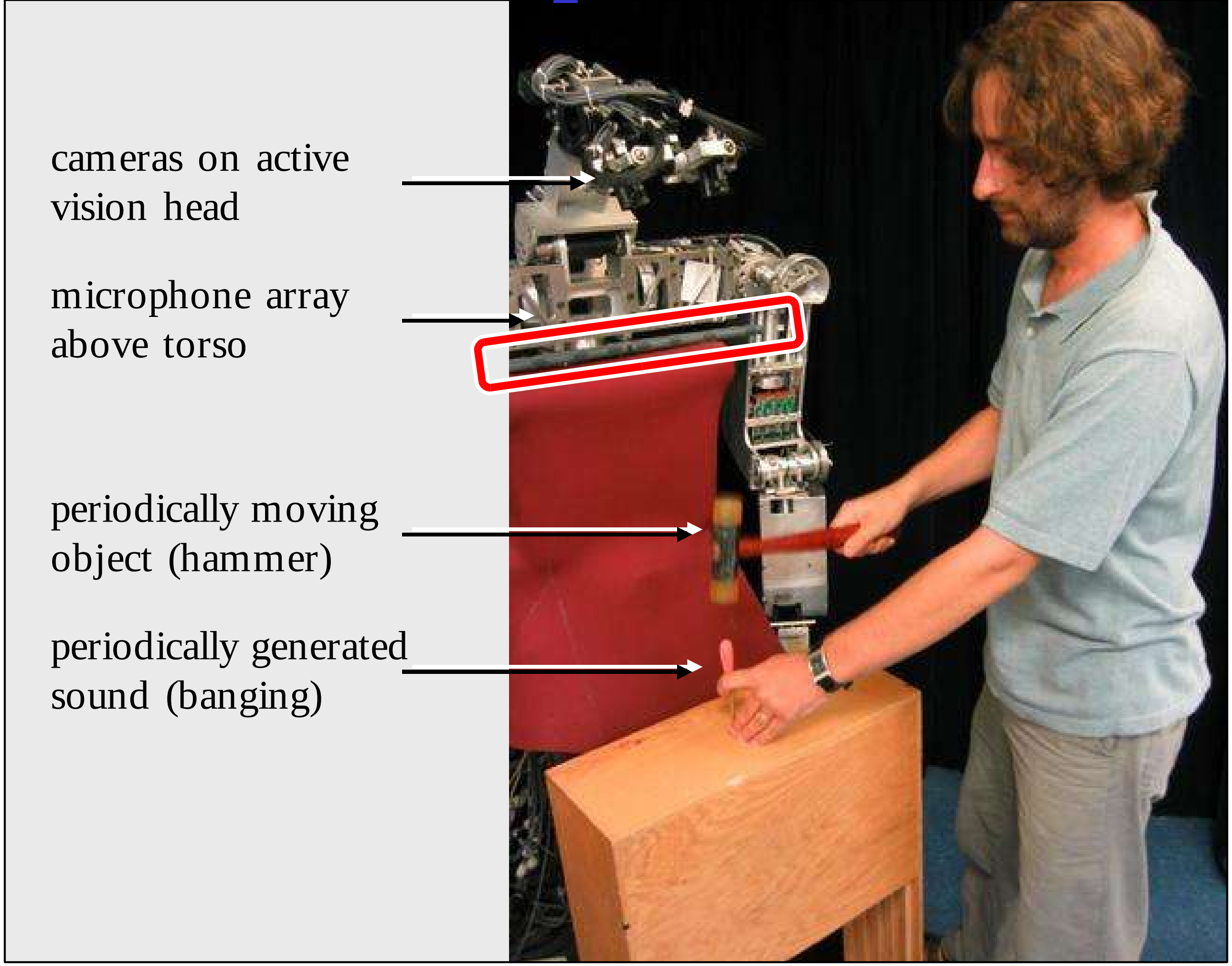
Exploiting cross-modal rhyth m for robot perception of objects

Artur M. Arsenio *Paul Fitzpatrick*

MIT Computer Science and Artificial Intelligence Laboratory

Cog – the humanoid platform



cameras on active vision head

microphone array above torso

periodically moving object (hammer)

periodically generated sound (banging)

Motivation

- Tools are often used in a manner that is composed of some repeated motion - consider hammers, saws, brushes, files, …
- Rhythmic information across the visual and acoustic sensory modalities have complementary properties
- Features extracted from visual and acoustic processing are what is needed to build an object recognition system

