Bioinspired Auditory Sound Localisation for Improving the Signal to Noise Ratio of Socially Interactive Robots

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

Hybrid Intelligent Systems
University of Sunderland
Sunderland, Tyne-and-Wear, SR6 0DD
www.his.sunderland.ac.uk
{john.murray,stefan.wermter,harry.erwin}@sunderland.ac.uk

***Abstract* – In this paper we describe a bioinspired hybrid architecture for acoustic sound source localisation and tracking to increase the signal to noise ratio (SNR) between speaker and background sources for a socially interactive robot's speech recogniser system. The model presented incorporates the use of Interaural Time Difference for azimuth estimation and Recurrent Neural Networks for trajectory prediction. The results are then presented showing the difference in the SNR of a localised and non-localised speaker source, in addition to presenting the recognition rates between a localised and non-localised speaker source. From the results presented in this paper it can be seen that by orientating towards the sound source of interest the recognition rates of that source can be increased.**

***Index Terms* – Sound Source Localisation, Signal to Noise Ratio, Speech Recognition, Sociable Interactive Robots, Human Robot Interaction.**

I. Introduction

Audition is a modality that is becoming more important in the role of developing socially interactive robots [1] to operating within a natural environment. Audition is the main communication method of humans and animals; therefore it is an important factor in the development of socially interactive robots. In order to integrate robots into society to interact and communicate with humans this process of interaction should be as natural as possible.

It has been shown that humans react and ultimately find it easier to communicate with robots in a manner in which they feel comfortable [2-4] and therefore any pre-requisite knowledge required for the control and interaction of such sociable robots should be where possible kept to a minimum. Thus, robotic human interaction is becoming an important aspect of robotic research [1-2] for developing socially interactive robots capable of communicating with humans.

The central auditory system (CAS) of the mammalian is extremely adept at using acoustics for communication purposes in both speech processing and auditory scene analysis (ASA). The CAS is also capable of localising sound sources within an acoustically cluttered environment enabling the mammalian to infer the relative position of any sources present. Due to the abilities and accuracy that has been demonstrated by that of the mammalian's CAS with respect to acoustics [5] the model presented within this paper draws its inspiration from the binaural cues known to be used by the CAS in addition to mechanisms believed to exist within biology to aid in the improvement of signal to noise rations for speech recognition.

II. Cocktail Party Effect

A. Listening to one speaker

One of the problems faced by both socially interactive robots and humans involves being able to intelligibly understand or interpret a source of interest within an acoustically cluttered environment. This is a problem that has been faced by most people at various points in their life, either at a crowded party, or in the middle of a busy shopping centre. This phenomenon is known as the *'Cocktail Party Effect'* as first noted by E.C. Cherry in 1953 [6]. The processing of the *'Cocktail Party Effect'* does not happen solely within the Auditory Cortex (AC) but begins at the pinnae of the ears. This phenomenon is of major interest to neuroscientists and roboticists [7-8].

B. Signal to Noise Ratio

One way in which this paper addresses the *'Cocktail Party Effect'* is by making changes to the signal to noise ratio (SNR) with respect to both a speaker and background source. The model presented in this paper demonstrates how a novel architecture for sound source localisation and predictive tracking can be used to increase the SNR in order to improve the accuracy of a speech recognition system by changing the position of the robot with respect to the desired sound source. The SNR described in this paper refers to the power ***P*** (or energy function) ratio of a meaningful signal and background noise (or unwanted / irrelevant noise) as shown in (1).

$$SNR = \frac{P_{SIGNAL}}{P_{NOISE}} \tag{1}$$

The SNR is given in units of decibel (dB), is usually expressed in terms of a logarithmic scale, and is 20 times the logarithmic of the power (or energy function) ratio (2).

$$SNR(dB) = 20\log_{10}\left(\frac{P_{SIGNAL}}{P_{NOISE}}\right) \tag{2}$$

As can be seen from (2) the greater the power difference between the two signals then the larger the ratio and thus more energy is therefore being received from the source of interest with respect to the background source(s).

