3D Localization of Audio-Visual Events with Binaural Hearing and Binocular Vision

Contributors:
Vassil Khalidov, Florence Forbes, Miles Hansard,
Elise Arnaud, Radu Horaud
INRIA Grenoble, France
Heidi Christensen, Yan-Chen Lu, Martin Cooke
University of Sheffield, UK

Work done under the POP project:
http://perception.inrialpes.fr/POP/

Audio-visual integration

- Steps towards audio-visual perception: How can we recognize an object that is both seen and heard?
- Strong & unambiguous audio OR visual stimuli do not require cross-modal integration
- In most natural situations it is difficult to extract unambiguous information from a single modality
- Visual data is dense, light sources are not relevant
- Auditory data is sparse, acoustic sources must be detected in the presence of reverberations

Approach

Related publications: Cooke at al. ISAAR'07, Christensen et al. Interspeech'07, Khalidov et al. MLMI'08, Khalidov et al. ICMI'O8, Arnaud et al. ICMI'08.

- Exploit the geometry associated with both binocular and binaural observations
- Cast the audio-visual localization problem in the framework of maximum likelihood with missing data:
- Maximize the expected complete-data log-likelihood:

$$E[\log P(\underbrace{\text{audio}, \text{vision}}_{\text{observed data}}, \underbrace{\text{audio-visual objects}}_{\text{missing data}})]$$

The CAVA corpus

CAVA: Computational audio-visual analysis

`http://perception.inrialpes.fr/CAVA_Dataset/`

Why a new audio-visual database?

The CAVA corpus has been designed to address the following question: How to perform an audio-visual analysis of what a person would hear and see while being in a natural environment, and moving the head naturally ?

Goal :

- gather **synchronized auditory and visual** dataset
- record the data form the **perspective of a person**

CAVA examples on meeting situations

fixed perceiver

(Loading M1.mov)

moving perceiver

(Loading M2.mov)

More CAVA examples

active hearing

(Loading AH1.mov)

moving sources

(Loading DCMS1.mov)

Experimental setup

Audio-visual equipment

- Stereoscopic camera pair fixed on a helmet
- Binaural microphone pair
- 6dof tracking device "Cyclope"
  `http://www.inrialpes.fr/sed/6doftracker/`
- Acquisition and synchronization software

Data recorded

- Stereo image pairs at 25 frames per second
- Binaural auditory signals at 44.1kHz
- Head position and orientation at 25 frames per second