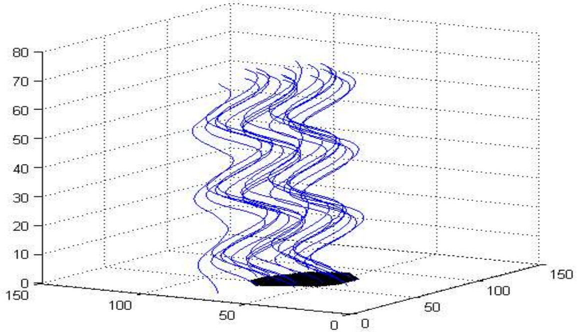
Interacting with the robot

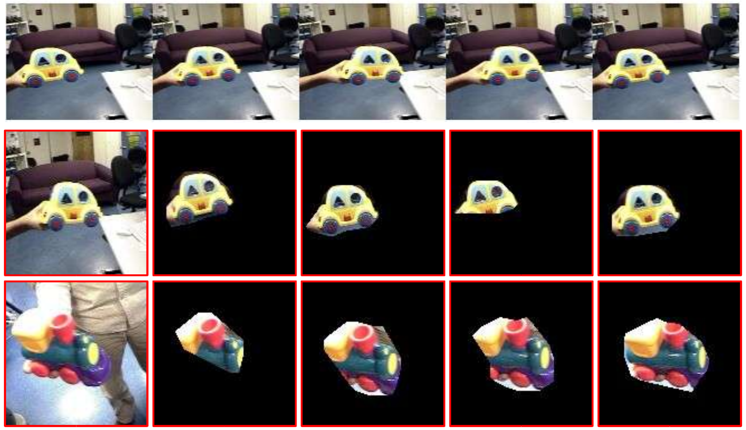
Talk outline

- Matching sound and vision
- Matching with visual distraction
- Matching with acoustic distraction
- Matching multiple sources
- Priming sound detection using vision
- Towards object recognition

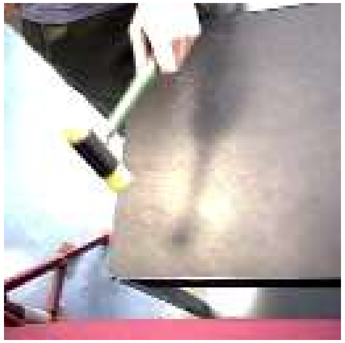
Detecting periodic events

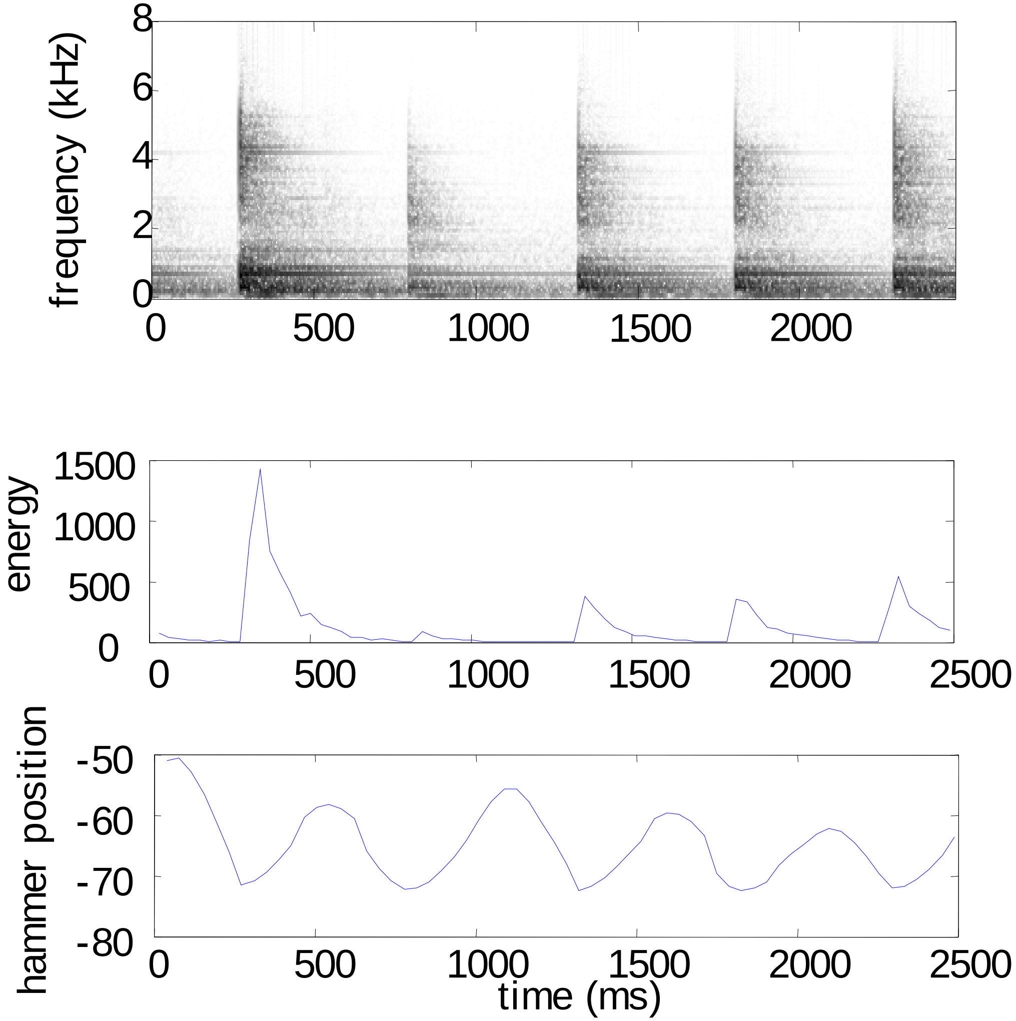
- Tools are often used in a manner that is composed of some repeated motion - consider hammers, saws, brushes, files.
- Points tracked using Lukas-Kanade algorithm
- Periodicity Analysis
    - FFTs of tracked trajectories
    - Periodicity Histograms
    - Phase verification





