TR - H - 128

Sound Localization in the Horizontal Plane: A Binaural Approach

Jérôme Amouyal *David Rainton*

1995. 2. 13

ATR人間情報通信研究所

〒619-02　京都府相楽郡精華町光台2-2　☎ 0774-95-1011

ATR Human Information Processing Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
Telephone: +81-774-95-1011
Facsimile: +81-774-95-1008

Sound Localization in the Horizontal Plane: A Binaural Approach

ATR Human Information Processing Research Laboratories

November 1994

Abstract

It is generally believed that humans localize sound in the horizontal plane using the Interaural Time Difference(ITD) and Interaural Amplitude Difference(IAD) between the signals received at the two ears. In this research the performance of an acoustic localization device composed of a front end binaural simulator and back end neural network classifier is evaluated.

First a small binaural database of speech, clapping, and music sounds was recorded using the ATR Head and Torso Simulator. Using the data, three different neural networks were trained to localize sounds using ITD, IAD and both ITD and IAD cues. Results presented indicate that the optimal window lengths for both ITD and IAD cue computation are of the order of 100ms. The network using both ITD and IAD cues outperformed those using just ITD or IAD information alone. Finally the construction of a parallel real time acoustic localization device is described.

Contents

List of Figures

Introduction

Most animals need to be able to localize the source of a sound. For example, rapid accurate sound localization can aid survival in a hostile environment or enable prey to be caught in the dark. An important question then is how does an animal determine the location of a sound source? Different animals solve this problem in different ways (for example cats have movable pinna, owls have asymmetric pinna, etc). In this report the sound source localization problem is examined through the construction of a human-like binaural localization device.

An obvious aid to sound source localization is the fact that humans have two directionally sensitive ears whose directions of maximum sensitivity are different. Thus sound from the right will sound louder in the right ear than the left. This interaural amplitude difference (IAD) is one potential sound source localization cue. Indeed, researchers have discovered cells in some animal brains that are sensitive to such level differences. Figure 1 shows the IAD for the signals received inside the left and right pinna of an artifical head simulator [11] for a sound source located on the right.

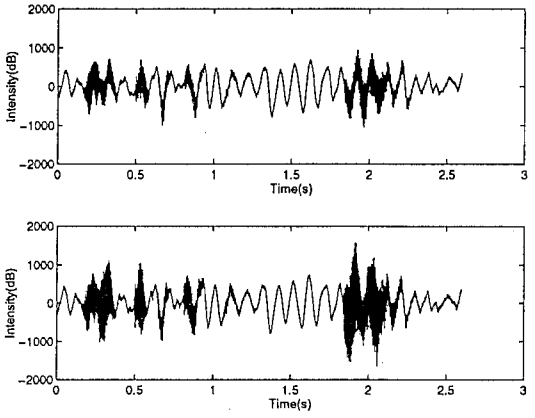


Figure 1: Left(top) and right(bottom) ear waveforms recorded inside an artificial head for a sound source located on the right.

Another important cue for sound source localization arises because of the spatial separation of the ears. This spatial separation results in a location

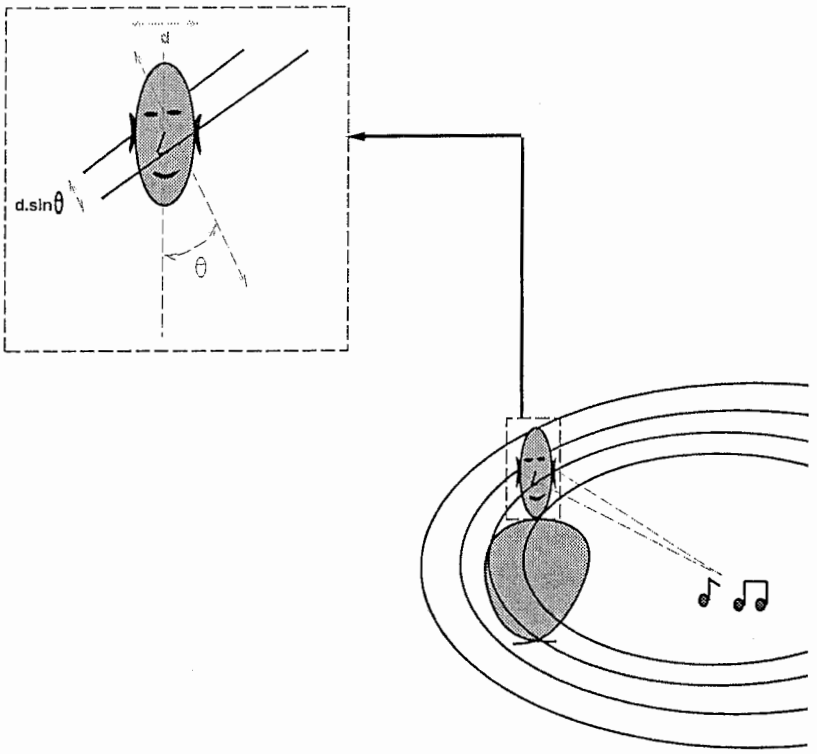


Figure 2: Interaural Time Difference.

dependent interaural time delay(ITD) between the signals received at the two ears. Thus a signal coming from the right will reach the right ear just before the left and visa versa (Figure 2). Again, as evidence for the use of ITD, biological studies have shown the existence of ITD sensitive coincidence detectors in the brains of some animals.