In order to determine the SNR from (2) the power or energy needs to be calculated for the signals P_{SIGNAL} and P_{NOISE}. Equation 3 shows the equation used for representing the amount of energy contained within a signal. The value returned from this energy function does not have any units as the values used to represent the signals are normalised by the systems DSP to ±1 and are thus dependent on the gain of the system.

$$\varepsilon = \sum_{i=1}^{n} [y_i]^2 \tag{3}$$

y_i represents the normalised amplitude value of the signal at sample *i*, *n* is the number of samples within the recording and ε represents the total energy summed over all samples.

III. Localisation

Sound source localisation plays an important part in audition with respect to signal levels of sources. In order to determine the location of an acoustic object within the environment and to either approach or track this object, the AC needs to have some method of determining the angle of incidence of that source i.e. the position along the horizontal plane.

The model presented in this paper shows a novel architecture for azimuth estimation and tracking for a mobile robot. This model is capable not only of estimating the angle of incidence of the source but also *maintaining* an acoustic track as the source traverses through the environment. Sections A and B describe the elements of the model that are responsible for the localisation and tracking functionality of the system.

The full model presented in this paper and in [11] enables a robotic system to estimate the angle and track a sound source with respect to background noise thus increasing the SNR levels in order achieve higher speech recognition accuracy. Fig. 1 shows a block diagram of the model presented.

Stage 1 contains the azimuth estimation algorithms required to determine the angle of incidence of the signals received by the microphones. Stage 2 uses the information provided by stage 1, i.e. azimuth angle, as input to the RNN for estimating the next position of the source within the environment. Stage 3 is the speech recognition system that interprets the information contained within the signals.

Several factors contribute towards the performance of such a system; these include the reduction in distance from the receivers to the source, therefore increasing the energy received and secondly, due to the pinnae of the mammalian ear, reflection of the background noise occurs as the head turns away from the background source and turns to face the speaker.

A. Azimuth Estimation

The azimuth estimation stage of the model is used to determine the current angle of incidence of the source with respect to the robots current frame of reference. As the sound signal propagates through the air it is detected by the microphones connected to the robot, digitally sampled and presented as two signal vectors *g(t)* and *h(t)*. These vectors contain the signals that are recorded from the environment; therefore the angle of incidence of the various sources with respect to the robot's position will alter the interaural phase difference (IPD) of the signal as it is received at either microphone, thus creating a lag or delay between the two signal vectors. It is this offset or *lag* that is used to calculate the angle of the source.

It has been shown that within the mammalian CAS the phase difference or *time delay of arrival* of the signals is used for angle estimation. This auditory cue is referred to as the Interaural Time Difference (ITD) [9]. The most notable mechanism used to explain the functionality of ITD is Jeffress 'coincidence detectors' which are described as the likely neuronal architecture within the CAS for coding the ITD cue [10].

The phase difference of the signals *g(t)* and *h(t)* is calculated using a method of cross-correlation, see (4). Cross-Correlation is used here to provide a measure of maximum similarity between the two vectors representing the signals detected by the two microphones. This similarity offset (or phase delay) is used to determine the ITD of the two signals and ultimately the angle of incidence of the sound source.

$$Corr(g,h)_j \equiv \sum_{k=0}^{N-1} g_{j+k} k_k \tag{4}$$

Fig. 2 shows the two signal vectors *g(t)* and *h(t)* being compared for maximum similarity by cross-correlation. The two signals are effectively *slid* across each other in the time domain creating a resulting product vector ***C*** containing the sum of the products of the values currently aligned at each time step, with a time step being equal to the time delay between sampling, i.e. $\Delta t = 1/f$. Fig. 3 shows the result of the *sliding product* with the highest peak value representing the point at which maximum similarity occurs**.**

