ARTICLE IN PRESS

Neural Networks ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2009 Special Issue

Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks

John C. Murray, Harry R. Erwin, Stefan Wermter *

Hybrid Intelligent Systems, University of Sunderland, Sunderland, Tyne and Wear, SR6 0DD, United Kingdom[1]

ARTICLE INFO

Keywords:
Robotics
Human–robot interaction
Sound-source localisation
Cross-correlation
Recurrent neural networks

ABSTRACT

In this paper we present a sound-source model for localising and tracking an acoustic source of interest along the azimuth plane in acoustically cluttered environments, for a mobile service robot. The model we present is a hybrid architecture using cross-correlation and recurrent neural networks to develop a robotic model accurate and robust enough to perform within an acoustically cluttered environment. This model has been developed with considerations of both processing power and physical robot size, allowing for this model to be deployed on to a wide variety of robotic systems where power consumption and size is a limitation. The development of the system we present has its inspiration taken from the central auditory system (CAS) of the mammalian brain. In this paper we describe experimental results of the proposed model including the experimental methodology for testing sound-source localisation systems. The results of the system are shown in both restricted test environments and in real-world conditions. This paper shows how a hybrid architecture using band pass filtering, cross-correlation and recurrent neural networks can be used to develop a robust, accurate and fast sound-source localisation model for a mobile robot.

1. Introduction

Robots within society are becoming more commonplace with their increased use in roles such as robotic vacuum cleaners (Asfour et al., 2008; Medioni, François, Siddiqui, Kim, & Yoon, 2007), library guides, surgery applications (Obando, Liem, Madauss, Morita, & Robinson, 2004) and pharmacy. However, interacting with such service robots causes problems for many users as specialised training or pre-requisite knowledge is usually required. In order to make the interaction process between humans and robots easier and more intuitive, it is necessary to make the interaction process as natural as possible to allow the robotic devices to be more easily integrated into the lives of humans (Severinson-Eklundh, Green, & Hüttenrauch, 2003). Utilising natural and intuitive interactions not only makes users feel more comfortable but reduces the time needed for the user to familiarise themselves with the interfaces and control systems of such robots.

Tour guide robots such as PERSES (Böhme et al., 2003) are now used to take visitors around various places such as museum buildings, pointing out interesting exhibitions, answering questions or just ensuring people do not get lost. One of the most natural methods of communication for this scenario would be in the acoustic modality. Therefore, when robots that are operating within human occupied environments it is necessary for the process of robot–human interaction to be as close as possible to that of human–human interaction.

Within the field of robotics, there is a growing interest in acoustics with researchers drawing on many different areas from engineering principles to biological systems (Blauert, 1997a). Previously, robotic navigation and localisation has been predominately supported by the vision modality (Wermter et al., 2004). Vision is widely used as a means for locating objects within the scene; however, in humans and most animals, the visual field-of-view is restricted to less than 180° due to the positioning of the eyes. Most cameras used for vision have an even narrower field-of-view, determined by the particular lens in use, of usually $<90°$. This restriction can be overcome in vision with the use of a conical mirror (Lima et al., 2001) to allow the full but distorted field-of-view of the scene to be seen or by using multiple cameras. However, to help overcome this limitation, humans and animals use the additional modality of hearing. That modality effectively gives them a full 360° field of 'view' of the acoustic scene.

As can be expected, the acoustical modality has both advantages and disadvantages to the use of vision. The major disadvantage with using acoustics to detect an object within the environment, is that the particular object must have an acoustical aspect, i.e. it must produce some form of sound that can be detected. Therefore,

* Corresponding author.
E-mail address: stefan.wermter@sunderland.ac.uk (S. Wermter).
[1] URL: http://www.his.sunderland.ac.uk.

0893-6080/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2009.01.013

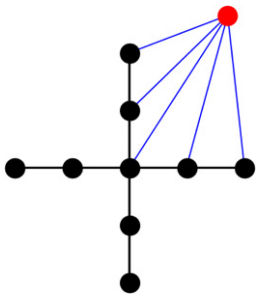

Fig. 1. TDOA for the most relevant microphone pairs within a cross-array matrix.

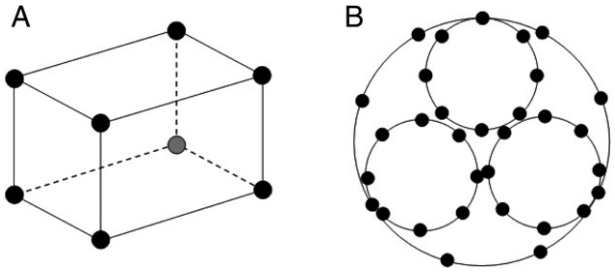

Fig. 2. Different multi-microphone configurations. (A) The system developed by Valin et al. (2003), and (B) The system developed by Tamai et al. (2005).

if a particular object that is being tracked does not emit sound it will be impossible to actively locate.

For the scenario proposed in this paper, the above limitation can be reduced to some extent with the use of the tracking prediction element of our model. However, as previously mentioned, the use of acoustics also has its advantages over vision. One such advantage is the ability to determine the direction of an object that may not even reside with the visual field-of-view. This supports the ability to locate objects that may be visually obscured by other objects or located around a corner (Huang et al., 1999). Hence, the choice of modality for the model presented in this paper is due to sound being a major part of the communication process in humans (Arensburg & Tillier, 1991).

There are some systems that already employ sound-source localisation on a mobile robot. These systems range from the multi-microphone array designs of Tamai, Kagami, Sasaki, and Mizoguchi (2005) and Valin, Michaud, Rouat, and Létourneau (2003), utilising engineering principles, to biologically plausible systems such as Smith's (2002). A common approach to sound-source localisation has been to use neural networks to determine the angle of incidence of the source (Datum, Palmieri, & Moiseff, 1996). The Robotic Barn Owl developed by Rucci, Wray, and Edelman (2000) also incorporates vision to improve the sound-source localisation by allowing the visual system to reinforce the acoustic tracking capabilities through training.

Of these methods, the most widely used is a multi-microphone array with a large number of microphones (usually 8 or more) arranged in a distributed configuration. One such matrix configuration for a multi-microphone array is shown in Fig. 1. One of the advantages of this approach is the microphone array structure which creates several multi-path vectors which are then used to determine the individual Time Delay of Arrival (TDOA) between each microphone pair.

Wang, Ivanov, and Aarabi (2003) demonstrates a different technique using a pair of '*distributed microphone arrays*'. Here the localisation of a robot is performed by two microphone arrays which are statically placed on the walls of the environment. The idea of this approach is to allow the microphone arrays to simultaneously triangulate the position of the sound source relative to the robot's position. When these two values are known the system can calculate the new required trajectory and update the robot's path. This approach, of course, relies on a preconfigured external array and cannot be applied to sound-source localisation in dynamic environments.

The systems of Tamai et al. (2005) and Valin et al. (2003) pursue a similar approach to that of Wang et al. (2003) but the microphones are mounted on the robot framework. Fig. 2 shows two different configurations. This particular approach allows the system to localise the sound source from the frame of reference of the robot as it dynamically moves through the environment. These methods still use the multi-path TDOA between successive microphone pairs to determine the location of the sound source.

Although fish and some amphibians use lateral line arrays (Corwin, 1992) in aquatic environments, non-aquatic vertebrates use binaural hearing with only two receivers, suggesting large receiver arrays provide no significant advantage in vivo.

While these array-based methods provide adequate localisation of a sound source within the environment, they have poor performance in several key areas that the model in this paper aims to address. Firstly, systems such as the ones described above would be insufficient for a dynamic socially interactive robotic system due to the constraints imposed by such systems, requiring microphone arrays to be placed in specific positions within the environment and then configured according to that environment (Wang et al., 2003).

When deploying the system in different locations it is necessary to adapt to the varying environments. Furthermore, the response and accuracy of the systems tend to decrease under acoustically cluttered environments, due to the need to analyse more acoustic data, in addition to sounds closer to the microphone arrays being louder as opposed to sounds closer to the robot. The response time and power consumption of these systems tend to make them useable only in certain scenarios where the robots can be tethered to larger machines and power sources i.e. not suitable for many search and rescue operations or free operation within a social context.

With multi-microphone arrays, there is much more data to be processed before a sound source can be localised. This excessive data processing slows down the response of the system. Additionally, especially with the system shown by Valin et al. the computing power required to process the signals is large due to the number of microphone pairs from the arrays, thus the robot needs to be tethered to a large computing cluster.

Therefore, the system presented within this paper looks at providing an accurate, socially interactable, fast response, sound-source localisation model for acoustically cluttered environments that can be implemented on a mobile robot.

Alternative systems rely predominantly on multi-microphone arrays to determine the azimuth position of the sound source. This increases the complexity of the signals that need to be processed. The model presented here requires only two microphones for localisation, performing adequately. The second component is the use of a recurrent neural network for tracking and predicting the location of a dynamic sound source, also enabling the system to maintain effective signal-to-noise ratio levels (Murray, Wermter, & Erwin, 2006), that is, to reduce the levels of unwanted (irrelevant) signals i.e. the noise, whilst increasing the levels of wanted (desired) signals as much as possible. In addition, with the increase in the SNR levels of the speaker versus the background, this paves the way for the incorporation of speech recognisers, enhancing the social capabilities of the system. The systems shown by Böhme et al. (2003), Datum et al. (1996), Smith (2002), Tamai et al. (2005) and Wang et al. (2003) amongst others are calibrated for certain environments, preventing the systems to be free to move

Please cite this article in press as: Murray, J. C., et al. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. Neural Networks (2009), doi:10.1016/j.neunet.2009.01.013

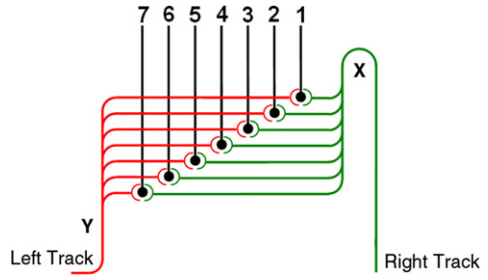

Fig. 3. Jeffress model of coincidence detectors for determining the ITD of a signal. The signals reach points X and Y at the same point in time.

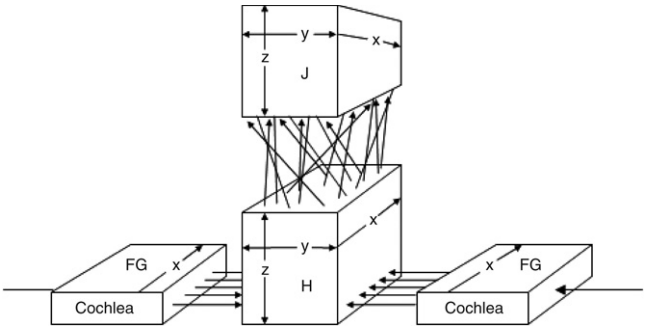

Fig. 4. Licklider's Triplex model showing the use of cross-correlation for measuring the IPD of signals received. As the signals arrive at the Cochlea the stimulus is mapped by frequency onto the spatial dimension '*x*'. The block H preserves the order in '*x*' but adds an analysis in the '*y*' dimension based on the ITD.

between different locations. Our model also incorporates a self-normalisation function, see Section 5.3, which allows the acoustic model to adapt to varying level conditions within the environment. Thus, if the robot moves to a louder area, it can attenuate the sounds preventing over modulation.

2. Biological inspiration

For the model presented within this paper, inspiration is taken from the known workings of the central auditory system (CAS) of the mammalian brain. The mammalian CAS has excellent accuracy in performing acoustic scene analysis and sound-source localisation (Blauert, 1997a). The auditory cortex (AC) can accurately localise a sound source within the environment to less than $\pm 5^\circ$ in azimuth and elevation (Blauert, 1997b), thus enabling the animal to orientate to the direction of the source. The accuracy of sound-source localisation in humans can reach $\pm 1^\circ$ in azimuth and $\pm 5^\circ$ in elevation (Blauert, 1997b). Thus the model presented within this paper draws on its inspiration, specifically from, the available acoustic cues and mechanisms that are believed to be used within biological systems.

Within this paper we build on two main biological models for localisation. Firstly, the system uses the interaural cues that are available in biology. The Jeffress model (Jeffress, 1948) details the use of coincidence detectors for localisation which are tuned in terms of timing in order to fire according to the Interaural Time Difference (ITD) between the two ears caused by the azimuth angle of incidence of the sound source, shown in Fig. 3.