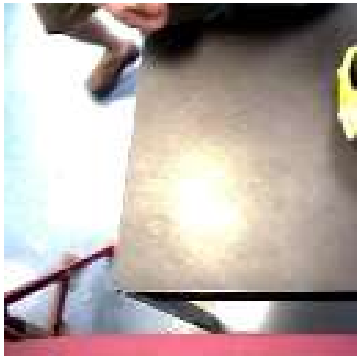
Matching sound and vision

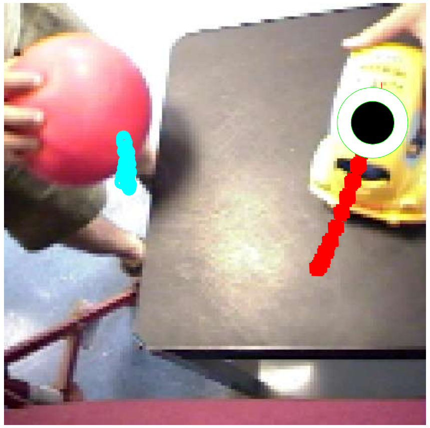




- The sound intensity peaks once per visual period of the hammer

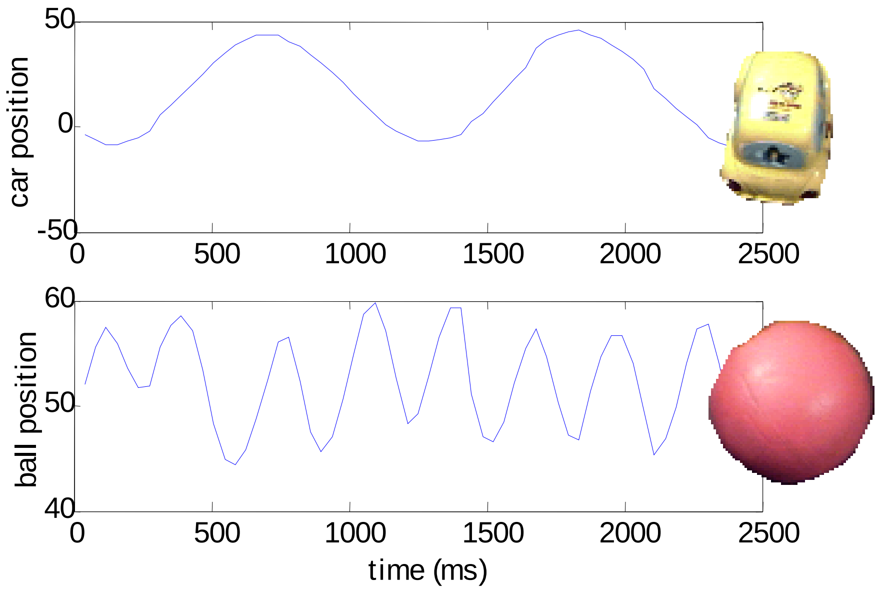
Matching with visual distraction

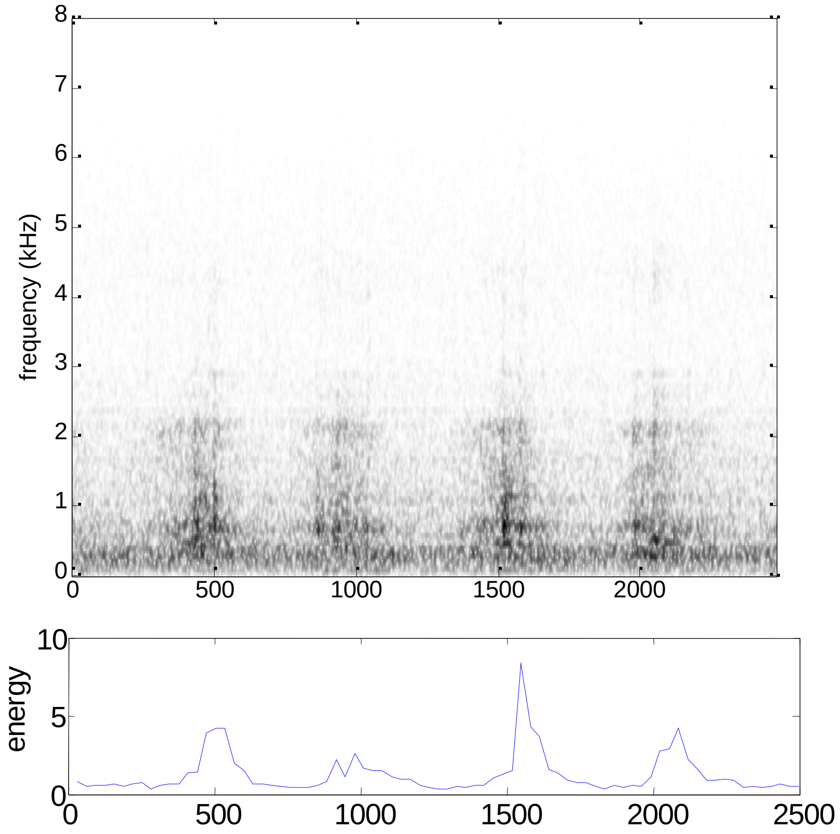




- One object (the car) making noise
- Another object (the ball) in view
  - – Problem: which object goes with the sound?
  - – Solution: Match periods of motion and sound

