Binaural Hearing for Robots

Fusion of Audio and Vision

Radu Horaud

Informatique Mathématiques
Inria

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

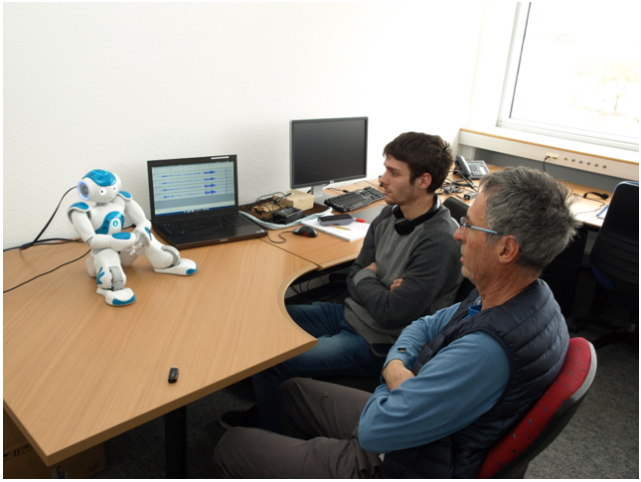
5. Fusion of Audio and Vision

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

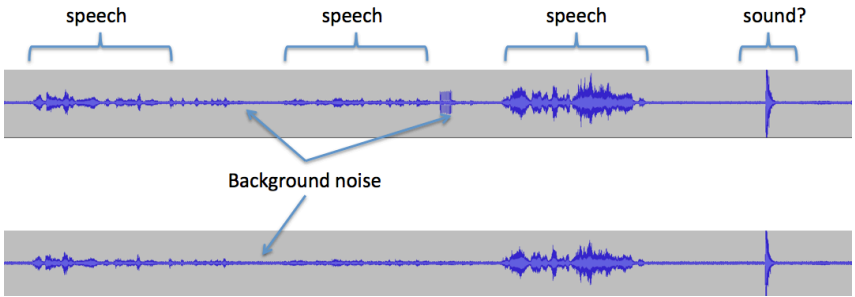
5. Fusion of Audio and Vision

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

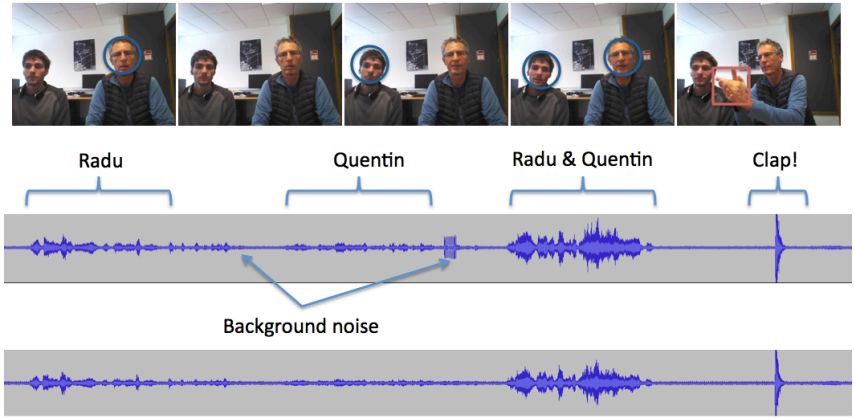
Quentin-Radu-Nao-a-6feb2015.wav

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.