Visual Attention Priming Based on Crossmodal Expectations

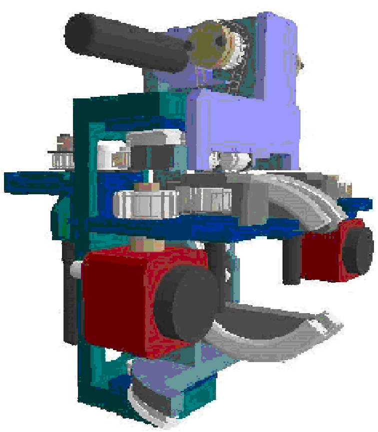
Carlos Beltrán-González
Giulio Sandini
Laboratory for Integrated Advanced Robotics
University of Genoa
Italy

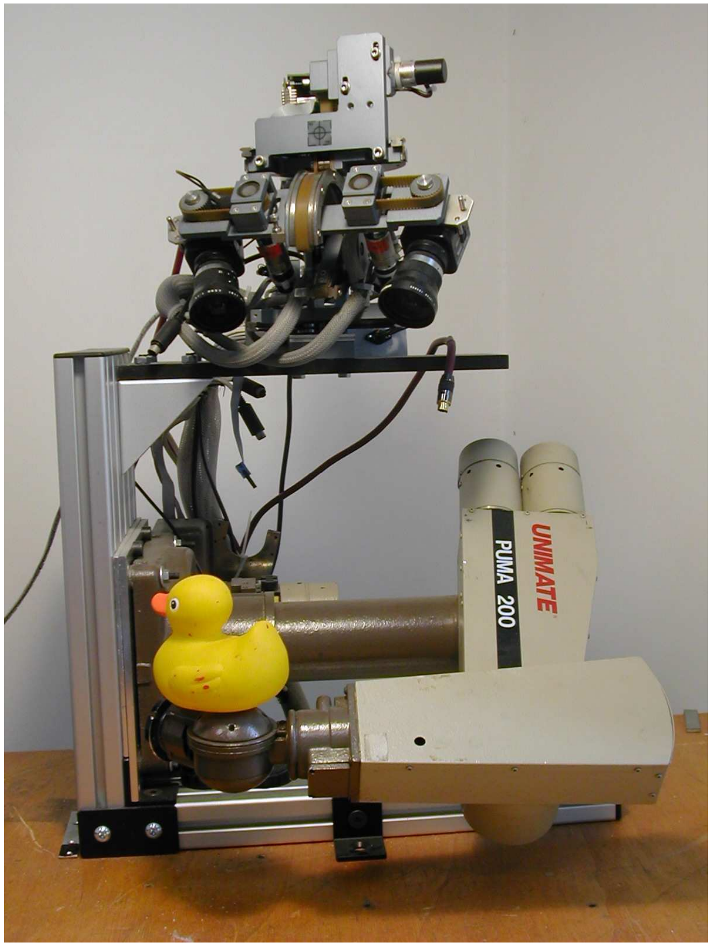
Introduction

- Group research:
  - Develomental/Epigenetics humanoids robotics.
  - Robots that interact with the world (manipulation, stereo vision, sound perception)
- My research:
  - How predictive and expectation mechanisms can help all the above
- This paper:
  - Crossmodal expectations. Sensors eliciting expectations in other sensors. **Goal: improve robot perception**

The experimental setup

- Eurobot
  - Upper torso humanoid robot
  - 10 degrees of freedom
    - 6 in the puma arm
    - 4 in the head
  - YARP/QNX architecture