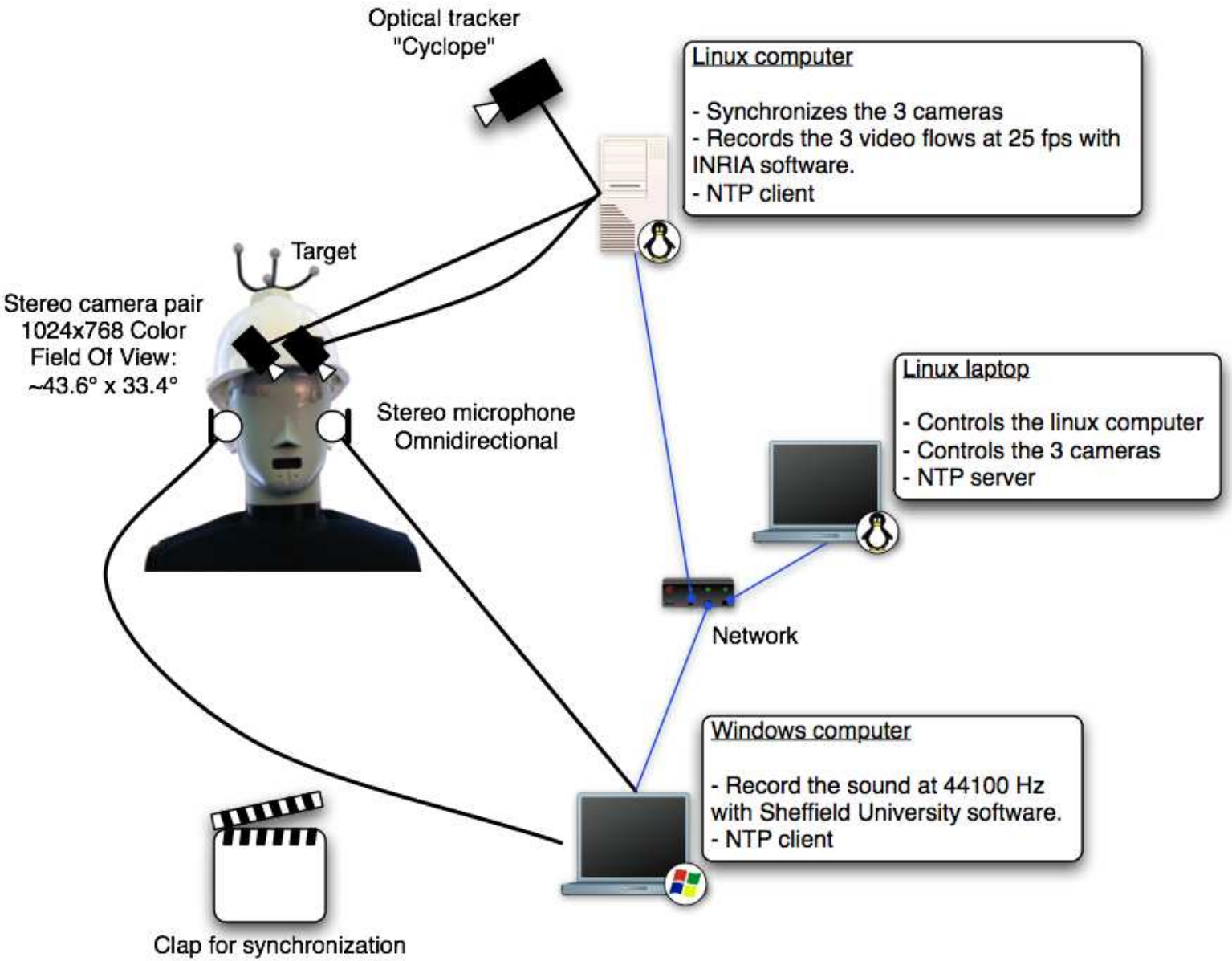
Experimental setup ... details







Experimental setup ... more details



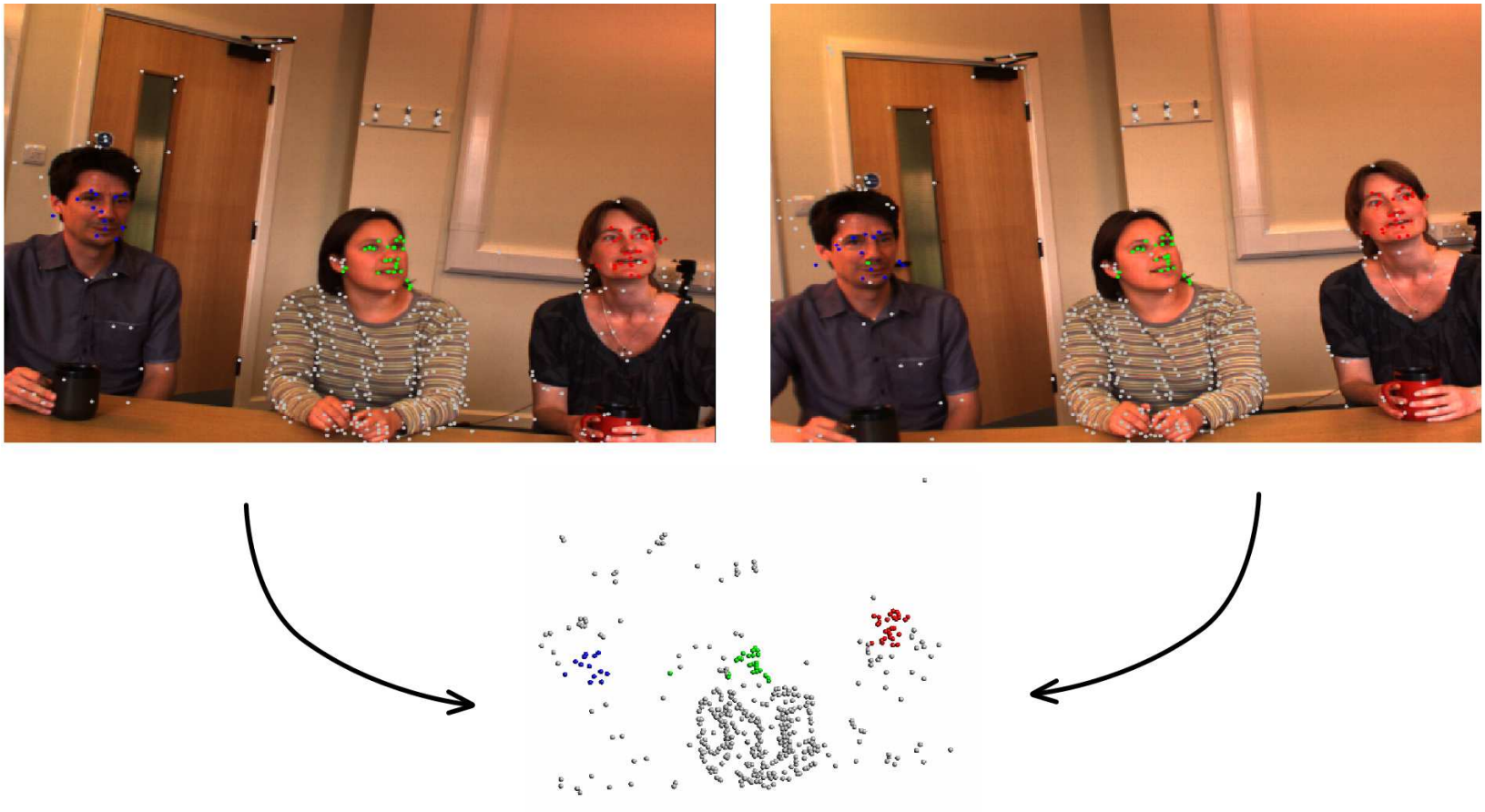
Scenarios

- Fixed perceiver observing speakers walking in and out of field of view, multiple & simultaneous sound sources.
