Visual Attention Priming Based on Crossmodal Expectations

Carlos Beltrán-González and Giulio Sandini
Laboratory for Integrated Advanced Robotics
University of Genova
Viale Causa 13, 16145 - Genova, Italy
{*cbeltran,sandini*}*@liralab.it*

***Abstract*— Humans perceive the world using five senses. Research results suggest that this multisensorial perception may be of fundamental importance for development and learning, as well as for creating cognitive capabilities. Moreover, humans have the capacity to create intersensorial expectations to guide attention and perception. We are interested in comprehending how these capabilities may improve robot perception. In this line of research, we present a cross-modal perceptual architecture that can segment objects based on visual-auditory sensorial cues, construct an associative sound-object memory, and create visual expectations of objects (attentional priming) using a sound recognition algorithm.**

***Index Terms*— Prediction, attention, MFCC, anticipation, mutual information, expectations, mixelgram**

I. Introduction

Though machine vision is a long and well established scientific discipline the several decades of intensive research were not enough to resolve a problem that humans seem to pass by almost intuitively. During the early days of computer vision, i.e. 1960's and 1970's, the research efforts were concentrated in the passive processing of single images. The visual cortex was believed to process all the information in the field of view and to do so in a sequential and increasingly complex process. This doctrine influenced the vision computational models of that time that tried to create general descriptions of the visible scene [7]. This period culminated with the publication of the influential Marr's work *Vision* [8], where computer vision was formalized as a pure information processing task.

In the late 1980's and early 1990's a new approach to computer vision appeared : *Active Vision* [1], [4]. During the late 1990's the concept of active vision was exploited, improved and expanded. A strong emphasis was made in the simplification of the early stage vision problems by exploiting the explorative capacities of vision systems. During this period, there was also an approach between "cognitive sciences" and robotics that yielded to epigenetic approaches to robotics and the investigation of the perception-action paradigm where the artificial system is able to act in the world and modify it (see [11] for an example).

More recently, the perception-action paradigm has been explored further in the area of humanoid robotics. For example, Metta and Fitzpatrick [10] have shown how to segment an ambiguous object from the background by active manipulation. Several researches stress that an agent could construct a self image by actively exploring and manipulating [2], [14].

However, though the *learn by doing* approach [5] is providing encouraging results, we think that further development is still necessary. Particularly, the exploitation of intersensorial relations for the improvement of perception has not been sufficiently explored, e.g. the interrelation between sound and vision. In this paper this problem is addressed studying audio-visual causal interrelations. In particular, it is studied how these interrelations may improve object perception and how they could be exploited to create intersensorial expectations.

The paper is organized as follows: first, we propose a conceptualization of crossmodal perception; second, we analyze the research in automatic sound recognition and show how traditional speech recognition techniques can be used to parametrize sounds produced by objects; third, we discuss how an approximation of a statistic called *mutual information* can be used to create a common intersensorial space for sound and vision; fourth, we present how the latter information can be used to segment an object from the background with the assistance of a color back-projection technique; and finally we show how the system: a) creates a sound-object associative memory, b) uses this memory to recognize sounds (through a dynamic time warping algorithm) and c) extracts from the memory a visual expectation associated with a sound auditory event.

II. Toward cross-modal perception

Neuroscience research is actively studying cross-modal relations in the human brain and several researchers suggest that perception is a multisensorial experience. However, still many questions remain unanswered, for example:

- How cognitive pathways may dominate perception (top-down approach).
- How different sensorial modalities are integrated.
- How these sensorial interrelations may guide development and learning.
- How sensorial modalities may influence each other.

There is little understanding on how these mechanisms may work in the human brain. However, some conclusions

can be advanced: a) sensorial interrelations seem to be fundamental for the development of high level cognitive abilities, b) perception seems to depend strongly on multisensorial cues.

In the robotics research field, we can categorize these assumptions in the context of the crossmodal perception paradigm. We conceive crossmodal perception as an *extension* of the active-vision/perception-action paradigms. The crossmodal perceptual agent uses multisensorial cues to reinforce its explorative perception and creates actively synchronized multisensorial inputs (e.g. by hitting repeatedly an object on the ground producing a change in both the visual field and the auditive input).

