

Active audition using the parameter-less self-organising map

Erik Berglund · Joaquin Sitte · Gordon Wyeth

Received: 18 November 2005 / Accepted: 3 January 2008 / Published online: 31 January 2008
© Springer Science+Business Media, LLC 2008

Abstract This paper presents a novel method for enabling a robot to determine the position of a sound source in three dimensions using just two microphones and interaction with its environment. The method uses the Parameter-Less Self-Organising Map (PLSOM) algorithm and Reinforcement Learning (RL) to achieve rapid, accurate response. We also introduce a method for directional filtering using the PLSOM. The presented system is compared to a similar system to evaluate its performance.

Keywords Active audition · Self-organisation

1 Introduction

For a robot communicating with humans or navigating the physical world, hearing is important. In communication, it contains a large portion of the information. In navigation, it informs of landmarks and warns of dangers which may be outside the field of view. Merely detecting sound is not enough, however—the direction of the source of the sound is also important. This is clear in the case of navigation but

equally true for communication where for example a message such as “come here” has no meaningful content unless the direction and distance to the source is, at the very least, estimated. Sound information is also relatively cheap in terms of computing power. Unlike vision, which uses sensors producing two-dimensional data, audition relies on sensors producing one-dimensional data. While vision offers great angular resolution, audition offers superior temporal resolution. One problem with audition is the microphones which, along with their associated circuitry, draw power, add weight and add a possible point of failure. Therefore, it is desirable to have as few microphones as possible. The system described in this paper uses two—this is the number used in nature and the smallest with which one can expect to detect direction.

Since the direction sensitivity of a two-microphone system is greatest along the plane perpendicular to the axis between the microphones (Bregman 1990), we can increase accuracy by moving this plane towards a sound source. This is the central idea of active audition (Reid and Milios 1999, 2003); to arrange the articulated microphone platform (such as a robot head) so that one maximises the information content available from sound.

This paper presents a method which solves the active audition task using the Parameter-Less Self-Organising Map (PLSOM) algorithm (Berglund and Sitte 2006) to find patterns in the high-dimensional features extracted from sound data, and associate these patterns with an appropriate motor action. This makes it unnecessary to have an explicit model or any *a priori* knowledge of the robot or environment acoustics. The presented method also utilises the short-term movement of the robot to achieve localisation of the sound source in three dimensions, a first for two-channel robotic systems. In order to do so the system incorporates a larger

E. Berglund (✉) · G. Wyeth
School of Information Technology and Electrical Engineering,
University of Queensland, Brisbane, Australia
e-mail: berglund@itee.uq.edu.au

G. Wyeth
e-mail: wyeth@itee.uq.edu.au

J. Sitte
School of Software Engineering and Data Communication,
Queensland University of Technology, Brisbane, Australia
e-mail: sitte@qut.edu.au

number of sound features than previous systems, including the new Relative Interaural Intensity Difference (RIID).

A system for separating sound signals from noise using two microphones based on direction is presented, implemented, tested, and compared to existing methods.

Section 2 will examine earlier works in the field of active audition and robot audition. Section 3 gives a brief system overview, before Sect. 4 deals with the system in more detail and Sect. 5 gives implementation details, experimental results and evaluation. Section 6 summarises the findings and Sect. 7 concludes the paper.

2 Earlier works

It is important to separate passive audition and active audition.

2.1 Passive audition

Passive audition is well understood, and has roots dating back to the works of Lord Rayleigh (J. W. Strutt 3rd Baron Rayleigh 1896) on human audition and the properties of sound. Since sound is a wave phenomenon many of the same principles are employed in radio transmission, radar, and sonar research. We will here give a brief overview of how it applies to robotic audition. Robotic audition, which is concerned with detecting sound and inferring the location and nature of the sound source, is distinguished from speech recognition, which is about retrieving the message encoded in the signal.

Most passive audition systems rely on three or more microphones. For example (Huang et al. 1995, 1997) presents a system for sound source localisation and separation based on three microphones. The system uses event detection and time delay calculations to provide a direction in the horizontal plane and can separate sources based on direction. Other examples are Rabinkin et al. (1996a) or Guentchev and Weng (1998). The latter is interesting in that it uses four microphones arranged as vertices in a tetrahedron to achieve three-dimensional localisation. Numerous other approaches to the direction detection problem have been proposed. Some examples are mask diffraction (Ge et al. 2003), formant frequencies (Obata et al. 2003), and support vector machines (Yamamoto et al. 2003). Use of passive audition to detect snipers during warfare has been suggested (Rabinkin et al. 1996b).

Some methods use integration of vision and audition in combination with Self-Organising Maps (SOMs). Prominent examples are those of Rucci et al. (1999) and Nakashima et al. (2002), which both rely on time delays and visual feedback to train the auditory system and pinpoint the location of the target.

The works of Rucci et al., Konishi (1993), and Day (2001) are inspired by the auditory system of the Barn Owl (*Tyto alba*). Rucci et al. recreated the brain pathways of the owl computationally with a high degree of fidelity. The resulting framework was connected to a robot which, assisted by visual cues, learned to orient towards the sound source with an error of approximately 1°. This work relied on visual feedback to train its motor response; the visual feedback was provided in the shape of a clearly distinguishable target (a small electric lamp that was always associated with the active speaker) and image processing software that determines whether the target was in front of the robot or off to the sides, issuing a 'reward' to the system based on this. Rucci et al. focus on the Interaural Time Difference (ITD) only, employing a relatively large distance between the microphones (300 mm) and pre-determined static networks. The mapping between the central nucleus of the inferior colliculus to the external nucleus of the inferior colliculus can be changed during training, but is partially pre-determined.

Elevation detection has long been the exclusive domain of multi-microphone systems. Detecting elevation by two microphones has long been assumed to be dependent on spectral cues produced by diffraction in the pinna (Blauert 1983; Shaw 1997; Kuhn 1987) and several studies in modelling these cues have been performed based on Head-Related Transfer Functions. Later research (Avendano et al. 1999) has identified elevation cues related to low-frequency reverberation in the human torso. These studies have in common that they focus on recreating a realistic spatial listening experience for humans, not on source location detection for robots.

Robotic work in this field has mainly concentrated on using multiple microphones with good results (for example, Tamai et al. 2004). Some recent applications (for example, Kumon et al. 2005) have created artificial pinnae with well-understood acoustics in order to use two microphones and spectral cues for elevation detection. This is thought to be analogous to how many animals, including humans, detect elevation.

2.2 Active audition

Active audition aims at using the movement of the listening platform to pinpoint the location of a sound source in the same way biological systems do. Active audition is therefore an intrinsically robotic solution to the problem of sound direction detection.

The term *active audition* appears to have originated in 1999 with Reid et al. (Reid and Miliotis 1999, 2003), who described a system using two omnidirectional microphones on a pan and tiltable microphone assembly. Two sets of possible directions to the sound source were computed using time delay and two different positions of the platform. The correct source direction was then the intersection of the two

sets; the system used its own movement to detect direction. Reid et al.'s approach was therefore different from approaches that only used the bearing to the sound source to direct movement. Apart from this early example there have been few instances of active audition research. One important exception was the SIG group, which started research into the topic around 2000 (Nakadai et al. 2000a).

The SIG method also relied on synergy of visual and auditory cues for direction detection (Nakatani et al. 1994; Nakadai et al. 2000a, 2000b, 2001, 2002a, 2002b, 2003a, 2003b, 2003c; Kitano et al. 2002). The output of this group represents the main body of active audition research to date, most of which has been conducted in the last five years. The SIG Humanoid project was (it has now been supplanted by the SIG 2 project) a large multidisciplinary cross-institution project aimed at creating a robotic receptionist capable of interacting with several humans at once in a realistic (noisy) office environment. The system used the Fast Fourier Transform (FFT) to extract interaural phase and intensity difference (IPD and IID, respectively) from a stereo signal.

Combinations of SOMs and Reinforcement Learning have been explored by (Sitte et al. 2000; Iske et al. 2000).

3 Overview of presented system

A robotic audition system should have the following qualities:

- Effective with just two microphones placed relatively close together; this reduces weight, power consumption, cost, and the risk of hardware failure. It also places fewer constraints on the size and shape of the robot, and the placement of the microphones.
- Distance estimation capabilities without relying on familiarity with the signal; so that the robot can determine the distance to a sound source it has not encountered before.
- Elevation estimation capabilities; in order that the robot can determine the elevation of a sound source.
- Robust to environmental changes; so that post-processing to compensate for environment is not necessary, thus reducing complexity and increasing flexibility.
- Flexible and easily adaptable to diverse architectures; so that the same software and/or hardware can be used on multiple platforms, thus reducing development time and cost.

This section gives a brief overview of such a system and discusses some of the obstacles that must be overcome in order to implement it. The system includes as many features as possible to maximise the available information and counteract the frequency-limited range of the Interaural Phase Difference (IPD) and Interaural Intensity Difference (IID). Therefore, both IPD and IID, as well as ITD and RIID, are

included. The calculation of the IPD, IID, and RIID features are extracted by the aid of the FFT. The ITD is calculated by cross-correlation of time-shifted copies of the right and left sound buffers. The IID, RIID, IPD, and ITD will be discussed in detail in Sect. 4.2.

The pre-processed features are combined into one feature vector which the data association strategy can use in order to make a decision about how to move the robot or convert into information about the location of the sound source.

By calculating the direction for each subband of the FFT, one can filter the entire signal based on the direction.

3.1 Feature selection rationale

Some of the features selected seem to duplicate the same information, yet on closer examination they are quite different.