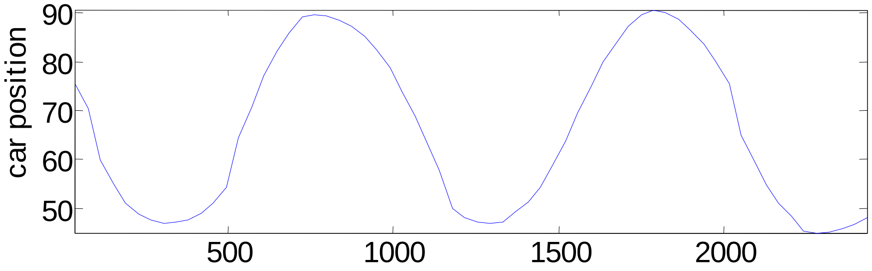
Comparing periods

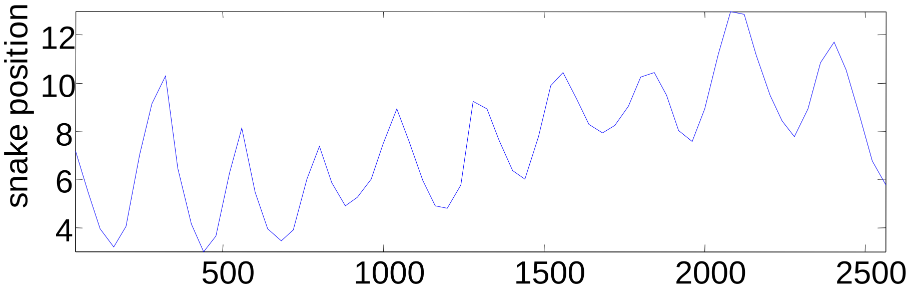


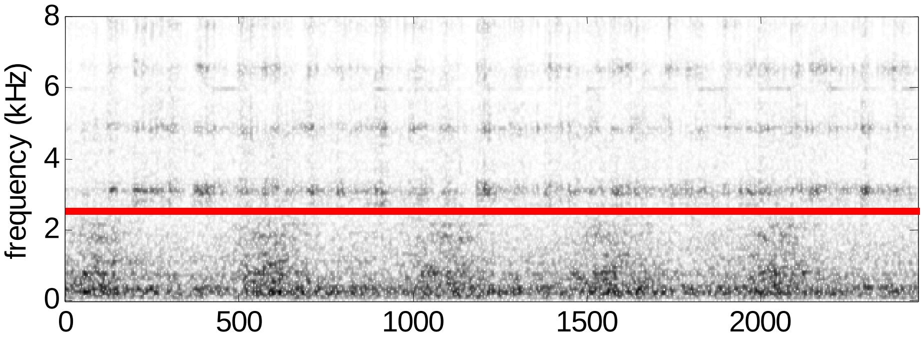


- The sound intensity peaks twice per visual period of the car

Matching with acoustic distraction

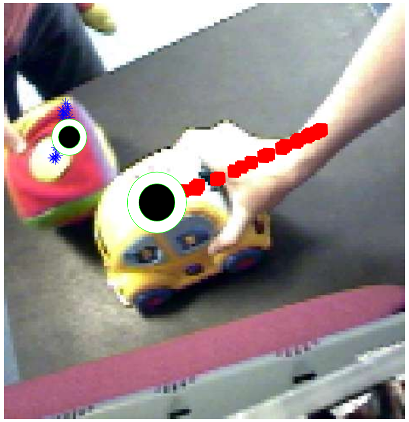






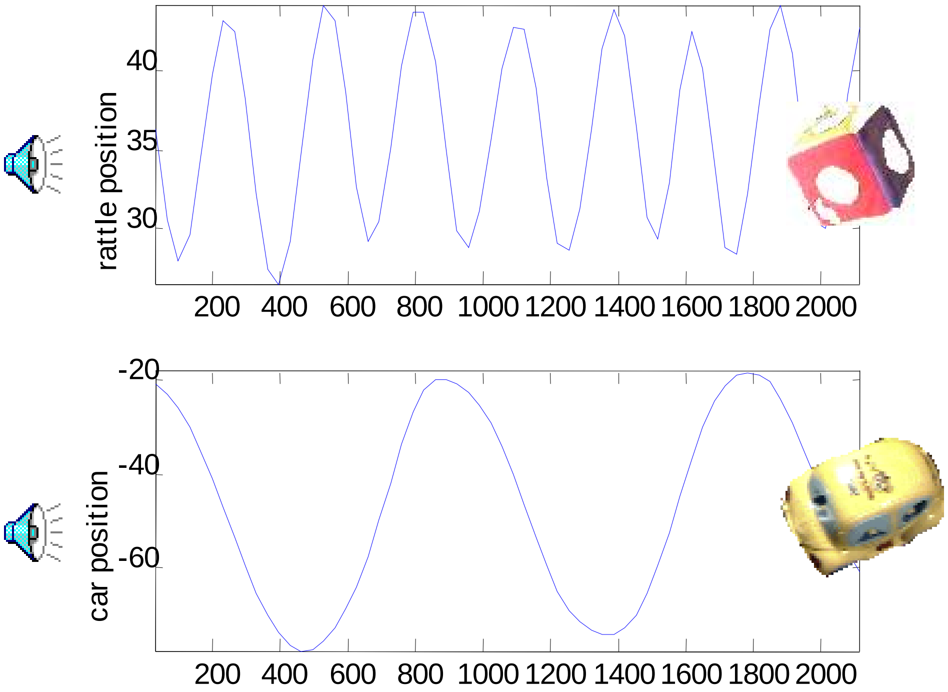
CIRAS 2003

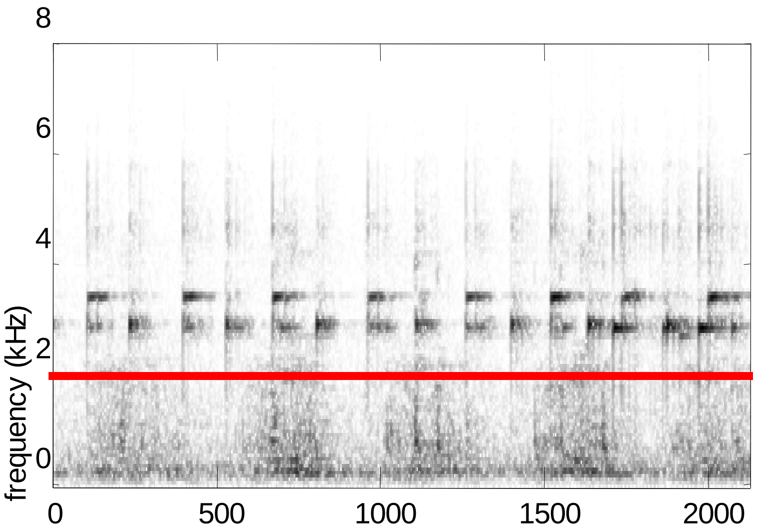
Matching multiple sources



- Two objects making sounds with distinct spectrums
  - – Problem: which object goes with which sound?
  - – Solution: Match periods of motion and sound

Binding periodicity features





- The sound intensity peaks twice per visual period of the car. For the cube rattle, the sound/visual signals have different ratios according to the frequency bands