The second cue for azimuth localisation that uses binaural cues is that of Licklider's triplex model (Licklider, 1959); this model as is shown in Fig. 4 gives an outline of the path of the auditory signals, from the cochlear nucleus to the Inferior Colliculus (IC). This model shows how cross-correlation (the main basis for the model described later within this paper) plays an important role in determining the azimuth of the sound source based on the Interaural Phase Difference (IPD) of the signals arriving at the two ears, caused by the angle of incidence of the sound source.

Our paper focuses specifically on the cues utilised by these models for the purpose of robotic sound-source localisation and tracking. Being able to accurately localise and track a desired sound source of interest within the environment not only allows the robotic system to determine the position of the source within the environment, in azimuth, but also helps to maintain a higher signal-to-noise ratio (SNR) between the robot and the target source, enabling better processing of the signals produced by the source in question. Section 5.1 discusses in more detail the advantages of improving the SNR. In addition, the model presented in this paper includes a normalisation and an energy function allowing the system to attend to only sounds that are considered to be of importance, see Section 5.3.

The first of our inspirations, as previously mentioned, is the biological acoustic cues for azimuth estimation (Hawkins, 1995) which are believed to be encoded in lower brainstem regions such as the medial superior olive (MSO) (Joris, Smith, & Yin, 1998). The second key biological inspiration comes from the decision to use two ears or 'microphones' for determining the azimuth angle of the sound source. Many of the robotic systems developed for the purpose of sound-source localisation, tend to utilise arrays of microphones (Tamai et al., 2005; Valin et al., 2003; Wang et al., 2003). However, using a large number of microphones generates much more data that needs to be analysed and processed in order to determine the direction of the source. While this paper demonstrates that with just two microphones the same accuracy (and better) can be achieved, but with much less processing requirements.

The third biological cue used is concerned with the attenuation, normalisation and content of the signal. The mammalian CAS has the ability to normalise (to some extent) the levels of the signals arriving at the ears. The outer hair cells of the cochlear can attenuate the signals to either increase quiet sounds, or decrease loud sounds (Géléoc & Holt, 2003). This allows for mammals to focus on quiet sounds or prevent damage to their ears from loud sounds. This functionality therefore translates well to the model presented in this paper, allowing for the robot to be able to adapt and normalise to the background levels in a particular environment. This is further discussed in Section 5.3.

3. Robotic framework

The robot platform used for the testing of our model is the PeopleBot, developed by ActivMedia, see Fig. 5. This particular robot is of upright design and based on the pioneer base. For the sound detection capabilities, two microphones are mounted onto the base of the robot. The microphones are separated by a distance of 30 cm to be close to that of the distance between human ears, and to give a slightly larger separation to account for the available sample rate of the sound card on the robot, see Fig. 6.

This particular robot is chosen due to its size, and footprint, with its height being between that of a child and an adult. In addition to being able to navigate environments that are occupied by humans. The processing of the acoustic data and corresponding control of the robot is provided via an on-board AMD K6-2 500 MHz processor and 512 MB RAM. This processor was found to be sufficient for processing the acoustic information, and providing control information for movement, of the robot when tracking sound sources within the environment.

4. Calculating the azimuth position

Many different approaches to sound-source localisation have been proposed in the past. One interesting method is the incremental control approach, which utilises the level difference of

Please cite this article in press as: Murray, J. C., et al. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. Neural Networks (2009), doi:10.1016/j.neunet.2009.01.013

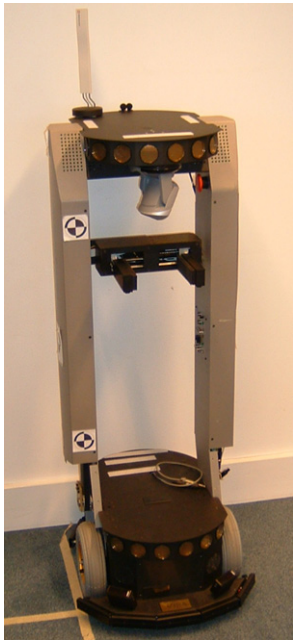

Fig. 5. PeopleBot used as the base for the acoustic model.

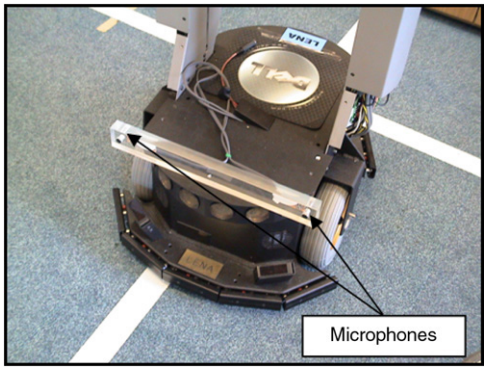

Fig. 6. The two microphones mounted on the PeopleBot-LENA.

the onsets of the signals received at the two microphones, shown by Smith (2002). Another model using an incremental approach is shown by Macera, Goodman, Harris, Drewes, and Maciokas (2004). Here the system uses the ILD cue to determine if the sound source is to the left or right of the robot and will instruct the system to turn in the required direction. The robot is instructed to turn a fixed amount, not related to the exact angle of the source, but determined beforehand by the operators. Fig. 7 shows how such an incremental approach would work to, localise a source. The model we present here shows a more direct method that reduces the time taken to localise the direction of the source, and focus in as accurately as possible on the first iteration. In addition, the model presented here maintains a track on the dynamic sound source as it traverses through the environment.

Fig. 7 shows two examples of a robot detecting a static sound source over several iterations. With each iteration the robot moves a distance of 1.5 m whilst in (a) turning 45° either left or right and in (b) turning 90°. As can be seen regardless of the starting position, distance, or angle increment, the robot takes several iterations to localise the source.

4.1. Calculating phase difference

To determine the azimuth position of the sound source the model uses a combination of the ITD and IPD cues. The interaural

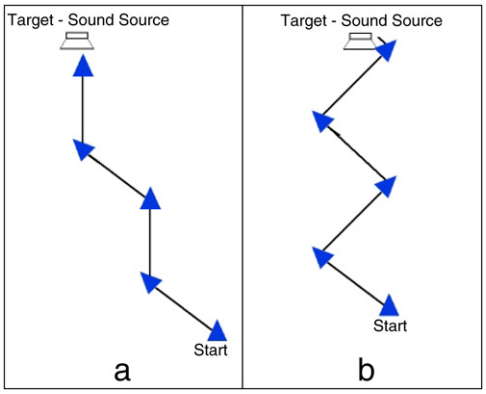

Fig. 7. Two example solutions for the incremental localisation model to locate a static source.

phase difference cue as described by Licklider (1959) determines how 'out of phase' a signal arriving at the two ears is. The interaural time difference cue as described by Jeffress (1948) determines the time delay between a signal arriving at the two ears. The model presented within this paper uses the IPD cue to determine the ITD and therefore the TDOA.

The first component of the model in determining the angle of incidence is the IPD cue. The IPD represents the phase difference between two similar signals received by the robot's two microphones. As the microphones are spatially separated by 30 cm, the angle of the source will determine which microphone detects the signal first. Thus, a temporal phase difference will be created between the separate recordings of the two microphones. It is this difference in 'phase' that is initially used in the model to help determine the azimuth of the sound source. Two signals $g(t)$ and $h(t)$ are sampled vector representations of the sounds that the microphones detect within the external environment, as a function of time. Each of the values within the vectors represents the detected amplitude of the signal at a specific point in time.

In order to determine the IPD of the recorded signal vectors $g(t)$ and $h(t)$, a method known as cross-correlation is used (Press, Flannery, Teukolskv, & Vetterling, 1992). Eq. (1) shows the formula for cross-correlation, this takes the incremental points of the vectors $g(t)$ and $h(t)$ as parameters. However, ultimately it is the time delay of arrival (TDOA) or ITD that is used to finally determine the azimuth angle. The ITD is calculated from the IPD results, as shown by Eq. (1), and is calculated from the offset in phase between the two signal data series.

Cross-correlation does not calculate the ITD between the two signals by using onset detection or signal timings *per se*; instead, cross-correlation uses the two signal vectors $g(t)$ and $h(t)$ to compute the IPD of the signals, within the vectors, as detected by the microphones.

$$\mathrm{Corr}(g, h)_j \equiv \sum_{k=0}^{N-1} g_{j+k} h_k \tag{1}$$

$$r = \frac{\sum_{i=1}^{n} (g_i - \bar{g})(h_i - \bar{h})}{\sqrt{\sum_{i=1}^{n} (g_i - \bar{g})^2 \sum_{i=1}^{n} (h_i - \bar{h})^2}}. \tag{2}$$

Eq. (2) shows the formula used to calculate the correlation coefficient of the two recorded signal vectors $g(t)$ and $h(t)$. This formula provides a measure of the correlation between the two data series normalised to the values of -1 to $+1$. This is known as Pearson's product moment correlation equation, where r is the correlation coefficient value of the data, g is the signal vector $g(t)$

Please cite this article in press as: Murray, J. C., et al. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. Neural Networks (2009), doi:10.1016/j.neunet.2009.01.013

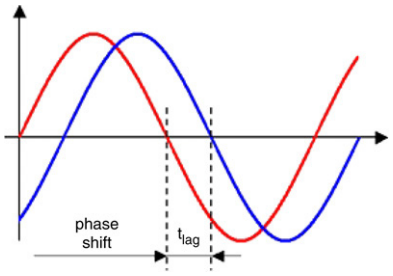

Fig. 8. Two arbitrary similar signals that are out of phase. The *x*-axis represents the angle that is increasing with time and the *y*-axis the amplitude of the signals.

and h is the signal vector $h(t)$. $\bar{g}$ and $\bar{h}$ both represent the mean of the respective data series.

Sound moves through the air at a speed determined by several physical conditions, namely the temperature, humidity and pressure of the environment. The distance of the source and the values of these variables will ultimately determine the time it takes for the sound to reach the ipsilateral ear. Once a signal is detected, the time the signal takes to reach the contralateral ear is what is used to ultimately determine the azimuth angle as this gives us our time and phase differences. Eqs. (3)–(5) show the propagation of sound through air.

$$c_{air} = (331.5 + (0.6 \cdot \theta))\,\mathrm{m/s} \tag{3}$$

where θ is the temperature in °C of the environment; however a more accurate equation can be seen in (4)

$$c_{air} = \sqrt{k \cdot R \cdot T} \tag{4}$$

$R = 287.05\,\mathrm{J/(kg\,K)}$ for air i.e. the universal gas constant for air with units of $\mathrm{J/(mol\,K)}$, T is the absolute temperature in Kelvin and k is the adiabatic index (1.402 for air), sometimes noted as γ as in Eq. (5).

$$\gamma = C_p/C_V. \tag{5}$$

The adiabatic index γ of a gas is the ratio of its specific heat capacity at constant pressure (C_p) to its specific heat capacity at constant volume (C_V).

Thus, in order to determine the time delay ITD of the signals arriving at the two microphones, it is necessary to ensure that the time delay measurement between the ipsilateral and contralateral ears is taken between the exact same components (or signal points) within the two recorded signal vectors $g(t)$ and $h(t)$. The main use of the cross-correlation function within our model is to determine the maximum point of similarity between the signals contained within the signal vectors $g(t)$ and $h(t)$. Finding the correlation point enables the delay or '*lag*' between the signals to be determined. Therefore, allowing the angle of incidence to be calculated. Fig. 8 shows two signals that are out of phase with each other and thus creating a time of t_{lag} lag between them.

As its input, the cross-correlation function takes two single row vectors representing a digitally sampled version of the signals recorded by the robot's microphones. These signal vectors are then analysed and an output is produced. The output is a single row vector containing the product sum of the values within the initial data series. Table 1 gives an example of the resultant output or correlation vector from a set of inputs.

To determine the point of maximum similarity, or highest correlation, the two recorded signal vectors are offset against each other. Then, using a sliding window the vectors are computed for maximum similarity. The result of this is a correlation vector $\boldsymbol{C}$, as shown in Table 1, whose size is determined by

$$C_{SIZE} = 2 \times N - 1, \tag{6}$$

Table 1
The correlation vector's values are created from the two signals $g(t)$ and $h(t)$ during the sliding-window process with the maximum point of correlation shown at position 12.