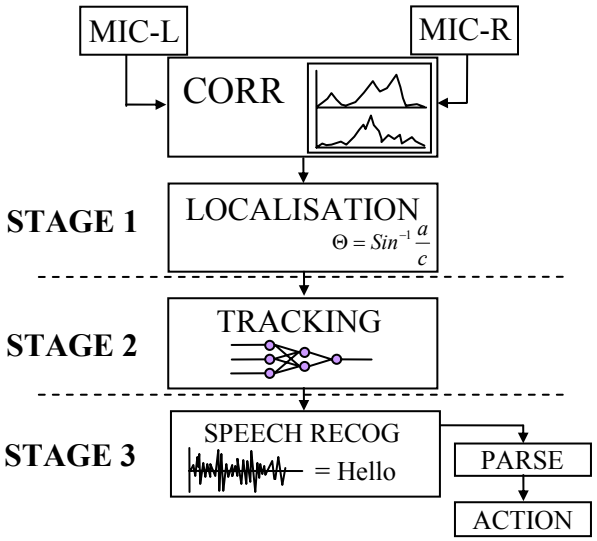

Fig. 1 – Block diagram of the sound source localisation and tracking system showing the two main stages coupled to the speech recognition system.

The position of the maximum value within *C* represents the point of maximum similarity and hence the *lag* between the two signals. The shaded area within Fig. 2 highlights a reference point showing a section of the signals in which similarity resides. In Fig. 3 it can be seen that maximum similarity between *g(t)* and *h(t)* occurs at a lag of -17 samples.

As previously discussed, within the vector ***C*** each lag or *step* has a time increment Δt of 22.67µs due to a sample rate of 44.1 kHz. Thus, depending upon the value of the lag within the correlation vector (-17 in Fig. 3) the ITD between the arrival of the signal at the two microphones can be determined.

Equations 5 through 10 show how the angle of incidence for the source can be determined from the cross-correlation of the two vector recordings *g(t)* and *h(t)* and ultimately the information from the resulting correlation vector *C*.

$$length(C) = (length(g) + length(h)) - 1 \qquad (5)$$

length(g) and *length(h)* gives the number of samples in the vectors *g(t)* and *h(t)* respectively.

$$\sigma = (length(g) - 1) - C_{MAX} \qquad (6)$$

C_{MAX} represents the position in ***C*** of the maximum value (i.e. maximum similarity) and σ is the delay or lag.

$$Sin\,\vartheta = \frac{a}{c} \qquad (7)$$

The function used for angle calculation; *a* = the length required, *c* = distance between the microphones = 0.30 meters.

$$ITD = \Delta \times \sigma \qquad (8)$$

ITD is the time delay of arrival of the signal at the two microphones, Δ is the time delay between samples and σ is the lag in phase between *g(t)* and *h(t)* as calculated from (5).

$$length(a) = ITD \times c_{air} = (\Delta \times \sigma) \times c_{air} \qquad (9)$$

length(a) is the extra distance the signal is required to travel to the contralateral microphone once detected by the ipsilateral microphone, c_{air} = speed of sound in air at 24°C = 345 m/s.

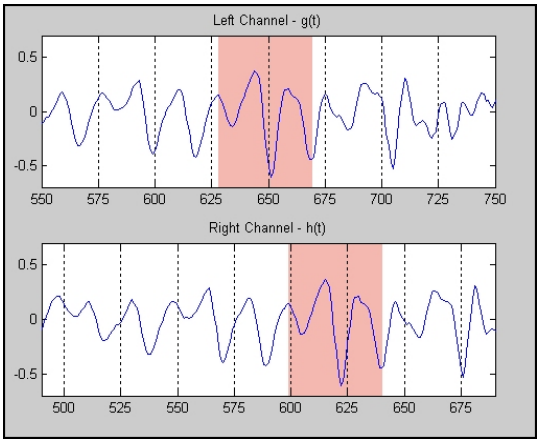

Fig. 2 – The sliding window effect of cross-correlation with the two shaded areas representing the correlating signals.