- Panning perceiver: **obtain recordings of controlled cues from an actively moving head**.
- Moving perceiver: provide challenging AV situations that **mimic situations that can appear in a real-life environment**

Binocular observations (1)



Binocular observations (2)

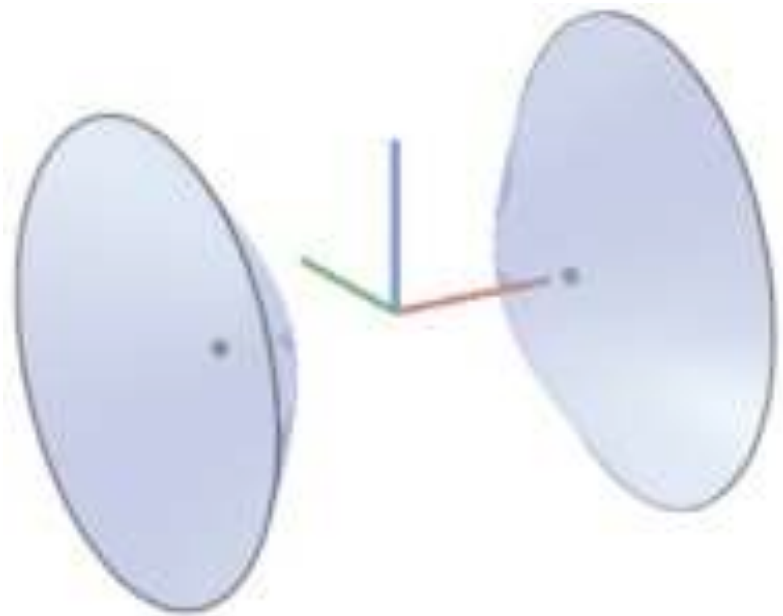


Binocular observations (3)