We attempt the first steps toward this kind of perception by trying to solve, in the context of a humanoid robot architecture, two problems: a) *object segmentation using multisensorial cues*, and b) *sound classification for attentional priming*. More formally, we suggest that these two problems could be conceptualized into two distinct phases:

- *Syn**a**esthetic Phase*: From *syn* "co" and *aisthanesthai* to perceive. This yields to an etymological interpretation as *joint perception* or *to perceive simultaneously*. In this particular experiment, this phase corresponds to an object segmentation based on the integration of sound and visual cues.
- *Synesthetic Phase*: A concomitant sensation; especially, a subjective sensation or image of a sense (as of color) other than the one (as of sound) being stimulated. In this experiment, this phase is formed by a sound classification algorithm that can *remember* the visual aspect of an object.

Notice that syn**a**esthetic and synesthetic are very similar (there is only a letter "a" difference), they have a common etymological origin, however, the meaning is slightly different in our interpretation. Moreover, it is worth noting that these words have been used interchangeably in the literature; particularly, *synaesthetic* is used in cognitive neuroscience to address an *unusual mixing of the senses* that affects certain people (see [18] for a review). This unusual mixing of senses interpretation stress the "strength" how senses interrelate, and also the *non relation with the real world*. For example, patients experience visual hallucinations (i.e. they *see* colors) when hearing a particular noise or they have smell hallucinations when they see a particular number.

We adopt a slightly different interpretation; we believe that most humans have *subjective* sensations activated by crossmodal interrelations. The differences with respect to the cognitive neuroscience point of view are: a) the "intensity" of the interrelations, and b) that they correspond to *real* sensorial experiences. In our view, the interrelations are related with sensorial *expectations* and not with sensorial hallucinations.

Thus, paraphrasing Fermüller and Aloimonos (see [7] chap.9), we may say that:

> Now, it has become clear that image understanding should also include the process of selective acquisition of data in space and time **from multisensorial cues**.

III. System architecture and experimental setup

Sound could be considered as important as vision. However, comparatively, little research has been done in the field of sound recognition. The research has been mainly concentrated in the recognition of speech and music and in the study of orienting behaviors [13]. More recently, the sound research has included works in scene analysis, detection of talking faces [6] and rhythm detection [3].

In a work related with the segmentation of objects by a humanoid robot, Arsenio and Fitzpatrick [3] address the problem of object detection based on the rhythm properties of movements, both in sound and vision streams. They address the recognition of toys designed for infants. We use a similar approach, but we do not exploit the rhythmic characteristics of movement but the intrinsic common information created in both sensorial streams when the toy is squeezed or shaked by the experimenter in front of the robot.

We will show that a combination of speech recognition techniques and statistics can be used to create a crossmodal perceptual architecture that can create associations between the images of toys and the sounds the toys produce; and, in a second stage, evocate the toy's visual image by recognizing the sound associated to the toy, and consequently, have the potential to exploit this visual expectation in explorative movements.

In Figure 1 we present the architecture of the crossmodal perceptual system. This system was implemented in YARP (Yet Another Robotic Platform) [9]. YARP is a framework for humanoid robotics development that provides support, among other things, for distributed computation and multi-operating system communications.

The proposed system was running in a rack of standard PC's, either with Microsoft Windows or QNX installed on the PC's. The system received its inputs from the environment through a PAL camera and two microphones. A standard PCI framegrabber, based on the Conexant Bt848 chip, digitalizes the images which are converted utilizing a log-polar mapping by a software conversion algorithm [19]. A standard audio PCI card digitalizes the sound signal obtained by the microphones. Both cards use a Direct Memory Access (DMA) mechanism to transfer the data streams into the computer main memory.

A fundamental problem was how to synchronize in time the video and sound streams. The standard acquisition cards do not employ any hardware synchronization line, so we developed a special device driver in the Windows operating system that controls the acquisition of both cards. The driver initializes the acquisition cards in a sequential manner using a

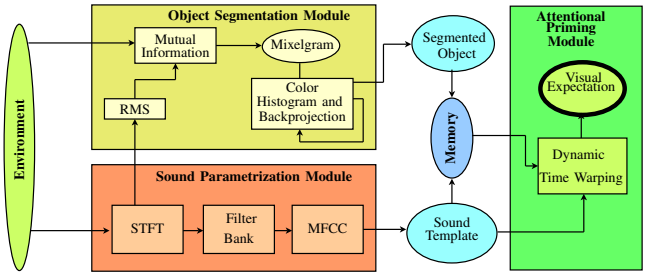


Fig. 1. The architecture of the crossmodal perceptual system

software *critical region* (i.e. a code execution flow that is not interrupted). Consequently, the acquisition software needs to run in a single process being executed by a single computer.