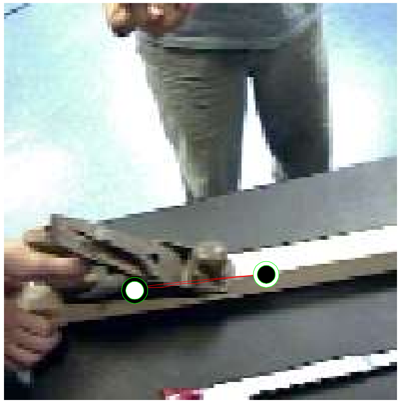
Statistics

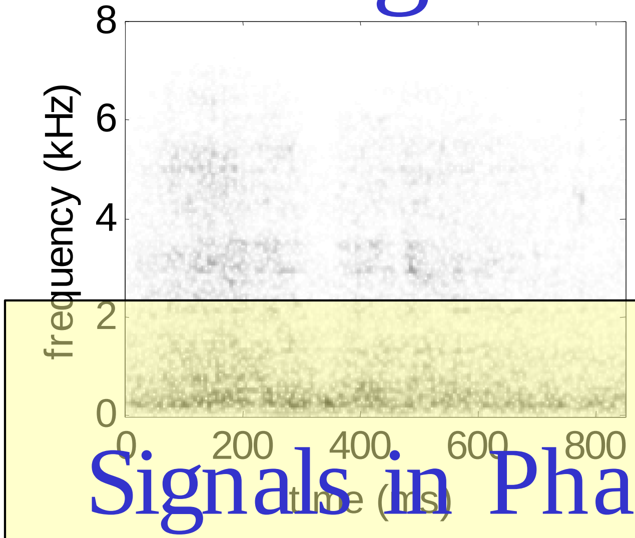
An evaluation of cross-modal binding for various objects and situations

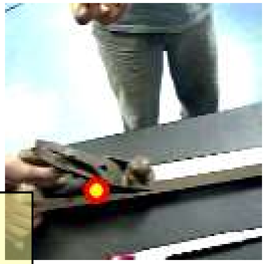
Experiment	*Visual period found*	*Sound period found*	*Bind made*	*Correct binds (%)*	*Incorrect Binds (%)*	*Missed Binds (%)*
hammer	**6**	**8**	**6**	**100**	**0**	**0**
Car & ball	**7**	**11**	**1**	**14**	**0**	**86**
Car	**6**	**6**	**5**	**83**	**0**	**17**
Car (Snake rattle in background)	**5**	**16**	**5**	**100**	**0**	**0**
Snake rattle (Car in Background)	**4**	**9**	**4**	**100**	**0**	**0**
Plane and Mouse	**23**	**27**	**16**	**46**	**41**	**13**

the sound generated by a periodically moving object can be much more complex and ambiguous