Vector element	$g(t)$ and $h(t)$	Vector element value
1	1111232111000000000 0000000001111112321	1
2	111123211100000000 000000001111112321	2
3	11112321110000000 00000001111112321	3
4	1111232111000000 0000001111112321	5
.....	.....	.....
11	01111232111 11111123210	21
12	**001111232111** **111111232100**	**22**
13	0001111232111 1111112321000	19
.....	.....	.....
17	00000001111232111 11111123210000000	6
18	000000001111232111 111111232100000000	3
19	0000000001111232111 1111112321000000000	1

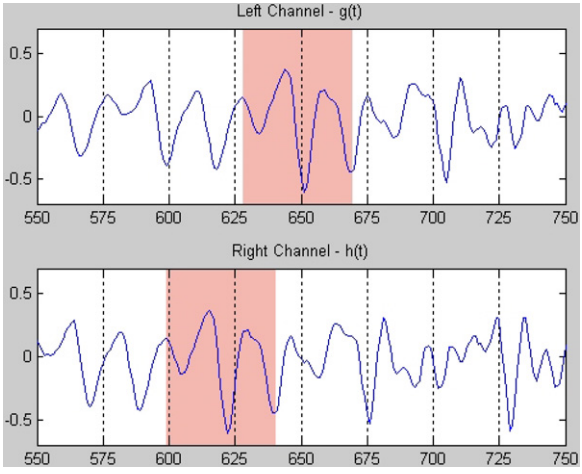

Fig. 9. The data series vector for signal $h(t)$ at $-t$ (lagged) position during the cross-correlation phase.

with N representing the length of the recorded signal vectors $g(t)$ and $h(t)$.

Figs. 9–11 show a signal recorded by the microphones, with the cross-correlation function being applied. Fig. 9 shows the signal vector $g(t)$ at a negative lag ($-t$) with respect to when it was detected and thus recorded by the microphone compared to when $h(t)$ was detected and recorded. The similarity points between $g(t)$ and $h(t)$ are shown in the shaded area for clarity. Fig. 10 shows the two signal vectors when they are in-phase (shown by highlight), and finally the vector $g(t)$ at a positive lag ($+t$) with respect to the signal $h(t)$. The resulting correlation vector is shown in Fig. 12.

4.2. Determining the angle from cross-correlation

The resultant cross-correlation vector represents the IPD of the signals in the recorded signal vectors. The correlation vector $\boldsymbol{C}$ gives the result as a number of time sample increments, which is the number of samples Δt, recorded by the robot, that the signals are out of phase. The values within the actual correlation vector locations, as shown in Fig. 12, correspond to how correlated the two signals are, at various steps of the cross-correlation process,

Please cite this article in press as: Murray, J. C., et al. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. Neural Networks (2009), doi:10.1016/j.neunet.2009.01.013

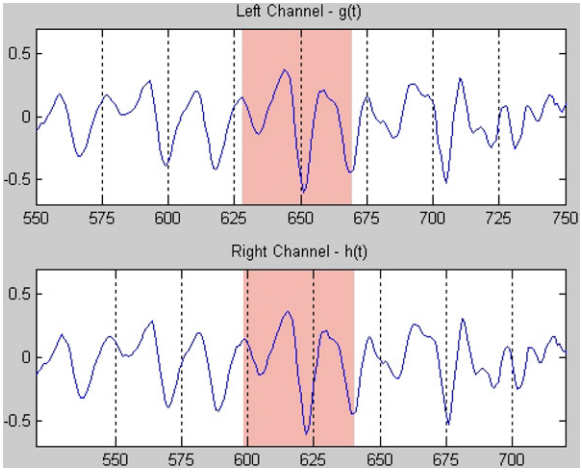

Fig. 10. The data series vectors $h(t)$ and $g(t)$ in phase and so at maximum point of correlation.

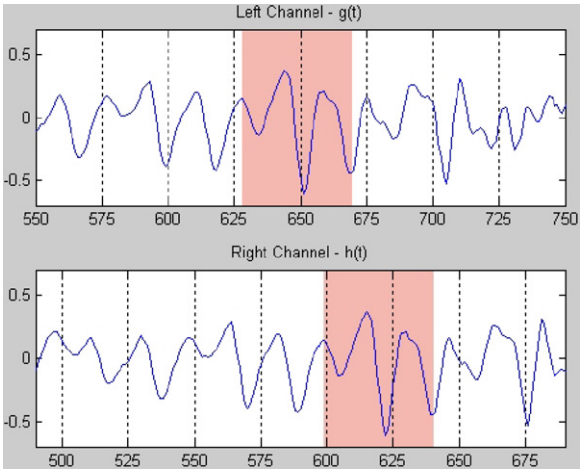

Fig. 11. The data series vector for signal $h(t)$ at $+t$ (leading) position during the cross-correlation phase.

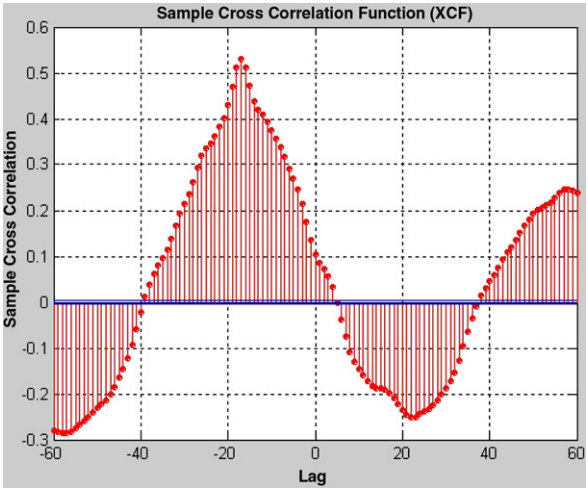

Fig. 12. The correlation vector $\boldsymbol{C}$ produced from the cross-correlation of signal vectors $g(t)$ and $h(t)$ with a lag or offset between maximum similarity of -17.

caused by the sliding-window effect. The smaller the value (in relation to the maximum value) the less similar the signal vectors at that correlation point. The larger the value the higher the similarity.

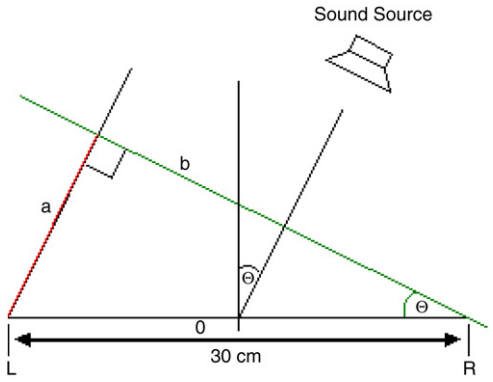

Fig. 13. The change in azimuth of the source will affect the length of 'a' i.e. the ITD.

The maximum value within $\boldsymbol{C}$, as shown in Fig. 12, represents the point at which maximum correlation occurs, or the point of maximum similarity: In this case with $h(t)$ at a lag of 17 with respect to $g(t)$. Thus, from the correlation vector $\boldsymbol{C}$ the angle of incidence along the azimuth place can be calculated.

$$\Delta t = \frac{1}{f}. \tag{7}$$

The sound system on the robot is capable of recording at a maximum sample rate of 44.1 kHz, therefore from Eq. (7) substituting in 44.1 kHz for f it can be seen that each of the samples within the recorded signal vectors are taken at time intervals of 22.6 μs.

To determine the ITD directly from the information in Fig. 12 Eq. (8) is used where σ represents the offset returned by the cross-correlation function and Δt the time between sound samples determined by Eq. (7). This gives us the time delay for the signal to travel from the ipsilateral microphone to the contralateral microphone. As there are 17 samples of phase difference between the two recorded signal vectors shown in Fig. 12, then substituting this data into Eq. (8) we find the ITD $=$ 384.2 μs. However, in order to determine the angle of incidence of the source, as shown in Fig. 13, the ITD value from Eq. (8) is substituted into Eq. (9).

$$\text{ITD} = \Delta t \times \sigma \tag{8}$$

$$\Theta = \sin^{-1} \frac{(c_{air} \times (\sigma \times \Delta t))}{c}. \tag{9}$$

Using Eq. (9) it can be seen that with a sound source that produces a correlation vector as shown in Fig. 12 and thus produces a phase lag of 17 samples, giving an ITD $=$ 384.2 μs, the angle of incidence is calculated to be approximately $+26.4°$.

5. Acoustic tracking

The acoustic model presented within this paper allows the robot to not only localise and attend to the sound-source angle of incidence, but also to track a dynamic source as it moves within the environment. This ability to track the source has two main uses; firstly, it allows the robot to be able to hone in on a target source, even if this target is moving from point to point, therefore providing improved performance over systems such as those proposed by Macera et al. (2004). This improvement is due to the position of the sound source being corrected in real-time allowing for an interception point; this is particularly useful for service robot scenarios where people may be moving around the environment. Secondly, acoustic tracking on the target will provide optimal signal-to-noise ratio (SNR) levels between the target and the robot, thus reducing the levels of uninteresting or background signals that may also be present within the environment.

Please cite this article in press as: Murray, J. C., et al. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. Neural Networks (2009), doi:10.1016/j.neunet.2009.01.013

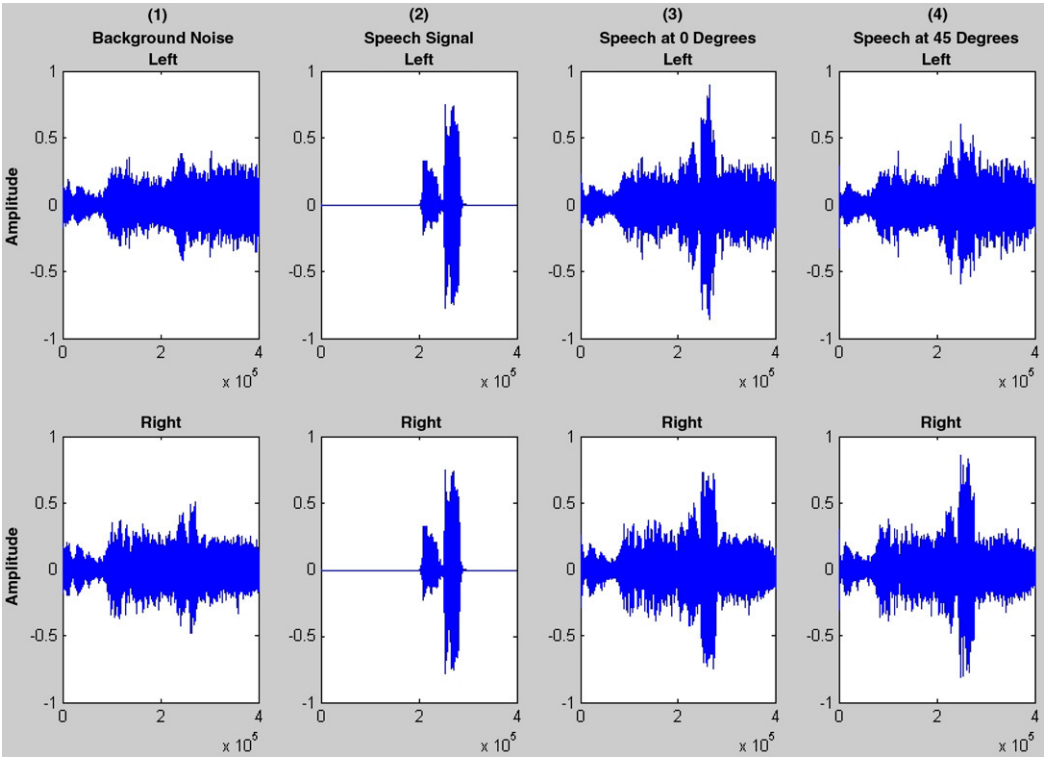

Fig. 14. (1) The left and right channels of the clutter source, (2) the speech signal, (3) both the background and the speech signals recorded at 0° azimuth, (4) the background signal at 0° and the speech signal at 45° azimuth.

5.1. Improving signal-to-noise ratio

The further the microphones are from the sound source the lower the amplitude levels of the received signal, due to degradation of the signal over distance. Therefore, to maximise the levels of the signal to be interpreted and tracked, it is helpful to keep the receivers as close to the sound source as possible, to maintain high amplitude levels of the incoming signal with respect to the background noise. When operating in an acoustically cluttered environment, it is somewhat difficult to detect and localise the source of interest due to the interference from other sources. This is seen in the phenomenon known as the 'Cocktail Party Effect' (Girolami, 1998; Newman, 2005). Hence, maintaining an acoustical track on a sound source within the environment is also useful in reducing the levels of background noise interference received by the system.