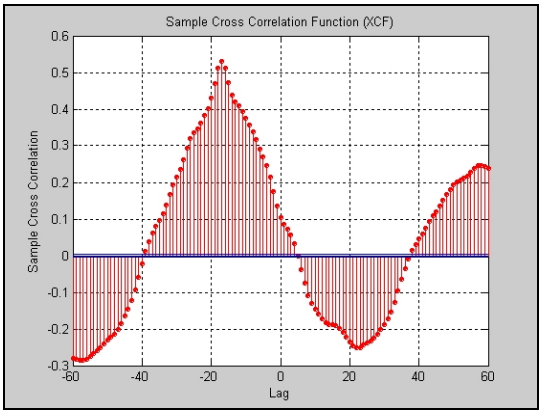

Fig. 3 – A plot of the resultant correlation vector ***C*** produced by the cross-correlation of the signal vectors *g(t)* and *h(t)*.

$$\vartheta = Sin^{-1}\frac{a}{c} = Sin^{-1}\frac{(\Delta \times \sigma) \times c_{air}}{c} \qquad (10)$$

ϑ represents the angle of incidence of the sound source on the horizontal plane.

B. Tracking and Prediction

Tracking and prediction is the second stage of the model presented in this paper and deals with maintaining a track on the sound source as it moves within the environment. This stage of the model incorporates the use of a recurrent neural network (RNN) for trajectory estimation prediction. The use of a RNN is required due to the temporal aspect of predicting a trajectory. In order to estimate the next position of the source it is necessary to know the previous positions that the source has taken in its movement.

Due to the restrictions of standard feedforward networks i.e. having no temporal aspect and therefore only being able to classify the pattern currently being presented, it is necessary to introduce recurrent connections into the network. The activations of the hidden layer units at time t_{n+0} from the RNN are copied via 1:1 projections to a context layer and are therefore available to the network at time t_{n+1} when the next pattern is presented; Fig. 4 shows the network architecture including number of units per layer used in the model.

Due to the networks recurrent nature, the system is able to retain a kind of *short-term memory* that can be used to provide additional 'previous' temporal information during pattern recognition. Equation 11 shows the formula for copying the activations from the hidden layer units to the context units.

The RNN adopts the standard backpropagation algorithm for adapting the weights in order to facilitate training; this is shown in (12). However, due to the temporal nature of a RNN as compared to that of standard feedforward networks, the equation in (12) is rewritten to include the temporal aspect of the network as provided via the context layer; this is referred to as backpropagation through time [12].

To ensure the network correctly identifies the temporal order, each pattern is stored in a separate sub-group. Fig. 5 shows an example of a collection of patterns for a sub-group.

The patterns within these groups are presented to the network in their sequential order whilst each sub-group is presented randomly. This ensures the temporal ordering is maintained for recognition purposes.

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

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

$$\Delta w_{ji} = \eta \sum_t \delta_i(t) a_j(t) \tag{13}$$

where, Δw_{ij} is the weight change between units i and j, n is the current pattern event, η is the learning rate = 0.25, δ_j is the error of unit j, o_i is the output of unit i, α is the momentum term to prevent weight oscillation by adding a small amount of the previous weight change to the current weight change.

V. Experimentation

Three separate environmental configurations were used to test the robustness and recognition rates of the system. The first experiment conducted involved creating two static sound sources i.e. a speaker and background source. In this experiment the sources were positioned at 0° azimuth with respect to the robot as shown in Fig. 6a.

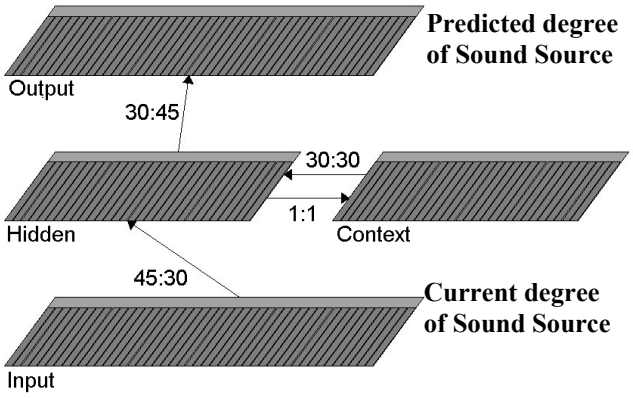

Fig. 4 – The architecture of the recurrent neural network.