However, the previous mechanism does not guarantee a perfect synchronization, and for this reason, the computer internal clock (*timestamps*) was employed to monitor the time alignment between the data streams. Using this technique we measured a time difference that in average of several tests was less than one millisecond with different CPU loads.

The sampling frequency for the sound was 44100 Hz, whilst the precision was 16 bits and the transfer memory was 1764 samples/frame. This produced a sound framerate of 25 frames per second at a rate of one frame each 40 milliseconds, exactly the same as the frame sequence rate of the PAL video images.

(a) (b) (c)

Fig. 2. Experiment objects as perceived by Eurobot: (a) A deformable plastic yellow duck, (b) a hollow hard plastic blue pig filled with plastic bottle caps, and (c) a hollow hard plastic red pig filled with chickpeas.

For the experiment, we used an upper torso humanoid robot called Eurobot and a set of three baby toys acquired in commercial stores. Figure 2 shows the group of toys as seen by the robot. Figure 2(a) is a deformable yellow plastic duck; it produces a high frequency sound when squeezed with the hand. The hollow hard plastic toy pigs shown in Figures 2(b) 2(c) are the same toy; the differences are: they have different colors and we have filled them with different materials. Therefore, the sound produced by each toy pig was slightly different.

IV. Sound Parametrization

The goal of the sound parametrization module was to obtain a low dimensional representation of sound. In the speech recognition literature this module is known as the signal-processing front-end. The idea is to have a sequence of measurements of the input signal, usually the output of some type of spectral analysis technique, that yields a "pattern" that represents the sound; though we prefer the term *sound template* for this representation. This sound template is a sequence of spectral vectors. Each of these vectors represents the frequency transformation of the sound in a short period of time; in our system, this period of time has a duration of 40 milliseconds. Therefore, the sound template is a representation of the sound both in time and in frequency.

To implement this sound parametrization module, we reviewed the most popular techniques used in speech recognition and, based on several research reports, we chose a technique called mel-frequency cepstral coefficients (MFCC). The MFCC algorithm can create a compact representation of sound into a vector of few parameters. We tested the MFCC algorithm in the Matlab environment using the auditory toolbox developed by Malcolm Slaney [21] and then we implemented a C++ version based on his algorithm for the YARP environment.

Algorithm 1 Calculate MFCC

loop
- Window the data with Hamming window
- Apply Fast Fourier Transform
- Compute the magnitude of the FFT
- Convert the magnitude into filter bank outputs
- Find the log_{10}
- Find the cosine transform to reduce dimensionality

end loop

Algorithm 1 shows the steps suggested by [21] to compute the MFCC transformation. In the next sections we explain in detail the parts of the algorithm.

A. Short-Time Fourier Transform (STFT)

The traditional approach to spectral analysis of the sound signal consists in applying a set of filter-banks (see [17]).

According to [17], the filter bank computation can be conveniently implemented by applying first a short-time fourier transform (STFT) to the incoming data:

$$S_n(e^{jw_i}) = \sum_m s(m)w(n-m)e^{-jw_i m} \quad (1)$$

where $s(m)$ is the sound sequence, and $w(n-m)$ is in our case a Hamming window.

The STFT produces a representation of the sound stream both in time and frequency domains that facilitates the application of the filter-bank in the frequency domain. Rabiner [17] proposes that the filter-bank can be implemented by varying adequately the frequency in the exponential term of equation (1); in the simplest case, this frequency has

an uniform distribution choosing $f_i = i(F_s/N)$, where F_s is the sampling frequency. However, non-uniform frequency distributions can be used; in particular, neurophysiological studies propose numerous models of the human auditory system. One of those is the mel-frequency scale where the filter-banks are distributed linearly in low frequencies and then they decrease logarithmically in higher frequencies. As suggested in [21], we constructed the filter-bank using 13 linearly-spaced filters (133.33 Hz between center frequencies) followed by 27 log-spaced filters (separated by a factor of 1.0711703 in frequency).

B. Mel-Frequency cepstral coefficients (MFCC)

The formula for the mel-frequency cepstral transform is as follows:

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

where c_i is the cepstral coefficient, and Y_k are the outputs of the filter-bank discussed in the previous section.

In our system, the MFCC transform reduces the dimensionality by transforming the output of 40 filter-banks into a compact representation of 13 cepstral coefficients. Figure 3 shows a graphical 3D representation of a MFCC transform applied to the sound produced by toy 2(a).