Fig. 14 shows the effects on the SNR when the robot is facing the sound source. The first two columns show the signals, from the left and right channels, for the independent signals (background clutter and speech) used to demonstrate the principle of improving the SNR. The plots also show the position in time when the sounds occur. The third and fourth columns show the left and right signatures for two test examples. The third column shows the background clutter and speech signal both positioned at 0° azimuth (directly in front of the microphones). The fourth column shows the background source remaining fixed at 0° with respect to the microphones, with the speech signal moved to an azimuth angle of 45° to the right of the microphones. The microphone and source setup is shown in Fig. 15.

In Fig. 14, the signal pairs, in column three, are fairly similar in their amplitude levels. This is due to the two sources being positioned at equal distances from the microphones and therefore the ILD will be similar. The signatures shown in the fourth column differ however, with the left channels' background clutter almost entirely obscuring the speech signal. It can be clearly seen that in the signature for the left microphone the speech signals' sound pattern is no longer distinguishable from that of the background source, whereas the speech signal is still clearly visible within the right microphones' plot. Therefore, one can clearly see the impact tracking has on maintaining an optimal SNR between the various signals, thus additionally showing the importance sound-source tracking plays in robot–human interaction.

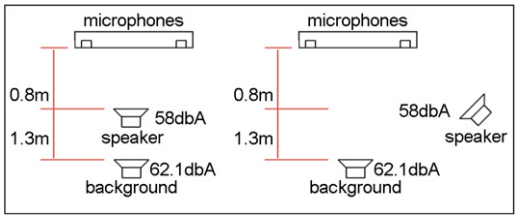

Fig. 15. The configuration for the microphones and sound sources used for SNR tests, with the levels of the signals shown in dBA.

5.2. Tracking with a recurrent neural network

In order to track the sound source, simply detecting the signal and computing the cross-correlation would prove insufficient due to the finite processing time required by recording the data, the cross-correlation algorithm and the instruction to move the robot to the desired position. This would result in the robot consistently lagging behind the source by some finite time determined by the above factors. Therefore, a predictor–corrector approach is taken in the model presented here. This approach employs the use of a recurrent neural network (RNN) to effectively learn the trajectory of the sound source by using current and previous positions to estimate the future location of the source allowing for a more accurate and real-time track. The RNN is able to perform this task using recurrent context layers, allowing the system to remember some iterations, determined by the hysteresis value (see Eq. (10)),

Please cite this article in press as: Murray, J. C., et al. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. Neural Networks (2009), doi:10.1016/j.neunet.2009.01.013

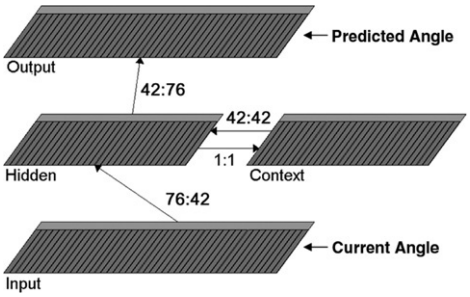

Fig. 16. The structure of the recurrent neural network.

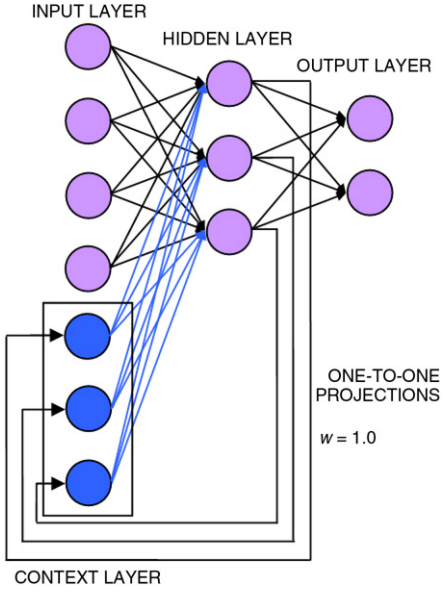

Fig. 17. The 1:1 projections of the context layer to the hidden layer of the simple recurrent neural network.

the previous positions of the sound source relative to the robot from the context layer and use these to predict an estimated future position.

$$\alpha C_i = (1.0 - \Phi) \times \alpha H_i + \Phi \times \alpha C_i \tag{10}$$

where Φ is the hysteresis value, αC_i is the activation of the context unit i, αH_i is the activation of the hidden unit i. With a larger hysteresis value the context units respond to the temporal events more slowly and are resistant to change. A smaller hysteresis value makes the context units incorporate information more quickly, holding it for a longer period of time.

Fig. 16 shows the overview of the RNN used in the acoustic model. The network has 76 input and 76 output units that represent the range of angles available from $-90°$ to $+90°$. There are 42 hidden and 42 context units, as each hidden unit is connected directly via 1:1 projections to the context units, see Fig. 17. Each of the input and output units map directly to the azimuth increments that the model is capable of detecting as determined by Eq. (9). Therefore, input unit 5 for example is responsible for mapping the azimuth angle 7.55° as can be ascertained from Eq. (9) substituting 5 for σ. Fig. 18 shows the angle representation of each of the input and output units.

As can be seen from Fig. 18 the system is more sensitive around the low angles, that is, angles less than $\pm 45°$ (with 0° being on the midline directly in front of the robot) as each sound sample increment σ represents a smaller range of angles due to the relation between angle and sound samples being plotted along a Tan curve as shown in Fig. 30. Therefore, enabling the robot to track

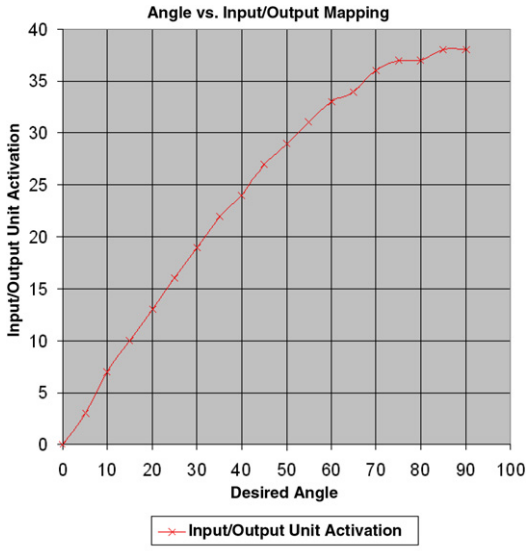

Fig. 18. The mapping between input/output unit and azimuth space.

the sound source will allow it to keep the source within this range and thus providing more accurate localisation results.

Eq. (11) shows the formula for determining the context layer activations at time t_{n+1}. The RNN used in this model determines the next location of the sound source at t_{n+1} along its trajectory. This network is based on a simple recurrent network (Elman, 1990).

$$a_i^C(t+1) = a_i^H(t). \tag{11}$$

The RNN is designed to accept input directly from the azimuth estimation stage of the model, with activation on the unit that represents the azimuth angle, calculated by the previous stage of the system, as shown in Fig. 18. The network is trained using the back propagation weight update rule shown in Eqs. (12) and (13) (Rumelhart, Hinton, & Williams, 1986). Eq. (12) represents the weight update rule for a standard multi-layer perceptron (MLP). However, as a RNN can be unfolded in time to represent a many layered MLP, this can be rewritten as shown in Eq. (13).

$$\Delta w_{ij}(n+1) = \eta \delta_j a_i + \alpha \Delta w_{ij}(n). \tag{12}$$

$\Delta w_{ij}(n+1)$ is the weight change to be applied to the connection between units i and j at pattern presentation $n+1$. This weight update is calculated during the presentation of pattern n and is affected by several factors. η represents the learning rate, and in experiments was set to 0.25 (derived from previous experimentation). This is multiplied by δ_j which represents the error of unit j and a_i the activation from unit i.

$$\Delta w_{ij} = \eta \sum_i \delta_i(t) a_j(t). \tag{13}$$

The weight update is determined by the learning rate η multiplied by the sum of the product of the error of unit i and the activation of unit j at each time iteration t.

The RNN is provided with several training sets that are used to allow the network to learn the temporal differences between the possible source speeds. The training sets provide the network with input activation and the desired output activation at the various time steps. For example, if the sound source is moving at a speed of 4° per sound sample, then a training set for this speed would have activation on input unit 1 for time t_0. For the second iteration, activation would be on input unit 3 at time t_1, and a target output activation on output unit 5, for time t_2, would then be set. Figs. 19–21 show examples of three different training

Please cite this article in press as: Murray, J. C., et al. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. Neural Networks (2009), doi:10.1016/j.neunet.2009.01.013

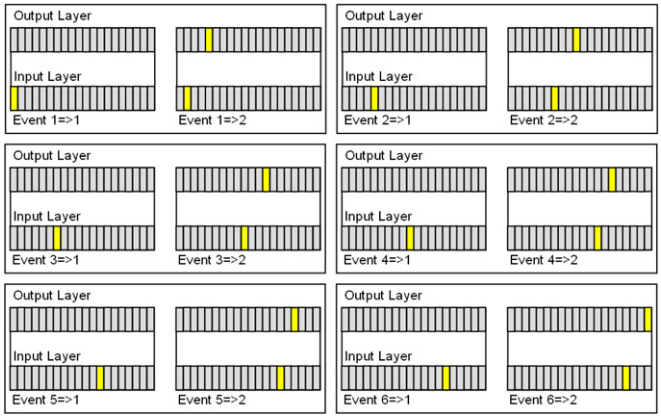

Fig. 19. The first 12 events of the training data for a speed of 4° per iteration.

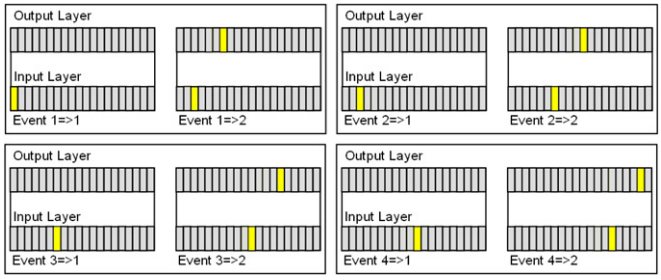

Fig. 20. The first 8 events of the training data for a speed of 6° per iteration.

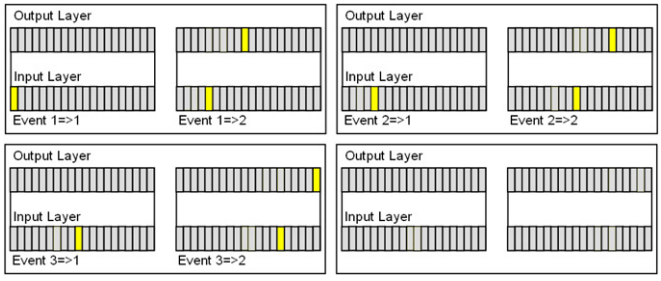

Fig. 21. The first 6 events of the training data for a speed of 8° per iteration.

sets, providing training data to the network, that resembles three different possible sound-source speeds.

As can be seen in Figs. 19–21 the input patterns are put into subgroups of two events each. These subgroups represent an angle speed of $x°$, starting at a particular input unit. It is necessary for the RNN to learn the sequential input patterns that a particular source may provide, however it is not necessarily required for the network to learn the temporal sequence of the events across subgroups. Fig. 19 shows some of the events for a speed of 4°, Fig. 20 for 6° and Fig. 21 for 8°.

The network is presented with the events of the subgroups in a sequential manner in order to maintain the temporal coding, i.e. with event x => 1 always preceding event x => 2. Once a particular subgroup has been presented to the network, constituting one epoch, then another subgroup is randomly chosen for presentation.