ITD and IPD seem identical on a cursory examination, but it is important to realise that ITD gives only *one* value per timestep where the IPD gives one value per timestep *and subband*. This is especially important because the profile of how the IPD changes relative to the frequency contains important clues to the direction. The ITD, on the other hand, is not limited to $\pm\pi$ radians, and gives a more robust yet less accurate value. In short, the ITD is frequency-independent; the IPD is frequency-dependent. The ITD is also better at localising transient sounds, while the IPD requires a sound to last the entire duration of the FFT input buffer to be reliably and accurately detected.

IID and RIID also seem identical at first glance but the reader should note that the RIID is necessary for distance detection but, since it is an absolute value, does not distinguish between left and right. The RIID must be an absolute value to prevent the distance resolution from inverting around the central axis.

Overall the aim is not to constrain or minimise the amount of data extracted from the sound signal but to provide sufficient data for the system to work with.

3.2 Data association strategy and possible obstacles to implementation

Once the features are extracted the task is to extract the salient information from them. In other words: how can the data be categorised into practical groups that can be used for decision making? This is a complicated task since the input contains large amounts of irrelevant data such as frequency distribution, intensity and intensity distribution. An ideal learning system will focus on those properties in the training examples that change between training examples and ignore those that stay constant. In that way it becomes easy to train the system by simply providing training data

that exemplifies change in the property the system should learn.

The learning system for this task must deal with the complexity of the data. Even at low sample rates and small FFT sizes, there will be hundreds of variables to be evaluated and correlated with each other. There is also the problem of providing sufficient training examples. In order to avoid recording samples from every possible location and every possible environmental condition, an ideal system would be able to generalise from few training examples. Although systems based on theoretical models have achieved good results, it would be preferable to have a model-less system or a system that develops its own model based on experience to avoid the need for a detailed analysis of the implementation platform's acoustics for each new implementation.

This leads in the direction of neural network based solutions. Traditional neural network algorithms (such as backpropagation) and architectures (multilayer perceptrons, feedforward networks) are very good at learning similar problems, but here the added constraint on the number of training examples must be taken into account. Training networks by using backpropagation also requires the training data to be correctly labelled.

It would therefore be a good idea to use a system that can self-organise, such as the Self-Organising Map (SOM). The SOM can reduce the dimensionality of a data set by finding low-dimensional manifolds embedded in the high-dimensional input space, which is exactly what is required here. The position of the sound source (three dimensions) is embedded in the input space (hundreds of dimensions). The input space is the set of all possible input combinations to a system. The data in a feature vector represents one point in the input space, the key to understanding the input space is understanding the feature vector. The contents of the feature vector is described in Sect. 4.2.

The problem with the SOM is that there is no firm theoretical basis for selecting the training parameters, so that the

search for the correct values becomes an empirical search through a four-dimensional space. The SOM also requires a large number of iterations to achieve a good mapping, which is a problem in this case since the preprocessing of each training example requires computation time. The combination of time-consuming training and the need for several training sessions in search for the optimal parameters means that the SOM-based approach will be slow.

Finally, the size of the input space and the relatively few training examples makes it unlikely that there is an even distribution of training examples, which will make the SOM clump weights in certain areas. This is undesirable since the system needs a continuous generalisation that covers the entire possible input space.

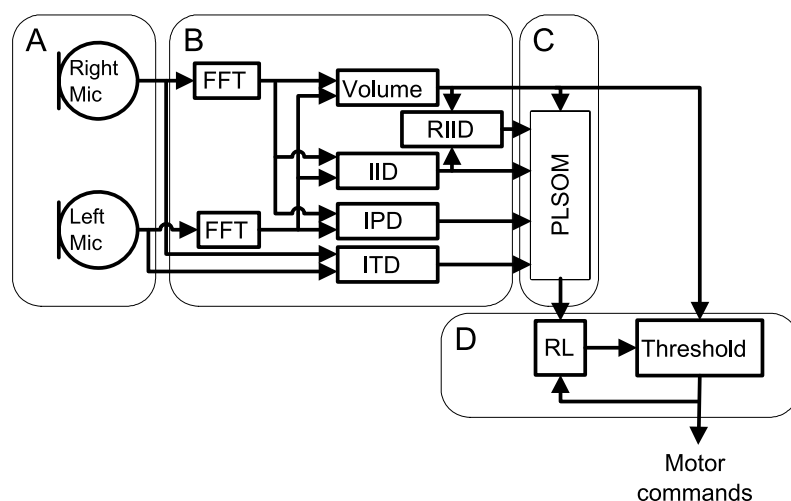
4 Active audition with PLSOM

This section describes a general active audition system capable of using any robot with two audio input channels for sound source localisation. To emphasise that the system is not tied to one specific robot architecture, the implementation details and experiments are discussed separately, in Sect. 5.

4.1 Brief system overview

The system needs an articulated robot platform with two microphones to function, subsystem **A** in Fig. 1. The signals are processed and sound features indicating the relationship between the stereo channels are extracted, see subsystem **B** in Fig. 1. These features are the IID, RIID, IPD, ITD, and volume, which will be discussed in more detail in the next section. These features are combined into a feature vector and used to train a Parameter-Less Self-Organising Map (PLSOM), see subsystem **C** in Fig. 1. The PLSOM learns to extract the low-dimensional manifold embedded in

Fig. 1 Active audition system layout. Sound information flows along the direction of the arrows



the high-dimensional feature vector, which corresponds to the position of the sound source. Once trained, the PLSOM starts passing the position information to a Reinforcement Learning (RL) algorithm, see subsystem **D** in Fig. 1. The RL subsystem uses the input from the PLSOM as feedback to learn to orient the robot towards the sound source.

4.2 Preprocessing

Four features are extracted from the sound signal, describing the difference in time, phase, intensity, and relative intensity of the two channels. In addition, the overall intensity is calculated.

For each sampling window, which is γ samples long, the system computes the FFT of the signal. The FFT divides the signal into its constituent frequency components, giving $\gamma/2$ components that approximately represent the original signal in the frequency domain. Each frequency component in the FFT output is here termed a subband—the FFT gives $\gamma/2$ subbands which can be averaged to fewer subbands to enable us to work with a smaller feature vector.

For each sampling window the system computes the ITD by simple offset matching on the original input signal in accordance with (1)

$$itd(L, R) = \arg \min_d \frac{\sqrt{\sum_{t=0}^{\gamma-2d} (L(t+d) - R(t))^2}}{(\gamma - 2d)} \tag{1}$$

where γ is the size of the FFT input buffer, L and R are array representations of γ samples of stereo sound, left and right channels respectively. d has a range that reflects the maximum possible time delay, measured in number of samples. The ITD could also be computed through cross-correlation.

For robots with low sample rates or small heads, and therefore small distances between the microphones, the ITD does not offer usable accuracy. As an example, consider a small robot with 70 mm between the microphones that uses a sample rate of 16000 Hz (an example of such a robot is given in Sect. 5). Since the speed of sound is 340 m/s, this gives a maximal time difference of only 200 μ s, this occurs when the sound source is at 90° to either side. At 16000 samples per second this is less than 3.3 samples difference between 0° and 90°, so the optimal theoretical accuracy available from ITD alone is roughly 30°.

As is well known, the FFT gives an array of complex numbers, one complex number $re(s) + i im(s)$ for each subband s . The real and imaginary parts are denoted $re_l(s)$ and $im_l(s)$ for the left channel, $re_r(s)$ and $im_r(s)$ for the right channel of subband s . The IID is computed according to (2).

$$ild(s) = \sqrt{re_l(s)^2 + im_l(s)^2} - \sqrt{re_r(s)^2 + im_r(s)^2}. \tag{2}$$

The IPD is computed according to (3)

$$ipd(s) = \begin{cases} a - b, & \text{if } |a - b| \leq \pi, \\ a - b + 2\pi, & \text{if } |a - b| > \pi \text{ and } a < b, \\ a - b - 2\pi, & \text{otherwise} \end{cases} \tag{3}$$

where a and b are given by (4) and (5), respectively

$$a = \arctan \frac{im_r(s)}{re_r(s)}, \tag{4}$$

$$b = \arctan \frac{im_l(s)}{re_l(s)} \tag{5}$$

and RIID according to (6):

$$rild(s) = \frac{ild(s)}{v(s)} \tag{6}$$

where $v(s)$ is the averaged volume of the left and right channel for subband s , computed according to (7):

$$v(s) = \frac{\sqrt{re_l(s)^2 + im_l(s)^2} + \sqrt{re_r(s)^2 + im_r(s)^2}}{2}. \tag{7}$$

IPD, IID, and RIID are based on the FFT so that there is one value for each of the $\gamma/2$ subbands. The ITD value is copied $\gamma/2$ times in order to lend it the same importance as the IPD, IID, and RIID in the PLSOM. Finally the ITD, IID, RIID, and IPD vectors are concatenated into one $2 \cdot \gamma$ -element feature vector x which, in addition to the volume v , is passed to the PLSOM.

4.3 PLSOM

The Parameter-Less Self-Organising Map has already been published independently (Berglund and Sitte 2003, 2006), therefore this section will only point out one modification that was applied relative to the standard PLSOM for the sake of improving sound processing performance. The PLSOM and other SOM algorithms depend on a distance measure to calculate the distance between each node and a given input; one example is the commonly used Euclidean norm. Unfortunately the Euclidean norm assigns equal importance to *all* parts of the input, which will be detrimental where subbands with low intensity have a high phase difference. To avoid this problem a slightly modified Euclidean norm was selected. When computing the distance between input and weight vectors, a weight based on the subband intensity is applied, substituting the Euclidean distance in the standard PLSOM algorithm with the norm:

$$\begin{aligned} \|x(t) - w_i(t)\|_v &= \sqrt{\sum_{s=1}^{2 \cdot \gamma} (x_s(t) - w_{i,s}(t))^2 v_{\text{mod}(s, \gamma/2)}(t)}. \end{aligned} \tag{8}$$

In (8) $w_{i,s}(t)$ is the weight of node i at time t in subband s , $x_s(t)$ is the input in subband s at time t . $v_{\text{mod}(s, \gamma/2)}(t)$ is the normalised (divided by the loudest volume up until time t) average volume of both channels in subband $\text{mod}(s, \gamma/2)$ at time t .