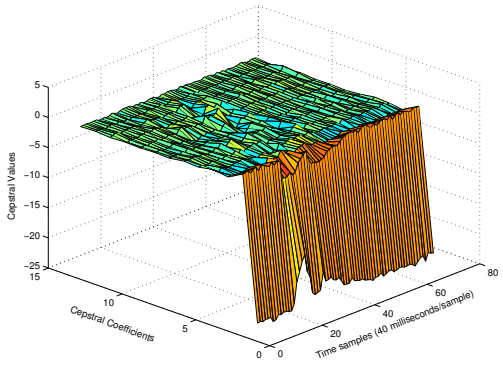


Fig. 3. Three dimensional representation of a MFCC Transform

After applying equation (2) we packed the cepstral coefficients in the sound template data structure. This template contains the cepstral coefficients associated to a sound produced by a toy. To detect the presence of an object producing a sound, we measure empirically the background sound level and we use it as a threshold to activate the template recording procedure.

V. Multisensory object segmentation (Synaesthesis)

Once the sound is parameterized, the level of *synchrony* of the sound and visual data streams needs to be measured. For this purpose, we use the method suggested by Hershey and Movellan based on the *mutual information* [6].

A. Mutual Information

Hershey and Movellan define the temporal synchronization of a video and sound channels as an estimate of the mutual information between both streams. Their algorithm was originally applied to the problem of finding a vocalizing person in a video sequence [6]. They consider that $a(t) \in R^n$ is a vector describing the acoustic signal at time t and that $v(x,y,t) \in R^m$ is a vector describing the video signal at the same time instant. They assume that these vectors form a set S of audio-visual vectors and that these vectors are independent samples from a joint multivariate Gaussian process. Under these assumptions, Hershey and Movellan affirm that an estimate of the mutual information can be calculated as

$$I(A(t_k);V(x,y,t_k)) = \frac{1}{2}\log_2 \frac{|\Sigma_A(t_k)||\Sigma_V(x,y,t_k)|}{|\Sigma_{A,V}(x,y,t_k)|} \qquad (3)$$

where $|\Sigma_A(t_k)|$ is the determinant of the covariance matrix of the audio stream, $|\Sigma_V(x,y,t_k)|$ is the determinant of the covariance matrix of a pixel of the image (e.g. the RGB values), and $|\Sigma_{A,V}(x,y,t_k)|$ is the joint covariance of both the audio and visual signals (see [6] and [20] for details about how to derive (3)).

To compute equation (3) different sound and images parametrizations can be used. In a first experiment, we calculated equation (3) using 13 mel-frequency cepstral coefficients (the parameters of covariance matrix $\Sigma_A(t_k)$) and three RGB values of the pixel (the parameters of the covariance matrix $\Sigma_V(x,y,t_k)$) during 0.6 seconds ($S = 15$). Consequently, the combined audio-vision covariance matrix $\Sigma_{A,V}(x,y,t_k)$ comprises 15x15 elements. The computation of the determinants of these matrices exhibits a considerable computational cost, because the determinants are calculated for each pixel in the image. This produces a considerable degradation of the system performance. Although this algorithm can be improved by having a distributed computation, we decided to use a simplified version of the mutual information as suggested by [6]. This is the special case when the data streams are in a one dimensional representation (i.e. $n = m = 1$). Then, the mutual information can be expressed as

$$I(A(t_k);V(x,y,t_k)) = -\frac{1}{2}(1-\rho^2(x,y,t_k)) \qquad (4)$$

where $\rho^2(x,y,t_k)$ is the Pearson correlation coefficient between $A(t_k)$ and $V(x,y,t_k)$ (see [15]). To obtain this one dimensional representation, we used for the sound the root mean square (RMS) of the short-time Fourier transform coefficients (see the arrow connection between the STFT box and the RMS box in figure 1) and a gray level value of the color RGB components. Notice that the MFCC transform was still used to form the sound template representation.

B. The Mixelgram

To conceptualize the output of the mutual information between sound and vision, Prince et al. [16] introduced the *mixel*; that is a combination of the words **m**utual and p**ixel**. They proposed that the mixels form a topographic representation called *mixelgram*. These can form shapes that are perceptually relevant for human observers [16]. Therefore, the mixelgram is to be considered a common space representation for both visual and audio sensorial channels.

Figure 4 depicts an example of the mixelgram of the toy 2(a). It is possible to distinguish the shape of the duck.

