Binaural Hearing for Robots

Fusion of Audio and Vision

Radu Horaud

Binaural Hearing for Robots

1. Introduction to Robot Hearing
2. Methodological Foundations
3. Sound-Source Localization
4. Machine Learning and Binaural Hearing
5. **Fusion of Audio and Vision**

Radu Horaud

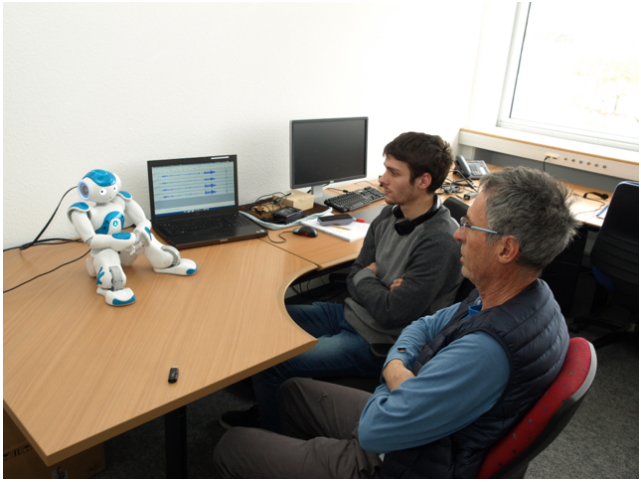
5. Fusion of Audio and Vision

1. Audio-visual processing challenges
2. Representation of visual information
3. The geometry of vision
4. Audio-visual feature association
5. Audio-visual alignment
6. Visually-guided audio localization
7. Audio-visual event localization
8. Audio-visual clustering
9. Conclusions

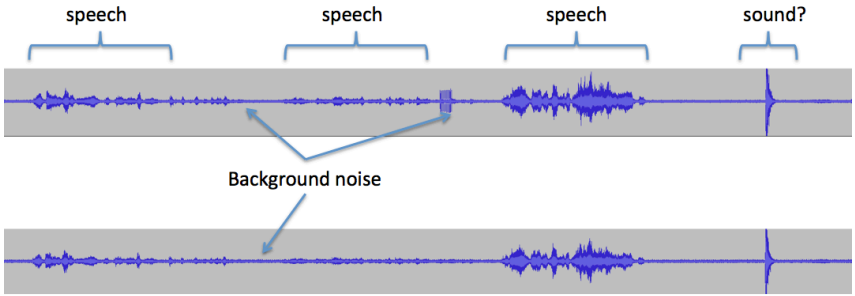
5. Fusion of Audio and Vision

1. **Audio-visual processing challenges**
2. Representation of visual information
3. The geometry of vision
4. Audio-visual feature association
5. Audio-visual alignment
6. Visually-guided audio localization
7. Audio-visual event localization
8. Audio-visual clustering
9. Conclusions

People and Robot



Auditory Data



Audio Only

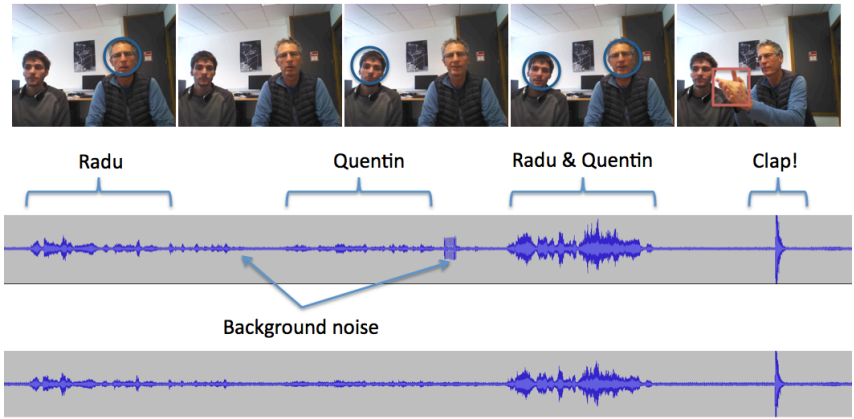
Quentin-Radu-Nao-a-6feb2015.wav

Visual Data

Vision Only

Quentin-Radu-Nao-v-6feb2015.mp4

Audio-Visual Fusion



Audio and Vision

Quentin-Radu-Nao-av-6feb2015.mp4

Audio and Visual Processing Examples

- **Auditory processing**: sound-source localization, sound-source separation, voice activity detection, acoustic-event recognition, etc.
- **Visual processing**: face detection, face recognition, face orientation, hand detection, gesture recognition, etc.

Audio and Vision Side-by-Side (I)

- **Audio challenge**: Identify acoustic sources in the presence of noise and reverberations.
- **Visual challenge**: Identify objects based on reflections of rays of light onto that objects.

Audio and Vision Side-by-Side (II)

Spatial and temporal resolutions:

- **Audio data**: sparse spatial resolution, high temporal resolution (44 000 samples per second).
- **Visual data**: dense spatial resolution (2MP), low temporal resolution (25 frames per second)

Audio and Vision Side-by-Side (III)

- **Visual data**: limited field of view, large variabilities in shape, texture, size, color, etc.
- **Audio data**: acoustic signals (voices, musical instruments, environmental sounds, etc.) are mixed.

Session Summary

- Audio data: sound localization, voice activity detection, etc.
- Visual data: face detection, face recognition, face orientation, etc.
- Audio-visual data have richer content.
- Audio-visual data fusion is challenging.

5. Fusion of Audio and Vision

1. Audio-visual processing challenges
2. **Representation of visual information**
3. The geometry of vision
4. Audio-visual feature association
5. Audio-visual alignment
6. Visually-guided audio localization
7. Audio-visual event localization
8. Audio-visual clustering
9. Conclusions

Visual Information for HRI

- Faces (identity, expression, focus of attention, eye gaze)
- Hand and body motions (actions and gestures)