Notice that since x is the $2 \cdot \gamma$ -dimensional concatenation of the $\gamma/2$ -dimensional ITD, IID, RIID, and ILD vectors the $\text{mod}(s, \gamma/2)$ operator must be applied to the subband index s . Thus, subbands with high volume influence the winning node more than subbands with a low volume, and there is no need for discrete volume based thresholding.

4.4 Reinforcement learning

The output of the PLSOM (the position of the winning node) is used by the RL subsystem to turn the robot head towards the sound source, based on the hypothesis (which will be tested in Sect. 5.2) that a node in the centre of the map is the winning node when the sound source is directly ahead of the robot.

Reinforcement Learning (RL) (Sutton 1992; Mitchell 1997; Sutton and Barto 1998; Gosavi 2003) is a machine learning paradigm based on the concept of an *agent*, which learns through interaction with its *environment*.

RL is particularly well suited to applications where a strategy balancing long-term and short-term gains (reflected in the value and reward functions) must be developed without any predetermined model of how the environment will respond to actions.

5 Experiments and results

To test the system described in Sect. 4 it was implemented using Sony Aibo ERS-210 robots (known simply as ‘Aibo’ hereafter) (Fig. 2). This section describes the details of the implementation, the experimental setup and the results. The Aibo robots used in the experiments were standard except for the addition of a wireless networking card.

The physical system consists of two parts:

1. The SONY Aibo robot—a dog-like quadruped robot with stereo recording capabilities. Several models exist, the experiments described in this paper were all carried out on ERS-210 models.
2. An ordinary desktop PC.

This robot was connected via a wireless link to the PC. Sound samples were transferred from the Aibo to the PC, on which the sound processing took place. Motor commands from the system were passed back to the Aibo.

The sound was transmitted as 16 bit/channel, 16000 samples/second linear Pulse Code Modulated (PCM) stereo data. The directional response of the Aibo can be seen in



Fig. 2 One of the Sony Aibo ERS-210 robots used in the experiments. Note that the artificial auricles are mounted separately from the microphones (the three horizontal slits)

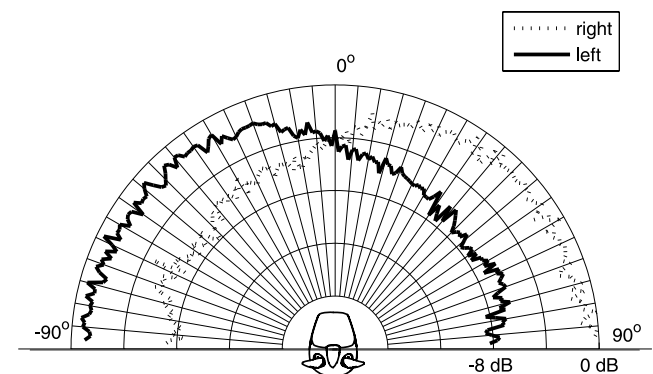


Fig. 3 Directional response of the Aibo microphones. The graph shows sensitivity as distance from the robot head; greater distance means greater sensitivity. Test conducted with speaker playing 79 dB white noise at 0.1 m in a room with a background noise of 50 dB and reverberation time (T_{60}) = 0.4 seconds. Note the increased sensitivity around ± 60 – 70°

Fig. 3. It is interesting to note that the directional response was not the same for all frequencies. At 800 Hz the directional sensitivity was almost completely opposite of that for white noise, as can be seen in Fig. 4. This would pose a serious problem for direction detection systems only based on intensity difference.

5.1 Experimental setup

During the experiments described below a PLSOM with 18×5 nodes and a neighbourhood size of 11 was used. The FFT input buffer size, γ , was equal to 512.

The 256-dimensional output was averaged to 64 dimensions to give us a smaller data vector and therefore smaller SOM. It was decided to not simply use a smaller γ since that would have constrained the frequency range of the system too much. Furthermore, the Aibo sends samples to the

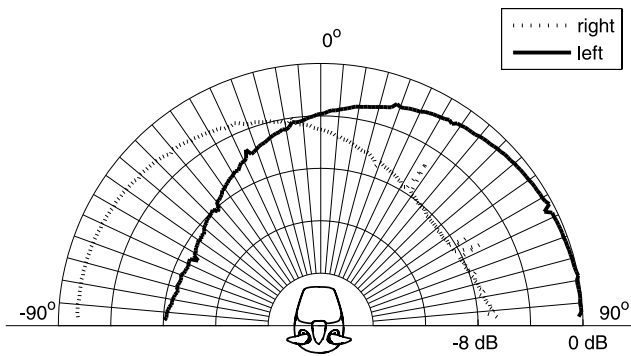


Fig. 4 Aibo sensitivity response to 800 Hz sine wave under the same conditions as in Fig. 3. Note that the sensitivity is switched along the central axis compared with that for white noise. This is caused by the sound source interfering with itself in the near field

PC in 512-sample packets. The maximum range of d in (1) was limited to 12.

Before application of the FFT the input signal was normalised over each 512-sample window. After application of the FFT the first subband was discarded, since it was almost pure noise. The signal was then downsampled before computing the IPD, IID, and RIID.

The logarithm of the ITD values was normalised by dividing by the largest ITD so far and then multiplied by 0.25. The IPD was normalised. The IID was normalised and multiplied by the volume of each subband. The RIID was normalised. The volume (defined in (7)) was normalised. The ITD, IPD, IID, RIID, and volume was combined into one feature vector, which was not normalised.

In order to train the PLSOM a number of sound samples were recorded; white noise samples from 38 locations in front of the robot were used. The samples were recorded with the speaker at 0.5 and 3 m distance from the robot, with 10° horizontal spacing, ranging from -90° to 90° . For each training step the PLSOM training algorithm presented 1.984 seconds of a randomly selected sample to the system, off-line. This was repeated for 10000 training steps. This sensitised each node to sound from one direction and distance, as shown in Fig. 5, and typically completed in less than eight hours on an entry-level (in 2005) desktop PC. The sound environment was a quiet office setting, noise sources included air-conditioning and traffic noises but no human voices. The floors were carpeted and the ceiling was ordinary office noise-absorbing tile. The two closest walls were hard concrete and glass. Reverberation time (T_{60}) was approximately 400 ms and background noise was close to 50 dB. This corresponds to data set A, to be discussed in Sect. 5.1.1. The physical setup was similar for all data sets, see Fig. 6 for an example.

The reinforcement learning was done online, with the robot responding to actual sound from a stationary speaker playing white noise in front of it. Initially the robot head was

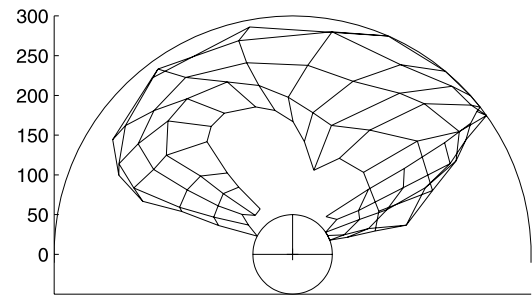


Fig. 5 Plot showing the input position each node was most sensitive to on average, relative to the head. Each line intersection represents a node. The semi-circle represents 3 m, the circle represents 0.5 m. Note how the largest PLSOM dimension was allocated to the most prominent feature of the sound

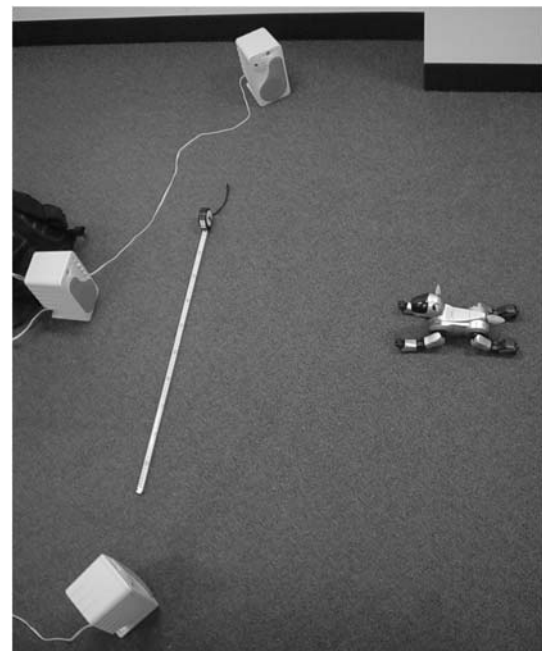


Fig. 6 A typical recording setup. The tape measure was extended to one meter for scale. This particular picture shows the recording of data set H, see Sect. 5.1.1

pointed in a randomly selected direction, and the RL training progressed until the robot kept its head steadily pointed towards the speaker, at which time a new direction was picked by random and the RL training continues. The RL training reached usable results in 2000 iterations, which typically took less than 20 minutes.

5.1.1 Data sets used throughout this paper

In Table 1 the reverberation time is given as T_{60} , which refers to the time it takes the reverberations of a sound to diminish by 60 dB. Data set A was used in all training, the other data sets were only used for testing.