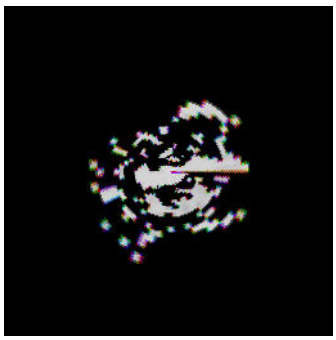


Fig. 4. The mixelgram of the duck toy. Notice that the mixelgram inherits the same log-polar geometry used in the original image.

C. Improved object segmentation

The original image and the mixelgram maintain a direct geometric correspondence, therefore the mixelgram can be used to segment the object by segmenting the pixel in the original image which position corresponds to an activated mixel. However, the segmentation obtained with this method has a low quality because many pixels of the object are not segmented at all. To improve the object segmentation, we use a technique based on color segmentation. We assume that the activated mixels belongs to a uniformly colored object. Then, we use a back-projection technique to improve the segmentation results.

The back-projection technique is implemented as follows: (i) the pixels segmented with the mixelgram are used to create a HS (Hue-Saturation) histogram, (ii) the HS histogram provides information about the object predominant color; then by applying a convenient threshold, the HS histogram can be used to segment by color the object in the original image (back-projection), (iii) the pixels segmented using the back-projection technique are then combined with the pixels segmented by the mixelgram to create an improved segmentation of the object.

Our implementation is similar to that described in [12]. However, in our system, the object is originally detected using the mutual information and we do not use a model of the background to segment the object.

As an example, figure 5 shows the HS histogram for the segmented object 2(b).

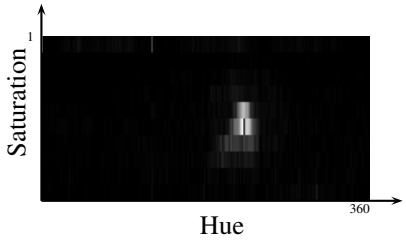
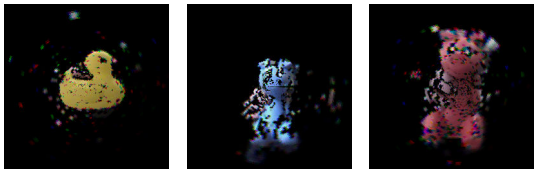


Fig. 5. The resulting HS histogram of the segmented blue pig toy

In figure 6 we present the results of the discussed segmentation process for the three toys used in the experiment. These were among the best segmentations obtained during the present experiment.



(a) (b) (c)

Fig. 6. The segmented toys

D. Associative memory

After an object is segmented, the segmentation results are stored in a dynamic lookup table. Each element in the lookup table contains the segmented image and the sound template associated to that object. To create the memory we appear the object in front of the robot several times squeezing or shaking the object with different speeds and strengths. This way, we produced slightly different sounds that were associated to the same object in the memory. This provided some robustness to the process of recognizing the sound.

VI. Attentional priming (Synesthesis)

This module performed basically a pattern classification for sound identification. When the system hears an unknown sound, the sound is parametrizated using the MFCC algorithm explained in section IV. Then, the sound template is compared with the memorized sound templates using a measure of similarity (distance).

To compare the sound templates it is necessary to compute both a local distance measure between the spectral vectors, and a global time alignment procedure [17]. To compute the local distance, we used the truncated cepstral distance $d_c^2(L)$ (see [17] page 195).

Experiment	Duck	Blue pig	Red pig
Segmentation (Synaesthesis)	64%	70%	75%
Classification (Synesthesis)	99%	88%	83%

TABLE I

SEGMENTATION AND RECOGNITION RESULTS FOR THE SYSTEM

A. Dynamic Time Warping

The global time alignment procedure is necessary because the automatic sound recognition system has to take into account: a) time alignment and b) time normalization. This can be done using a Dynamic Time Warping (DTW) algorithm [17]. We used the DTW algorithm to compare the heard sound to those stored in the associative memory discussed in previous section. During the experiment we produced these sounds outside the robot field of view. The system was able to recognize the sound and *remember* the object image associated with the sound. Then, the recovered toy image was presented to the experimenter for verification.

VII. RESULTS AND DISCUSSION

Table I presents the empirical results obtained during the presented experiment. In the case of the segmentation results, the table shows the percentage of segmentation trials with similar results of those presented in figure 6. Because a color segmentation is used, lighting conditions influence the segmentation. The results presented were obtained with good lighting conditions.

In the case of the sound recognition results, we did the experiment in a quiet laboratory environment with only some computer generating background noise. The results in both cases degraded significantly when we performed the experiment in noisy conditions, as for example, with people talking in the room.