Example: People engaged in a conversation

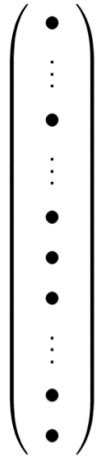
Vision Only

Quentin-Radu-Nao-v-6feb2015.mp4

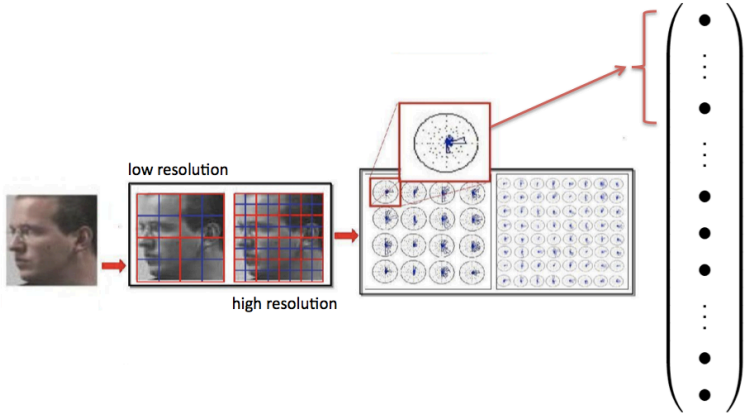
Facial Features

- Face localization
- Facial features (lips, eyes, etc.)
- Eye gaze
- Lip movements
- Head movements
- Head orientation
- Temporal tracking

Visual Feature Vectors

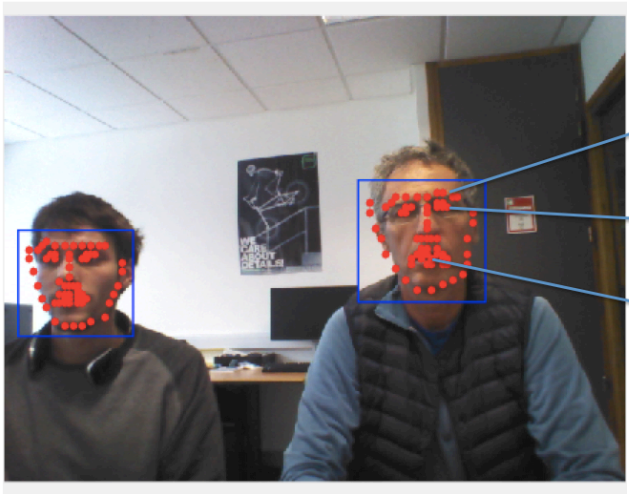


Example of Feature Vector



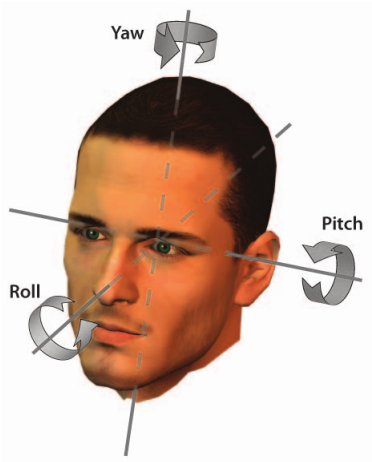
HOG: histogram of oriented gradients

Face Landmarks

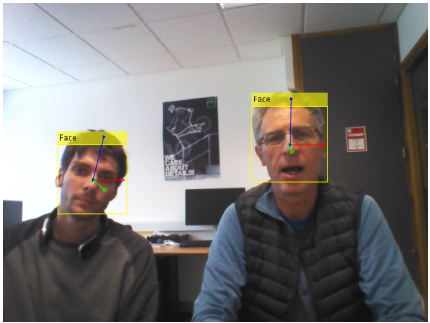


Examples of Face Landmarks

Head Orientation



Examples of Head Orientation



Estimation of head orientation must be robust to variabilities in face appearance and face expressions.

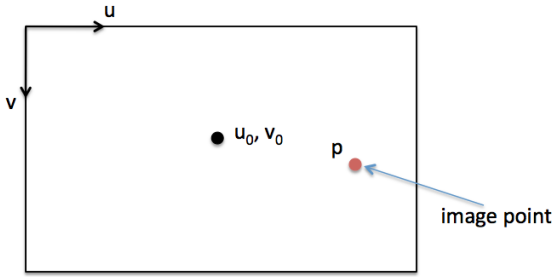
Session Summary

- Faces contain rich visual information
- Face features can be detected and tracked
- Features detectors are not always reliable
- Visual features complement auditory features

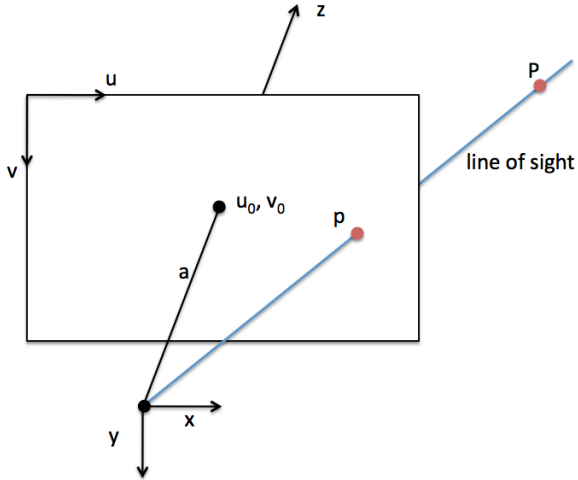
5. Fusion of Audio and Vision

1. Audio-visual processing challenges
2. Representation of visual information
3. **The geometry of vision**
4. Audio-visual feature association
5. Audio-visual alignment
6. Visually-guided audio localization
7. Audio-visual event localization
8. Audio-visual clustering
9. Conclusions

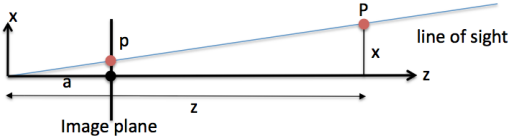
The Image Model



The Camera Model



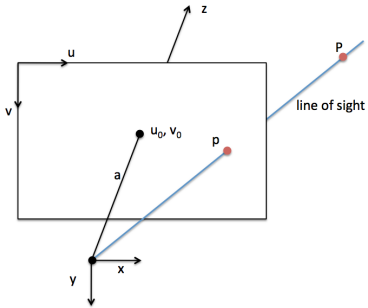
Perspective Camera Model



- Camera parameters: focal length a and image center $(u_0,\ v_0)$
- Coordinates of p in the image plane: $\boldsymbol{p} = (u - u_0,\ v - v_0)^\top$
- Coordinates of P in the camera frame: $\boldsymbol{P} = (x,\ y,\ z)^\top$
- From similar triangles we obtain:

$$\frac{u - u_0}{a} = \frac{x}{z} \quad \text{and} \quad \frac{v - v_0}{a} = \frac{y}{z}$$