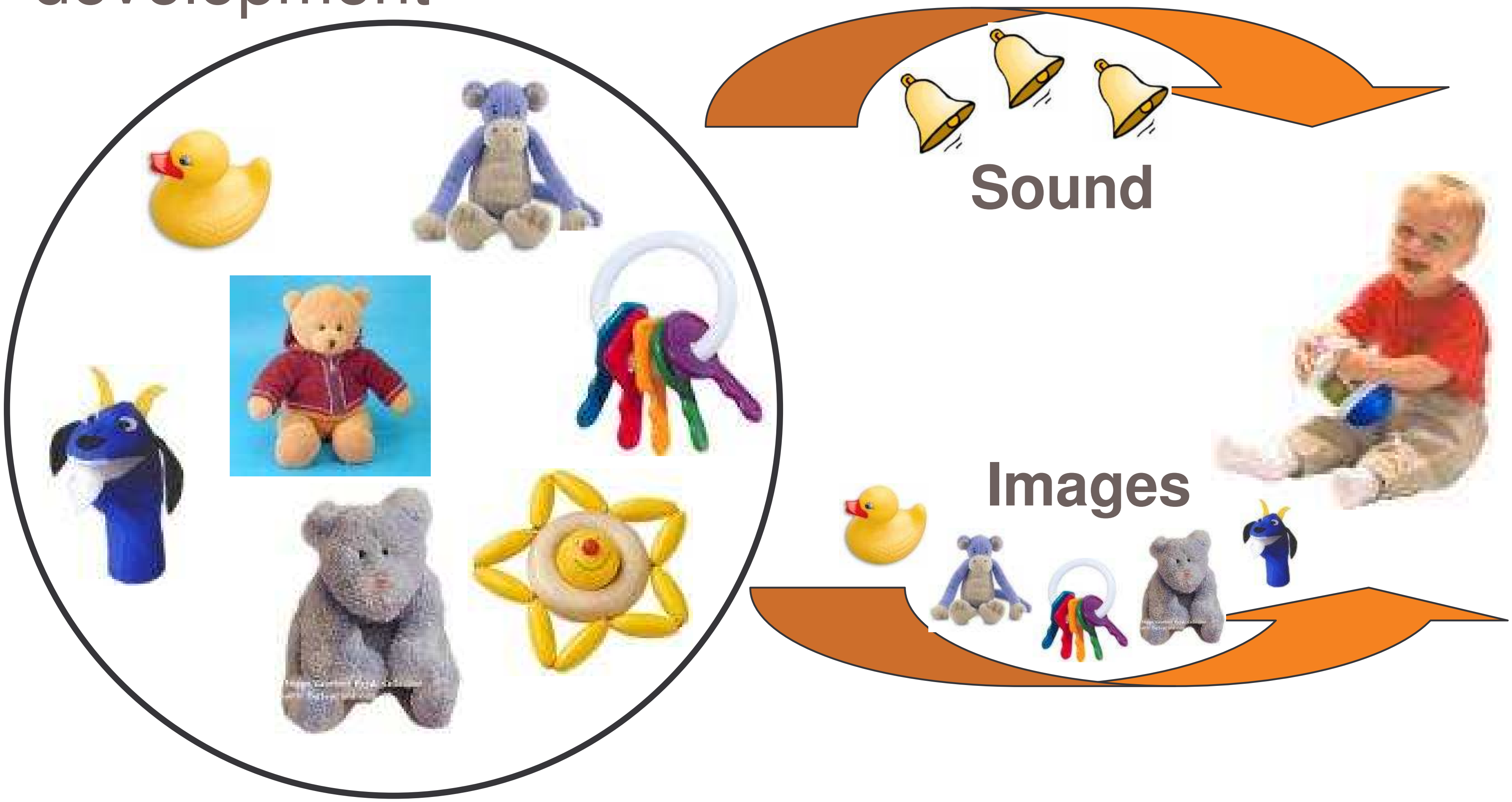


Question

- Can we have the robot to segment noisy objects learning audio-visual associations and use only sound to create visual expectations?

