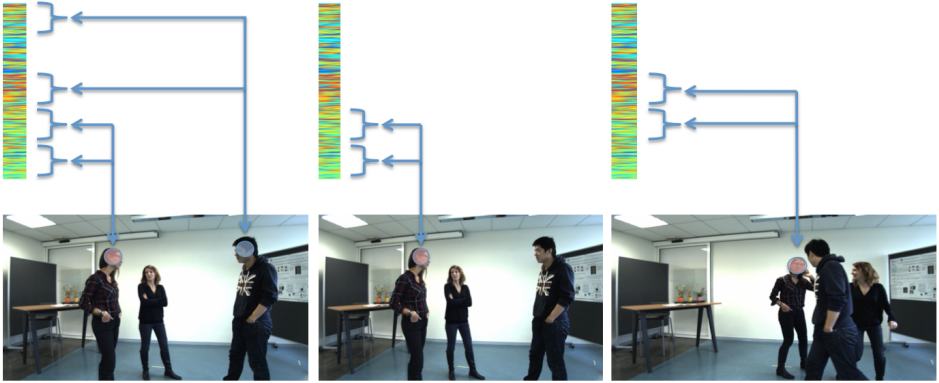
Learning sound-source localization

- The learning data is room-independent, but *robot-head dependent* (shape, material, microphone locations, and number of microphones).
- A chirp signal (high-pitched sound) is used for learning, e.g. `https://upload.wikimedia.org/wikipedia/commons/0/0f/Expchirp.ogg`.
- Speech signals are very sparse in the STFT domain – not suitable for training.
- We use *Gaussian mixture of inverse regression* as it only needs a small training set.[5]

`http://perception.inrialpes.fr/Free_Access_Data/ICMI2015/ICMI_Demo.mp4`

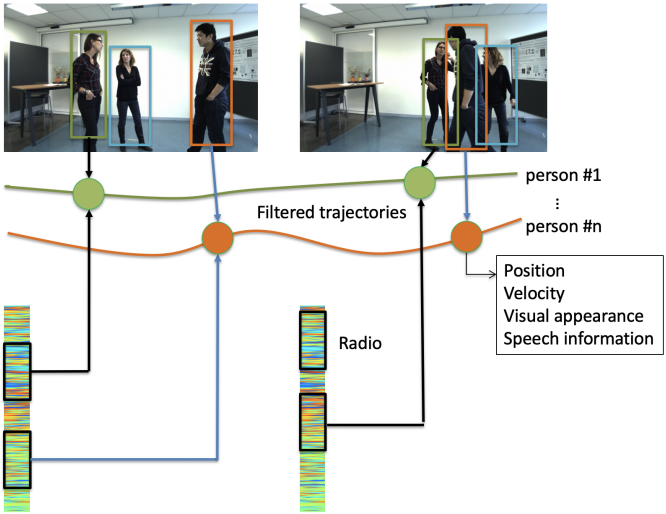
[5] Deleforge, Forbes & Horaud, "High-dimensional regression with Gaussian mixtures and partially-latent response variables," *Statistics and Computing*, 2015

Audio-based Localization of Multiple Moving Speakers[6]



[6]X. Li, Y. Ban, L. Girin, X. Alameda-Pineda, R. Horaud, "Online Localization and Tracking of Multiple Moving Speakers in Reverberant Environments," IEEE Journal of Selected Topics in Signal Processing, Jan. 2019

Audio-Visual Speaker Tracking



Intractability

- N continuous latent variables (speakers) at frame t: $\boldsymbol{x}_t = (\boldsymbol{x}_{t1}, \dots \boldsymbol{x}_{tn}, \dots \boldsymbol{x}_{tN})$,
- M (audio and visual) observations at frame t: $\boldsymbol{o}_t = (\boldsymbol{o}_{t1}, \dots \boldsymbol{o}_{tm}, \dots \boldsymbol{o}_{tM})$,
- M discrete latent variables assigning an observation m at t to speaker n: $Z_t = (Z_{t1}, \dots Z_{tm}, \dots Z_{tM})$ with:
  - assigned to speaker n: $Z_{tm} = n$, and
  - assigned to *nobody*: $Z_{tm} = 0$.