Table 1 Datasets used for training and testing

Dataset	Source elevation	Source distance	Source direction	Source direction increments	Signal type	Signal intensity	Noise intensity	T60	Head inclination
A	0°	0.5 m, 3 m	−90° to 90°	10°	White noise	65 dB	50 dB	0.4 s	0°
B	0°	1 m	−88° to 88°	1°	White noise	77 dB	58 dB	0.5 s	0°
C	0°	1 m	−88° to 88°	1°	Speech	65 dB	58 dB	0.5 s	0°
D	0°	1 m	−88° to 88°	1°	800 Hz sine wave with two harmonics	77 dB	58 dB	0.5 s	0°
E	0°	0.2, 0.7, 1, 2.5, and 5 m	−90° to 90°	10°	White noise	63 dB	50 dB	0.4 s	0°
F	0°, 40°	1 m	−70° to 70°	10°	White noise	65 dB	50 dB	0.4 s	0°
G	0°, 40°	1 m	−70° to 70°	10°	White noise	65 dB	50 dB	0.4 s	0°, 40°
H	0°	1 m	−88° to 88°	1°	White noise	65 dB	50 dB	0.4 s	0°

5.2 Horizontal direction detection

This experiment was intended to measure the accuracy of the direction detection property of the system. The presented system was expected to map the direction to the source along the longest axis of the PLSOM. The horizontal index of the winning node in the map should therefore be an indication of the direction to the source in a way such that similar directions (for example 10° and 0°) were mapped to nodes close together along the horizontal axis (for example node number five and seven). This must be the case in order for the reinforcement learning algorithm to work. For the same reason it was important for the presented system that sounds from straight ahead are mapped to the middle of the PLSOM. Two experiments were carried out:

1. A speaker playing white noise was placed in front of the robot at a distance of approximately one meter. The sound intensity was approximately 70 dB (C) with background noise of 58 dB and reverberation time (T60) is 0.4 seconds. The setting was a normal office with carpet floors and air conditioning running. There were hard reflective surfaces to the left and right of the robot at 1.8 metres and behind and in front of it at 2.5 metres. The output of the system was recorded. The speaker was moved slowly to the left (as seen from the robot) until the system output stabilised at a new value. The angular displacement was recorded. This was repeated five times, and the mean average difference was computed.
2. A set of white noise sound samples was played in the same setting as described above. The angle was varied from −88° to 88° in 1° increments through turning the robot head, resulting in 177 samples, corresponding to data set B. The position of each sample was evaluated by the system, and the output was plotted, along with standard deviation, versus the actual angle in Fig. 7.

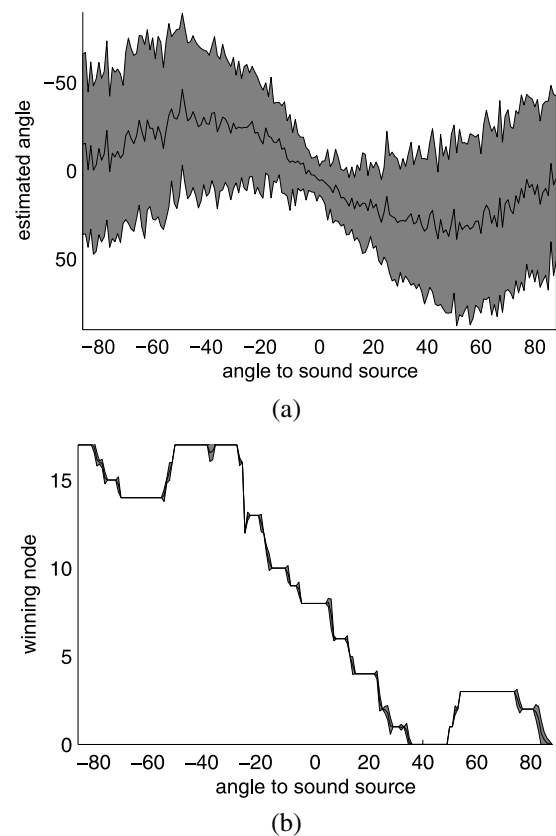


Fig. 7 The grey area is one standard deviation scaled to match the axis units (a) Estimated angle vs. actual angle using the SIG method. Note the large deviation. (b) Average winning node vs. actual angle using the PLSOM method with 18×5 nodes. Note the small deviation

This was performed with the presented system and a well-known active audition system, the SIG system. The first experiment is designed to test the sensitivity of a active audition system in its most sensitive area, while the second experiment is designed to test a system's ability to distinguish a wide selection of sound directions.

5.2.1 SIG system

In order to provide a performance baseline for system evaluation, a version of the SIG system as described by Nakadai et al. (2002a) was implemented. The SIG group reported an accuracy of 5° but when the published method was implemented on an Aibo robot a somewhat lower accuracy of 10° was achieved in experiment one. It was hard to measure this accuracy exactly given the large standard deviation of the estimated angle, which made it hard to note when the system settled on a new stable value. This discrepancy can have several explanations:

- The SIG Humanoid robot had high resolution input; the Aibo used a low resolution sampling rate of 16 kHz.
- The shape of the head of the Aibo robot was highly non-spherical, making calculating time delays (and therefore the idealised IPD) between the ears difficult (Nakadai 2004).

In experiment two the SIG system performed as indicated by Fig. 7(a), note the large standard deviation.

5.2.2 PLSOM system

The PLSOM system was trained with data set A, and consistently managed an accuracy of approximately 5° in experiment one. This is comparable to human acuity (Moore 1997), especially when considering the low resolution data available. The results of experiment two are displayed in Fig. 7(b). The system showed a clear correspondence between winning node and source direction.

These results indicate that in the -20° to 20° window, the PLSOM goes through eight different winning nodes, giving about 5° per node.

5.2.3 Comparison of SIG and PLSOM systems

The PLSOM was almost free of deviation. This enabled the system to estimate the direction using a small number of samples, in other words; quickly. The SIG method on the other hand had to average over a larger number of samples to achieve a reliable estimate, thus increasing response time. This occurs despite both systems measuring angle in discrete steps.

5.2.4 Further tests with the PLSOM system and discussion

Labelling the PLSOM nodes in order to give an estimated direction in human-readable terms (angles or radians) is not necessary and indeed not desirable for robotic applications. Nevertheless, in order to give an intuitive idea of the performance of the direction finding algorithm a labelling step was

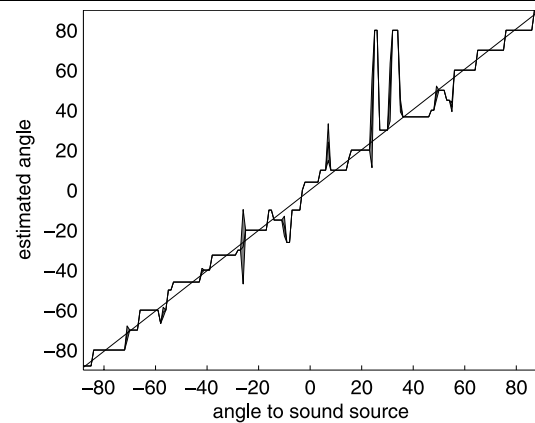


Fig. 8 The estimated angle of a labelled PLSOM vs. the actual angle. An idealised straight line has been inserted for reference. The grey areas (barely visible) represent one standard deviation

performed in which each node in the PLSOM was associated with the direction it was most sensitive to, as described below. The result can be seen in Fig. 8.

The sound data set used in the labelling was the same as the training set, which was different from the test set. The labelling was performed in the following way. The PLSOM was trained as in previous experiments, using data set A, and training was then disabled. Each of the samples in data set A was then played to the system in turn. For each sample the winning node was noted. At the end of the session the nodes were labelled with the mean average of the angles for which it was the winning node. Since the training set only contains samples for every 10° , some nodes were never winning nodes. These nodes were given labels formed from interpolating between the two closest nodes with labels. The accuracy was then tested as before, using data set B.

Note that while it seemed like the labelling resolved the ambiguity around large angles seen in Fig. 7(b), this was only because some of the horizontal displacement of the source is mapped along the vertical axis in the PLSOM. Indeed, the labelling introduced a new inaccuracy around 30° since there were not enough labelling examples to label all nodes correctly (there were 90 nodes, but only 38 labelling examples).

The PLSOM was also tested with speech signals (data set C) and generated sine waves (data set D) to see what effect sound type has on accuracy. As can be seen from Fig. 9 the PLSOM performed only marginally worse with the speech signal despite the speech signal's intermittent nature and lower volume (65 dB instead of 77 dB).