Line of Sight

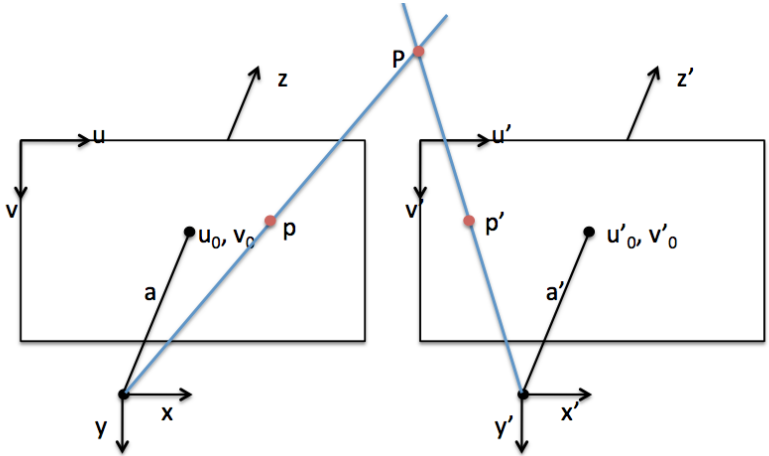


- The line of sight $\mathcal{L}$ through image point p is defined by **camera parameters**, $(a,\ u_0,\ v_0)$:

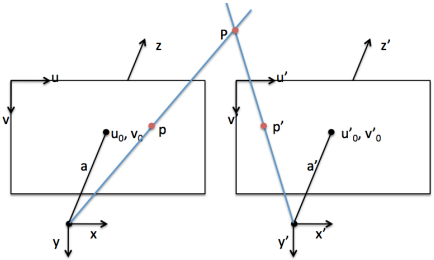
$$P \in \mathcal{L} := \begin{cases} ax - (u - u_0)z = 0 \\ ay - (v - v_0)z = 0 \end{cases}$$

- $(a,\ u_0,\ v_0)$ estimated using calibration

Binocular Camera Pair



Binocular Reconstruction Principle



- Point P is at the intersection of two lines of sight, through p (left) and p' (right)
- Point P in the left camera frame: $P = (x,\ y,\ z)^\top$
- Point P in the right camera frame: $P = (x',\ y',\ z')^\top$
- The two lines of sight must be represented in the same frame (left or right)

Binocular Calibration

- Let $\mathbf{R}$ and $\boldsymbol{t}$ be the rotation matrix (3×3) and translation vector between left and right:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \mathbf{R} \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} + \boldsymbol{t}$$

- $\mathbf{R}$ and $\boldsymbol{t}$ are the parameters of the binocular system, they are estimated by calibration.
- $\mathbf{R} = [r_{ij}]_{i,j=1}^{i,j=3}$, $\boldsymbol{t} = (t_1,\ t_2,\ t_3)^\top$.

Point Reconstruction

- Point P is the intersection of two lines of sight represented in the same coordinate frame (left camera):

$$P := \begin{cases} ax - (u - u_0)z = 0 \\ ay - (v - v_0)z = 0 \\ a'(r_{11}x + r_{12}y + r_{13}z + t_1) - (u' - u'_0)(r_{31}x + r_{32}y + r_{33}z + t_3) \\ a'(r_{21}x + r_{22}y + r_{23}z + t_2) - (u' - u'_0)(r_{31}x + r_{32}y + r_{33}z + t_3) \end{cases}$$

- This enables reconstruction of P from the left/right pair of image points p, p'.

Session Summary

- Image model
- Camera model
- Binocular model
- Camera calibration
- Binocular reconstruction

5. Fusion of Audio and Vision

1. Audio-visual processing challenges
2. Representation of visual information
3. The geometry of vision
4. **Audio-visual feature association**
5. Audio-visual alignment
6. Visually-guided audio localization
7. Audio-visual event localization
8. Audio-visual clustering
9. Conclusions

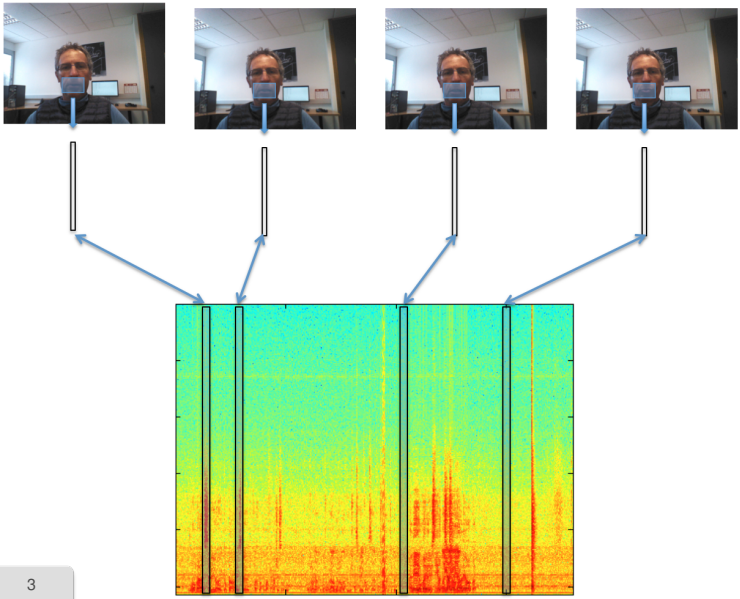
Audio-Visual Features

- **Visual features**: location in the image plane, feature vectors, spatial orientation, etc.
- **Auditory features**: head-related transfer function, binaural features, source directions, voice activity, etc.
- **Challenge**: How to associate these features for more robust sound-source identification (localization, separation, recognition)?