5.3. Attending only to the target source

As mentioned in Section 2 the third biological constraint is the adaption to sound levels, in addition to attending to a desired source. When the system is deployed in a real-world environment it is necessary to ensure that the robot does not attend to every

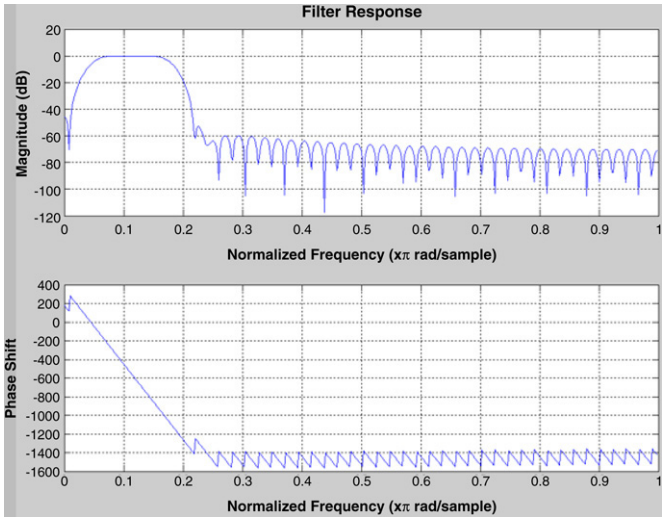

Fig. 22. The response of the bandpass filter for allowing only speech to pass.

sound in the room, that is detected, as this would create very sporadic behaviour, and prevent the robot from easily tracking the target sound source. Therefore, within the model three methods have been introduced to help reduce this effect. Firstly, a bandpass filter has been introduced to the system, filtering out any sounds that do not fall within the frequency range of human speech. This ensures that anything outside the range of 1 kHz to 4 kHz (de Boer, 2005) is not analysed by the system. Fig. 22 shows the response of the bandpass filter. Secondly, the use of an energy function is employed to gain a measure of the energy contained within the sound by analysing the duration and amplitude of the signal. This prevents short signal bursts, that may fall within the bandpass filter range, from influencing the robot to attend to its source location.

Eq. (14) shows the energy function used in this model; the function integrates the square of the amplitude at each sample point within the recorded signal vectors thus giving a relative measure of how much energy exists within the signal. This function is used to help decide if the signal may be of interest due to a combination of its duration, frequency band and energy.

$$\epsilon = \sum_{i=1}^{n} [y_i]^2. \tag{14}$$

Fig. 23 shows two signals detected within the environment and located at various azimuth positions. The first signal shown in the figure represents background clutter within the environment, i.e. sounds that are not intended for localisation. The second signal is that of an '*interesting*' sound source that is intended to be detected by the system and ultimately localised. Therefore, Fig. 23 shows the respective amounts of energy contained within the two samples, with the '*interesting*' signal having much higher amounts of energy than the clutter, especially when the durations of the signals are taken into account. The background clutter sources were generated by things such as, doors opening and closing and sounds outside coming through the window. As can be expected the amount of energy contained within these recorded sounds is very low in comparison to the energy contained within a spoken word as shown in the second plot.

Thirdly, the model uses the local sound levels and feedback to normalise to the baseline level of the sounds in the environment. The mammalian CAS has the ability to attenuate or amplify the sounds it hears (Géléoc & Holt, 2003) in order to increase or decrease the sound levels. This is particularly useful within a robotic scenario as the levels within the environment are

Please cite this article in press as: Murray, J. C., et al. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. Neural Networks (2009), doi:10.1016/j.neunet.2009.01.013

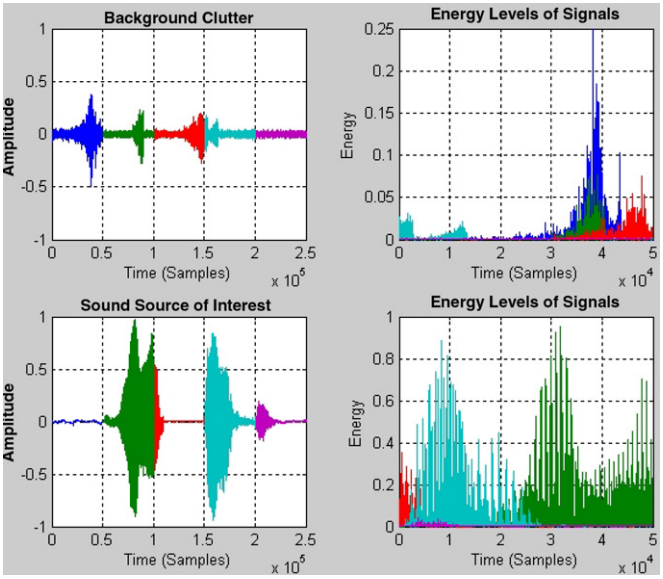

Fig. 23. Compares the energy signatures of an interesting sound source vs. background clutter. Note: The plot for the energy levels of the sounds generated by the background clutter has the *y*-axis magnified by a scale of 4 for display purposes.

subject to constant change. In addition, different environments themselves have different acoustic levels. Therefore it is useful for the model to adapt to this and allow an increase in sensitivity in quiet environments and a decrease in environments that are too loud, over modulating and causing distortion. If the sampled environmental levels are below a particular baseline then the system increases the gain incrementally by +3dB or if the levels are overmodulating then they are incrementally reduced by −3dB. Fig. 24 shows how this affects the sampling of a signal.

Together, these three factors constitute the third biological constraint, allowing the model to selectively determine the type of sound source that it attends to, in addition to ensuring correct sound levels within the environment. This increases the accuracy and robustness of the model in attending to sound sources of interest such as a human voice.

6. Localisation and tracking model

The model within this paper has three main components with a fourth control stage. Which include the filtering and attention stage, azimuth estimation stage, and the tracking and prediction stage with a fourth stage providing motion control. Each of the different stages have several sub components that together provide the functional model.

The first stage of the system starts from the signals recorded at the microphones and begins by filtering any unwanted signals. Then a normalisation against the background is performed to ensure that the robot performs optimally within that environment. Finally the model waits for signals and computes the energy function to ensure that they contain the correct properties. The system here is the first to incorporate self-normalisation and energy functions to determine the source of interest.

The second stage, azimuth estimation, receives the recorded signal vectors from the previous stage. These vectors are then processed in accordance with the description given in Section 4. This allows for the azimuth position of the sound source to be determined and then passed on to the next stage which is the tracking and prediction stage.

The third stage of the model is responsible for the motion tracking of the source as it traverses along its trajectory. It is necessary for the model to be able to, not only track the source as it moves, but also to be able to estimate the position of the source at time t_{n+1}. This allows the system to be able to provide a faster response to the tracking of the source, as opposed to continually sampling–estimating, sampling–estimating, etc. Once the RNN is estimating the future position of the source through the environment the system is instructed to attend to this estimated future position, thus enabling the system to maintain a closer track on the sound source.

The tracking and prediction stage contains a simple recurrent neural network and is used for estimating future positions along the trajectory of the source as it moves within the environment.

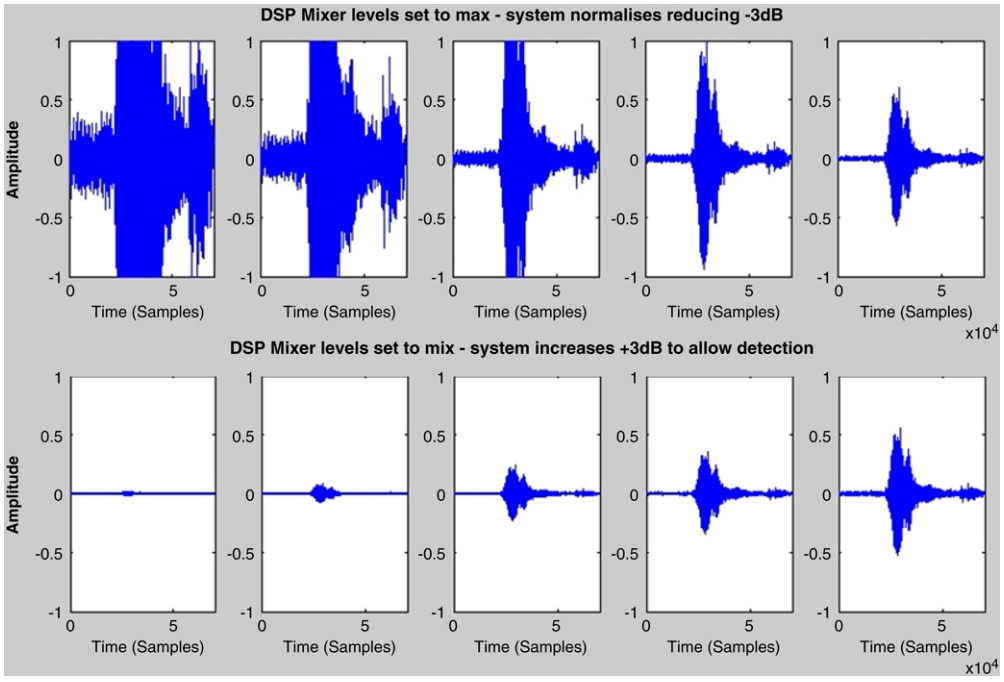

Fig. 24. The normalisation stages for reducing by −3dB or increasing by +3dB the gain on the digital signal processor mixer within the acoustic model.

Please cite this article in press as: Murray, J. C., et al. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. Neural Networks (2009), doi:10.1016/j.neunet.2009.01.013

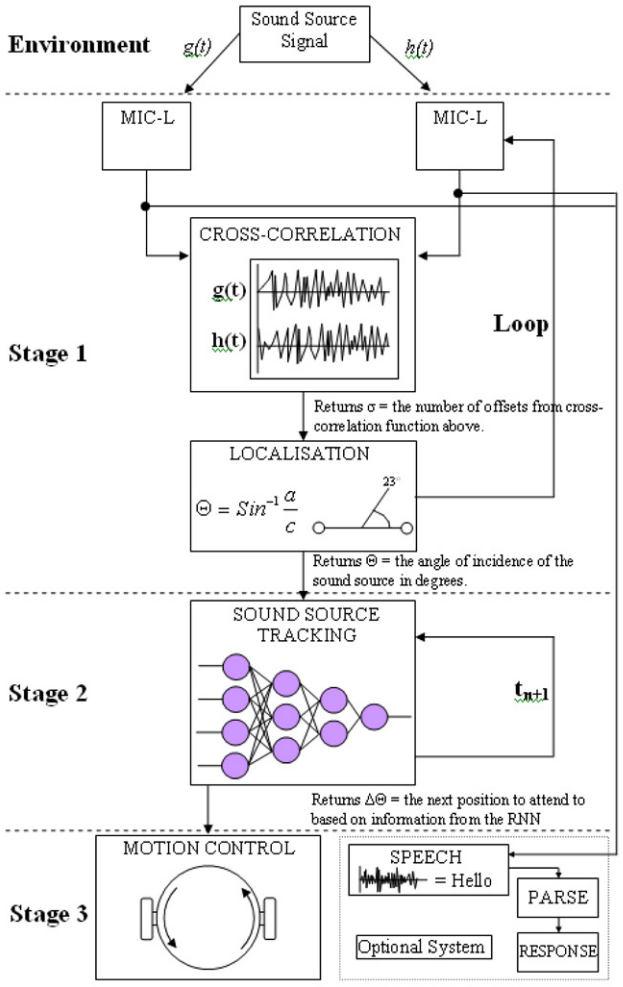

Fig. 25. The complete system model developed for sound-source localisation and tracking.

The recurrent network estimates the angle of the source at time t_{n+1} based on the previous measurements. With the input to this stage coming from the previous azimuth estimation stage, as shown in Fig. 25, in the form of an angle value representing the source position at t_n. This unit activation represents the angle calculated by the network to represent the estimated position of the source at t_{n+1} which is then presented to the next stage within the model for processing. Stages two and three provide a novel hybrid architecture for sound-source localisation and tracking by combining cross-correlation with recurrent neural networks to provide a robust and accurate model, which provides increased performance and response times over existing models.

The final stage of the model is the motor control stage and as such its main purpose is to control the position (direction and forward movement) of the robot based on the input received from the hierarchical stages discussed above. The input to this stage comes directly from the recurrent network, in the third stage used for estimating the trajectory of the system, and is therefore in the form of an angle value instructing the robot to turn a specified amount. The output of the motor control stage comes in the form of direct movement of the robot itself, positioning to the required angle of incidence of the source.

7. Experimental design

The first method of testing the azimuth estimation stage of the model is to see if cross-correlation is able to determine the angle of incidence of a source, within the environment, represented by two vectors $g(t)$ and $h(t)$, containing the digital representation of the signals detected. Simulated waveforms are created that represent a signal at varying azimuth angles of incidence, to test the cross-correlation method. To create the waveforms it is necessary to record a mono signal at 44.1 kHz. This frequency generates samples at intervals of approx. 22.67 ms as determined by Eq. (7). The sound file is recorded in mono to ensure that when the signal is duplicated to create the stereo file the left and right channels are identical; the mono signal is then copied into both the left and right channels. From this file it is then possible to create any required simulated angle of incidence within the constraints of Eq. (9), which shows the angle represented by a particular delay offset σ.