$$P(\boldsymbol{x}_t|\boldsymbol{o}_{1:t}) = \sum_{\tau=1}^{t} \sum_{n=0}^{N} \sum_{m=1}^{M} \int_{\boldsymbol{x}_1, \dots \boldsymbol{x}_{t-1}} P(\boldsymbol{x}_1, \dots \boldsymbol{x}_t, Z_{\tau m} = n|\boldsymbol{o}_{1:t}) d\boldsymbol{x}_1 \dots d\boldsymbol{x}_{t-1}$$

$$= \underbrace{\sum_{c=1}^{C} \pi_c P(\boldsymbol{x}_t; \boldsymbol{\theta}_c)}_{\text{mixture}}, \text{with} \quad C = (N+1)^{tM} \text{mixture components!}$$

Variational inference of multiple speaker tracking[8]

- The joint distribution is replaced by a factorized distribution:

$$P(\boldsymbol{x}_t, Z_t^v, Z_t^a | \boldsymbol{o}_{1:t}) \approx q(\boldsymbol{x}_t) q(Z_t^v) q(Z_t^a)$$

- This variational approximation provides the best factorized estimate of the true posterior distribution in the sense of the Kullback-Leibler divergence.
- Unsupervised and efficient variational expectation-maximization (VEM):
  - estimate the observation-to-object assignment probabilities $q(Z_t^v), q(Z_t^a)$, and
  - update the model parameters (speaker position, velocity, covariance matrix, etc.).
- audio-observation-to-speaker assignments provide diarization information.[7]

[7] I. Gebru, S. Ba, X. Li, R. Horaud, "Audio-Visual Speaker Diarization Based on Spatiotemporal Bayesian Fusion." IEEE Trans. PAMI, May 2018

[8] Y. Ban, X. Alameda-Pineda, L. Girin, R, Horaud, "Variational Bayesian Inference for Audio-Visual Tracking of Multiple Speakers," IEEE Trans. PAMI, 2019

Audio-visual Dataset: AVDIAR

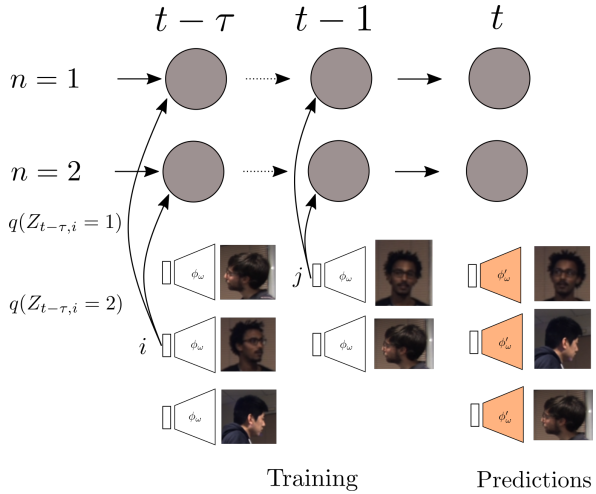
https://team.inria.fr/perception/avdiar/

- People wander around, turn their faces towards the speaker (and not facing the camera!).
- Casual dialogue, speech turns with overlap, background noise.
- Fully annotated (location and trajectory for each person, speech activity of each person over time).