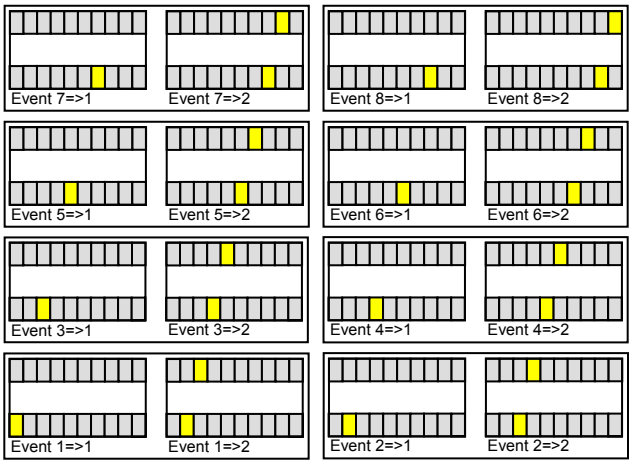

Fig. 5 –A training sub group for one of the possible source speeds.

The second configuration within this experimental test also involved the use of static sources. However, the configuration saw the speaker source remain fixed at 0° with respect to the robot whilst the background 'clutter' source was positioned at an angle of 45° to the left of the robot as shown in Fig. 6b.

With the second experimental test the background source was fixed at a position of 0° whilst the speaker source is positioned 45° left of the robot. The first configuration within the second experimental test represented that of Fig. 7a. However, this experiment required three separate steps as opposed to two for that of experiment one. In the second and third configurations of this test the robot attended to the 45° angle of the speaker source as opposed to remaining stationary.

In order for the robot to attend to the speaker position without *missing* any important information, the system is provided with what we deem an *attention* word. Attention words can also be seen in communication between humans, e.g. with the use of "Excuse me", "garçon" or "waiter!" which are used to simply gain the attention of someone you wish to speak to before actually providing them with the important information you wish them to interpret. Once the system has orientated then the flow of information may continue.

Due to the system's ability to track the speaker as it moves through the environment this attention word is therefore only required once per tracking of a speaker source. Fig. 7 shows the sound source and microphone positions for the orientating stage of the second experimental testing phase.

Finally, the third experimental setup involves the use of a static background source positioned at 0° with a dynamically moving speaker source. Here the system is initialised to the configuration shown in Fig. 6a and the speaker source is then positioned at increments of 15° along a circle of radius 1.5m around the robot. Fig. 8 shows the incremental positions for the dynamic source.

VI. Results

The results presented in this paper are only concerned with the changes in SNR in addition to the increase in accuracy and repeatability this brings to the speech recogniser. The results for azimuth estimation and sound source tracking are not presented here, for more information on these results please refer to Murray *et al*. [11].

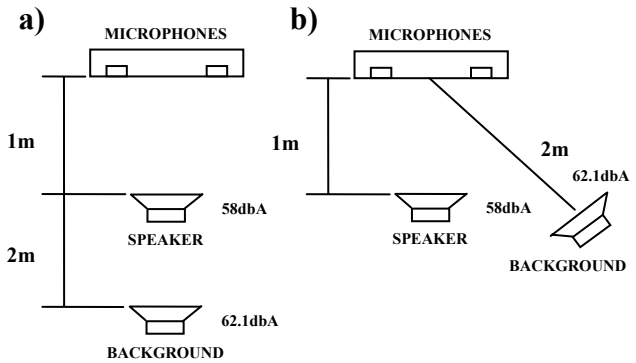

Fig. 6 – a) Shows the first configuration of the first experimental setup with both sources at 0° azimuth, b) Shows the second configuration with the speaker source at 0° and background source at 45°L of the robot.

Figure 9 shows the individual background and speaker signals to allow a comparison between these and the combined signals at the various azimuth positions. Fig. 10 shows the results of the static test experiment displaying how changes in azimuth can affect the SNR of the signals. Within Figs. 9 and 10 the y-axis represents the amplitude of the signals whilst the x-axis shows the number of samples Δ. Each sample is taken at a time interval of 22.67μs. Therefore, the duration of the background source is $22.67\mu s * 4x10^5 \approx 9$ seconds and the speaker duration approximately $22.67\mu s * 1x10^5 \approx 2.3$ seconds.