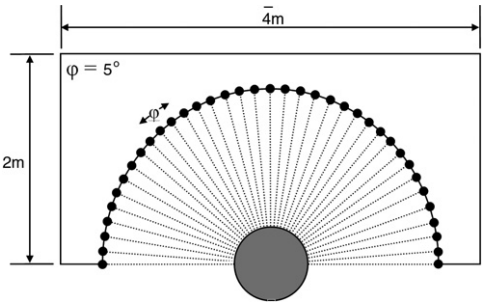

Fig. 26. The experimental setup showing the sound-source locations relative to the robot.

To provide the azimuth estimation stage with a signal whose source appears to originate from the left hemisphere of the robot, the right channel within the wave file needs to be shifted back in time, thus delayed (offset) from the left channel by a specific number of samples (dependent on the effective angle of incidence desired). This provides a simulated Interaural Time Difference between the left and right vectors in the wave file creating a delay between the ipsilateral and contralateral microphones. The offset required to achieve the desired effect involves shifting the right channel backwards a specific number of samples. That is, in '$-t_n$' as, if the source is to the left of the robot in real-world conditions, then the right (or contralateral) microphone would be delayed in receiving the signal. The size of the delay needed for a particular angle of incidence is calculated by transposing Eq. (9) to give

$$\sigma = \frac{\mathrm{Sin}\,\theta \times c}{c_{air} \times \Delta t} \tag{15}$$

where σ is the number of offsets required, θ is the angle that is to be simulated, c is the distance between the two microphones, c_{air} is the speed of sound in air and Δt is the time delay between samples.

In order to determine the accuracy of the model, with free form sounds, several experiments were set up to see if the robot could orientate itself to face the direction of the sound source. The robot is positioned at a specific location and a sound source is placed at varying positions within the lab to determine if the robot can localise the source at these positions and to what accuracy. Firstly, the sound source is positioned at a distance of 1.5 m from the robot at 5° increments giving a total of 37 individual source positions, as shown in Fig. 26. The initial source angle is −90° and the final position is +90°.

Secondly, the robot is placed at six different locations within the lab with four separate statically placed sound sources used to generate background clutter. These are placed in the corners of a 5 m × 5 m square. A dynamic source is then placed within the environment at increments of 15° around the robot as shown in Fig. 27.

The second individual component of the model for solitary testing is that of the tracking and prediction stage, whose flow

Please cite this article in press as: Murray, J. C., et al. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. Neural Networks (2009), doi:10.1016/j.neunet.2009.01.013

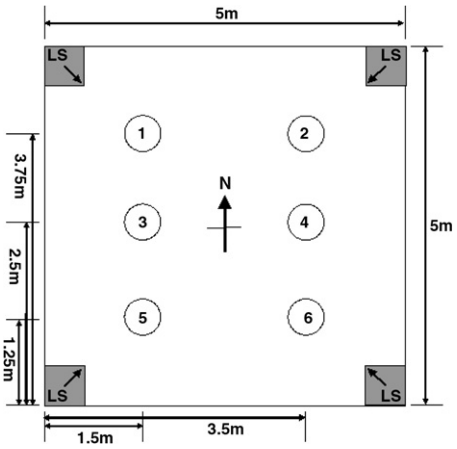

Fig. 27. The experimental configuration for the trials conducted under *real-world* conditions, showing starting positions (1–6) for the robot and locations of static background sources (LS).

of control follows on from the azimuth estimation stage. For experimentation purposes due to this stage being tested as an individual component, the input is not provided by direct output from the azimuth estimation stage, but rather is provided via several other methods as described below.

First, test data is systematically created by a pattern generator which provides the input to the network. Second, randomly generated test data is used to produce varying valid and invalid sequential patterns for presentation to the network and finally test data recorded from the azimuth estimation stage during its testing phase are used for presentation to the tracking and prediction algorithm.

Figs. 28 and 29 show a 270° panoramic view of the test environment used to measure the performance of the model. Fig. 28 shows the motion of the sound source as it moves through the environment (represented by the blue line). It also shows the motion and response of the robot as it tries to track the source 'without' the use of motion prediction via the RNN (shown by the red line). In Fig. 29 the green line represents the motion and response of the robot in its attempt to track the source. This time the acoustic model is equipped with the tracking and prediction RNN described earlier.

8. Results

As previously mentioned, the azimuth estimation stage of the model was initially tested with simulated waveforms of varying angles determined by changing the number of offsets in which the two channels are shifted relative to each other. Table 2 shows angle ranges of $-90° \rightarrow +90°$ in azimuth with increments of 5°. It also shows the calculated number of offsets required for the specified angle in addition to the actual physical number of offsets that can be used with the associated azimuth angle. Due to the range of angles and samples between 0° to +90° and 0° to −90° being symmetrical only one range is shown in Table 2.

The simulated waveforms are presented to the model five times to ensure repeatability and accuracy. Therefore, the values shown in Table 2 are the average results over the total number of trials. Also shown is the actual angle a specific number of offsets will produce from the cross-correlation function from the simulated angles generated by Eq. (15), using the number of physical offsets available. Looking at Table 2, we can see that the azimuth angles ±90° and ±85° both fall within a range of 38.2–38.5 samples. However, it is not possible to measure an increment of time beyond the highest sample rate available as determined by Eq. (7).

Only whole samples can be used during the cross-correlation phase when the system is determining the angle of incidence of the source. Therefore, a lag or delay of 38 samples in Table 2

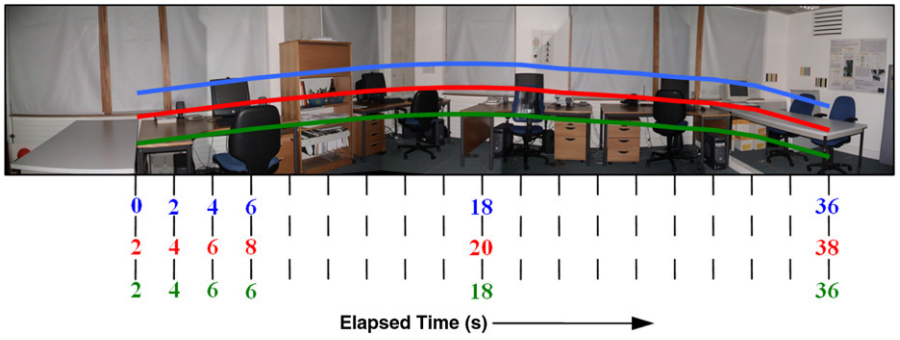

Fig. 28. Localising and attending to the sound source within the environment without the use of predictive tracking RNN. Red is the position of Robot as it tracks the source, Blue is showing a 2 s lag. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

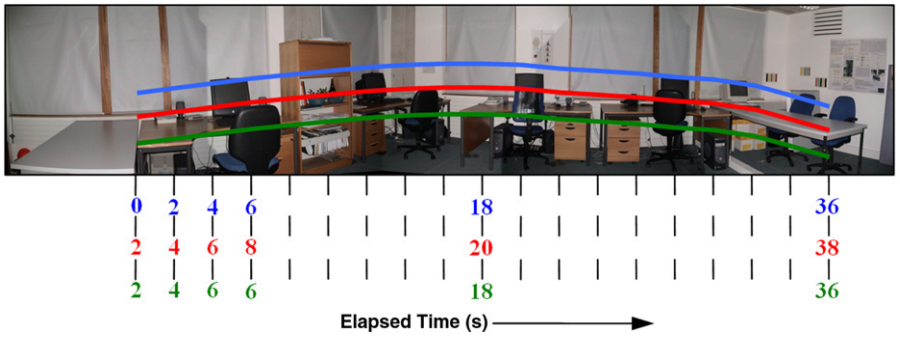

Fig. 29. The response of the complete model tracking the sound source (Blue). Green = Position of the robot when equipped with the predictive tracking RNN. Showing a 2 s delay for the first three time steps and then the system catches up to the source. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: Murray, J. C., et al. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. Neural Networks (2009), doi:10.1016/j.neunet.2009.01.013

Table 2
The calculated offset required for a particular angle vs. the actual offsets determined by the robot.

Actual angle (°)	Calculated offsets	Physical offsets	Angle from offsets (°)
±0	0	0	0
±5	±3.34	3	4.49
±10	±6.66	7	10.52
±15	±9.93	10	15.12
±20	±13.12	13	19.82
±25	±16.21	16	24.66
±30	±19.17	19	29.70
±35	±21.99	22	35.01
±40	±24.65	24	38.74
±45	±27.12	27	44.75
±50	±29.38	29	49.13
±55	±31.41	31	53.94
±60	±33.21	33	59.38
±65	±34.76	34	62.45
±70	±36.04	36	69.85
±75	±37.04	37	74.76
±80	±37.77	37	74.76
±85	±38.20	38	82.28
±90	±38.35	38	82.28

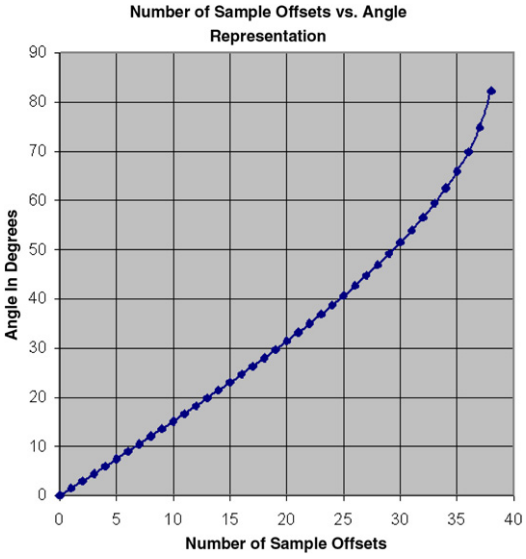

Fig. 30. The relation between number of sample offsets and the represented angle of incidence of the sound source.

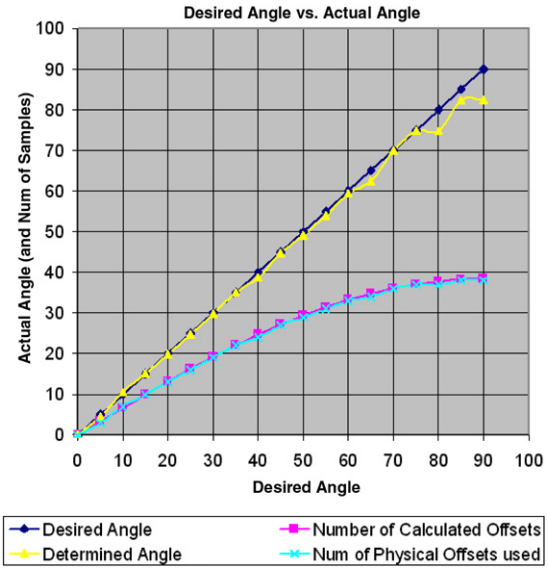

Fig. 31. A plot of desired simulated waveforms versus the actual physical angle, also shown is a comparison between required number of offsets and physical number of offsets.

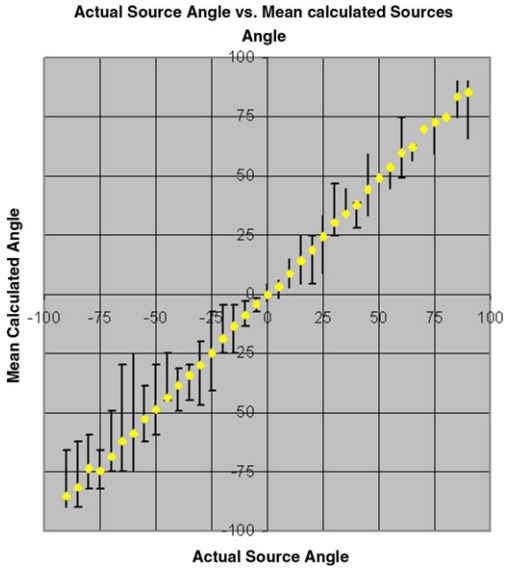

Fig. 32. The average azimuth estimation results including error bars.

would represent the angles ±85° and ±90°. However, the total angle range for sample number 38 encompasses 82.28°–90° and therefore the current system shown in this paper would always attend to an angle of 82.28° for this number of delay samples.