With the generated sine wave, on the other hand, the results were not so good. This is not surprising on a largely frequency-domain based method, as the spectrum of a nearly pure tone contains little information. It should also be noted that the recording distance (one meter) was very close to the near field (depending on which definition one uses). The

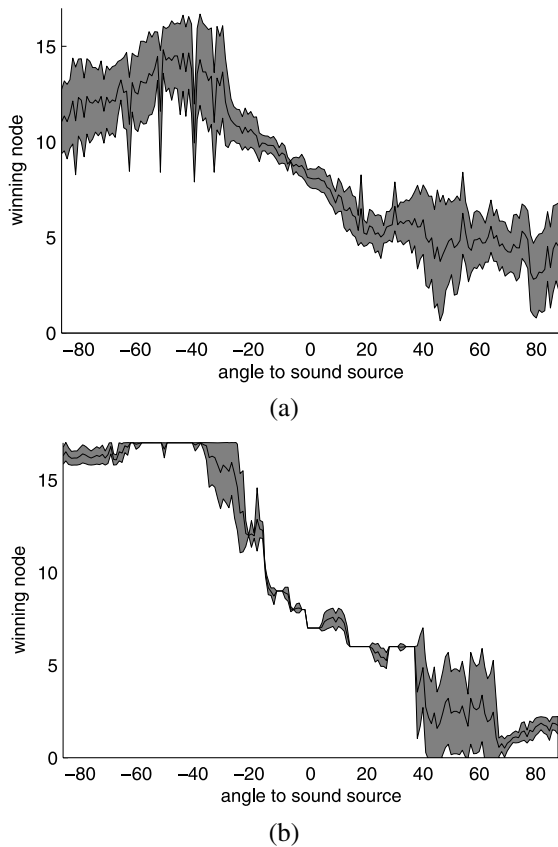


Fig. 9 The grey area was one standard deviation, scaled to match the axis units. **(a)** Average winning node vs. actual angle using the same PLSOM as in Fig. 7(b). Here the test data was speech, the sentence “She had your dark suit in greasy wash water all year” from the TIMIT database. **(b)** Average winning node vs. actual angle using the same PLSOM as in Fig. 7(b). Here the test data was a sine wave with two harmonic overtones. Base frequency is 800 Hz, intensity circa 77 dB at one meter

near field is the area close to a sound source where constructive and destructive interference can give rise to differences in amplitude not consistent with the inverse square law, especially when the signal consists of only one or a few frequencies. One factor in estimating the extent of the near field is the size of the transducer, which in this case was a CDSONIC JC-999 speaker with a maximum dimension of 210 mm. No clear consensus of the correct estimate of the extent of the near field seems to exist, with estimates ranging from 0.2 m to 1.28 m at 800 Hz. Some authors (e.g. Brungart and Rabiowitz 1996) give the near field as the region of space within a fraction of a wavelength away from a sound source. This would give a near field boundary at less than 0.425 m at 800 Hz. The measurements (see Fig. 3 and Fig. 4) indicated significant interference at this distance and frequency, which would go a long way towards explaining the relatively poor performance displayed in Fig. 9(b).

Still, comparing the plots for the three different recordings, as in Fig. 10, it was clear that the difference was small, especially in the most sensitive areas.

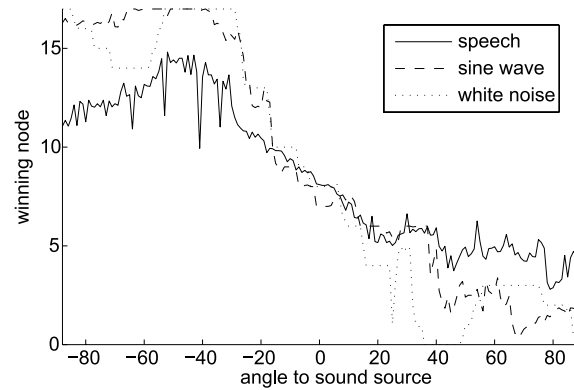


Fig. 10 The direction detection results for the three different sounds presented above overlaid in one graph. The PLSOM used was trained with white noise only

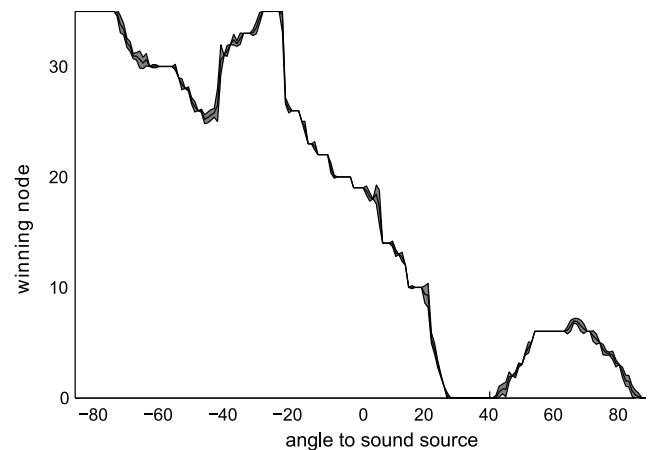


Fig. 11 Average winning node vs. actual angle using the PLSOM method with 36×10 nodes. The grey area is one standard deviation

The experiments show that the PLSOM system achieves greater levels of accuracy and speed than a well researched and documented method, the SIG group algorithm. The experiment was also run with larger networks to see the effect of size. Networks up to 36 nodes wide were trained without problems, as seen in Fig. 11. This gives a theoretical accuracy of 2.2° , although one should note the added deviation around large angles. One important observation from Fig. 7(b) and Fig. 11 was the area of great inaccuracy where the sound source is at approximately $\pm 60^\circ$ angle to the sagittal plane. These areas of low directional sensitivity coincided with the areas of highest distance sensitivity, see Sect. 5.5. The reason was that the PLSOM allocated most weights along the dimension that contained the most information. This hypothesis was supported by the arrangement of the weights in Fig. 5. Unfortunately, changing the shape of the PLSOM grid did not appear to alleviate the problem. Another cause for this deviation was the shape of the Aibo head. The microphones were placed relatively close to the rear of the head, so that when the angle between the head

roll axis and the direction to the sound source increased to $\pm 45^\circ$ the IID started decreasing instead of increasing because sound waves passed more easily behind the head than in front of it, see Fig. 3.

5.3 Horizontal source localisation

Since there is a correspondence between the winning node and the direction to the sound source, a reinforcement learning system should be able to orient towards the sound source. The system was expected to learn to orient the robot head towards the sound source using only the output from the PLSOM as feedback.

To determine how well the system outlined in Sect. 4.4 was able to orient the robot head towards the sound source the following experiment was conducted. The PLSOM was trained in the same way as described in Sect. 5.1. Then PLSOM training was switched off and the RL subsystem was turned on. A speaker is placed one meter in front of the robot playing white noise at approximately 77 dB. The RL subsystem was allowed to move the robot head. Whenever the PLSOM winning node stabilised in the middle for more than five iterations, the robot head was rotated to a randomly selected horizontal angle and the learning continued. The absolute difference between the direction of the robot head and the direction of the sound source was termed the *directional error* in the following. The directional error was recorded for each iteration only when the robot was not facing directly towards the sound source (as indicated by the output of the PLSOM). The rewards were as follows:

- 0.5 for when the output (the location of the winning node) of the PLSOM moved towards the centre.
- 0.1 for when the output of the PLSOM remained in the centre.
- -0.1 for output of the PLSOM did not change.
- -0.5 for all other cases.

These reward were designed to be positive for actions that were desirable (contributing to correct head orientation) and negative otherwise. The exact values were arrived at through trial and error. The robot had 14 actions to choose from; turning the head 70, 40, 30, 20, 15, 10 or 5 degrees to either side. The robot could also choose to keep its head still. This test was carried out off-line using data set B. In the off-line experiment, the system was running as it would in a real-world application, but instead of controlling a real robot it controlled a virtual robot that returned sound samples drawn from data set B based on the direction of the sound source and the position of the virtual robot.

The total averaged error of the learning algorithm was less than 0.26 radians and it was able to orient the head to within $\pm 5^\circ$.

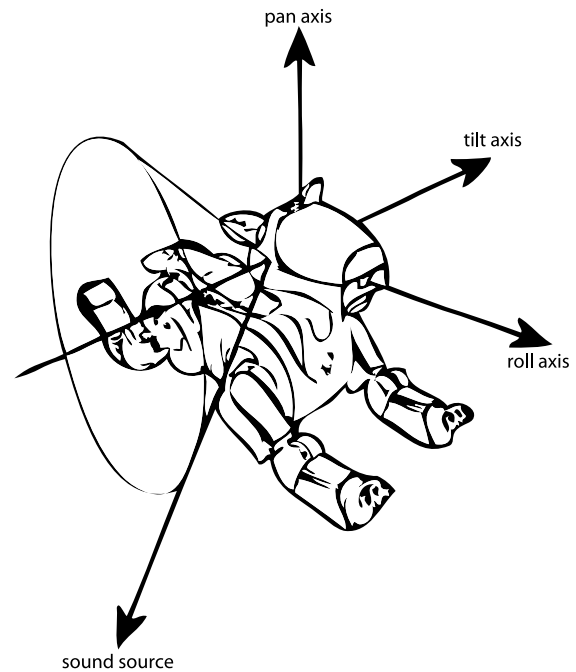


Fig. 12 The cone of confusion can be seen as a cone with its vertex between the microphones obtained by rotating the vector to the sound source around the tilt axis. The actual position of the sound source may lie anywhere on the cone

5.4 The cone of confusion

Since IID, IPD, RIID or ITD with two microphones can only give an angle relative to the median plane,¹ the actual position of the sound source could lie on a cone with its vertex between the microphones, see Fig. 12. Humans solve this through spectral cues, caused by our asymmetric head and ears, and through active audition. The Aibo, being completely symmetric, must rely on active audition alone. By tilting the robot head the system was able to work out the elevation, thus eliminating the cone of confusion in cases where the source was in front of the robot but elevated. The algorithm for doing this is outlined below:

1. Pan the Aibo head left/right until the PLSOM winning node is in the middle.
2. Roll the head to one side. The head is rolled 23° , or as close to 23° as permitted by the safety limits described in (Sony 2005).
3. Use the same RL system as used to pan the head, but instead of connecting the output to the pan motor, connect it to the tilt motor.
4. Tilt the head up/down until the PLSOM winning node is in the middle.
5. Roll the head back.

¹The median plane is the plane perpendicular to the line through both microphones and equidistant to them.