Constrained/Unconstrained Audio-Visual Association

- **Single audio-visual object**: Binaural and visual vector are concatenated into a unique audio-visual vector.
- **Several audio-visual objects**: One-to-one audio-to-visual association is complex:
  - Several visual objects that vary over time
  - Several audio sources that are mixed together

Single Audio-Visual Object



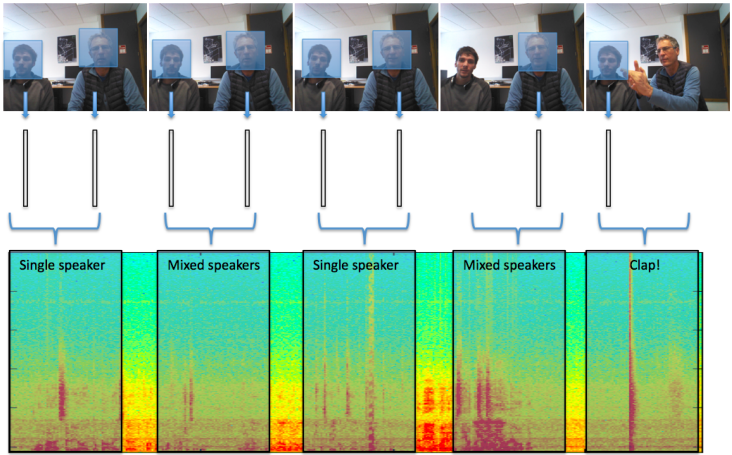
Audio and Visual Features Side-by-Side

- Binaural features (ILPD vectors): $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_l, \ldots, \boldsymbol{y}_L$
- Visual features (HOG vectors): $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_l, \ldots, \boldsymbol{v}_L$
- Audio-visual features: $(\boldsymbol{y}_1, \boldsymbol{v}_1), \ldots, (\boldsymbol{y}_l, \boldsymbol{v}_l), \ldots, (\boldsymbol{y}_L, \boldsymbol{v}_L)$
- Visual features increase the discriminative power of audio features, example: **lip movements for speech recognition**

Multiple Audio-Visual Objects

Simple audio-visual feature association is ineffective!

Multiple Audio-Visual Object Analysis



Session Summary

- Constrained audio-visual analysis
- Unconstrained audio-visual analysis
- Audio-visual feature association
- Multiple audio-visual objects

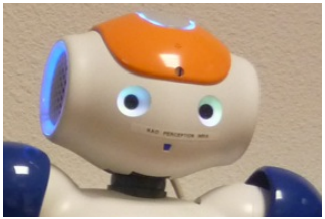
5. Fusion of Audio and Vision

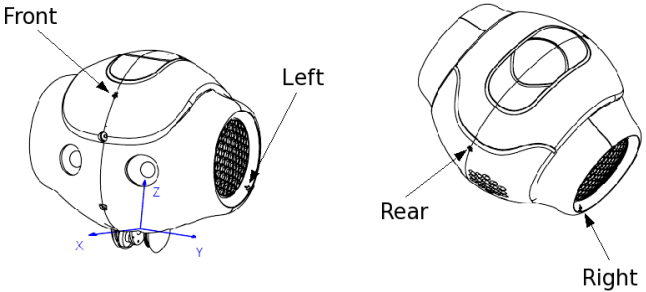
1. Audio-visual processing challenges
2. Representation of visual information
3. The geometry of vision
4. Audio-visual feature association
5. **Audio-visual alignment**
6. Visually-guided audio localization
7. Audio-visual event localization
8. Audio-visual clustering
9. Conclusions

Aligning Audio and Visual Data

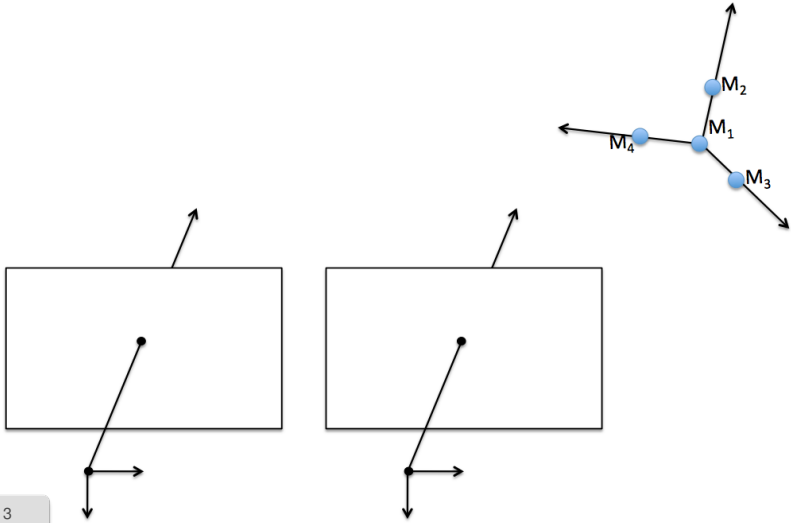
1. Alignment based on microphone-camera geometry
2. Alignment based on learning a mapping between sounds and images

NAO Robot

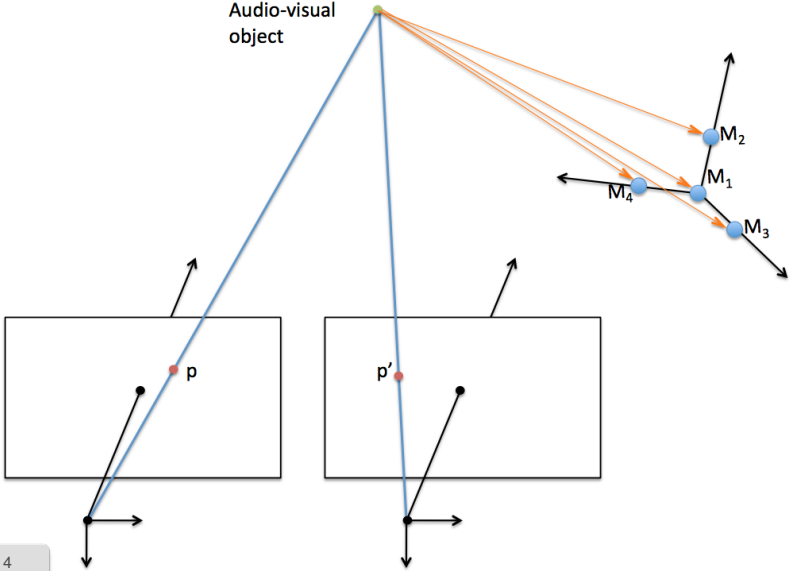




Four Microphones and Two Cameras



Four Microphones and Two Cameras



Calibration Principles (I)

- Visual localization:

  The position of the audio-visual *object* is estimated from the left-image and right-image point-pair (p, p').
  Let $\boldsymbol{S}_v = (x_v, y_v, z_v)$ be the object position in the camera-centered frame.

- Audio localization:

  The position of the audio-visual *object* is estimated from three time-delays $\tau_{1,2}, \tau_{1,3}, \tau_{1,4}$ and from the microphone configuration.
  Let $\boldsymbol{S}_a = (x_a, y_a, z_a)$ be the object position in the microphone-centered frame.