Fig. 30 shows the angle representation of a particular number of sample offsets used by the cross-correlation function. Using the plot in Fig. 31 the number of samples required for any desired angle within the range ±90° can be determined. It can also be seen that, from approximately the last ten offset samples, the angle representation begins to increase more dramatically in its gradient. This shows that with each sample offset increment, the difference in angles between successive sample offsets increases. This provides less accuracy for the localisation of the sound source due to a particular sample representing a larger range of angles on the azimuth plane.

Fig. 32 shows the distribution of the calculated (or measured) angles from each of the source position increments used in the sound signals experimental trials described in the above experimentation section. The graph shows the angles determined by the azimuth estimation component of the model, versus the actual position of the source, within the environment. As can be seen from this graph the system's minimum and maximum errors are at their maximum at ±90° and at their minimum at 0°.

The plots in Figs. 33–35 show that the estimated (mean) angle of the source versus the actual source position was fairly accurate, with no large noticeable errors in the estimation of the source position. The results show that when the robot attends to the estimated azimuth position, provided by the cross-correlation stage, the actual attended angle is in some cases between 0° and 4° less than the actual estimated position. This induced error is due to several factors, namely friction between the floor and the robot wheels, the accuracy of the motors when moving the robot in small discrete angles, due to lack of momentum.

The results shown in Tables 3 and 4 show the readings from the six separate starting points for the azimuth tests shown in Fig. 27. For each of the starting positions, several trials are run and the results averaged over all trials. The recordings made from the tests are actual source angle (relative to the start position of the source taken to be 0°), recorded angle (as estimated by the azimuth component of the model) and attended angle (the position

Please cite this article in press as: Murray, J. C., et al. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. Neural Networks (2009), doi:10.1016/j.neunet.2009.01.013

Table 3
The calculated trials for positions 1, 2 and 3 for the azimuth estimation stage.

Start point 1		Start point 2		Start point 3	
Estimated angle	Attended angle	Estimated angle	Attended angle	Estimated angle	Attended angle
−90	−87	−90	−87	−90	−88
−74	−74	−74	−74	−69	−69
−59	−57	−59	−58	−59	−58
−46	−45	−44	−43	−46	−44
−29	−27	−31	−29	−29	−28
−15	−15	−15	−13	−15	−13
0	0	0	1	1	0
16	15	15	14	15	14
31	31	29	27	29	29
42	40	42	41	42	40
59	57	59	58	59	57
74	72	74	71	74	72
90	88	82	80	90	87

Table 4
The calculated trials for positions 4, 5 and 6 for the azimuth estimation stage.

Start point 4		Start point 5		Start point 6	
Estimated angle	Attended angle	Estimated angle	Attended angle	Estimated angle	Attended angle
−90	−88	−82	−80	−90	−84
−74	−72	−69	−68	−69	−68
−62	−60	−62	−59	−62	−60
−46	−44	−46	−45	−46	−44
−29	−29	−33	−30	−33	−30
−13	−12	−18	−16	−16	−15
0	0	0	1	0	− 1
15	14	16	14	15	15
27	26	31	29	31	30
46	44	42	40	44	43
59	56	59	57	59	58
74	70	74	72	74	73
90	82	90	85	82	82

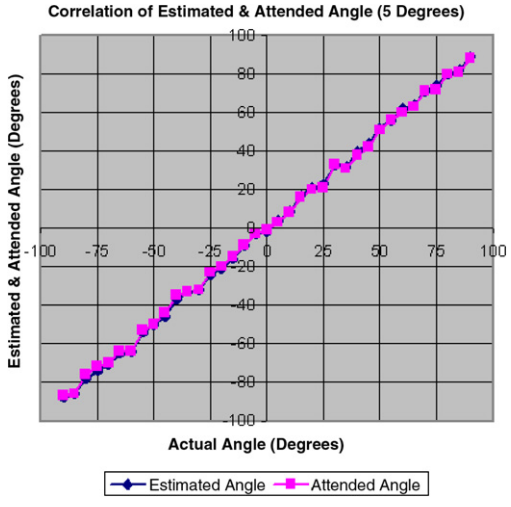

Fig. 33. Correlation of estimated and attended angle, 5° increments.

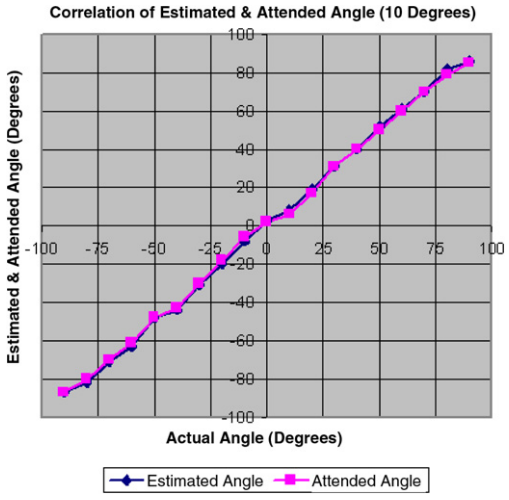

Fig. 34. Correlation of estimated and attended angle, 10° increments.

the robot actually moves to). These different starting positions are chosen to ensure that the static environmental conditions do not affect the results and performance of the system.

The results shown in Tables 3 and 4 demonstrate the difference between the actual position of the sound source and the estimated position of the source, calculated from the azimuth estimation stage of the model. In addition, the tables show the attended angle, that is, the angle the robot turns to when it is instructed by the model in order to face the sound source. It can be seen from the tables that the attended angle was generally the same as, or less than 4° out, than the estimated angle of the source. It is concluded that this angle error is due to motion factors such as friction between the robot and the flooring within the environment in addition to lack of inertia and calibration factors.

It can been seen that the overall results of the systems match up closely. That is, the estimated angle of the source, in comparison to the attended angle, is fairly close to the position of the actual source. In addition, it can also be seen that, regardless of the starting position or starting orientation, the results of the system remain the same. This shows the robustness of the system when positioned at various locations within the environment at multiple orientations.

Please cite this article in press as: Murray, J. C., et al. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. Neural Networks (2009), doi:10.1016/j.neunet.2009.01.013

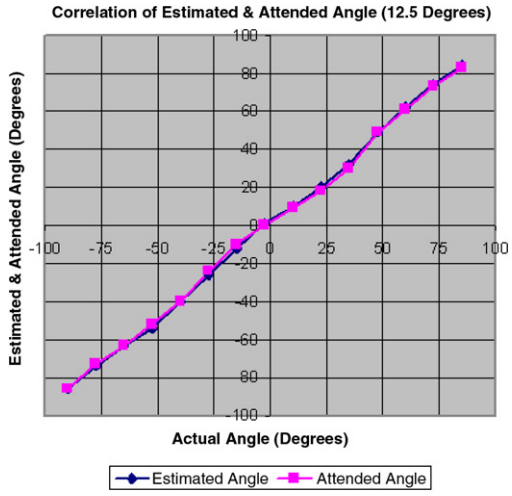

Fig. 35. Correlation of estimated and attended angle, 12.5° increments.

Trial	Step	Input Pattern	Expected Output	Network Output
1	$t0$			
	$t1_1$			
	$t2_1$			
	$t3_1$			
2	$t0$			
	$t1_2$			
	$t2_2$			
	$t3_2$			
3	$t0$			
	$t1_1$			
	$t1_2$			
	$t1_3$			
	$t1_4$			

Fig. 36. Compares input patterns with the expected output patterns for the sequentially created training data.

Once the network has been trained to recognise all the patterns required, experiments on the functioning of the network when applied to the model itself is carried out. The results from the first stage of testing for the RNN requires the use of sequentially created test data. A sample of this test data is shown in Fig. 36. Here the input patterns with the expected output patterns are provided and the network tested to ensure it provides the required desired output.

The results from Fig. 36 show the temporal order of the patterns as they are presented to the network. The time step column in the figure depicts whether it is the first, or second pattern etc. in the temporal sequence to be presented to the network. As many trials begin with the same initial starting point shown at t_0 and the second pattern at t_1 being the variable speed pattern of the source then only the second temporal patterns need to be shown with the t_0 only once per trial.

Fig. 37 shows the response of the network after presentation of pattern 'Trial $1-t_0$' shown in Fig. 36 and following directly after the presentation of pattern 'Trial $1-t1_1$'. As can be seen after the initial input pattern is presented to the network, no output activation is seen on the output layer until presentation of $t1_1$ when the next sequential pattern is provided. Fig. 38 shows the output of the network after presentation of four sequential patterns 't_n+0, t_n+1, t_n+2 and t_n+3'. Here the output from the network can be shown for tracking over a longer period of time than is shown in Fig. 37.

The two main stages of the model, the azimuth estimation stage and the tracking and predictor stage, are coupled together in order

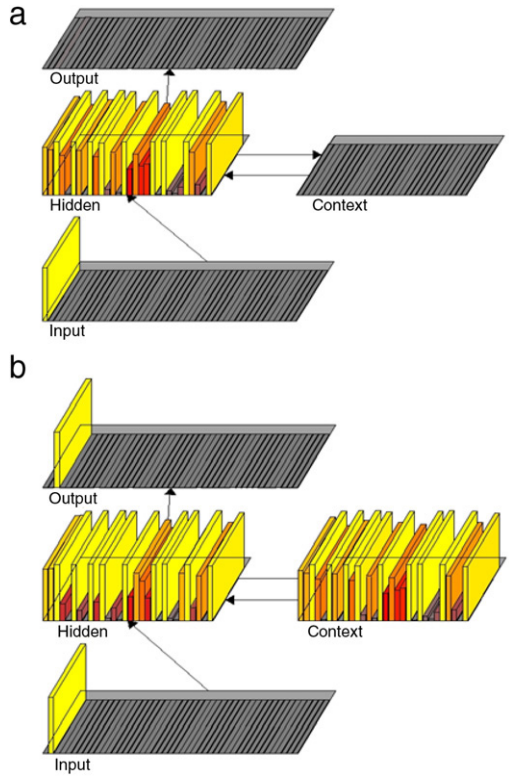

Fig. 37. (a) The network response after presentation of pattern t_0. (b) The network response after presentation of the second pattern t_1.

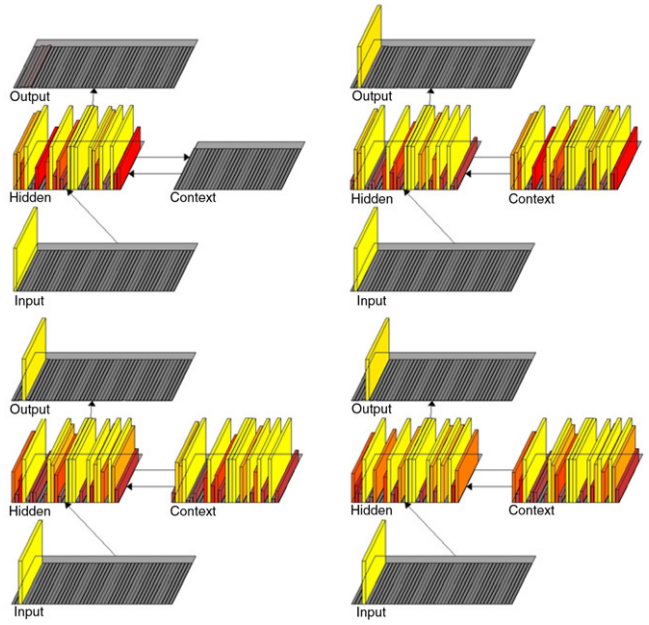

Fig. 38. Four sequential inputs to the network t_{n+0}, t_{n+1}, t_{n+2}, t_{n+3} demonstrating continuous output.

to test the reaction of the system as a whole. Firstly, the combined system is tested off-line, that is, the robot is not incorporated into the system. Table 5 shows the results (over the full set of trials) from the data acquired from the azimuth estimation stage tests.

Special attention should be drawn to the 'Relative Angle' column within Table 5. Due to the network being trained to recognise increasing sequential patterns, when a source is within a negative region of azimuth the system calculates the position as if in the positive range but remembers the negative sign (this

Please cite this article in press as: Murray, J. C., et al. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. Neural Networks (2009), doi:10.1016/j.neunet.2009.01.013

Table 5
Azimuth estimation data set applied as input to RNN stage.

