Exploiting cross-modal rhythm for robot perception of objects

A. Arsenio and P. Fitzpatrick
CSAIL, MIT, Cambridge, MA 02139, USA
{arsenio,paulfitz}@csail.mit.edu

Abstract

This paper presents an approach to perceiving rhythmically moving objects that generate sound as they move. The work is implemented on the humanoid robot Cog [2]. We show selectivity and robustness in the face of distracting motion and sounds. Our method does not require accurate sound localization, and in fact is complementary to it. We are motivated by the fact that objects that move rhythmically are common and important for 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, where 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. Our goal is to build the perceptual tools required for a robot to learn to use tools and toys through demonstration.

1 Introduction

Tools are often used in a manner that is composed of some repeated motion – consider hammers, saws, brushes, files, etc. This repetition can potentially aid a robot to perceive these objects robustly. Our approach is for the robot to detect simple repeated events at frequencies relevant for human interaction, using both visual and acoustic perception. The advantage of combining rhythmic information across these two modalities is that they 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.

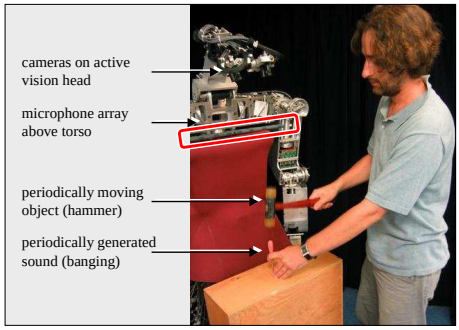

Figure 1: The experimental platform. The humanoid robot Cog [2] is equipped with cameras in an active vision head, and a microphone array across the torso. A human demonstrates some repetitive action to the robot, such as using a hammer.

Our work is implemented on Cog [2], an upper-torso humanoid robot (see figure 1). Previous work on Cog has shown the ability to extract a perceptual benefit (object segmentation) from known motion (object tapping) [4]. Cog has previously been applied to performing basic repetitive behaviors [8] including turning a crank, hammering, sawing, playing with a slinky, swinging a pendulum, and so on. In a sense, the current work is a perceptual analog of this motor ability, which had no sensory component other than direct feedback from the robot's joints.

2 Detecting rhythmic motion

Perhaps the most direct way to detect periodicity is to use the Short-Time Fourier Transform (STFT).

This transform maps a signal into a two-dimensional function of time and frequency. A STFT is applied to each input signal available,

$$I(t, f_t) = \sum_{t=0}^{N-1} i(t')h(t' - t)e^{-j2\pi f_t t'} \qquad (1)$$

where h is a windowing function, and N the number of samples. Periodicity is estimated from a periodogram determined for all signals from the energy of the windowed FFTs over the spectrum of frequencies. These periodograms are then processed to determine whether they correspond to a clear unambiguous period. A periodogram is accepted if its energy is concentrated over a narrow peak.

This is a very general method, and is similar to that adopted in [7]. There are some difficulties with applying it. We must choose a time span over which to perform periodicity inference. The longer this span, the larger the number of signal repetitions available, and the larger the set of frequencies that can be processed, increasing the range of oscillating objects with which the robot can interact. But visual tracking performance decreases with the increase of the time window, and the assumption of constant period over this window is weakened for larger intervals, for both audio and visual signals. We attempted to address this tension by adopting a flexible compromise between the spatial and frequency based views of a signal. The periodicity detection is applied at multiple scales, with long spatial intervals (and hence small window sizes) providing more precise spatial, local information, while larger windows increase the frequency resolution. For objects oscillating during a short period of time, the movement might not appear periodic at a coarser scale, but does appear as such at a finer scale. If a strong periodicity is found, the points implicated are used as seeds for object segmentation (discussed in section 4). Otherwise the window size is halved and the procedure is repeated for each half.

The STFT based-method demonstrated good performance after extensive evaluation for *visual* signals from periodic objects, but it is not appropriate for periodicity detection of *acoustic* signals. These signals may vary considerably in amplitude between periods, which – particularly when combined with variability in the length of the periods – suggests that Fourier analysis is not appropriate. This led us to the development of a more robust method for periodicity detection, which was applied to both acoustic and visual signals. We construct a histogram of the durations between successive instances of particular values of the signal, and search for the most common duration, which we take to be the signal's period. To make this process efficient, we use a hash table indexed by the signal values (combined with their derivatives) discretized relative to the minimum and maximum value of the signal. Each index of the hash table corresponds to distinct values for the signal and its derivative, and the information stored in the table corresponds to the time instants when each value/derivative pair last occurred. After building the table, a period histogram is constructed by the differences found during assignments to the table. The histogram maximum corresponds to the signal period.