6. If the PLSOM winning node is not in the middle, go to 1.

It was expected that vertical detection is less accurate than horizontal direction detection since the PLSOM was trained for horizontal, not vertical, directions. The restrictions on how much the neck joint of the Aibo could move might also make this approach difficult to use in some cases.

In order to test the algorithm the system was subjected to the following test. A speaker playing white noise at 68 dB was placed 30° to the right of the robot at one meter distance and 0° elevation. The Aibo head started out pointing straight ahead with 0° elevation. The system was then left to run until it did not alter the head position any more. The horizontal deviation was then noted. This was repeated 10 times each for three elevations: 20°, 30°, and 40°.

The system determined the correct tilt within $\pm 10^\circ$. The system was not very useful because of the limited mobility of the Aibo's neck joints which restricts how much it is possible to tilt and roll the head without damaging the robot, as described in the Aibo ERS-210 manual (Sony 2005). The cone of confusion can also be advantageous for three-dimensional sound source localisation, as will be discussed in Sect. 5.7.

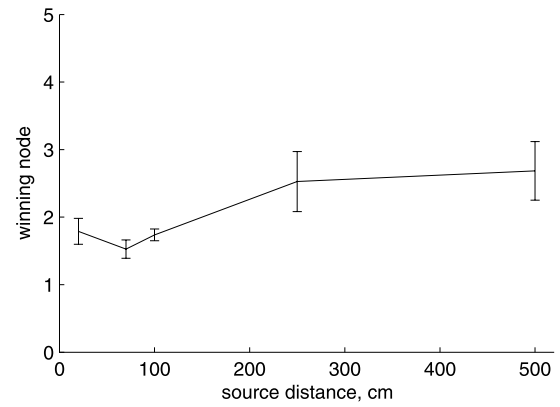
5.5 Distance resolution

Distance detection of sound, as horizontal direction detection, is a composite task. Humans do this mainly through two cues (Mershon and Bowers 1979):

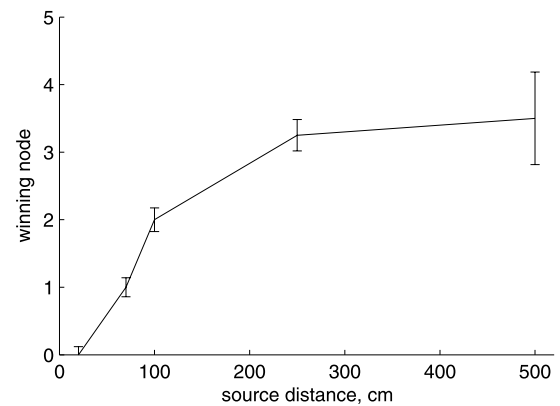
1. Level of familiar sounds. This is of limited usefulness under some applications, but in speech applications it is highly useful since human speech falls into a narrow level range.
2. Ratio of reverberation to direct sound levels. This is only applicable to intermittent sounds, and requires a familiarity with the reverberation characteristics of the room.

Instead of trying to reproduce these two heuristics we selected a metric that is robust and works for both intermittent and continuous sounds as well as unfamiliar sounds of unknown source intensity; the RIID. See (6) for details. The central idea behind the RIID, discussed in Sect. 4.2, is that sound, being a wave phenomenon, decreases in intensity in accordance with the inverse square law, see (Brungart and Rabiowitz 1996). Thus, one can get an idea of the distance by comparing the intensity falloff between the two microphones to the absolute intensity recorded at one of the microphones, which is exactly what the RIID does. Theoretically this should give a metric which is independent of the intensity of the sound source.

Since the RIID is part of the input of the PLSOM, the distance to the sound source should map to one of the output dimensions of the PLSOM.



(a)



(b)

Fig. 13 Average winning node vs. source distance with one standard deviation for various angle ranges. The stronger the correlation between distance and winning node, the better. Note that this map had only been trained with inputs from a distance of 0.5–3 m, which explains the poor correspondence over three meters. (a) -90° to 90° angle between sound source and median plane. (b) $\pm 60\text{--}70^\circ$ angle between sound source and median plane

In order to test this hypothesis the following experiment was carried out. The system was trained with the standard training set, data set A. Then a speaker playing white noise was placed at five different distances from the robot; 0.2, 0.7, 1, 2.5, and 5 m. The volume of the transmitted signal was adjusted so that it gave the same intensity of 63 dB measured at the robot to prevent the algorithm from simply estimating the distance based on intensity. For each distance, the robot's head was turned from -90° to 90° in 10° increments. The result was recorded in data set E. For each angle/distance position the system recorded the position of the winning node of the PLSOM (along the shortest axis of the PLSOM output space). The average winning node was plotted with standard deviation in Fig. 13(a). Clearly, this did not provide a great deal of accuracy. Upon examination of the data it became evident that the distance sensitivity was virtually zero close to the median plane, which corresponds well with what one would expect from a distance metric based on the RIID,

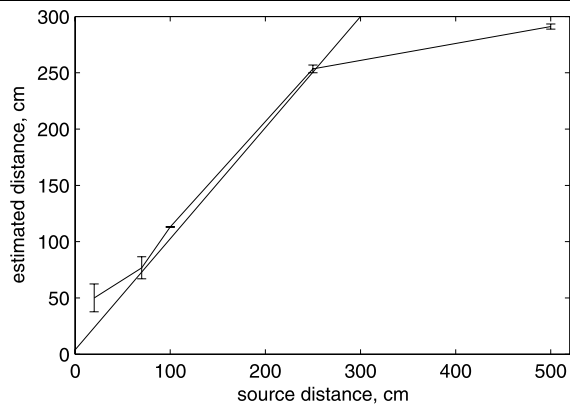


Fig. 14 Estimated source-robot distance vs. actual source-robot distance with one standard deviation. These points represent cases where the source was located $60\text{--}70^\circ$ to the left or right of the median plane of the robot head. A linear fit has been inserted for reference

since the IID (and hence the RIID) decreases with decreasing angle to the median plane.

Theoretically one would be able to discern distance better when the sound source is at greater angles to the median plane. This was supported by the experimental data, which showed good correspondence between winning node and source distance for $\pm 60\text{--}70^\circ$ angle between the median plane and the sound source, see Fig. 13(b). One would expect the distance sensitivity to be at a maximum at $\pm 90^\circ$, but because of the shape of the Aibo head sound passes easily around the back, altering the RIID at angles greater than $\pm 70^\circ$. Also, note that the area of high distance sensitivity largely coincided with the areas of lowest directional sensitivity and vice versa. Although labelling the output of the system to provide human-readable distance estimates is not always practical for robotic applications this step was performed to give the reader a more intuitive impression of the system's accuracy, shown in Fig. 14.

This experiment showed that within the range it was trained in the system performed remarkably well. It should be noted that the distances of 0.7, 1, and 2.5 m were accurately labelled despite the system never having encountered sounds from these distances before, since the training set (data set A) only contained samples recorded at 0.5 and 3 m.

5.6 Limits to response time

In order to test the response time of the system, the following two experiments were conducted:

1. The robot was filmed with a digital camera. A set of four keys were dropped onto a wood surface one meter in front of the robot from a height of 150 mm. This was repeated 20 times. The video recording was analysed with video editing software, and the time from the keys hit the wood surface to the robot begins turning was measured.

2. Three sounds of different durations were played to the left and right of the robot: keys dropping (duration 60 ms), hands clapping (20 ms), and fingers snapping (20 ms). This was repeated 10 times for each sound. The number of times the robot turns in the correct direction was noted, and a success rate calculated.

The first experiment gave a mean response time of 0.5 seconds. The second experiment gave a success rate of 80% for the 60 ms sound but only 50% for the 20 ms sounds.

5.7 Taking advantage of the cone of confusion for elevation estimation

The cone of confusion might seem like a problem at first, but the principles of active audition can be used to turn it into an advantage. Consider the following scenario: the system detects a sound source at 60° to the right of the sagittal plane. It can be deduced that the actual position of the sound source is on a cone with axis parallel to the head tilt axis and vertex between the microphones with an opening angle of 60° .

Now the system turns the head 45° to the right in response to the stimuli. This changes the angle between the sound source and the sagittal plane but the change depends on the elevation of the sound source. If the source is in the horizontal plane, the new angle is $(60^\circ - 45^\circ) = 15^\circ$. If, on the other hand, the source is above or below the horizontal plane, the new angle will be greater than 15° . The relationship between the change in sagittal/source angle, the pan movement of the head and elevation of the source is shown in (9)–(11)

$$\varpi = \psi - \psi', \quad (9)$$

$$\eta = (\sin(v))^2 (\cos(\varpi))^2 - 2 \sin(v) \cos(\varpi) \sin(v') + (\sin(v'))^2 + (\sin(v))^2 (\sin(\varpi))^2, \quad (10)$$

$$\mu = \arccos\left(\frac{\sqrt{\eta}}{\sin(\varpi)}\right). \quad (11)$$

Here, μ is the elevation angle of the source, seen as rotation around the tilt axis. v is the apparent angle between the source and the sagittal plane, ψ is the angle the head is at before movement, and ψ' is the angle after movement. v' is the apparent angle between source and sagittal plane after the pan movement. This is clearly independent of distance, but is unable to distinguish between negative and positive μ . Therefore the following experiments will be conducted with only negative μ , which is to say the source is in or above the horizontal plane. It might be feasible to implement the explicit algorithm as described in (9)–(11), but it requires knowing the explicit values of v and v' .