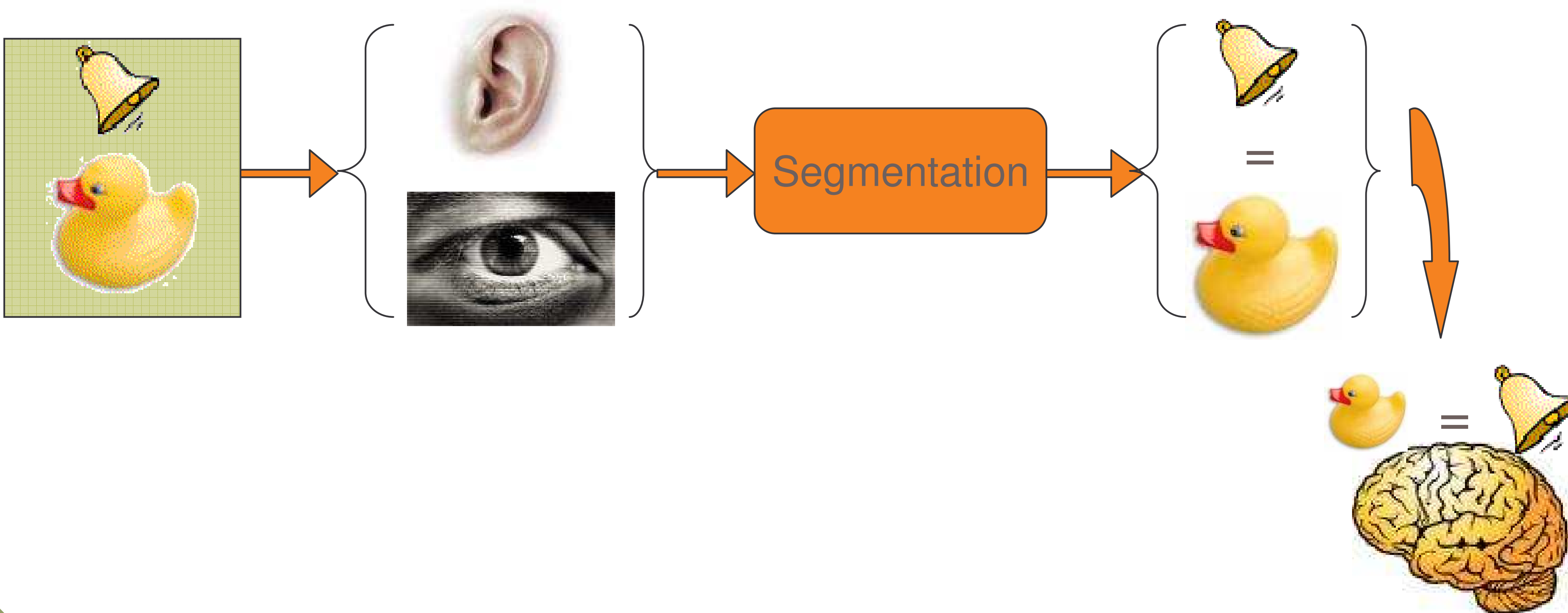
Motivation

- The importance of audio-visual cues in early development



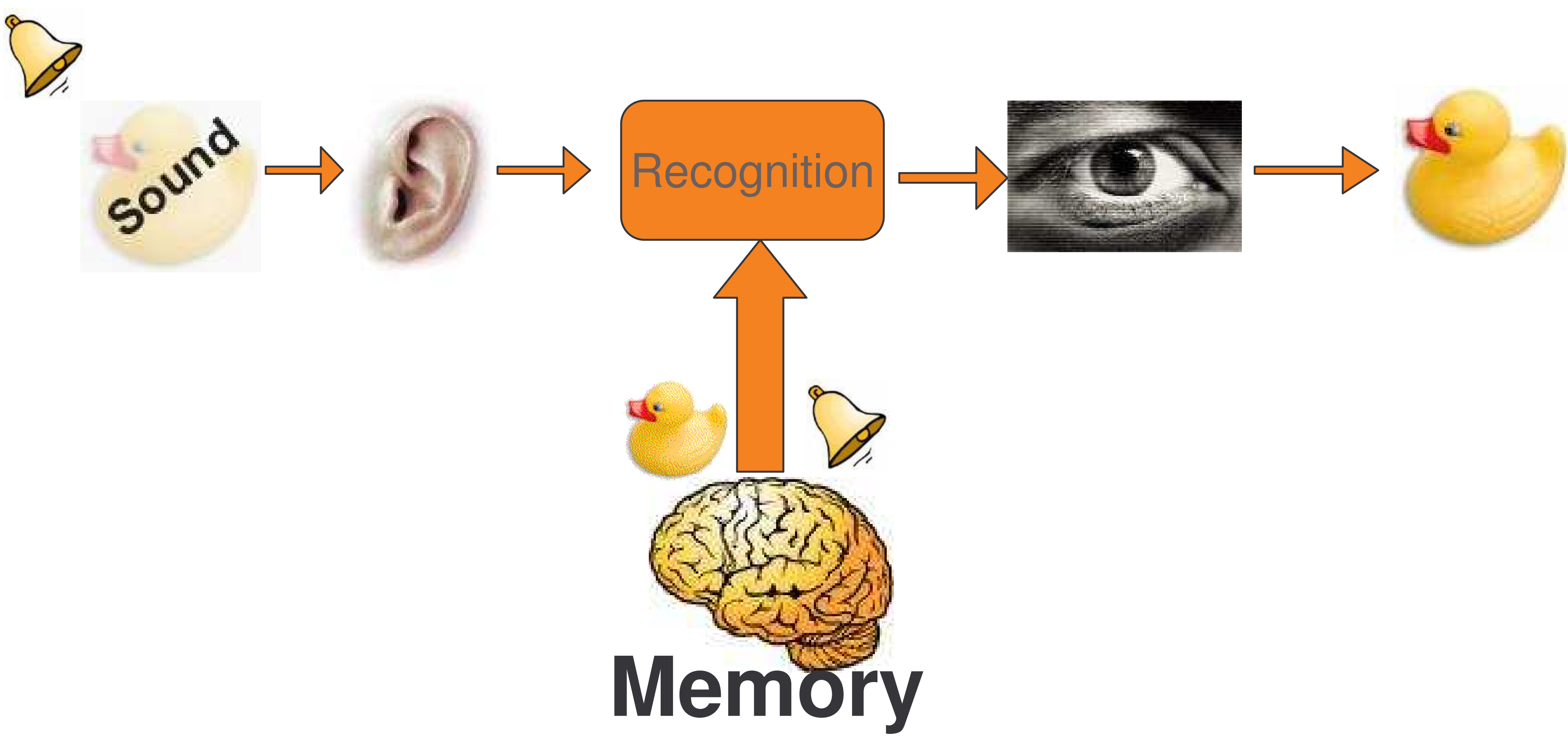
Phase 1 (segmentation)

1) How to create the **audio-visual association**



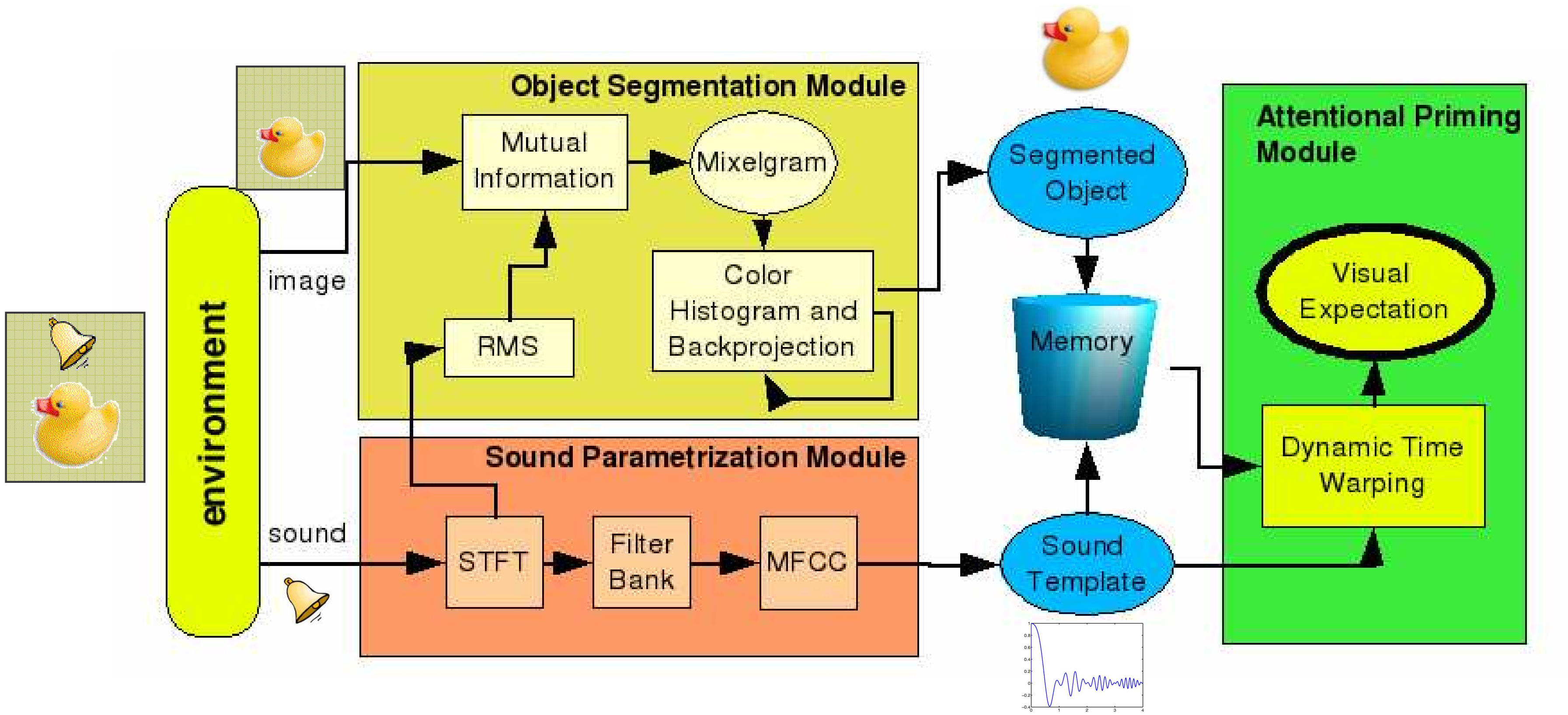
Phase 2 (recognizing)