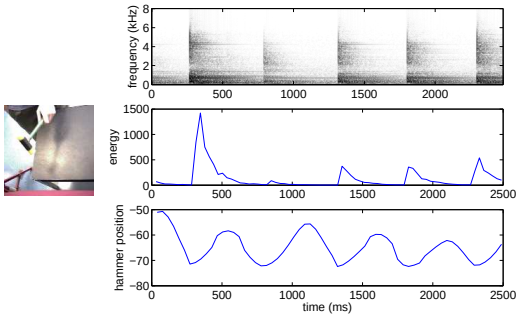

Figure 2: Audio and visual information for a hammer hitting a table. The robot monitors the spectrum of the sound signal (the top graph shows the complete spectrogram, the middle graph shows the overall energy signal) while tracking the trajectory of objects visible to its camera (bottom).

3 Matching sound and vision

Due to physical constraints, the set of sounds that can be generated by manipulating a particular object is often quite small. For tools and toys which are suited to one specific kind of manipulation – as hammers encourage banging – there is even more structure to the sound they generate. We expect that for such objects, when sound is produced through motion, the audio signal will be highly correlated both with the motion of the object and the identity of the tool.

This concept can be illustrated with a basic example: hammering (see figure 2). Whenever the hammer bangs in the table, a distinctive audio signal is produced, spread all over the frequency bands, with a sharp rise of energy at the instant of impact and rapid fall-off thereafter. If we monitor the visual trajectory of the hammer along the main axis of motion, it oscillates at the same frequency as the sound, with approx-

imately zero phase-shift at the moment of impact. The details of this sound may vary with the surface that the hammer bangs against, but the overall pattern is consistent.

For this example and all other experiments described in this paper, our system tracks moving pixels in a sequence of images from one of the robot's cameras using a multiple tracking algorithm based on the a pyramidal implementation of the Lukas-Kanade algorithm. A microphone array sampled the sounds around the robot at 16kHz. The Fourier transform of this signal is taken with a window size of 512 samples and a repetition rate of 31.25Hz. The Fourier coefficients are grouped into a set of frequency bands for the purpose of further analysis, along with the overall energy. 'Binding' or grouping of simultaneous audio and visual signals takes place if the periods of both signals match within a tolerance of approximately $60ms$. No binding is carried out if a moving object is silent, or if a noise-making object lies outside of the robot's field of view.

We now work through three cases of cross-modal binding of increasing complexity, beyond the simple situation already described for the hammer. The first case 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 show that multiple sound and visual sources can be bound together appropriately.

3.1 Matching with visual distraction

In this section we consider the case of multiple objects moving in the robot's visual field, only one of which is generating sound. The robot uses the sound it hears to filter out uncorrelated moving objects and determine a candidate for cross-modal binding. This is a form of context priming, in which an external signal (the sound) directs attention towards one of a set of potential candidates.

Figure 3 shows measurements taken during an experiment with two objects moving visually, at different rates, with one - a toy car - generating a rolling sound, while the other - a ball - is moving silently. The acoustic signal is linked with the object that generated it (the car) using period matching. The movement of the ball is unrelated to the period of the sound, and so that object is rejected. In contract, for the car there

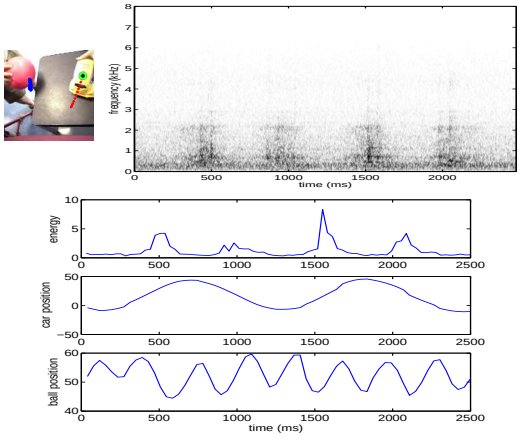

Figure 3: The top left image shows a car and a ball moving simultaneously, with their trajectories overlaid. The spectrogram during this event is shown to the right. Sound is only generated by the rolling car – the ball is silent. A circle is placed on the object (car) with which the sound is bound. The sound energy and the visual displacements of the objects are given.