The first column in Fig. 10 shows both the background and speaker sources positioned at 0°, thus it can be seen that both the left and right channels have similar SNR levels. However, if the background source is at an angle of 0° with the speaker moved to 45° right of the microphones then it can be seen that the SNR between speaker and background increases for one of the channels (right channel) and the signal level rises higher above the background level. The right microphone still detects the same levels of background signal (due to the speakers and background source remaining fixed).

The difference in the SNR of the signals shown in Fig. 10 can be determined if we calculate the energy in the signals between the overlapping sample positions – samples 200000 to 300000. These samples are the points at which the speaker signal is generated during background generation and are compared against the same points from the background source. Table 1 shows the SNR differences from experiment one. For this experiment ten trials where conducted using two different background signals and two speaker signals.

Table 1 gives the energy (based on the energy function shown in (3)) of the signals shown in Figs. 9 and 10 in addition to the SNR differences. As the results in table 2 show, the positions of the two sources with relation to the robots 0° reference point changes the SNR between them and therefore changes the relative amount of energy received from each source. Fig. 11 shows these results plotted with angle change vs. SNR difference. The signals detected by the robot are stored to file and presented to the speech recogniser to determine if the signal can be interpreted. Table 3 shows the results of the experiments for all the speaker words of varying duration and amplitude.

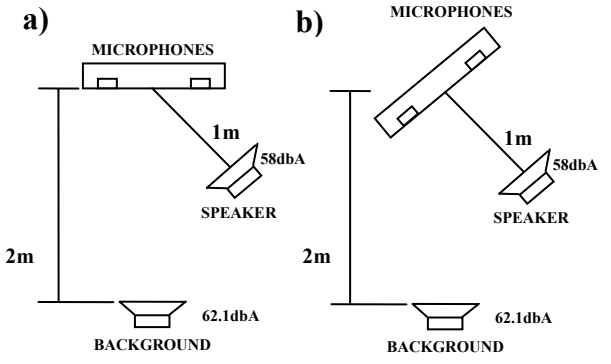

Fig. 7 – a) The second configuration of the experiment with the speaker source placed at 45°L, b) shows the third configuration once the 'orientation' has taken place and the robot is facing the speaker source.

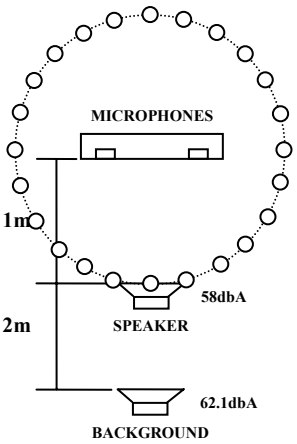

Fig. 8 – The dynamic configuration and increments for the speaker source as it moves in increments of 15° around the robot with a radius of 0.8m.

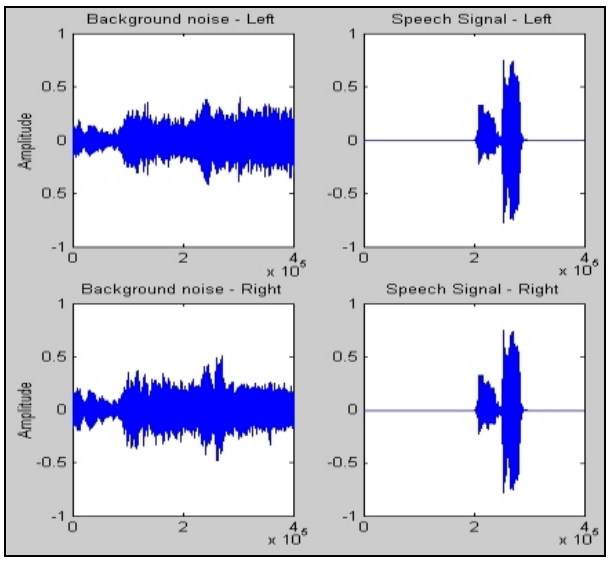

Fig. 9 – Shows the separate signals for the background and speaker source for both the left and right channels as recorded by the robots microphones.

