Exploiting cross-modal rhythm for robot perception of objects

Artur Arsenio and Paul Fitzpatrick

What: We are developing a cross-modal binding mechanism for a humanoid robot (Figure 1) to recognize rythmically moving objects such as tools or toys. Our approach uses the motion of the objects, the sound they make, and the relation between the motion and sound. We have demonstrated selectivity and robustness in the face of distracting motion and sounds. Our method does not require prior sound localization or sound-source separation.

Why: Objects that move rhythmically are common and important in the kinds of workplaces where we might employ a humanoid robot. The humanoid form is often argued for so that the robot can interact well with tools designed for humans, and such tools are typically used in a repetitive manner, whether the sound is generated by physical abrasion or collision: hammers, chisels, saws etc. We also work with the perception of toys designed for infants – rattles, bells etc. – which could have utility for entertainment/pet robotics.

The advantage of combining rhythmic information across acoustic and visual senses is that these senses have complementary properties. Since sound waves disperse more readily than light, vision retains more spatial structure – but for the same reason it is sensitive to occlusion and the relative angle of the robot's sensors, while auditory perception is quite robust to these factors. The spatial trajectory of a moving object can be recovered quite straightforwardly from visual analysis, but not from sound. However, the trajectory in itself is not very revealing about the nature of the object. We use the trajectory to extract visual and acoustic features – patches of pixels, and sound frequency bands – that are likely to be associated with the object. Both can be used for recognition. Sound features are easier to use since they are relatively insensitive to spatial parameters such as the relative position and pose of the object and the robot.

How: For each object moving visually, fragments of the concurrent sound input are taken for periods of that object, aligned, and compared. If the fragments are consistent, with sound and vision in phase with each other, then the visual trajectory and the sound are considered bound.

The relationship between object motion and the sound generated varies in an object-specific way. The hammer causes sound when changing direction after striking an object. The bell typically causes sound at either extreme of motion. A toy truck causes sound while moving rapidly with wheels spinning; it is quiet when changing direction.

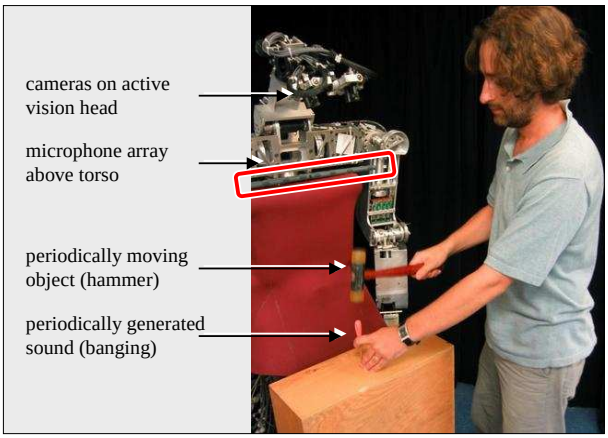

Figure 1: The experimental platform. The humanoid robot Cog is equipped with cameras in an active vision head, and a microphone array across the torso. A human demonstrates some repetitive action to the robot, such as using a hammer.

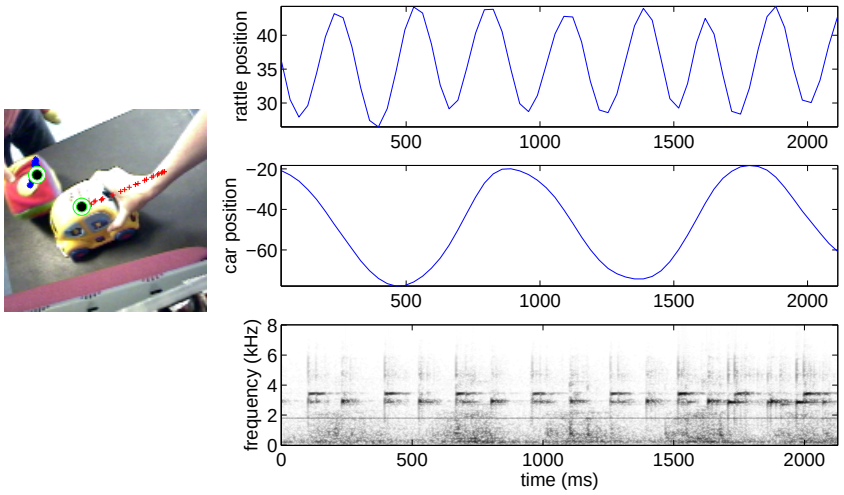

Figure 2: The car and the cube, both moving, both making noise. The line overlaid on the spectrogram (bottom) shows a cutoff determined automatically between the high-pitched bell in the cube and the low-patched rolling sound of the car. The frequencies of both visual signals are half those of the audio signals.

When sound is produced through motion, for such objects the audio signal is highly correlated both with the motion of the object and the tools' identity. The spatial trajectory is used to extract visual and acoustic features – patches of pixels, and sound frequency bands – that are associated with the object.

Progress: We have worked through three cases of cross-modal binding of increasing complexity. The first is when multiple moving objects are visible, but only one repeating sound is heard. If the sound matches the motion of one of the objects, it will be bound to that one and not the other. Similarly, if two repeating sounds with different periods are heard, and a single moving object is visible, the sound with matching period can be bound with the visible object – this is the second case examined. Finally, we have shown that multiple sound and visual sources can be bound together appropriately (see Figure 2).

Future: Features extracted from the visual and acoustic segmentations are what we need to build an object recognition system (in the visual domain see [3, 2], and [4] has looked at the recognition of sound generated by a single contact event). Each type of feature is important for recognition when the other is absent. But when both are present, then we can do better at recognition by looking at the *relationship* between visual motion and the sound generated, and use such data for object recognition.

Research Support: Funds for this project were provided by DARPA as part of the "Natural Tasking of Robots Based on Human Interaction Cues" project under contract number DABT 63-00-C-10102, and by the Nippon Telegraph and Telephone Corporation as part of the NTT/MIT Collaboration Agreement. Artur Arsenio supported by Portuguese grant PRAXIS XXI BD/15851/98.

References:

[1] A. Arsenio and P. Fitzpatrick. Exploiting cross-modal rhythm for robot perception of objects. In *Proceedings of the Second International Conference on Computational Intelligence, Robotics, and Autonomous Systems*, Singapore, December 2003.

[2] Artur Arsenio. Embodied vision - perceiving objects from actions. *IEEE International Workshop on Human-Robot Interactive Communication*, 2003.

[3] Paul Fitzpatrick. *From First Contact to Close Encounters: A Developmentally Deep Perceptual System for a Humanoid Robot*. PhD thesis, Massachusetts Institute of Technology, Department of Electrical Engineering Computer Science, Cambridge, MA, 2003.

[4] E. Krotkov, R. Klatzky, and N. Zumel. *Robotic perception of material: Experiments with shape-invariant acoustic measures of material type*. O. Khatib and K. Salisbury, editors, Experimental Robotics IV. Springer-Verlag, 1996.