Calibration Principles (II)

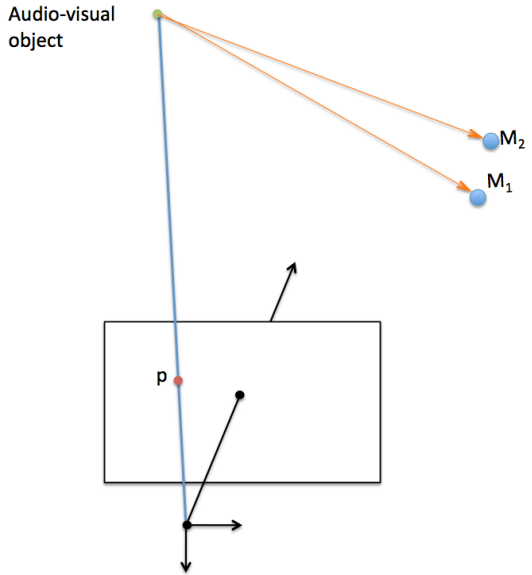
- Let $\mathbf{R}_{av}$ and $\boldsymbol{t}_{av}$ be the unknown rotation matrix and translation vector between the camera-centered and microphone-centered coordinate frames.
- For an audiovisual object $S^{(k)}$, we have:

$$\boldsymbol{s}_a^{(k)} = \mathbf{R}_{av}\boldsymbol{s}_v^{(k)} + \boldsymbol{t}_{av},\ \forall k, 1 \leq k \leq K.$$

- The rotation and translation are estimated by solving these K equations using least squares:

$$(\hat{\mathbf{R}}, \hat{\boldsymbol{t}}) = \underset{\mathbf{R},\boldsymbol{t}}{\text{argmax}} \sum_{k=1}^{K} \|\mathbf{R}_{av}\boldsymbol{s}_v^{(k)} + \boldsymbol{t}_{av} - \boldsymbol{s}_a^{(k)}\|^2$$

Two Microphones and One Camera



Calibration Principles

- The audio-visual object lies along the line of sight defined by the image point p with coordinates (u, v).
- The TDOA $\tau_{i,j}$ between two microphones, M_i and M_j can be estimated using the maximum of the NCC function between the two microphone signals.
- By placing an audio-visual object in several positions, one can learn function $\tau_{i,j} = f_{ij}(u, v)$ using **linear regression**.
- With more than two microphones, it is possible to learn a regression function f_{ij} for each microphone pair.

Mapping a Sound-Source onto the Image Plane

- Consider now a source that emits a sound in front of the robot.
- This sound-source can be mapped onto the image plane in the following way.
- The NCC function can be combined with $\tau_{i,j} = f_{ij}(u, v)$, such that:

$$(\hat{u}, \hat{v}) = \underset{u,v}{\operatorname{argmax}} \operatorname{NCC}\Big(x_i(t)\, x_j(t - f_{ij}(u, v))\Big)$$

Extension To Several Microphones

- This formulation can be extended to an arbitrary number of microphones.
- For each microphone pair a linear regression function f_{ij} is trained.
- The image location of the sound-source is estimated with:

$$(\hat{u}, \hat{v}) = \underset{u,v}{\operatorname{argmax}} \sum_{i \neq j} \mathrm{NCC}\Big(x_i(t)\, x_j\big(t - f_{ij}(u, v)\big)\Big)$$

Session Summary

- Combining stereo reconstruction with audio-source localization
- Calibration procedure for two cameras and four microphones
- Alignment procedure for one camera and two microphones
- Mapping a sound source onto an image

5. Fusion of Audio and Vision

1. Audio-visual processing challenges
2. Representation of visual information
3. The geometry of vision
4. Audio-visual feature association
5. Audio-visual alignment
6. **Visually-guided audio localization**
7. Audio-visual event localization
8. Audio-visual clustering
9. Conclusions

Combining visual and audio data

- The camera-microphone calibration method that we studied allows reconstruct a visual object and to predict its associated TDOA.
- We consider a setup composed of **two cameras** and **two microphones**.
- The microphone positions are known in the camera coordinate system (calibration process).
- This allows to associate a visual object to an audio source.

Associating a TDOA with a Visual Object

- A point $\boldsymbol{S}$ that belongs to a visible object can be reconstructed from its images in the left and right cameras.
- If there is a sound source located at $\boldsymbol{S}$, then one can predict the TDOA associated with a microphone pair:

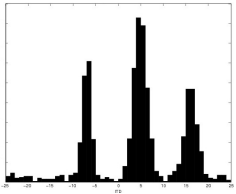
$$\tau(\boldsymbol{S}) = \frac{\|\boldsymbol{S} - \boldsymbol{M}_1\| - \|\boldsymbol{S} - \boldsymbol{M}_2\|}{\nu}$$

- The TDOA of a sound-source can be estimated with:

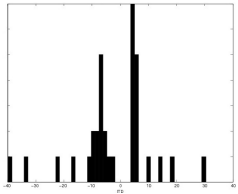
$$\hat{\tau} = \underset{t'}{\operatorname{argmax}}\, \mathrm{NCC}\Big(x_i(t)\, x_j(t - t')\Big)$$

- If $\hat{\tau} \approx \tau(\boldsymbol{S})$, then an audio-visual object may be at $\boldsymbol{S}$.