2) How to generate a **visual expectations** from sound information

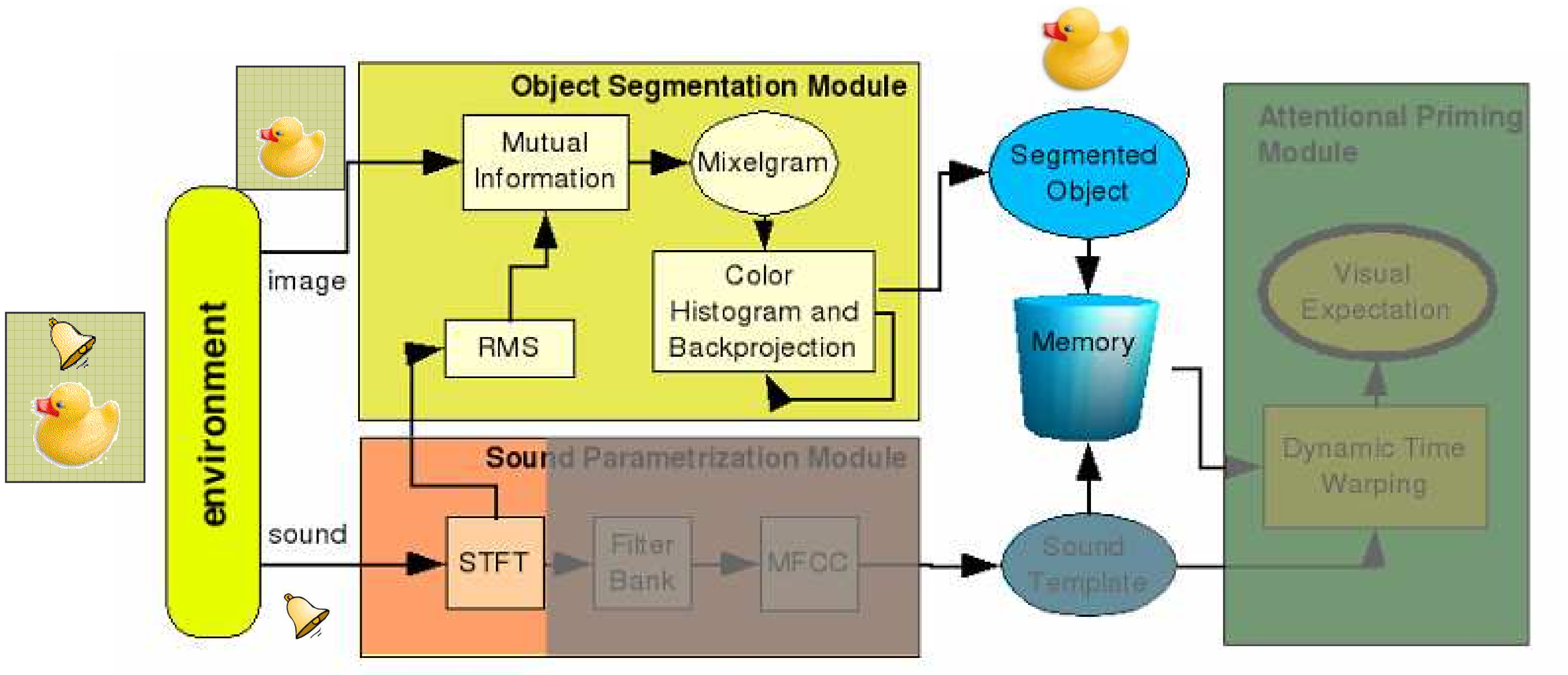


Attention Modulation based on Crossmodal Expectations

Software Architecture



Object Segmentation



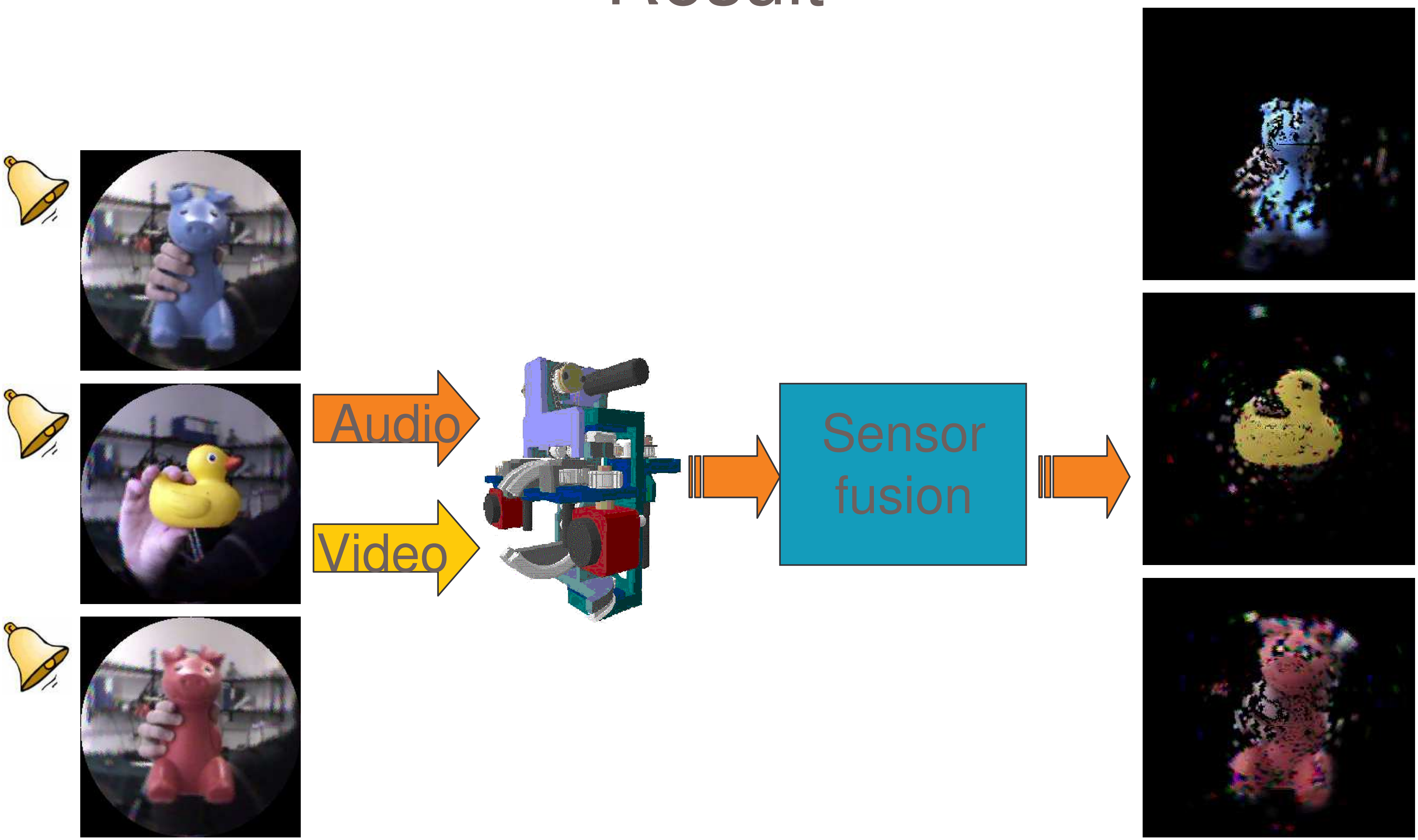
Mutual information
Hershey & Movellan (2000)