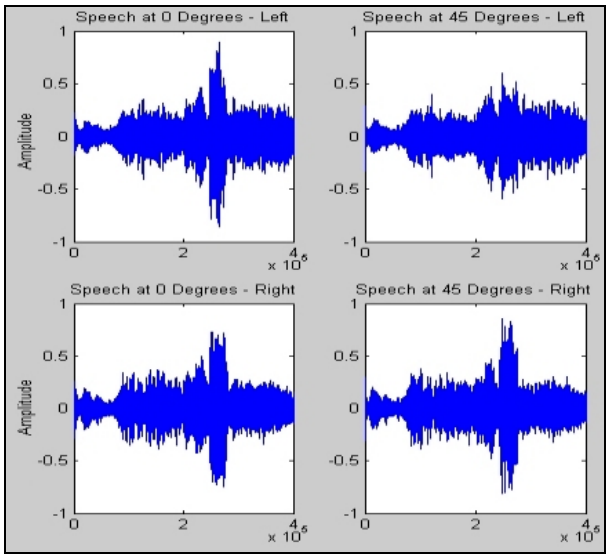

Fig. 10 – Shows the combined background and speaker signals for both the left and right channels as recorded by the robot at 0°, 0° and 0°, 45° respectively.

TABLE 1 – THE CHANGE IN SNR FOR THE SIGNALS IN FIG. 10 - LEFT CHANNEL.

Source	Energy	SNR	SNR(dB)
Background	31.618	N/A	N/A
Speaker at 0°	38.327	1.2122	1.6714
Speaker at 45°	33.523	1.0603	0.5082

TABLE 2 – THE SNR CHANGES FROM VARYING BACKGROUND AND SPEAKER SOURCE POSITIONS WITH RESPECT TO THE ROBOT.

Source Angle		Energy		Gains	
Speak	Backgnd	Speaker	Backgnd	SNR	SNR(dB)
0	0	39.795	30.963	1.2853	1.09
0	15	39.795	29.764	1.3370	1.261
0	30	39.795	27.453	1.4495	1.6123
0	45	39.795	26.601	1.496	1.7493
0	60	39.795	26.112	1.524	1.8298
0	75	39.795	25.239	1.5767	1.9776
0	90	39.795	24.361	1.6336	2.1313
15	0	39.028	30.963	1.2604	1.0053
30	0	38.523	30.963	1.2442	0.9488
45	0	37.312	30.963	1.2051	0.8100
60	0	36.597	30.963	1.182	0.7260
75	0	35.201	30.963	1.1369	0.5571
90	0	34.362	30.963	1.1098	0.4523

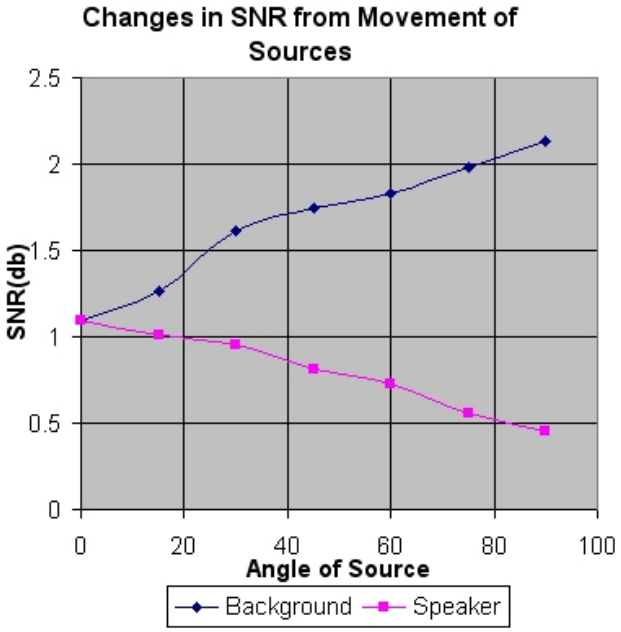

Fig. 11 – The effect of angle position on the SNR of two sources

TABLE 3 – RECOGNITION RATES TAKING INTO ACCOUNT SNR DIFFERENCES.