Mapping People Onto Sounds



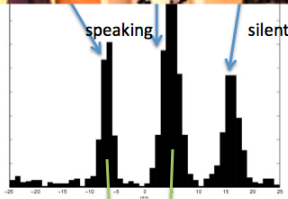
Histogram of visual data



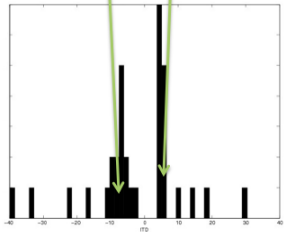
Histogram of estimated TDOAs

Audio-Visual Association

visual data

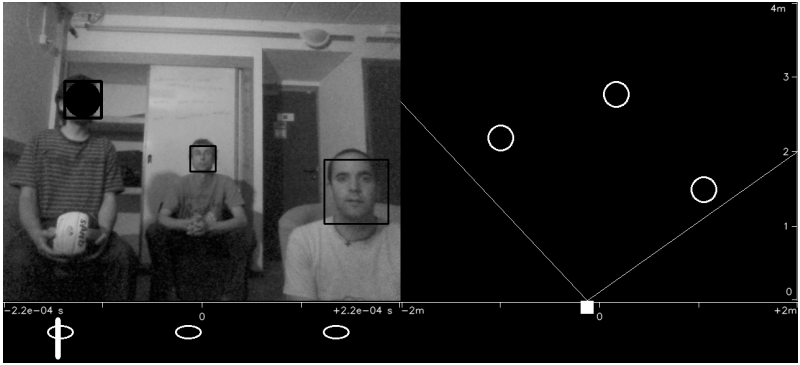


visual-data histogram



TDOA histogram

Faces That Speak



Video

play humanoids_video.avi

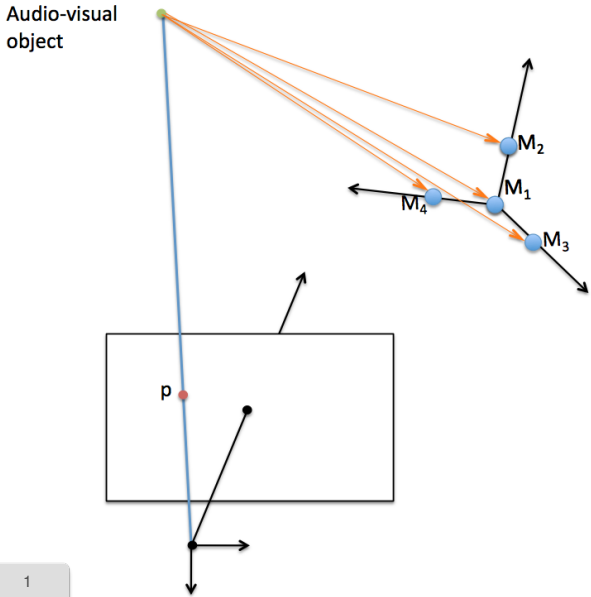
Session Summary

- Two cameras and two microphones
- Estimate the 3D position of a visual object from the two cameras
- Map this 3D position onto the TDOA axis
- Combine the mapped visual object with the estimated TDOA
- associate a person with each TDOA

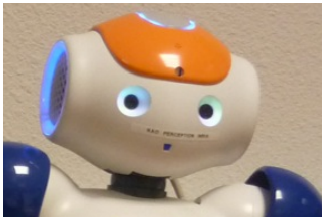
5. Fusion of Audio and Vision

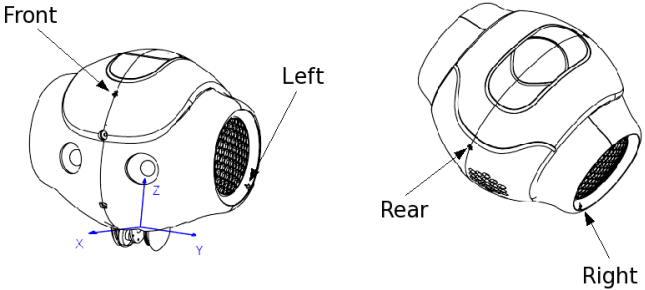
1. Audio-visual processing challenges
2. Representation of visual information
3. The geometry of vision
4. Audio-visual feature association
5. Audio-visual alignment
6. Visually-guided audio localization
7. **Audio-visual event localization**
8. Audio-visual clustering
9. Conclusions

One Camera and Four Microphones



Nao Robot



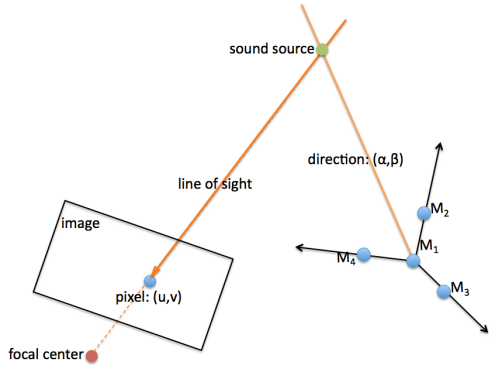


Microphone Configuration

- With four non-coplanar microphones it is possible to estimate both the horizontal (azimuth) and vertical (elevation) direction of a sound source.
- The accuracy of TDOA estimation depends on the between-microphone distances, larger the distance, more accurate the TDOA estimation:

  left-right microphone distance: 12 cm.
  front-rear microphone distance: 9 cm.
- We can learn a sound propagation model for four microphones instead of using an acoustic propagation model.

Projective Camera Model



Audio-Visual Alignment

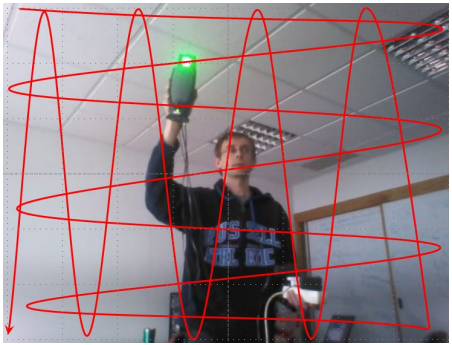
- There a one-to-one mapping between the direction of a sound source and its position in the image plane
- Instead of explicitly calibrating the camera-microphone setup, we can use learning techniques to estimate a regression function that maps a pixel position onto a TDOA (week #3):

$$\tau_{i,j} = f_{ij}(u, v), \ \forall i \neq j$$

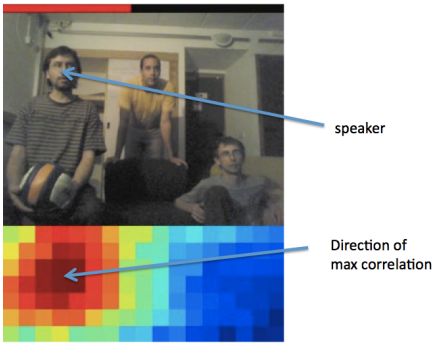
- The sound-source can now be mapped onto the image with:

$$(\hat{u}, \hat{v}) = \underset{u,v}{\operatorname{argmax}} \sum_{i \neq j} \mathrm{NCC}\Big(x_i(t)\, x_j(t - f_{ij}(u, v))\Big)$$