- **Problem: which pixels in the image are generating the sound??**

- Sensor integration
- Computes mutual information between **two sensory** channels over a time window (length *S*)
- Assumes Gaussian distributed sensory signals
- Synchrony defined as mutual-information between sensory channels

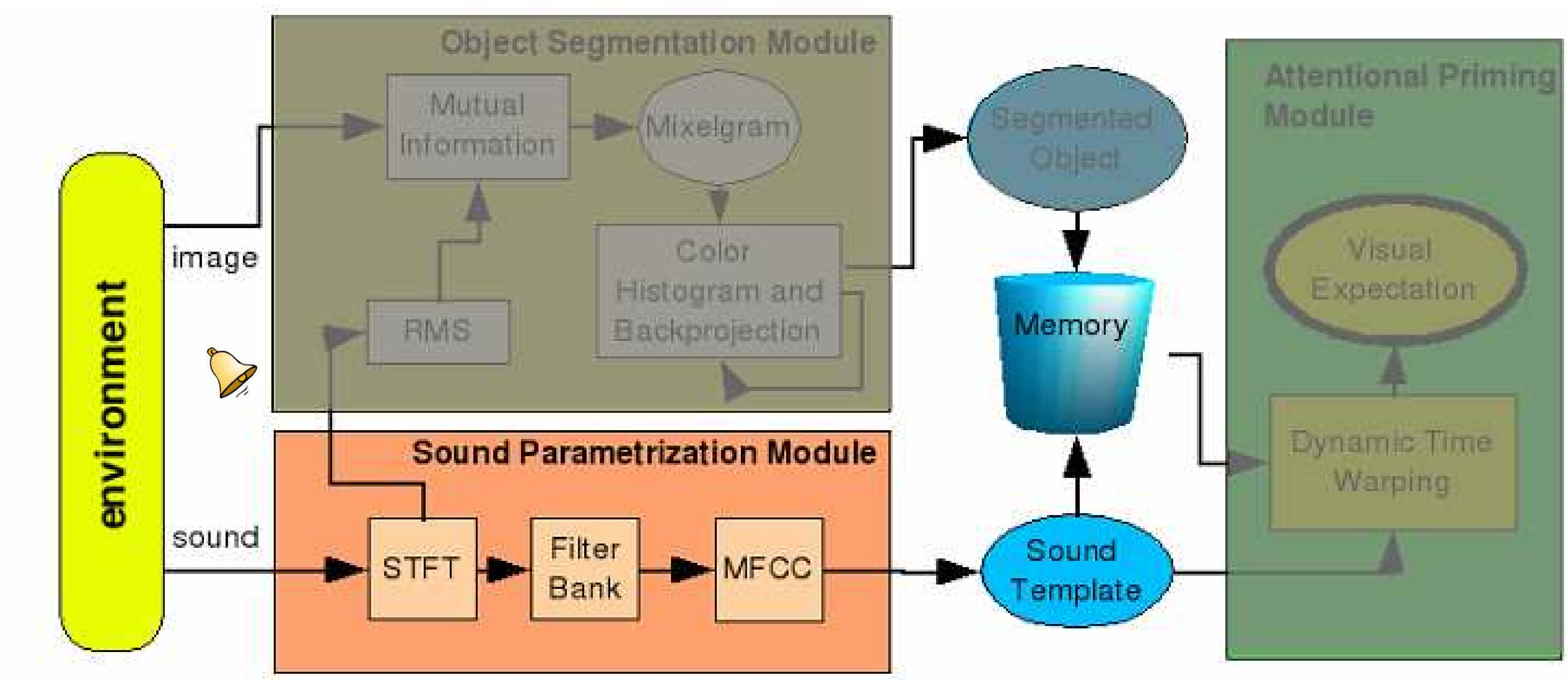
$$M(x, y, t_k) = \frac{1}{2}\log_2 \frac{|\sum A(t_k)||\sum V(x, y, t_k)|}{|\sum A, V(x, y, t_k)|}$$