Examples

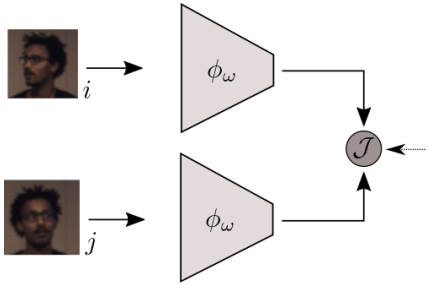
- Informal conversation, people take speech turns, move around, appear and disappear, etc.:
  `http://perception.inrialpes.fr/Free_Access_Data/BAN_TPAMI2018/FFOV/Seq13-4P-S2M1.mp4`
- Whenever a person goes outside the camera field of view, audio tracking takes over:
  `http://perception.inrialpes.fr/Free_Access_Data/BAN_TPAMI2018/PFOV/Seq19-2P-S1M1.mp4`

Variational Bayesian Tracking with Online Person Re-identification[9]



[9] G. Delorme, Y. Ban, G. Sarrazin, X. Alameda-Pineda, "ODA-Track: Online Deep Appearance for Robotic Multiple Person Tracking," Submitted, 2019

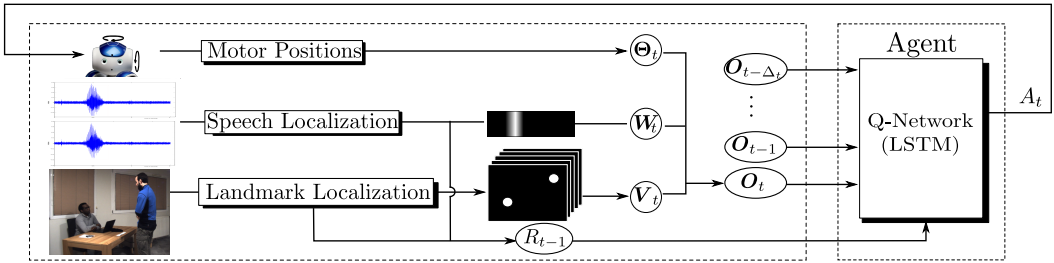
Online Learning with a Siamese Architecture



$$\gamma_{ij} = \begin{cases} p_{ij}, & \text{if } p_{ij} \geq \eta \\ (p_{ij} - 1), & \text{otherwise} \end{cases}$$

$$p_{ij} = \sum_{n=1}^{N} q(Z_{t_i i} = n) q(Z_{t_j j} = n)$$

Audio-visual Gaze Control[10]



http://perception.inrialpes.fr/Free_Access_Data/dRL/prletters.mp4

[10] S. Lathuillière, B. Massé, P. Mesejo, R. Horaud, "Deep Reinforcement Learning for Audio-Visual Gaze Control," IEEE/RSJ IROS'18

Discussion

- Audio-visual human-robot interaction versus state-of-the-art Q&A spoken dialog systems.
- Supervised method for sound-source localization and audio-visual alignment – robot-head dependent training methodolody.
- Bayesian framework for multiple person tracking using multimodal observations (geometric, photometric, acoustic, etc.) with provable efficient solvers.
- Online deep appearance learning.
- Deep reinforcement learning for audio-visual gaze control.

Future Work

- Combine audio-visual machine learning with conversational speech and with situated dialog.
- Address more thoroughly the link between audio-visual machine perception and robot motion control:
  - learning how to look and how to listen to people,
  - audio-visual-based robot control,
  - non-holonomic robots, redundant robots, etc.

Social Robots in the Near Future

- H2020 project SPRING (01/2020 - 12/2023),
- EU funding: 8.3 million euros
- `https://spring-h2020.eu/`
- Social robots for the elderly in hospitals and retirement facilities
- Inria Grenoble (coordinator),
- Heriot Watt U., Czech Technical U.,
- Bar Ilan U., Trento U.,
- Broca Hospital,
- Pal Robotics and
- ERM Ltd.

ARI by Pal Robotics

The Team

- `https://team.inria.fr/perception`
- Computer vision and machine learning:
  Xavi Alameda-PIneda, Sileye Ba, Stéphane Lathuillière, Benoît Massé, Israel Gebru, Yutong Ban, Guillaume Delorme.
- Audio and speech processing:
  Laurent Girin, Xiaofei Li, Sharon Gannot, Dionyssos Kounades-Bastian, Simon Leglaive.
- Statistical machine learning:
  Florence Forbes, Antoine Deleforge.
- Robotics:
  Soraya Arias, Bastien Mourgue, Guillaume Sarrazin, Fabien Badeig.