is a very definite relationship. In fact, the sound energy signal has two peaks per period of motion, since the sound of rolling is loudest during the two moments of high velocity motion between turning points in the car's trajectory. This is a common property of sounds generated by mechanical rubbing, so the binding algorithm takes this possibility into account by testing for the occurrence of frequencies at double the expected value. Section 3.4 develops another approach to dealing with these and similar situations. Section 4 looks at using the very richness of the relationship between audio and visual signals for the purposes of object recognition.

3.2 Matching with acoustic distraction

This section considers the case of one object moving in the robot's field of view, and one 'off-stage', with both generating sound. Matching the right sound to the visible object is achieved by mapping the time history of each individual coefficient band of the audio spectrogram (see figure 4) to the visual trajectory of the object. We segment the sound of the object from the background by clustering the frequency bands with the same period (or half the period) as the visual target, and assign those bands to the object.

Within the framework being described, visual information is used to prune the range of frequency bands of the original sound - the coefficient bands of the au-

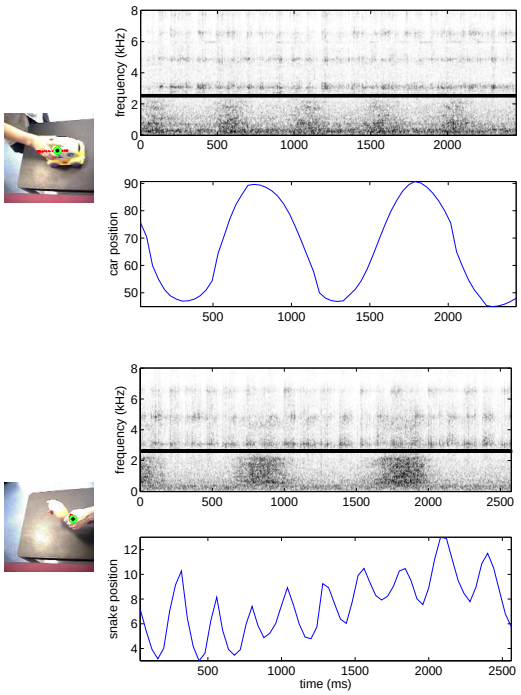

Figure 4: The two spectrogram/trajectory pairs shown are for a shaking toy car and snake rattle. The top pair occurs with only the car visible, and the lower pair occurs with only the snake visible. The line in each spectrogram represents the cutoff pitch frequency between the car and snake.

dio visual are segmented into clusters of bands that characterize the sound of an object. For the experiment shown in the upper part of figure 4, the coefficients ranging from 0 to 2.6Hz are assigned to the object. Afterwards, a band-pass filter is applied to the audio-signal to filter out the other frequencies, resulting the clear sound of the car with the sound of the rattle removed or highly attenuated. For the experiment shown in the lower part of figure 4 the roles of the car and snake were switched. A band-pass filter between 2.6-2.8Hz is applied to the audio-signal to filter out the frequencies corresponding to the car, resulting the snakes' sound.

3.3 Matching multiple sources

This experiment considers two objects moving in the robot's field of view and both generating sound, as presented in Figure 5. Each temporal trajectory of a coefficient group is mapped into one of the visual trajectories if coherent with its periodicity. For each object, the lower and the higher coefficient band are

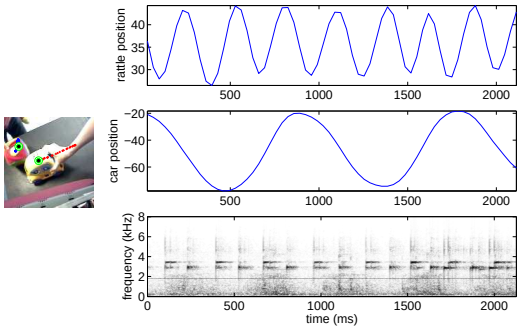

Figure 5: The car and the cube, both moving, both making noise. The line overlaid on the spectrogram (bottom) shows the cutoff determined automatically between the high-pitched bell in the cube and the low-patched rolling sound of the car. A spectrogram of the car alone can be seen on Figure 3. The frequencies of both visual signals are half those of the audio signals.

labelled as the lower and higher cut-off frequencies, respectively, of a band-pass filter assigned to that object. The complex sound of both the moving car-toy and the cube-rattle are thus segmented into the characteristic sound of the car and sound of the rattle through band-pass filtering. Multiple bindings are thus created for multiple oscillating objects producing distinct sounds.