Speaker Angle		Recognition % Trials	
Speaker	Background	Left	Right
0°	0°	17	20
0°	45° Left	10	24
0°	90° Left	12	45
45° Left	0°	21	17
90° Left	0°	51	14

VII. CONCLUSIONS

This paper highlights a need for robotic sound source localisation when dealing with sociable robotic systems that are to communicate with humans in a natural environment. It has been shown in this paper that when a robot orients itself towards a speaker, recognition rates improve. Table 3 shows that with both sources positioned at 0° recognition rates are at approximately 20%. However, when the system orients itself towards the speaker a higher recognition accuracy of 51% is achieved. It must be noted that these results are not intended to be a *best* measure from a speech systems recognition point of view, but are designed to show an orientation performance increase. Therefore, the source levels and distances are chosen to provide reduced recognition rates from the system. Thus with a reduced recognition accuracy it could be demonstrated how orientating towards the source can increase the performance of the system. It can be noted that by simply orientating towards the source a gain of more than 2.5-fold can be achieved in the recognition rate and thus improve the robustness and reliability of such a system.

VIII. FURTHER WORK

Further work can be performed to increase the recognition rates of the system. This would include the use of pinnae that can help to further filter out and reflect unwanted signals giving higher recognition rates. Incorporating a blind source separation (BSS) algorithm would also be of benefit to such a system as this would allow the speaker signal to have increased SNR whilst using BSS to separate the unwanted signals from the desired signal allowing more of the signal of interest to be acquired from the convolved sources.

REFERENCES

[1] Breazeal, C. Towards sociable robots. Robotics and Autonomous Systems, Volume 42, Issues 3-4, 31, Pages 167-175, March 2003.

[2] Breazeal, C. Emotion and sociable humanoid robots. International Journal of Human-Computer Studies, Volume 59, Issues 1-2, Pages 119-155, July 2003.

[3] Fong, T., Nourbakhsh, I. and Dautenhahn, K. A survey of socially interactive robots. Robotics and Autonomous Systems, Volume 42, Issues 3-4, Pages 143-166, 31 March 2003.

[4] Severinson-Eklundh, K., Green, A. and Hüttenrauch, H. Social and collaborative aspects of interaction with a service robot. Robotics and Autonomous Systems, Volume 42, Issues 3-4, Pages 223-234, 31 March 2003.

[5] Blauert, J. Spatial Hearing – The Psychophysics of Human Sound Localization. Page 39, Table 2.1, 1997

[6] Cherry, E.C. Some experiments on the recognition of speech, with one and two ears. Journal of the Acoustic Society of America, Volume 25, Pages 975-979, 1953.

[7] Newman, R.S. The Cocktail Party Effect in Infants Revisited: Listening to One's Name in Noise. Developmental Psychology, Volume 41, Issue 2, Pages 352-362, March 2005.

[8] Zhou, Y and Xu, B. Blind source separation in frequency domain. Signal Processing, Volume 83, Issue 9, Pages 2037-2046, September 2003.

[9] Yin, T.C.T. Neural Mechanisms of Encoding Binaural Localisation Cues in the Auditory Brainstem. In: Integrative Functions in the Mammalian Auditory Pathway, Donata Oertel, Richard R. Fay and Arthur N. Popper (Eds.). Page 99, ISBN: 0-387-98903-X, 2001.

[10] Jeffress, L.A. A place theory of sound localization. J. of Comp. Physiol. Psychol. Volume 41, Pages 35-39, 1948.

[11] Murray J., Erwin H., Wermter S. A Hybrid Architecture using Cross-Correlation and Recurrent Neural Networks for Acoustic Tracking in Robots. Biomimetic Neural Learning for Intelligent Robots, *Springer-Verlag*, 2005.

[12] Elman, J. L., Finding Structure in Time. Cognitive Science, Volume 14, Number 2, Pages 179-211, 1990.