Result

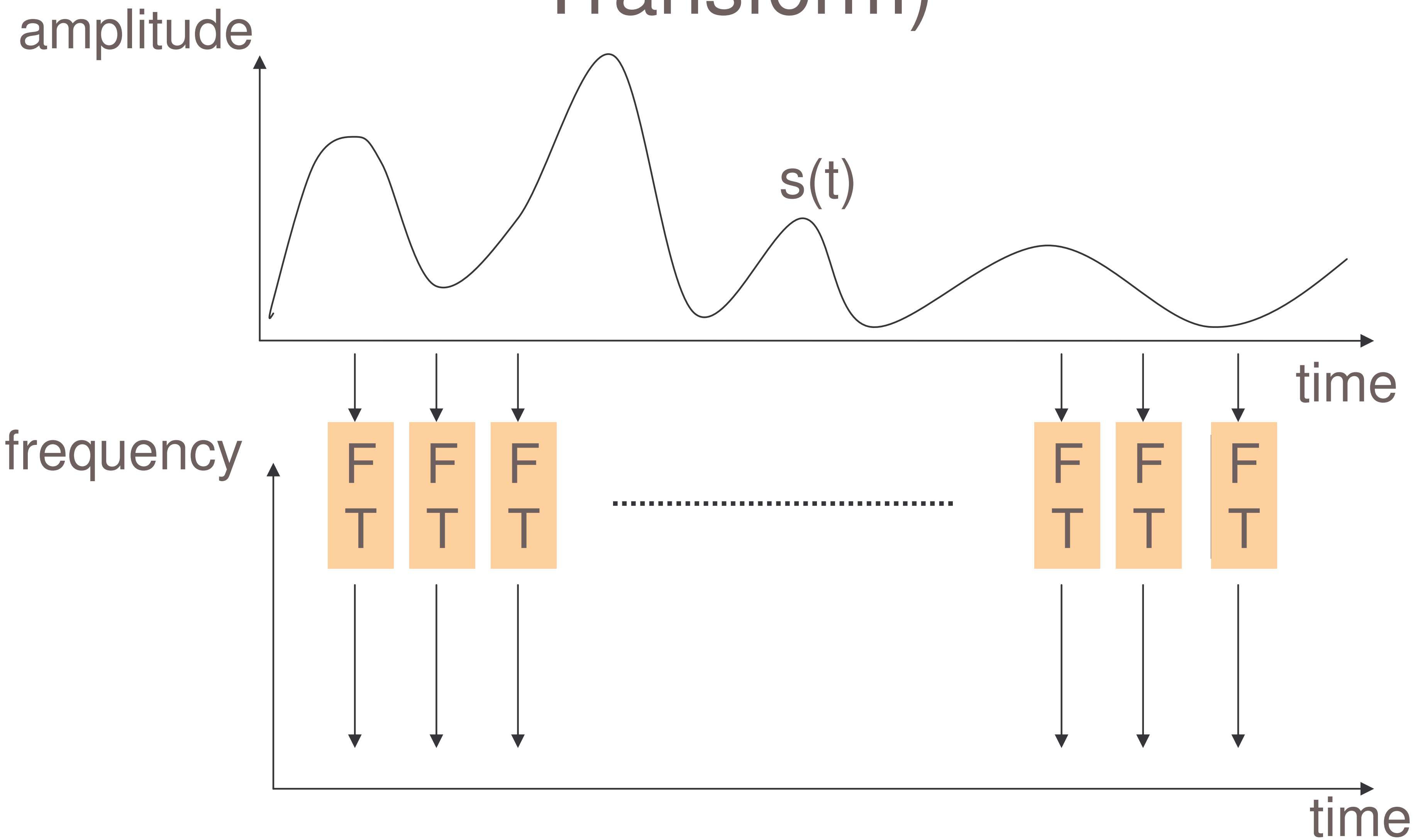


Sound Parametrization

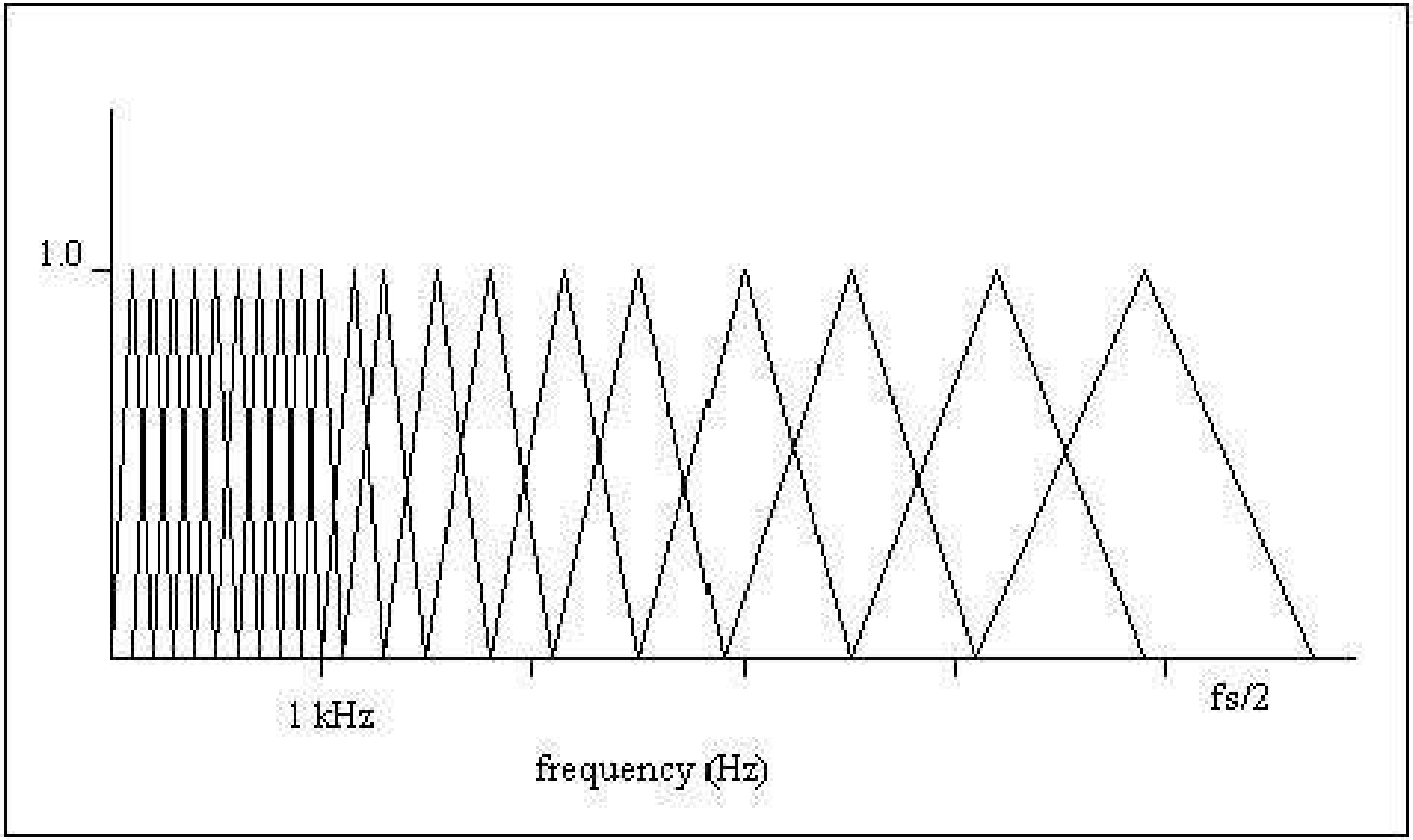
Problem: How can we obtain a low dimensional representation of sound??