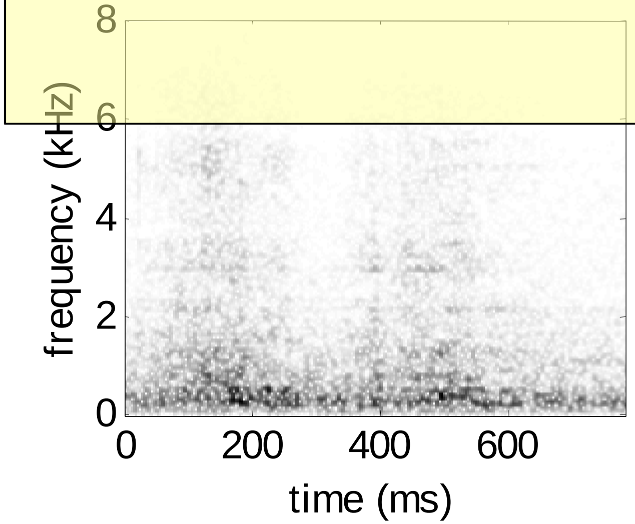
Priming sound detection using vision



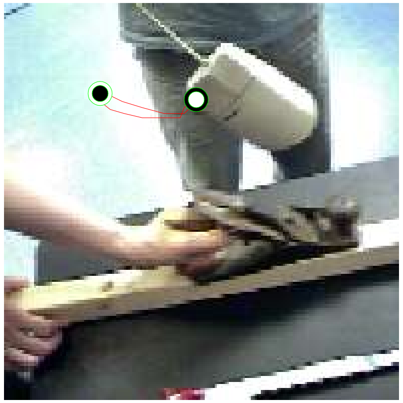


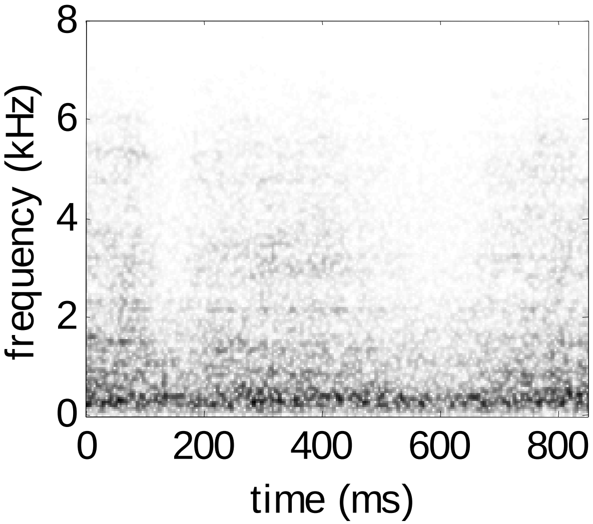


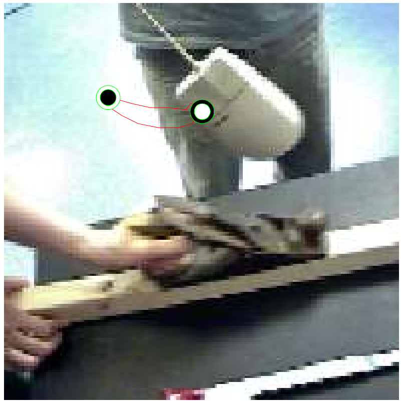
Signals in Phase

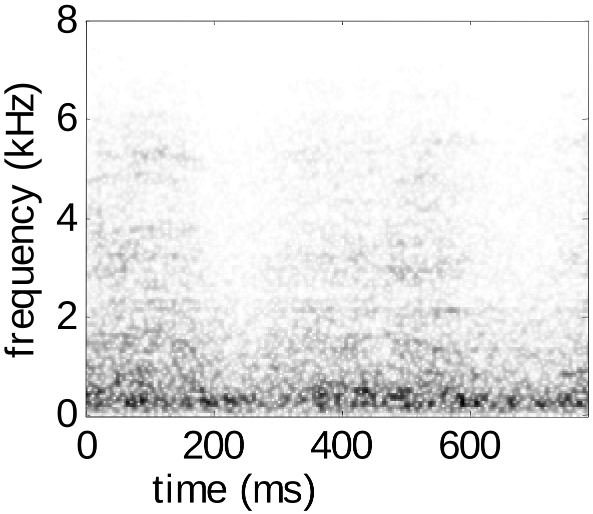


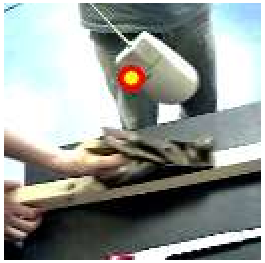
Signals out of phase!





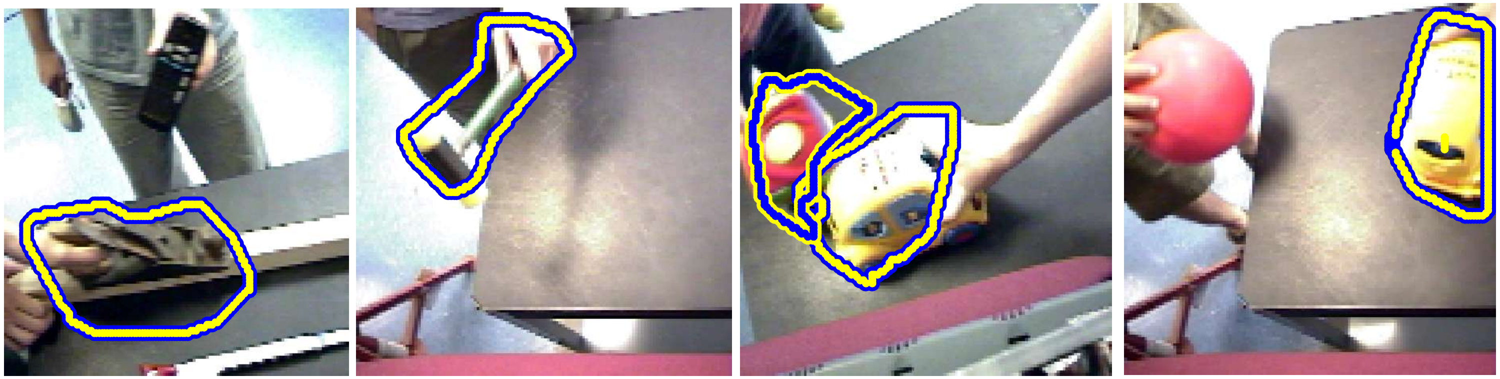






Object recognition

- Visual object segmentation



- Cross-modal object recognition
  - Ratio between acoustic/visual fundamental frequencies
  - Phase between acoustic and visual signals
  - Range of acoustic frequency bands