- $\mathbf{F} = \{\boldsymbol{F}_1, \ldots, \boldsymbol{F}_m, \ldots, \boldsymbol{F}_M\} \in$ 3-D
- $\boldsymbol{F}_m = (u, v, d)$: $(u, v) =$ pixel, $d =$ disparity
- $s = (x, y, z)$: 3-D coordinates of an audio-visual object (speaker, ...)
- 3-D visual disparity is linked to 3-D coordinates : $\boldsymbol{F}_m = (u, v, d) = f(\boldsymbol{s}) = \left(\frac{x}{z}, \frac{y}{z}, \frac{b}{z}\right)$

Binaural observations

- $\boldsymbol{G} = \{G_1, \ldots, G_k, \ldots, G_K\} \in$ 1-D
- $G_k =$ interaural time difference (ITD)
- 1-D ITD:
  $G_k = g(\boldsymbol{s}) = \frac{1}{c}\left(\|\boldsymbol{s} - \boldsymbol{s}_{M_1}\| - \|\boldsymbol{s} - \boldsymbol{s}_{M_2}\|\right)$

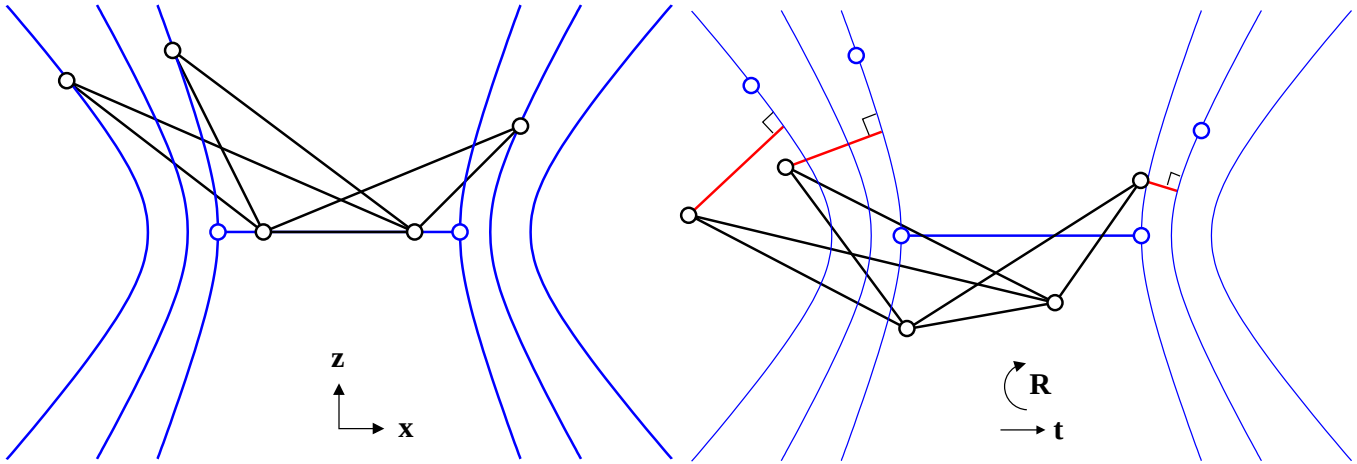


The geometry of 3-D audio-visual fusion

- $s = (x, y, z)$: 3D coordinates of an audio-visual object (speaker, ...)
- 1-D auditory disparity (ITD):
  $G_k = g(\boldsymbol{s}) = \frac{1}{c}\left(\|\boldsymbol{s} - \boldsymbol{s}_{M_1}\| - \|\boldsymbol{s} - \boldsymbol{s}_{M_2}\|\right)$
- 3-D visual disparity (projective space):
  $\boldsymbol{F}_m = (u, v, d) = f(\boldsymbol{s}) = \left(\frac{x}{z}, \frac{y}{z}, \frac{b}{z}\right)$
- Auditory and visual data are put on an equal footing
- These equations can be used for camera-microphone calibration, i.e., determine where are the microphones (M_1 & M_2) in stereo-sensor space.

Audio-Visual Calibration

- Inter-aural and inter-ocular axes will not be exactly aligned.
- Suppose that the locations of N sound-sources are reconstructed in each modality.
- This gives a N points (binocular) and N hyperboloids (binaural).
- The binocular/binaural transformation could be estimated by minimizing the distance of each point to the corresponding hyperboloid.
- Transformation could be Euclidean, affine, projective. . .



Multiple-speaker model

- $\boldsymbol{s} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n, \ldots, \boldsymbol{s}_N\} \in$ 3-D
- $\boldsymbol{s}_n$: 3D coordinates of speaker n
