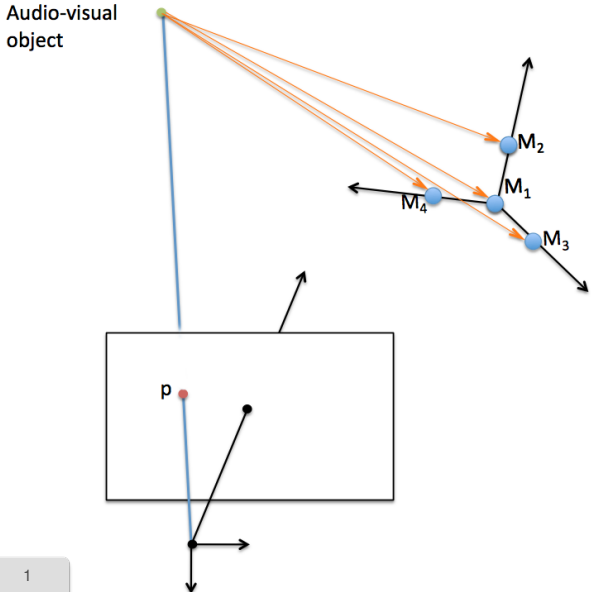
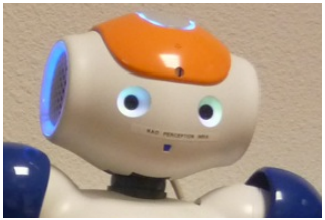
5. Fusion of Audio and Vision

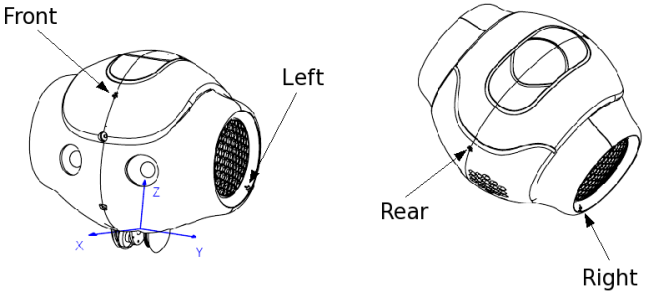
1. Audio-visual processing challenges
2. Representation of visual information
3. The geometry of vision
4. Audio-visual feature association
5. Audio-visual alignment
6. Visually-guided audio localization
7. **Audio-visual event localization**
8. Audio-visual clustering
9. Conclusions

One Camera and Four Microphones



Nao Robot



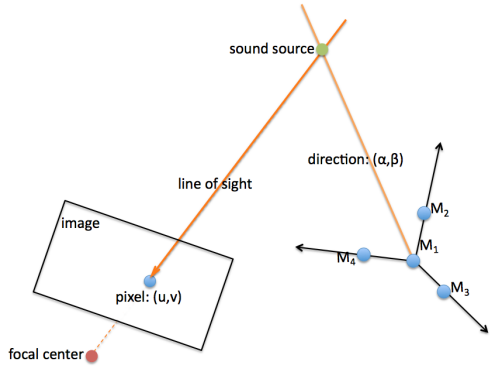


Microphone Configuration

- 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

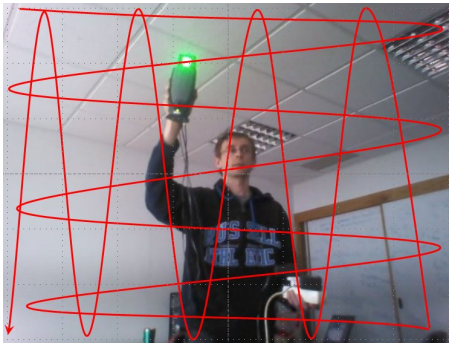
- There a one-to-one mapping between the direction of a sound source and its position in the image plane
- Instead of explicitly calibrating the camera-microphone setup, we can use learning techniques to estimate a regression function that maps a pixel position onto a TDOA (week #3):

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

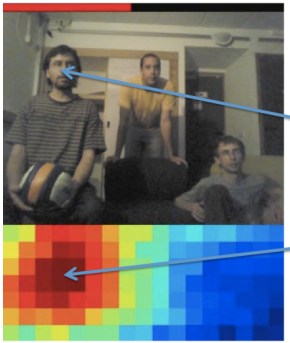
- The sound-source can now be mapped onto the image with:

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

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