Sound parametrization (STFT – Short Time Fourier Transform)



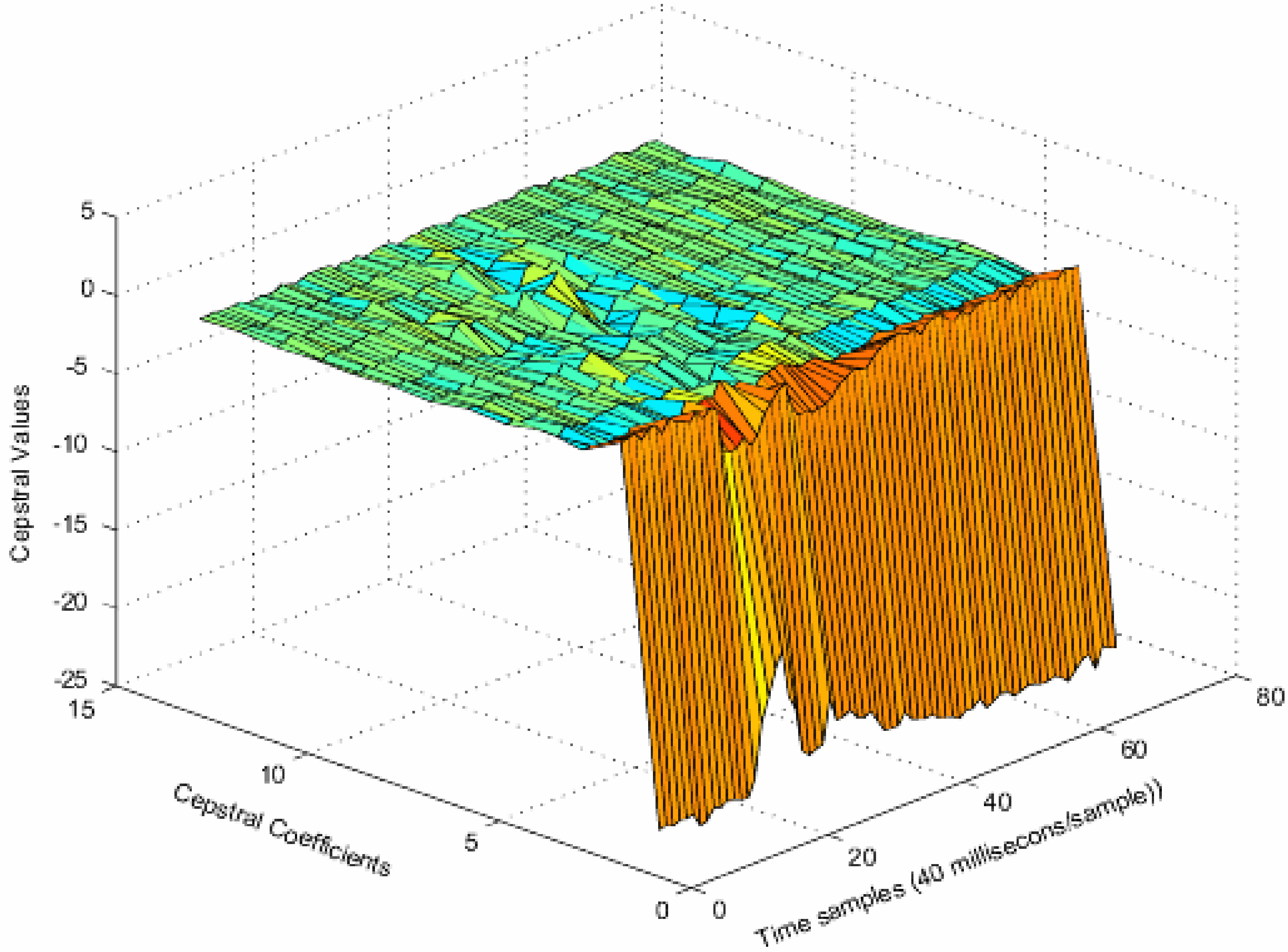
Sound parametrization (Filter bank – Mel frequency filters)



Sound Parametrization (MFCC – Mel Frequency Ceptral Coefficients)