Localization of a Sound in the Image Plane

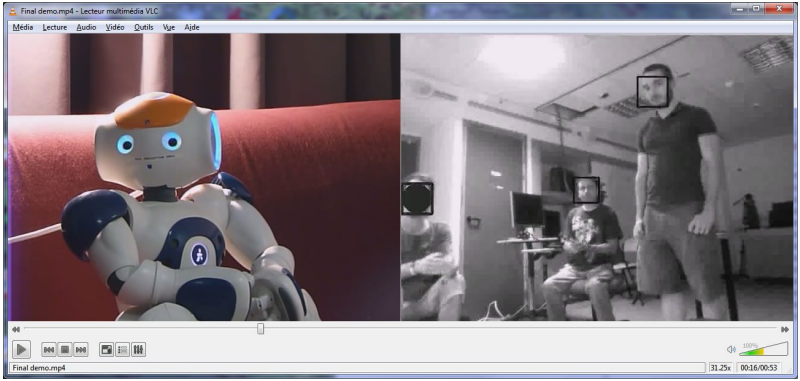


Learning



Prediction

Localizing Speaking Faces with NAO



Video

play final_demo.mp4

Session Summary

- One camera and four microphones
- Camera-microphone calibration not needed
- Sound directions correspond to lines of sight
- Combination of audio and visual features in the image plane

5. Fusion of Audio and Vision

1. Audio-visual processing challenges
2. Representation of visual information
3. The geometry of vision
4. Audio-visual feature association
5. Audio-visual alignment
6. Visually-guided audio localization
7. Audio-visual event localization
8. **Audio-visual clustering**
9. Conclusions

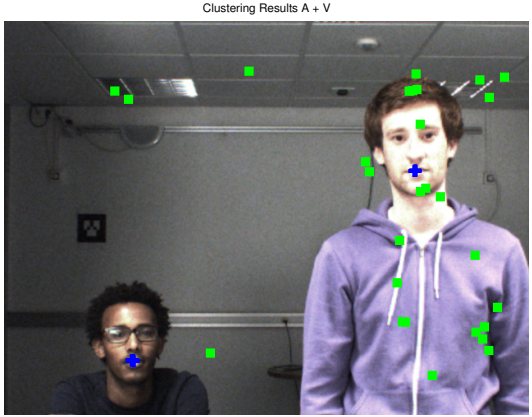
Audio-Event Localization

- An audio event that is recorded with four microphones can be localized in the image plane.
- The cross-correlation function consider a short time interval to estimate the TDOA values.
- Over time, there are many such **audio features** available in the image plane.
- Because of background noise and reverberations, the localization is corrupted by errors.

Visual Feature Localization

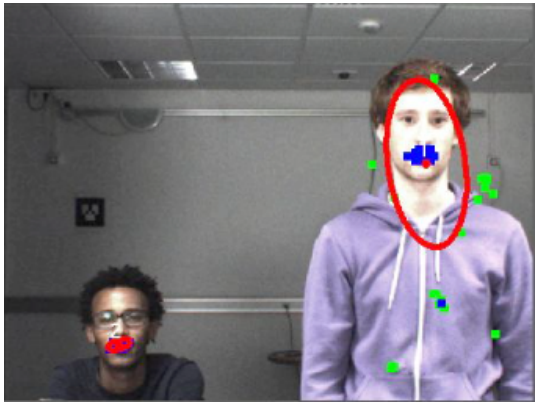
- Face detection and localization
- Face landmarks: lips, eyes, etc.
- In particular, lip detection and localization is quite reliable.

Auditory and Visual Features Side-by-Side



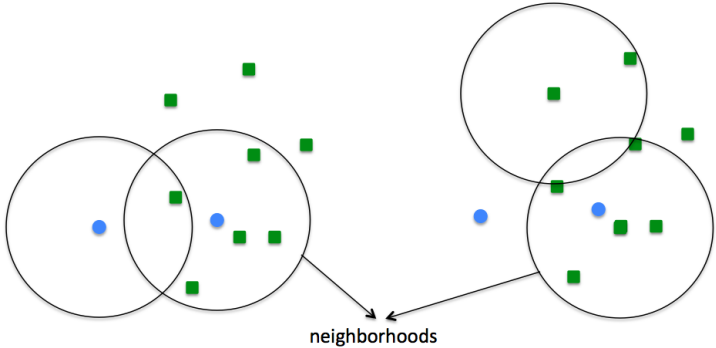
- Visual features (lips): **blue**.
- Audio features (audio events): **green**.

Audio-Visual Clustering



- First cluster contains only visual features (silent person).
- Second cluster contains both visual and audio features (speaking person).

Audio-Visual Weights



- The weight of a visual feature i:

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

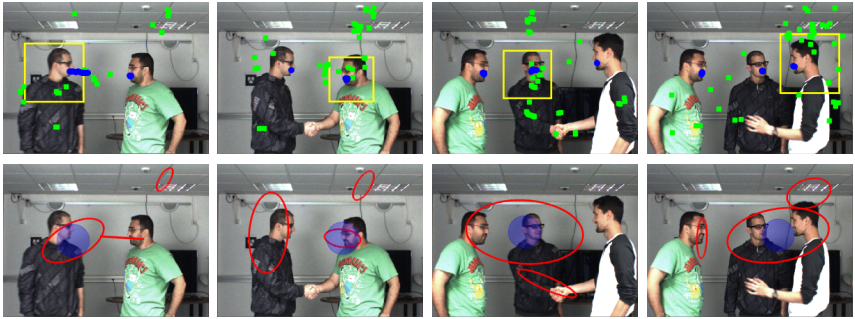
Weighted-Data Gaussian Mixture Model

- Each feature $\boldsymbol{x}_i$ (audio or visual) has a weight w_i and this can be plugged into a GMM:

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

- A weighted-data expectation-maximization algorithm can be used to find **audio-visual clusters**.

Example



Video

Play cocktail-party.mp4

Session Summary

- Audio and visual features in the image plane
- Weighting the features
- Weighted-data Gaussian mixture
- Audio-visual clustering

Week Summary

- Auditory analysis, visual analysis, audio-visual analysis.
- audio-visual feature association.
- Cameras and camera-microphone arrangements.
- Audio-visual alignments.