Actual angle (°)	Estimated angle	Relative angle	Input unit activation	Expected output	Actual output
−90	−87.1	2.9	2	0	0
−85	−86.7	3.3	2	0	0
−80	−80.2	9.8	5	8	8
−75	−73.5	16.5	9	13	13
−70	−74.4	15.1	8	0	0
−65	−61.3	28.7	15	22	22
−60	−61.4	28.6	15	0	18
−55	−55.5	34.5	18	21	21
−50	−48.9	41.1	21	24	24
−45	−45.1	44.9	23	25	25
−40	−38.2	51.8	26	29	29
−35	−36.2	53.8	27	28	28
−30	−31.7	58.3	30	33	33
−25	−27.3	62.7	32	34	34
−20	−18.9	71.1	36	40	40
−15	−15.2	74.8	38	40	40
−10	−9.6	80.4	41	44	44
−5	−4.1	85.9	43	45	45
0	1.2	N/A	1	0	0

reduces the size of the network). So therefore $+0°$ to $+90°$ is used as opposed to $-90°$ to $0°$. As can be seen from Table 5 there are anomalies within the classification for the RNN. Upon closer inspection of the network it was found that activation on unit 15 followed by another 15 failed to provide the desired output of 0 and in fact predicted an 18.

The next stage of testing the combined components of the system, is to incorporate the model on the robot itself, and using environmental data, allowing the robot to be free to attend to the source. Table 6 shows the experimentation data and the various inputs and outputs of the cross-correlation and RNN stages of the model. The initial azimuth detection angle when the system begins attending to the sound source is set to be perceived as a relative angle of $0°$.

Table 6 contains several readings. The Source Angle represents the actual position that the source is placed at within the environment during the experimentation, with the Estimated Angle representing the angle returned from the azimuth estimation stage of the model. The Relative Angle shows the estimated angle of incidence in terms of its relation to the previous position and the robot's frame of reference. Attended Angle shows the actual azimuth value that the robot attended to. As previously discussed the angle differences in terms of error are possibly due to factors such as motion friction. The last two columns within Table 6 are related to the RNN aspect of the model, with the Input Activation column showing the unit activated on the input layer of the RNN which represents the relative angle. The final column gives the output activation of the network based on the inputs.

9. Discussion

This paper presents a novel hybrid approach to sound-source localisation and tracking on a mobile robot, deployed within an acoustically cluttered environment. The system has shown that by drawing on inspiration from the mammalian auditory system, including mechanisms such as the Jeffress model (Jeffress, 1948) and Licklider's triplex model (Licklider, 1959), it is possible to create an effective model for the localisation and tracking of sound sources within the environment, with respect to background clutter.

If robots are to become more prevalent within society then social interaction between both robot and human is an important aspect of the robot integration. Therefore, equipping a robotic system with an acoustic modality brings the interaction between humans and robots closer, as it enables humans to interact with robots in a more effective manner. The model here enables a human operator to interact with a robot within a cluttered environment, allowing the robot to know the position and continuously orientate towards and track the location of the controller.

Comparing the results of the model presented in this paper with the results of other similar models and systems is a difficult task to accomplish. This difficulty is due to several factors. The first of these is that the majority of systems that have been developed report the accuracy of their systems in terms of degrees from the desired target. However, the algorithmic details are not given, thus making it hard to simulate the system for use as a benchmark. The second, and possibly the most important factor, when performing a quantitative analysis of other systems, is the lack of a defined and established experimental methodology. This lack therefore makes it difficult to have a standard benchmarking experimental set-up, which would ensure that all models of sound-source localisation and tracking are conducted in a similar fashion, giving rise to benchmark experiments.

Experiments on this model have shown that the system is capable of azimuth estimation to an accuracy of $\pm 1.5°$ azimuth around the centre $0°$ point of the system up to an accuracy of $\pm 7.5°$ azimuth around the $-90°$ or $+90°$ ranges. In conclusion, it has been shown that robotic sound-source tracking with a mobile robot drawing inspiration from the mammalian auditory system is a viable and effective way to develop an acoustic sound-source localisation model, which performs close to human accuracy (Blauert, 1997b) and operates within acoustically cluttered environments.

10. Further work

Currently, the system presented within this paper is restricted to localising within the azimuth plane. This is due to the system only using the ITD, IPD cues and TDOA for localisation as the system lacks pinnae on the receivers of the model, i.e. the microphones. Due to this lack of pinnae, the system is unable to take advantage of other auditory cues that are used by mammals such as notch filters and shadowing (Spezio, Keller, Marrocco, & Takahashi, 2000). Such cues are useful for enabling estimation of not only the azimuth plane but also elevation, therefore allowing the localisation of a sound source within two dimensions.

Furthermore, the introduction and use of a HRTF would make greater use of the ILD or IID cues in addition to shadowing which would provide yet another localisation dimension, namely that of distance thus, ultimately allowing a sound-source's position to be

Please cite this article in press as: Murray, J. C., et al. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. Neural Networks (2009), doi:10.1016/j.neunet.2009.01.013

Table 6
Data acquired from experimentation combining azimuth estimation stage and RNN stage running on the robot.

Source angle (°)	Estimated angle	Relative angle	Attended angle	Input activation	Expected output
−90	−90	0	−87	0	0
−75	−72.6	17.4	−70	9	18
−60	−59.1	13.5	−58	7	14
−45	−46.9	12.2	−45	7	14
−30	−31.4	15.5	−30	8	16
−15	−14.4	17	−14	9	18
0	2.6	17	1	9	18
15	15.3	12.7	13	7	14
30	31.8	16.5	29	9	18
45	43.4	11.6	42	6	12
60	63.7	20.3	62	11	22
75	75.8	12.1	73	7	14
90	90	14.2	88	8	16

estimated within 3D space. However, the system presented within this paper is concerned with orientating and facing the sound source to increase the SNR of the speaker, thus highest accuracy is desired around the midline, i.e. 0°. The ILD gives its highest accuracy around the −90° and +90° positions due to maximum difference in the signal levels received at the ears, whereas the ITD and IPD cues are more accurate around the 0° position as the gradient of the angle change is at a minimum.

To achieve sound-source localisation within a 3D space other cues such as shadowing of the signals received at the ears, in addition to notch filters are used. Notch filters are the changes in the spectra of the signals arriving at the ears. These are created by the shape of the ear or pinna and vary according to the elevation of the sound source. This would give the model the ability to not only know if the source is to the left or right of the robot's current position but also to determine if the source originates from above or below the current level of the robot.

Acknowledgements

We would like to thank Jindong Liu for his comments on an earlier version of this paper and Chris Rowan for technical support of the robot equipment and experiments.

References

Arensburg, B., & Tillier, A.-M. (1991). Speech and the Neanderthals. *Endeavour*, *15*(1), 26–28.
Asfour, T., Azad, P., Vahrenkamp, N., Regenstein, K., Bierbaum, A., Welke, K., et al. (2008). Towards humanoid manipulation in human-centred environments. *Robotics and Autonomous Systems*, *56*(1), 54–65.
Blauert, J. (1997a). *Spatial hearing—The psychophysics of human sound localization*. The MIT Press.
Blauert, J. (1997b). *Spatial hearing the psychophysics of human sound localization*. p. 39 (Table 2.1).
de Boer, B. (2005). *The evolution of speech in encyclopedia of language and linguistics* (2nd ed.). Elsevier.
Böhme, H., Wilhelm, T., Key, J., Schauer, C., Schröter, C., Groß, H., et al. (2003). An approach to multi-modal human–machine interaction for intelligent service robots. *Robotics and Autonomous Systems*, *44*(1), 83–96.
Corwin, J. T. (1992). Regeneration in the auditory system. *Experimental Neurology*, *115*(1), 7–12.
Datum, M. S., Palmieri, F., & Moiseff, A. (1996). An artificial neural network for sound localisation using binaural cues. *The Journal of the Acoustical Society of America*, *100*, 372–383.
Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.
Girolami, M. (1998). A nonlinear model of the binaural cocktail party effect. *Neurocomputing*, *22*(1–3), 201–215.
Géléoc, G. S. G., & Holt, J. R. (2003). Auditory amplification: Outer hair cells *pres* the issue. *Trends in Neuroscience*, *26*(3), 115–117.
Hawkins, H. L. (1995). Models of binaural psychophysics. In *Auditory computation* (pp. 366–368). Springer.
Huang, J., Supaongprapa, T., Terakura, I., Wang, F., Ohnishi, N., & Sugie, N. (1999). A model-based sound localization system and its application to robot navigation. *Robotics and Autonomous System*, *27*(4), 199–209.
Jeffress, L. A. (1948). A place theory of sound localization. *Journal of Comparative and Physiological Psychology*, *41*, 35–39.
Joris, P. X., Smith, P. H., & Yin, T. C. T. (1998). Coincidence detection in the auditory system: 50 years after Jeffress. *Neuron*, *21*(6), 1235–1238.
Licklider, J. C. R. (1959). Three auditory theories. In E. S. Koch (Ed.), *Psychology: A study of a science* (pp. 41–144). Study 1, Vol. 1. New York: McGraw-Hill.
Lima, P., Bonarini, A., Machado, C., Marchese, F., Marques, C., Ribeiro, F., et al. (2001). Omni-directional catadioptric vision for soccer robots. *Robotics and Autonomous System*, *36*(2–3), 87–102.
Macera, J. C., Goodman, P. H., Harris, F. C. Jr., Drewes, R., & Maciokas, J. B. (2004). Remote-neocortex control of robotic search and threat identification. *Robotics and Autonomous Systems*, *46*, 97–110.
Medioni, G., François, A. R. J., Siddiqui, M., Kim, K., & Yoon, H. (2007). Robust real-time vision for a personal service robot. *Computer Vision and Image Understanding*, *108*(1–2), 196–203.
Murray, J. C., Wermter, S., & Erwin, H. R. (2006). Bioinspired auditory sound localisation for improving the signal to noise ratio of socially interactive robots. In *Proceedings of the international conference on intelligent robots and systems* (pp. 1206-1211).
Newman, R. S. (2005). The cocktail party effect in infants revisited: Listening to one's name in noise. *Developmental Psychology*, *41*(2), 352–362.
Obando, M., Liem, L., Madauss, W., Morita, M., & Robinson, B. (2004). Robotic surgery in pituitary tumors. *Operative Techniques in Otolaryngology—Head and Neck Surgery*, *15*(2), 147–149.
Press, W. H., Flannery, B. P., Teukolskv, S. A., & Vetterling, W. T. (1992). Correlation and autocorrelation using the FFT. In *Numerical recipes in C: The art of scientific computing* (2nd ed.). Cambridge University Press.
Rucci, M., Wray, J., & Edelman, G. M. (2000). Roboust localisation of auditory and visual targets in a robotic barn owl. *Robotics and Autonomous Systems*, *30*(1–2), 181–193.
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In *parallel distributed processing: Explorations in the microstructure of cognition* (pp. 318–362). Cambridge, MA: MIT Press (Chapter 8).
Severinson-Eklundh, K., Green, A., & Hüttenrauch, H. (2003). Social and collaborative aspects of interaction with a service robot. *Robotics and Autonomous Systems*, *42*(3–4), 223–234.
Smith, L. S. (2002). Using IIDs to estimate sound source direction. In B. Hallam, D. Floreano, J. Hallam, G. Hayes, J. A. Meyer (Ed.), *From animals to animals 7*. MIT Press.
Spezio, M. L., Keller, C. H., Marrocco, R. T., & Takahashi, T. T. (2000). Head-related transfer functions of the Rhesus Monkey. *Hearing Research*, *144*(1–2), 73–88.
Tamai, Y., Kagami, S., Sasaki, Y., & Mizoguchi, H. (2005). Three rings microphone array for 3D sound localization and separation for a mobile robot audition. In *IEEE IRS/RSJ international conference on intelligent robots and systems* (pp. 903–908).
Valin, J., Michaud, F., Rouat, J., & Létourneau, D. (2003). Robust sound source localization using a microphone array on a mobile robot. In *IEEE/RSJ international conference on intelligent robots and systems* (pp. 1228–1233).
Wang, Q. H., Ivanov, T., & Aarabi, P. (2003). Acoustic robot navigation using distributed microphone arrays. *Information Fusion*, *5*(2), 131–140.
Wermter, S., Weber, C., Elshaw, M., Panchev, C., Erwin, H., & Pulvermüller, F. (2004). Towards multimodal neural robot learning. *Robotics and Autonomous Systems Journal*, *47*(2–3), 171–175.

Please cite this article in press as: Murray, J. C., et al. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. Neural Networks (2009), doi:10.1016/j.neunet.2009.01.013