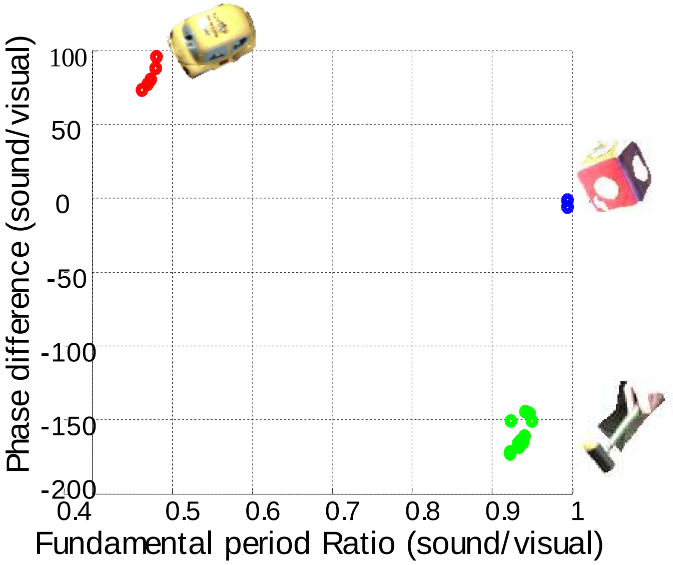
Cross-modal object recognition

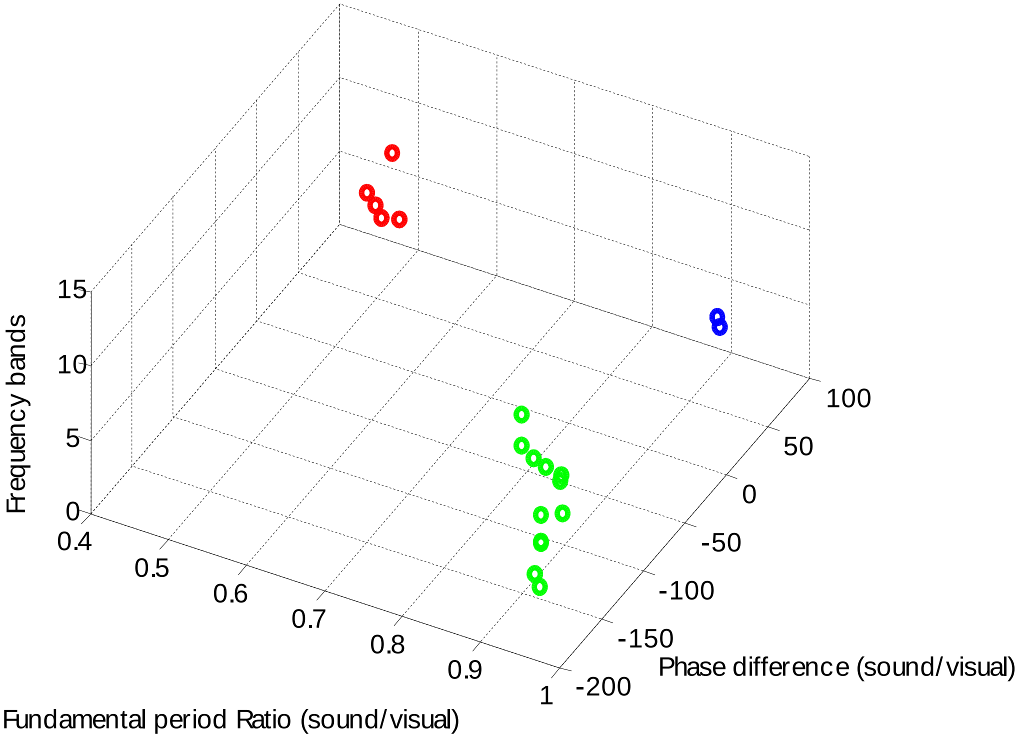
Causes sound when changing direction after striking object; quiet when changing direction to strike again

Causes sound while moving rapidly with wheels spinning; quiet when changing direction

Causes sound when changing direction, often quiet during remainder of trajectory (although bells vary)

Clustering


100
50
0
-50
-100
-150
-200
0.4
0.5
0.6
0.7
0.8
0.9
1
Phase difference (sound/visual)
Fundamental period Ratio (sound/visual)


15
10
5
0
Frequency bands
0.4
0.5
0.6
0.7
0.8
0.9
1
Fundamental period Ratio (sound/visual)
100
50
0
-50
-100
-150
-200
Phase difference (sound/visual)

Conclusions

- Different objects = distinct acoustic-visual patterns which are a rich source of information for object recognition. Object differentiation from both its visual and acoustic backgrounds by binding pixels and frequency bands that are oscillating together
- Cognitive evidence that, for humans, simple visual periodicity can aid the detection of acoustic periodicity
- More feature can be used for better discrimination, like the ratio of the sound/visual peak amplitudes
- Each type of features are important for recognition when the other is absent. But when both are present, then we can do better by looking at the relationship between visual motion and the sound generated.

CSAIL

Questions?