Instead, the existing implicit representation of the sound source position can be re-used. This can be done without converting the sound source position to explicit coordinates:

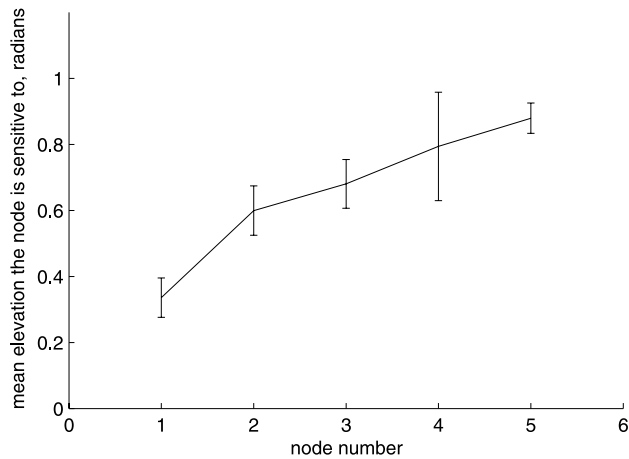


Fig. 15 Elevation each node was sensitive to vs. node index with one standard deviation. The map was five by two nodes and the neighbourhood size was 4.27

The processing is done using the implicit position representation by utilising the dimension-reducing properties of the PLSOM. Therefore, the system was extended with a second PLSOM, which took a two-dimensional input vector:

1. The current estimated position in the form of the horizontal index of the winning node divided by the width of the entire map.
2. The difference in head direction since the last input divided by the difference in winning node since the last input. The head direction was given by the neck joint angle in this application, in a mobile implementation it would need to be calculated from inertial sensors or similar in order to get the change in direction of the head relative to the source.

This second PLSOM was also of two output dimensions, giving a representation of horizontal angle along one axis and elevation along the other axis.

Training took place in two stages. First, the PLSOM was trained with data set A, as before. Then, the RL was trained with data set B. During training of the RL subsystem the elevation of the sound source was simulated by altering the angle of the sound source in accordance with (9)–(11) solved for the apparent angle, ν . The elevation was varied from 0 to 63°. The RL subsystem was trained in this way for 1000 weight updates. During this training the output of the direction detecting PLSOM was fed to the elevation detecting PLSOM, as outlined above. The elevation detecting PLSOM was five by two nodes and the neighbourhood size was set to 4.27, settings that were arrived at empirically.

As can be seen in Fig. 15, the elevation detection PLSOM became sensitive to different elevations along its longest axis. The system was tested by selecting 100 source positions at elevations ranging from 0° to 63° in 6.3° increments and horizontal positions ranging from -88° to 88° in 17.7°

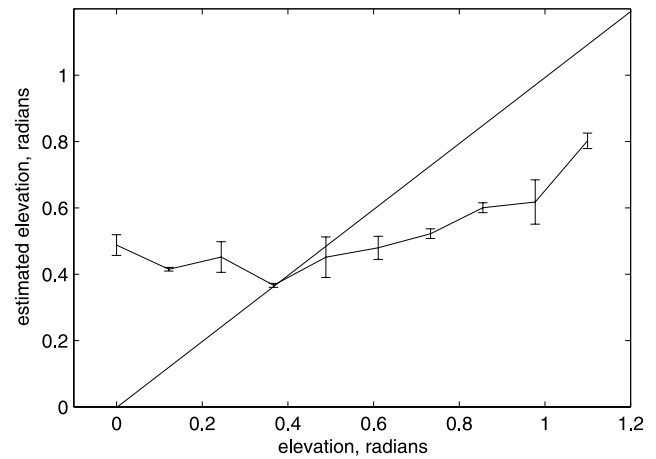


Fig. 16 Estimated elevation vs. actual elevation with one standard deviation. A linear fit has been added for reference. The map was five by two nodes and the neighbourhood size was 4.27

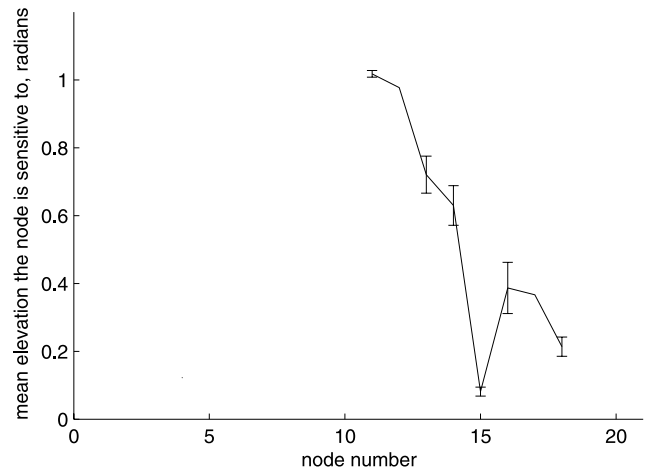


Fig. 17 Elevation each node was sensitive to vs. node index with one standard deviation. The map was 20 × 8 nodes and the neighbourhood size was nine

increments and letting the robot orient towards them while recording the output of the elevation PLSOM. The nodes of the PLSOM were then labelled with the elevation in the test data set which they were most sensitive to, and the experiment was run again with a different data set (data set H). The mean estimated elevation and standard deviation was noted for each elevation and plotted in Fig. 16. As can clearly be seen, this was not particularly sensitive at all. Increasing the map size to 20 × 8 nodes with a neighbourhood size of nine gives a better result, see Fig. 18.

Unfortunately, the larger map size meant there were insufficient numbers of training examples to label all nodes, so some nodes simply never got selected as winning nodes and it was not possible to calculate the mean elevation they were sensitive to, as seen in Fig. 17.

For the large map, the Pearson correlation coefficient r was 0.62, which compares favourably, given the difference

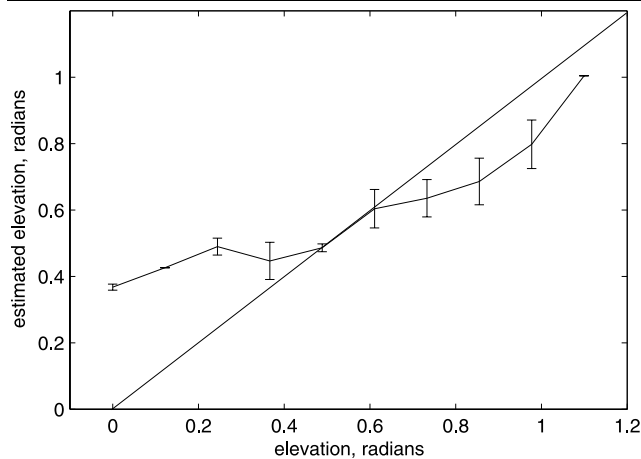


Fig. 18 Estimated elevation vs. actual elevation with one standard deviation. A linear fit has been added for reference. The map was 20×8 nodes and the neighbourhood size was nine. Pearson correlation coefficient $r = 0.62$

in sampling rate and resolution, with human accuracy in simulated tests by Avendano et al. (1999) where r up to 0.77 was achieved. Most of the inaccuracy was around low elevations, where the granularity of the horizontal angle detection map should become more evident. The method of approximating the elevated sound should also give more inaccuracies around small elevations because of the 1° granularity of the samples used for approximation. It should be noted that Avendano's experiments were performed using headphones, so the subjects did not have the option of using active head movements to locate the sound.

Although this result was less impressive than the 6° accuracy exhibited by multi-microphone systems (Tamai et al. 2005), this should be seen in relation to the number of microphones and consequently amount of processing required.

6 Discussion

This paper has analysed the problem of sound direction detection and active audition for robots using two-channel audio. The goal was to create a system that can detect the position of a sound source in three dimensions and orient towards it. The system should also be robust to changes, learn from experience without the need for supervision and react quickly. The system should be able to react to both transient and continuous sounds, as well as sounds with both broad and narrow spectra. Finally the system should be able to filter sound from noise based on direction without having to rely on assumptions about the nature of the sound or the noise.

The presented system uses unsupervised learning exclusively. The system was compared with the SIG humanoid system and human hearing for horizontal localisation and

human hearing for elevation detection. Human hearing can detect changes in source position down to 1° with absolute positional error approximately 5° , the SIG system has a reported accuracy of approximately 5° , but on the test platform the accuracy was approximately 10° in the best areas with a standard deviation of $\pm 40^\circ$ in some areas. The presented system achieves an accuracy of 5° with negligible deviation on this platform. Correlation between elevation and estimated elevation is $r = 0.62$, when humans have achieved up to $r = 0.77$ in simulated tests and up to $r = 0.98$ in tests where the test subjects can move their heads freely. Sound source distance estimation approaches generally apply more than two sound channels and are not directly comparable to the presented system, which achieves an estimation error of less than 150 mm and a standard deviation of less than ± 0.1 m. The orienting behaviour is accurate to within $\pm 5^\circ$ and responds in 0.5 s, even when network lag is factored in.

The system has been tested with white noise, human speech and pure frequencies with two harmonics. Its response has been tested with continuous sounds and transient sounds such as keys dropping, hands clapping and fingers snapping, giving 80% success rate for sounds of 60 ms duration and 50% success rate for 20 ms duration.

7 Conclusion

We have described a system for letting a robot determine the direction, distance and elevation of a sound source and learn to orient towards a sound source without supervision. The system performs well compared to other approaches we have implemented. The system determines a model of the acoustic properties of the robot using a PLSOM, and automatically associates the output of the PLSOM with the correct motor actions. Interestingly, this is done without any explicit theoretical model of the environment; there is no information built into the system about the shape of the robot head, the distance between the robot ears or the physics of sound. All these constants and their internal relationships are implicitly worked out by the system. This makes the system flexible since it can be implemented on any robot with stereo microphones without alteration.