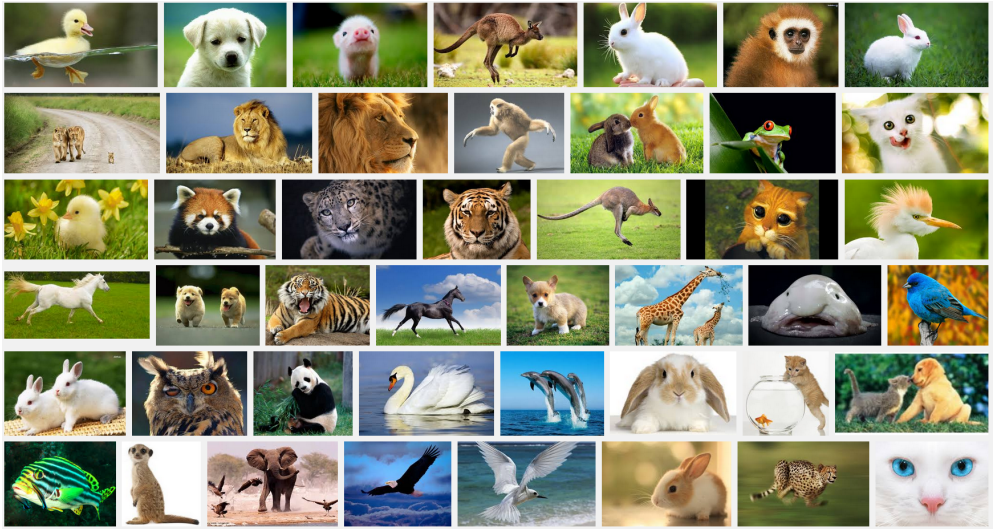
Week Summary (Continued)

- Visually-guided audition.
- Audio-visual event localization.
- Audio-visual clustering.
- Example of solving a complex audio-visual task.

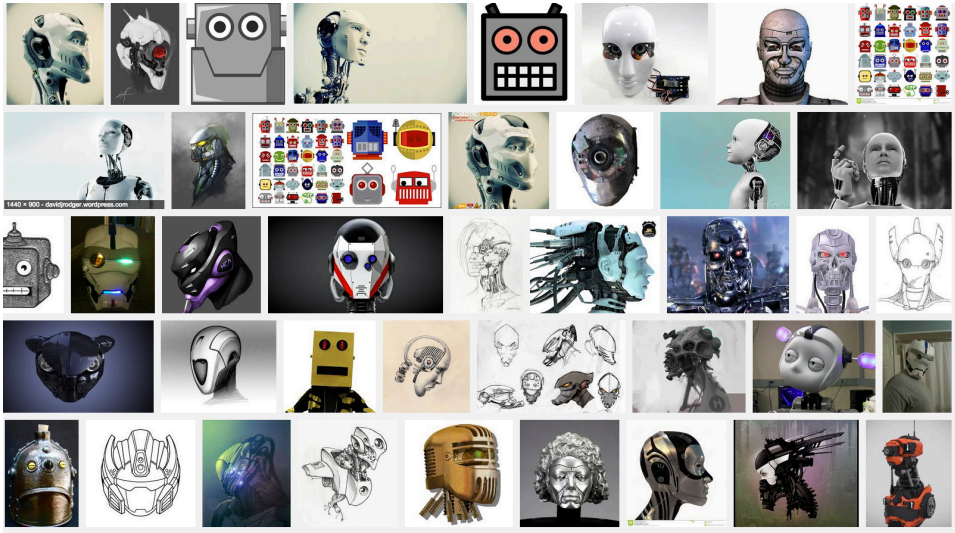
5. Fusion of Audio and Vision

1. Audio-visual processing challenges
2. Representation of visual information
3. The geometry of vision
4. Audio-visual feature association
5. Audio-visual alignment
6. Visually-guided audio localization
7. Audio-visual event localization
8. Audio-visual clustering
9. **Conclusions**

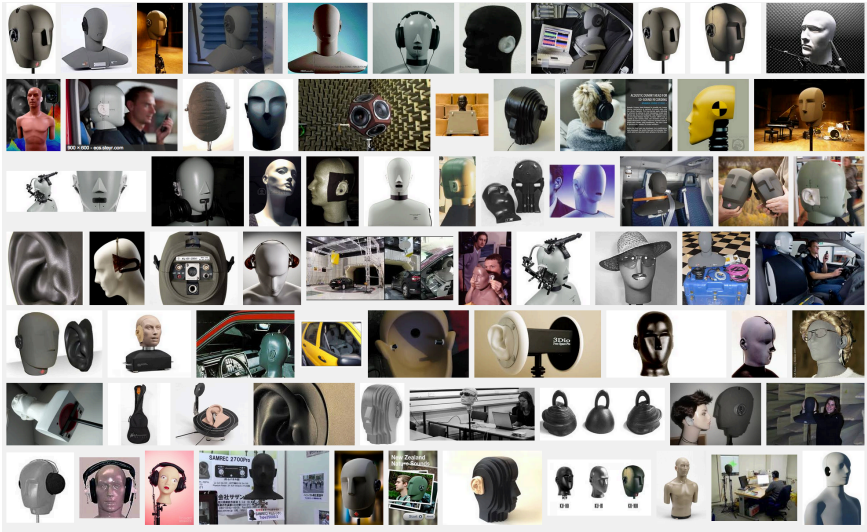
Binaural Hearing in Nature



Robot Heads



Binaural Acoustic Heads



HCI versus HRI

- Hearing allows interaction with highly rich content.
- Human-computer interaction is restricted to a single user.
- Human-robot interaction enables unconstrained multimodal communication between people and robots.

Audio Signal Processing

- Discrete auditory signals
- Discrete short-time Fourier transform
- Audio Spectrograms
- Binaural feature extraction

Sound-Source Localization

- Direct propagation model
- Geometry of two or more microphones
- Implicit propagation models
- Learning a propagation model

Machine Learning for Binaural Hearing

- Binaural features for localization
- Supervised and unsupervised learning methods
- Manifold learning
- Piecewise linear regression
- Supervised sound-source separation and localization

Audio-Visual Analysis

- Audio-visual feature association
- Camera models and Camera-microphone setups
- Audio-visual alignment
- Audio-visual event localization

Acknowledgements

- Xavier Alameda-Pineda,
- Soraya Arias,
- Antoine Deleforge,
- Vincent Drouard,
- Florence Forbes,
- Sharon Gannot,
- Israel-Dejene Gebru,
- Laurent Girin,
- Xiaofei Li,
- Jordi Sanchez-Rieira,
- and many others...

Sponsorship

- EU FP7 project HUMAVIPS
  `http://humavips.inrialpes.fr/`
- EU FP7 project EARS
  `http://robot-ears.eu/`
- ERC Advanced Grant VHIA
  `https://team.inria.fr/perception/vhia/`
- Aldebaran Robotics