It is worth stressing that the real world is full of objects making all kinds of noise. However, the system is robust to such disturbances. On the experiments presented throughout this paper, people were speaking occasionally while interacting with the robot, while other people were making everyday sounds while working. If the distracting sound occurs at the same range of frequencies as the sound of the oscillating object, then a binding might just not occur for that specific time, but occur after a few seconds when the interference noise switches to other frequencies or disappears.

3.4 Priming sound detection using vision

Figure 6 shows how well the methods described for binding sounds with objects work on a series of experiments. False bindings occur in only one case, where two objects are moving, a mouse and a plane. Only the plane is generating sound. The sound is a rough noise with silence at the two extrema of the plane's motion, and hence appears to have a frequency of double that of the trajectory. This is close to the frequency of oscillation of the mouse, so simple period matching occasionally gives false results. This is symptomatic of a more general problem: the sound generated by a

Experiment	visual period found	sound period found	bind made	good bind (%)	bad 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 background)	5	16	5	100	0	0
snake (car background)	4	9	4	100	0	0
plane and mouse	23	27	16	46	41	13

Figure 6: An evaluation of the efficacy of cross-modal binding for various objects and situations.

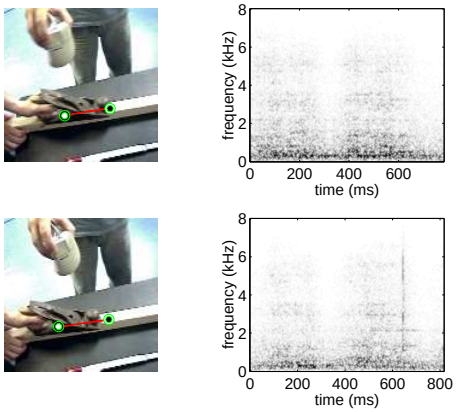

Figure 7: An experiment in which a plane is being pushed across wood while a mouse is shaken in the background. Shown are the highest quality acoustic matches for this sound (right) and the object with which they correspond (left). Matches against the mouse are much lower and below threshold.

periodically moving object can be much more complex and ambiguous than its visual trajectory. The extrema of an approximately repeating trajectory can be found with ease, and used to segment out single periods of oscillation within an object's movement. Single periods of the sound signal can be harder to find, since there is more ambiguity – for example, some objects make noise only at one point in a trajectory (such as a hammer), others make noise at the two extrema (some kinds of bell), others make noise during two times of high velocity between the extrema (such as a saw), and so on. For cases where periodicity detection is difficult using sound, it makes sense to define the period of an action in the visual domain based on its trajectory, and match against this period in the sound domain – instead of detecting the period independently in each domain. We have developed an approach, where for each object moving visually, fragments of the sound 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 bound. This is a more stringent test than just matching periods, yet avoids the problem of determining a period reliably from sound information. Figure 7 shows results for the plane-and-mouse example described at the beginning of this section, showing that it does in fact rectify the problem. The periodicity detection method in section 2 is still necessary when only sound information is available.

There is evidence that, for humans, simple visual periodicity can aid the detection of acoustic periodicity. If a repeating segment of noise is played, the repetition can be detected for much longer periods if a light is flashing in synchrony with some point in the period [1]. More generally, there is evidence that the cues used to detect periodicity can be quite subtle and adaptive [5], suggesting there is a lot of potential for progress in replicating this ability beyond the ideas already described.