Acknowledgements The authors would like to thank Dr. Kazuhiro Nakadai of the Japan Science and Technology Corporation and Dr. Frederic Maire of the Smart Devices Lab at the Queensland University of Technology for their valuable input. This project is supported, in part, by a grant from the Australian Government Department of the Prime Minister and Cabinet. NICTA is funded by the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council.

References

Avendano, C., Algazi, V. R., & Duda, R. O. (1999). A head-and-torso model for low-frequency binaural elevation effects. In *Proceed-*

- ings of workshop on applications of signal processing to audio and acoustics (pp. 179–182), October 1999.
- Berglund, E., & Sitte, J. (2003). The parameter-less SOM algorithm. In *ANZIHS* (pp. 159–164).
- Berglund, E., & Sitte, J. (2006). The parameter-less self-organising map algorithm. *IEEE Transactions on Neural Networks*, 17(2), 305–316.
- Blauert, J. (1983). *Spatial hearing*. Cambridge: MIT Press.
- Bregman, A. (1990). *Auditory scene analysis*. Massachusetts: MIT Press.
- Brungart, D. S., & Rabiowitz, W. R. (1996). Auditory localization in the near-field. In *Proceedings of the ICAD, international community for auditory display*.
- Day, C. (2001). Researchers uncover the neural details of how Barn Owls locate sound sources. *Physics Today*, 54, 20–22.
- Ge, S. S., Loh, A. P., & Guan, F. (2003). Sound localization based on mask diffraction. In *ICRA '03* (Vol. 2, pp. 1972–1977), September 2003.
- Gosavi, A. (2003). *Simulation-based optimization: parametric optimization techniques and reinforcement learning*. Dordrecht: Kluwer.
- Guentchev, K., & Weng, J. (1998). Learning based three dimensional sound localization using a compact non-coplanar array of microphones. In *AAAI spring symposium on international environments*.
- Huang, J., Ohnishi, N., & Sugie, N. (1995). A biometric system for localization and separation of multiple sound sources. *IEEE Transactions on Instrumentation and Measurement*, 44(3), 733–738.
- Huang, J., Ohnishi, N., & Sugie, N. (1997). Building ears for robots: sound localization and separation. *Artificial Life and Robotics*, 1(4), 157–163.
- Iske, B., Rueckert, U., Sitte, J., & Malmstrom, K. (2000). A bootstrapping method for autonomous and in site learning of generic navigation behaviour. In *Proceedings of the 15th international conference on pattern recognition* (Vol. 4, pp. 656–659), Barcelona, Spain, September 2000.
- Kitano, H., Okuno, H. G., Nakadai, K., Matsui, T., Hidai, K., & Lourens, T. (2002). *SIG, the humanoid*. <http://www.symbio.jst.go.jp/symbio/SIG/>.
- Konishi, M. (1993). Listening with two ears. *Scientific American*, 268(4), 34–41. Deals with how the owl locate its prey by hearing. Of special interest to me is the layout of the owl's ears and neural pathways. A lot of the information on the biology of owls is redundant.
- Kuhn, G. F. (1987). Acoustics and measurements pertaining to directional hearing. In *Directional hearing* (pp. 3–25). New York: Springer.
- Kumon, M., Shimoda, T., Kohzawa, R., Mizumoto, I., & Iwai, Z. (2005). Audio servo for robotic systems with pinnae. In *International conference on intelligent robots and systems* (pp. 885–890).
- Mershon, D. H., & Bowers, J. N. (1979). Absolute and relative cues for the auditory perception of egocentric distance. *Perception*, 8, 311–322.
- Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.
- Moore, B. C. J. (1997). *An introduction to the psychology of hearing* (4th ed.). New York: Academic Press.
- Nakadai, K. (2004). Private communication.
- Nakadai, K., Lourens, T., Okuno, H. G., & Kitano, H. (2000a). Active audition for humanoid. In *AAAI-2000* (pp. 832–839).
- Nakadai, K., Okuno, H. G., Laurens, T., & Kitano, H. (2000b). Humanoid active audition system. In *IEEE-RAS international conference on humanoid robots*.
- Nakadai, K., Hidai, K., Mizoguchi, H., Okuno, H. G., & Kitano, H. (2001). Real-time auditory and visual multiple-object tracking for humanoids. In *IJCAI* (pp. 1425–1436).
- Nakadai, K., Okuno, H., & Kitano, H. (2002a). Realtime sound source localization and separation for robot audition. In *Proceedings IEEE international conference on spoken language processing* (pp. 193–196).
- Nakadai, K., Okuno, H. G., & Kitano, H. (2002b). Exploiting auditory fovea in humanoid-human interaction. In *Proceedings of the eighteenth national conference on artificial intelligence* (pp. 431–438).
- Nakadai, K., Matsuura, D., Okuno, H. G., & Kitano, H. (2003a). Applying scattering theory to robot audition system: robust sound source localization and extraction. In *Proceedings of the 2003 IEEE/RSJ international conference on intelligent robots and systems* (pp. 1147–1152).
- Nakadai, K., Okuno, H. G., & Kitano, H. (2003b). Robot recognizes three simultaneous speech by active audition. In *ICRA '03* (Vol. 1, pp. 398–405).
- Nakadai, K., Okuno, H. G., & Kitano, H. (2003c). Robot recognizes three simultaneous speech by active audition. In *ICRA '03* (Vol. 1, pp. 398–405).
- Nakashima, H., Mukai, T., & Ohnishi, N. (2002). Self-organization of a sound source localization robot by perceptual cycle. In *Proceedings of the 9th international conference neural information processing* (Vol. 2, pp. 834–838).
- Nakatani, T., Okuno, H. G., & Kawabata, T. (1994). Auditory stream segregation in auditory scene analysis with a multi-agent system. In *AAAI-94* (pp. 100–107).
- Obata, K., Noguchi, K., & Tadokoro, Y. (2003). A new sound source location algorithm based on formant frequency for sound image localization. In *Proceedings 2003 international conference on multimedia and expo* (Vol. 1, pp. 729–732), July 2003.
- Rabinkin, D., Renomeron, R., Dahl, A., French, J., Flanagan, J., & Bianchi, M. (1996a). A DSP implementation of source location using microphone arrays. *Proceedings of the SPIE*, 2846, 88–99.
- Rabinkin, D., Renomeron, R., French, J., & Flanagan, J. (1996b). Estimation of wavefront arrival delay using the crosspower spectrum phase technique. In *Proceedings of 132nd meeting of the ASA*.
- Reid, G., & Milios, E. (1999). *Active stereo sound localization*.
- Reid, G., & Milios, E. (2003). Active stereo sound localization. *Journal of the Acoustical Society of America*, 113(1), 185–193.
- Rucci, M., Edelman, G., & Wray, J. (1999). Adaptation of orienting behavior: from the barn owl to a robotic system. *IEEE Transactions on Robotics and Automation*, 15(1), 15.
- Shaw, E. A. G. (1997). Acoustical features of the human external ear. In *Binaural and spatial hearing in real and virtual environments* (pp. 49–75). Mahwah: Lawrence Erlbaum Associates.
- Sitte, J., Malmstrom, K., & Iske, B. (2000). Perception stimulated generation of simple navigation behaviour. In *Proceedings of SPIE: Vol. 4195. Mobile robots*, Boston, MA, USA (pp. 228–239).
- Sony (2005). *Open-R and Aibo documentation*. <http://openr.aibo.com/openr/eng/index.php4>.
- Strutt, 3rd Baron Rayleigh, J. W. (1896). *The theory of sound* (2nd ed.). London: Macmillan.
- Sutton, R. S. (Ed.). (1992). *Reinforcement learning*. Dordrecht: Kluwer Academic.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. Cambridge: MIT Press.
- Tamai, Y., Kagami, S., Amemiya, Y., Sasaki, Y., Mizoguchi, H., & Takano, T. (2004). Circular microphone array for robot's audition. In *Proceedings of IEEE* (pp. 565–570).
- Tamai, Y., Sasaki, Y., Kagami, S., & Mizoguchi, H. (2005). Three ring microphone array for 3D sound localization and separation for mobile robot audition. In *Proceedings of international conference on intelligent robots and systems* (pp. 903–908).
- Yamamoto, K., Asano, F., van Rooijen, W. F. G., Ling, E. Y. L., Yamada, T., & Kitawaki, N. (2003). Estimation of the number of sound sources using support vector machines and its application

to sound source separation. In *ICASSP '03* (Vol. 5, pp. 485–488), April 2003.



Erik Berglund received the equivalent of a Bachelor's degree in Computer Engineering from Østfold University College in 2000 and a Ph.D from the University of Queensland in 2006. He is currently a senior research fellow at UQ. Research interests include neural networks, face recognition and implicit data processing.



Joaquin Sitte is an Associate Professor at the School of Software Engineering and Data Communications, Queensland University of Technology, Australia, where he leads the Smart Devices Lab. Joaquin received his Licenciado degree in physics from the Universidad Central de Venezuela in 1968 and his Ph.D. degree in quantum chemistry from Uppsala University, Sweden, in 1974. Until 1985 he was an Associate Professor at the Universidad de Los An-

des, Merida, Venezuela, where he also headed the Surface Physics Research Group. Since 1986 he is on the faculty of Queensland University of Technology. He has a special interest in the use of neural networks for sensing, thinking, learning and actuation in autonomous robots.



Gordon Wyeth is a senior lecturer in robotics at the University of Queensland, Australia. He received his Bachelor in Engineering (Honours) from the same university in 1989, and his PhD from the same university in 1997. He is President of the Australian Robotics and Automation Association. His research centres on the development of robotic systems based on neuroethological data. This research is being applied in humanoid robotics, consumer robot applications and robot soccer.