- Two sets of assignment variables (the hidden data):
  $\boldsymbol{A} = \{A_1, ..., A_M\}$ for video, and
  $\boldsymbol{A}' = \{A'_1, ..., A'_K\}$ for audio
- The notation $A_m = n$, $n = 1, \ldots, N, N+1$ means that the visual observation m is assigned to speaker n or is is assigned to an outlier class $N+1$.
- Similarly, $A'_k = n$, $n = 1, \ldots, N, N+1$ means that the audio observation k is assigned to speaker n or is is assigned to an outlier class $N+1$.
- Both $\boldsymbol{A}$ and $\boldsymbol{A}'$ are missing: the *observation-to-speaker assignments* are unknowns.

Likelihood model

- The observed data $\boldsymbol{F}$ and $\boldsymbol{G}$ are considered to be realizations of the random variables $\boldsymbol{f}$ and $\boldsymbol{g}$.
- Probability of an observation to belong to a speaker (inlier):
  - $P(\boldsymbol{f}_m|A_m = n) = \mathcal{N}\left(\boldsymbol{f}_m|f(\boldsymbol{s}_n), \Sigma_n\right)$
  - $P(g_k|A'_k = n) = \mathcal{N}\left(g_k|g(\boldsymbol{s}_n), \sigma_n^2\right)$
  - Alternatively, we can also use the t-distribution.
  - The methodology is independent of the normal/t-distribution choice.
- Likelihood of an observation to be an outlier:
  - $P(\boldsymbol{f}_m|A_m = N+1) = \mathcal{U}\left(V_{3D}\right) = \frac{1}{V}$
  - $P(g_k|A'_k = N+1) = \mathcal{U}\left(U_{1D}\right) = \frac{1}{U}$
- The model's parameters:

$$\boldsymbol{\theta} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N, \Sigma_1, \ldots, \Sigma_N, \sigma_1, \ldots, \sigma_N\}$$

The observed-data log-likelihood

We assume that the variables are independent and identically distributed, then:

$$\log P(\boldsymbol{f}, \boldsymbol{g}) = \sum_{n=1}^{N+1} \log \left(\pi_n \sum_{m=1}^{M} P(\boldsymbol{f}_m | A_m = n) + \pi'_n \sum_{k=1}^{K} P(\boldsymbol{g}_k | A'_k = n) \right)$$

where π_n and π'_n are the prior probabilities.

The maximization of the log-likelihood is not tractable because of the presence of the hidden variables.