4 Object segmentation and recognition

Different objects have distinct acoustic-visual patterns which are a rich source of information for object recognition. Our approach can differentiate objects from both their visual and acoustic backgrounds by finding pixels and frequency bands (respectively) that are oscillating together. We deal with a fundamental problem in computer vision - *object segmentation* - by detecting and interpreting natural human behavior such as waving or shaking objects. This is important, because object segmentation on unstructured, non-static, noisy, real-time and low resolution images is a hard problem. Results for object-background separation are shown in figure 8 and were obtained with varying light conditions. The environment was not manipulated to improve natural occurring elements, such as shadows or light saturation from a light source. All the experiments were made while a human or the robot were performing an activity.

Features extracted from the visual and acoustic segmentations are what we need to build an object recognition system (in the visual domain see [3], and [6] 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 (see figure 9), to which the method in section 3.4 gives us easy access.

Figure 8: Examples of object segmentations.

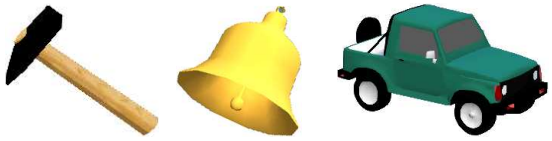

Figure 9: The relationship between object motion and the sound generated varies in an object-specific way. The hammer causes sound when changing direction after striking an object. The bell typically causes sound at either extreme of motion. A toy truck causes sound while moving rapidly with wheels spinning; it is quiet when changing direction.

5 Discussion and conclusions

We described techniques to detect periodicity of signals, identifying their strengths and limitations. Through the detection of visual periodic events, we were able to localize an object in the visual field and extract information concerning its trajectory over time, as well as to segment a visual representation of an object from an image. In addition, sound segmentation - the identification of the frequency bands that best characterize an object - was also possible from just acoustic information. A cross-modal strategy to period detection proved necessary and advantageous, being more robust to disturbances either in sound or vision, and providing a better characterization of objects. We discussed how to reliably bind the visual appearance of objects to the sound that they generate, and to achieve selectivity: a visual distractor was filtered out using sound, and sound was also used to prime the visual field. In addition, we argued that the cross-modal strategy is well suited for integration with object recognition strategies for searching visually for tools and toys and finding/recognizing them whenever their sound is perceived by the robot.

A lot about the world could be communicated to a humanoid robot through human demonstration. The robot's learning process will be facilitated by sending it repetitive information through this communication channel. If more than one communication channel is available, such as the visual and auditory channels, both sources of information can be correlated for extracting richer pieces of information. We demonstrated in this paper a specific way to take advantage of correlating multiple perceptual channels at an early stage, rather than just by analyzing them separately - the whole is truly greater than the sum of the parts.

Acknowledgements

This work was funded by DARPA DABT 63-00-C-10102 ("Natural Tasking of Robots Based on Human Interaction Cues"), and by NTT under the NTT/MIT Collaboration Agreement. Arsenio was supported by Portuguese grant PRAXIS XXI BD/15851/98.

References

[1] J. A. Bashford, B. S. Brubaker, and R. M. Warren. Cross-modal enhancement of repetition detection for very long period recycling frozen noise. *Journal of the Acoustical Soc. of Am.*, 93(4):2315, 1993.

[2] R. A. Brooks, C. Breazeal, M. Marjanovic, and B. Scassellati. The Cog project: Building a humanoid robot. *Lect. Notes in Comp. Sci.*, 1562:52–87, 1999.

[3] P. Fitzpatrick. Object lesson: discovering and learning to recognize objects. October 2003. Accepted for publication at the 3rd International Conference on Humanoid Robots, Karlsruhe, Germany.

[4] P. Fitzpatrick and G. Metta. Grounding vision through experimental manipulation. *Philosophical Transactions of the Royal Society: Mathematical, Physical, and Engineering Sciences*, 2003. In press.

[5] C. Kaernbach. Temporal and spectral basis of the features perceived in repeated noise. *Journal of the Acoustical Soc. of Am.*, 94(1):91–97, July 1993.

[6] E. Krotkov, R. Klatzky, and N. Zumel. Robotic perception of material: Experiments with shape-invariant acoustic measures of material type. In *Experimental Robotics IV*. Springer-Verlag, 1996.

[7] R. Polana and R. C. Nelson. Detection and recognition of periodic, non-rigid motion. *International Journal of Computer Vision*, 23(3):261–282, June/July 1997.

[8] M. M. Williamson. Exploiting natural dynamics in robot control. In *14th Eur. Meet. on Cybernetics and Systems Research*. Vienna, Austria, 1998.