For the recognition module we used only the $c_1 \dots c_{12}$ cepstral coefficients. The use of the c_0 cepstral coefficient degraded the capacity of the system to distinguish between similar objects. This was the case with the two pig toys that are made of the same material. This result make us suggest that the c_0 cepstral coefficient could be used to implement an algorithm to distinguish *classes* of objects. This may be convenient when the classification needs to be done among a big number of different sounds.

VIII. FUTURE WORK

This paper has presented the first steps to endow a humanoid robot with an attention priming mechanism based on crossmodal expectations. The final goal is to have an active robot with robust segmentation and classification algorithms that can actively create its own associations. Therefore, future work will include : robustness improvement, integration with sound orienting behavior algorithms and integration with other sensor modalities (e.g. touch).

ACKNOWLEDGMENT

We thank the support of QSSL through the QNX Educational support program. The work presented in this paper has been partially supported by the european FP6 project RESCUER (IST-2003-511492).

REFERENCES

[1] J.Y. Aloimonos, I. Weiss, and A. Bandopadhay. Active vision. *International Journal on Computer Vision*, pages pp. 333–356, 1987.

[2] A. Arsenio. *Cognitive-Developmental Learning for a Humanoid Robot: A Caregiver's Gift*. PhD thesis, Massachusetts Institute of Technology, 2004.

[3] A. Arsenio and P. Fitzpatrick. Exploiting cross-modal rhythm for robot perception of objects. In *2nd International Conference on Computational Intelligence, Robotics, and Autonomous Systems*, pages p 15–18, Singapore, 2003.

[4] D. Ballard. Animate vision. *Artificial Intelligence*, 48(1):pp. 1–27, February 1991.

[5] Paul Fitzpatrick, Giorgio Metta, Lorenzo Natale, Sajit Rao, and Giulio Sandini. Learning about objects through action - initial steps towards artificial cognition. In *Proceedings of the 2003 IEEE International Conference on Robotics & Automation*, 2003.

[6] J. Hershey and J. Movellan. Audio-vision: Using audiovisual synchrony to locate sounds. *Advances in Neural Infromation Processing Systems*, 12, 2000.

[7] M.S. Landy, L.T. Maloney, and M. Pavel. *Exploratory Vision: The Active Eye*. Springer Series in Perception Engineering. Springer, 1996.

[8] D. Marr. *Vision: a computational investigation into the human representation and processing of visual information*. W. H. Freeman, San Francisco, 1982.

[9] Metta, Fitzpatrick, and al. *Yet Another Robotic Platform (YARP). http://yarp0.sourceforge.net/*.

[10] G. Metta and P. Fitzpatrick. Better vision through manipulation. *Adaptive Behavior*, 11(2):109–128, 2003.

[11] Giorgio Metta. *Babyrobot: A Study on Sensori-motor development*. PhD thesis, University of Genova, 1999.

[12] Giorgio Metta, Antonios Gasteratos, and Giulio Sandini. Learning to track colored objects with log-polar vision. *Mechatronics*, 14:989–1006, 2004.

[13] L. Natale, G. Metta, and G. Sandini. Development of auditory-evoked reflexes: Visuo-acoustic cues integration in a binocular head. *Robotics and Autonomous Systems*, 39(2):pp. 87–106, 2002.

[14] Lorenzo Natale. *Linking Action to Perception in a Humanoid Robot: A Developmental Approach to Grasping*. PhD thesis, University of Genova, 2004.

[15] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in C : The Art of Scientific Computing*. Cambridge University Press, second edition, October 1992.

[16] Christopher G. Prince, George J. Hollich, Nathan A. Helder, Eric J. Mislivec, Anoop Reddy, Sampanna Salunke, and Naveed Memon. Taking synchrony seriously: A perceptual-level model of infant synchrony detection. In *Proceedings of the Fourth International Workshop on Epigenetic Robotics*, 2004.

[17] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall Signal Processing Series. Prentice Hall, 1993.

[18] Anina N. Rich and Jason B.Mattingley. Anomalous perception in synaesthesia: A cognitive neuroscience perspective. *Nature Reviews — Neuroscience*, 2002.

[19] Giulio Sandini and Giorgio Metta. Retina- like sensors: motivations, technology and applications. *In Sensors and Sensing in Biology and Engineering*, 2002.

[20] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:pp. 379–423, 623–656, July, October 1948.

[21] Malcolm Slaney. Auditory toolbox. version 2. Technical Report 1998-010, Interval Research Corporation, 1998.