Maximum likelihood with hidden data

- $\max E_{\boldsymbol{a},\boldsymbol{a}'}[(\log P(\mathbf{f},\boldsymbol{g},\boldsymbol{a},\boldsymbol{a}'))|\mathbf{f},\boldsymbol{g}]$ is the expected complete-data log-likelihood
- complete data: observed and hidden data
- The expectation-maximizaton (EM) algorithm maximizes this function
- There are several version of EM, such as expectation conditional maximization (ECM) and generalized expectation maximization (GEM).
- These algorithms (EM, ECM, GEM) converge to a local minimum of the observed-data log-likelihood

Posterior probabilities

α_{mn} and α'_{kn} are the posterior probabilities of a visual observation m and of an auditory observation k to belong to the speaker n (or to be an outlier), conditioned by the observations:

$$\alpha_{mn} = P(A_m = n|\boldsymbol{f}_m) = \frac{\pi_n P(\boldsymbol{f}_m|A_m = n)}{\sum_{i=1}^{N+1} \pi_i P(\boldsymbol{f}_m|A_m = i)}$$

and

$$\alpha'_{mn} = P(A'_k = n|\boldsymbol{g}_k) = \frac{\pi'_n P(\boldsymbol{g}_k|A'_k = n)}{\sum_{i=1}^{N+1} \pi'_i P(\boldsymbol{g}_k|A'_k = i)}$$

Maximizing the expectation

- $\max_{\boldsymbol{\theta}}(E[\log P(\mathbf{f}, \boldsymbol{g}, \boldsymbol{a}, \boldsymbol{a}')|\mathbf{f}, \boldsymbol{g}])$

$$
\begin{aligned}
& E[\log P(\mathbf{f}, \boldsymbol{g}, \boldsymbol{a}, \boldsymbol{a}')|\mathbf{f}, \boldsymbol{g}] = \\
- \quad & \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} \alpha_{mn} \left(\boldsymbol{f}_m - f(\boldsymbol{s}_n)\right)^T \Sigma_n^{-1} \left(\boldsymbol{f}_m - f(\boldsymbol{s}_n) + \log|\Sigma_n|\right) \\
- \quad & \frac{1}{2} \sum_{k=1}^{K} \sum_{n=1}^{N} \alpha'_{kn} \left((g_k - g(\boldsymbol{s}_n))^2/\sigma_n^2 + \log \sigma_n^2\right) \\
+ \quad & \text{other constant terms}
\end{aligned}
$$

The Generalized EM algorithm (GEM)

E-step: Compute the posterior probabilities α_{mn} and α'_{kn} conditioned by the current values of the parameter set $\boldsymbol{\theta}$;

M-step: Maximize (minimize) the expected complete-data log-likelihood over $\boldsymbol{\theta}$ conditioned by the current values of the posteriors. This involves non-linear minimization (Newton-Raphson) instead of a closed-form solution as in the standard EM algorithm:

$$\boldsymbol{\theta}^{(q)} = \boldsymbol{\theta}^{(q-1)} + \gamma^{(q)} \Gamma^{(q)} \frac{dE}{d\boldsymbol{\theta}}$$

Implementation

- Initialization of the number of clusters and the cluster centers (currently we use a face detector)
- Each sequence of input data is split into time-intervals, each interval is of 1/8 seconds
- There are 1000 visual observations and 10 auditory observations per time-interval

	time-int	*AV-int*	*AV-OK*	*AV-missed*	*AV-false*
M1	166	89	75	0.16	0.14
TTOS1	76	69	60	0.13	0.43

Example: five speakers, three visible

(Loading M1.mov)

Example 1 (two speakers)

(Loading frame1440-3.avi)

Example 2 (one speaker)

(Loading frame1450-5.avi)

Popeye: an audiovisual head

(Loading frame popeye-sheffield.mpg)

Conclusions

- We introduced a framework for audio-visual fusion in the spatial (3D) domain.
- The method is based on unsupervised clustering using the GEM algorithm
- We plan to extend the method to deal a varying number of AV objects
- Binocular/binaural perception allows strong links with neurophysiology
- A more sophisticated model should include eye and head motions, i.e., active perception and attention.