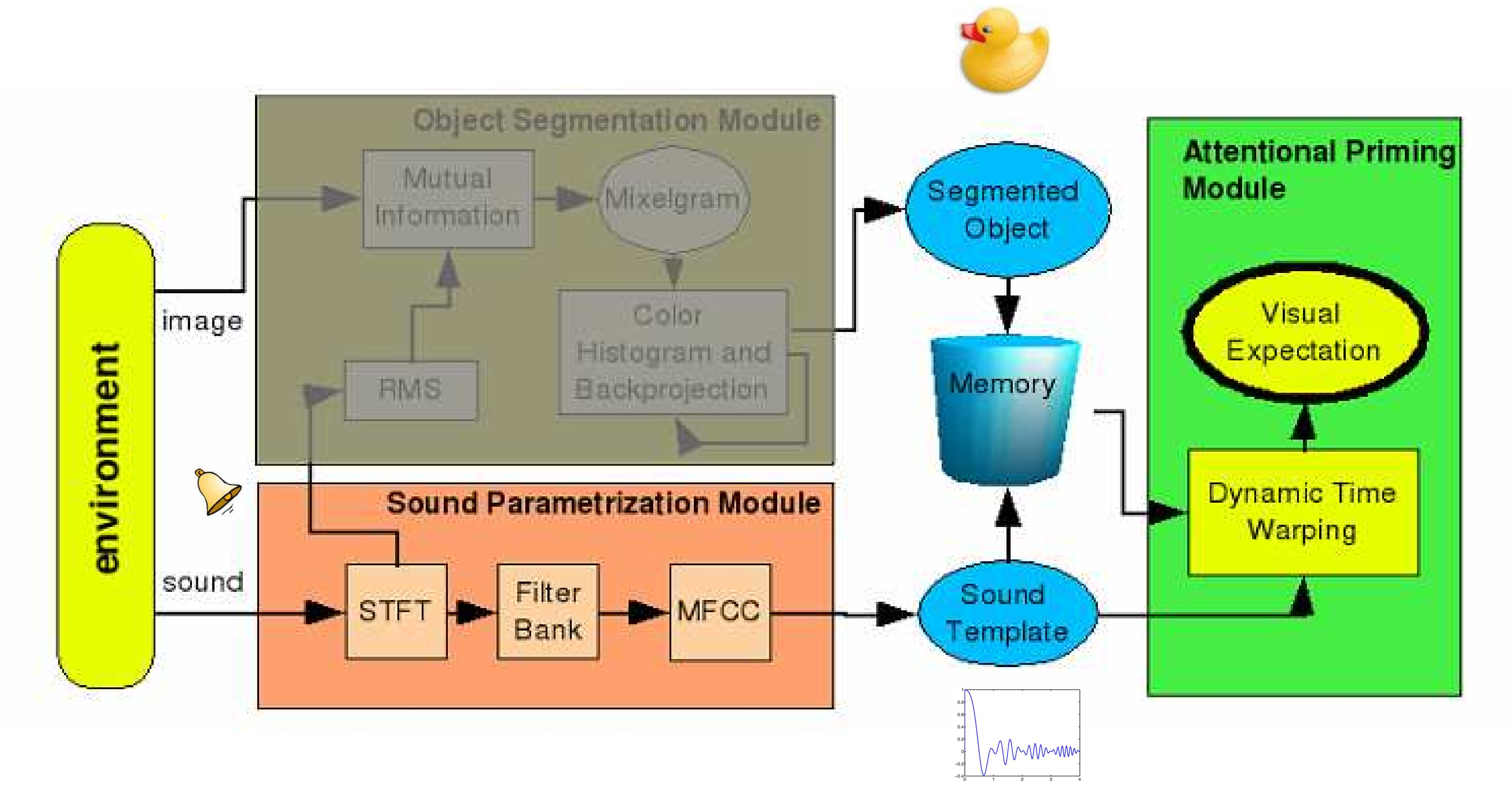
$$c_i = \frac{2}{N} \sum_{k=1}^{N} Y_k \cos\left[i(k+0.5)\frac{\pi}{N} \right], i = 1, 2, \ldots, M$$

Sound Parametrization
(the sound template)

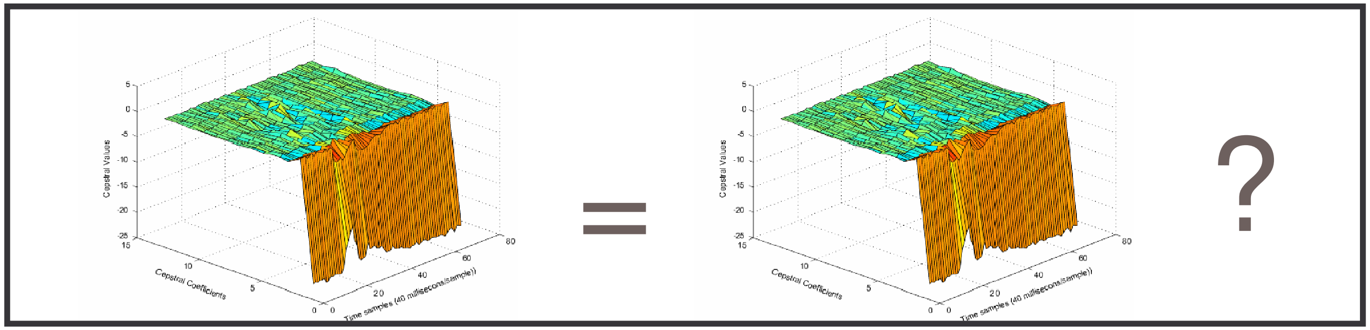


Visual expectation

Problem: How can we recognize a sound??



Visual expectation (DTW- Dynamic Time Warping)



- Local distance measurement among the spectral vectors (MFCC)
- Global distance between all the vectors
  - Time alignment between the two sound utterances
  - Time normalization
- Dynamic programming problem

Results

	Duck	Blue Pig	Red Pig
Segmentation	64%	70%	75%
Recognition	99%	88%	83%

- A set of 100 trials for each object
- The experimenter supervised visually the success or failure

Drawbacks

- Segmentation can be improved
- Algorithms seriously affected by environmental sound
- High computational cost

Future work

- Integrate these algorithms with sound orienting behaviours on an active vision system
- Integrate with traditional attention algorithms.
- Use similar techniques to integrate tactile information.

Questions?

Thank you
for your
attention!