It should also be noted that humans are able to localize sounds from just the signal received at a single ear [5]. It is believed that this monaural localization capability arises from the shape of the pinna which produces spatially dependent spectral notches in the resulting auditory signal. Humans are presumably capable of separating such pinna induced spectral notches

from those arising naturally due to the spectral characteristics of the source. However, a study of monaural localization was outside the scope of this research, which was concerned only with binaural localization.

Using an artificial human binaural simulator, the aim of this work was to construct and evaluate an acoustic localization device using both ITD and IAD information. Envisaged applications for such work include the construction of cameras capable of automatically focussing on selected sound sources, alarm systems capable of detecting the location of noisy intruders and directionally selective microphones.

1 ITD and IAD - Further Considerations

1.1 Interaural Time Delay(ITD)

Assuming far field (i.e. that the acoustic wavefront is planar at the ears) then the relationship between the ITD(τ) and the wavefront angle of incidence(θ) can be approximated by the formula,

$$\tau = \frac{d}{c} \sin\theta \tag{1}$$

where d is the head diameter and c the sound velocity of sound in air (see Figure 2). If the distance of the sound source from the head is greater than 1m, then any error introduced into τ by the far field assumption will be less than 1%. It should be obvious that the ITD is theoretically independent of both the sound spectral characteristics and the distance of the sound from the head (assuming far field). However it should also be noted that any estimate of ITD may well depend on both.

Although ITD is clearly an important cue for sound source localization it does have several major practical limitations. First, it is unable to distinguish between sounds arising from in front of and behind the head. As show in Figure 3 it is not possible distinguish between two sounds arising at angles of incidence θ and $\pi - \theta$. Without apriori knowledge of sound source location, or additional cues, it is not possible to resolve this front back ambiguity using ITD alone.

Secondly, the value of ITD is very small(the maximum delay depends on the size of the head, but is typically around 670μs). Thus high sampling

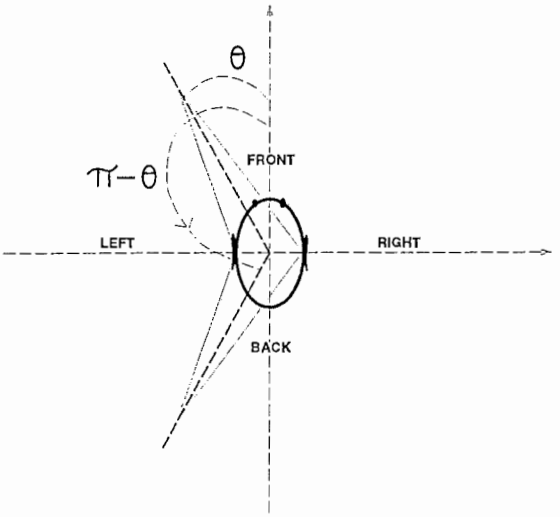


Figure 3: The front back ITD ambiguity

frequencies are required to obtain reasonable angular resolution. Also, ITD does not vary linearly as a function of angle (see Figure 4). Figure 5 shows the angular error, corresponding to a time delay estimation error of $\pm 10\mu$s. If the time delay is known modulo $\pm 10\mu$s, then θ will be between θ_{min} and θ_{max}. The value of θ_{max} - θ_{min} for different values of θ is shown in Figure 5. The error decreases between $\pm(80-90)$ degrees because of the a priori knowledge of the maximum possible delay; if the maximal time delay is 670 μs, and the time delay estimate is 665 μs, then the actual time delay must lie between 655 and 670 μs, not 655 and 675 μs.

Clearly angular resolution is a function of the angle of incidence, being greatest directly in front of the head (0 degrees) and progressively decreasing towards the side of the head (80 degrees). For the construction of a fixed acoustic localization device uniform resolution, independent of sound source location, is preferable(although obviously if the device is rotatable then this is not a problem).

A third problem with ITD lies with its estimation. Typically ITD is extracted from the cross-correlation function of the signals received at the

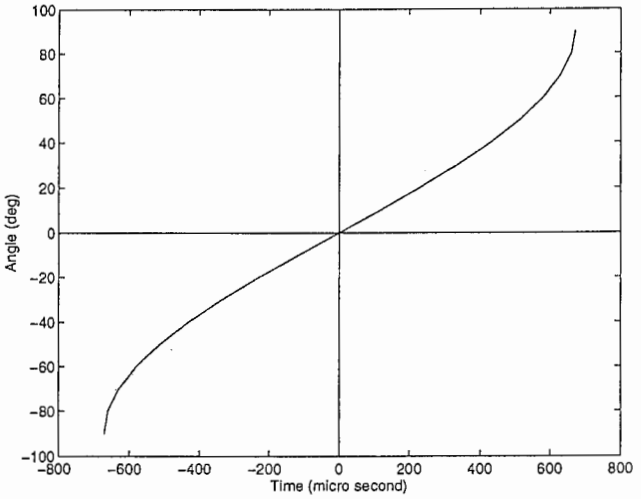


Figure 4: The waveform angle of incidence θ as a function of the ITD

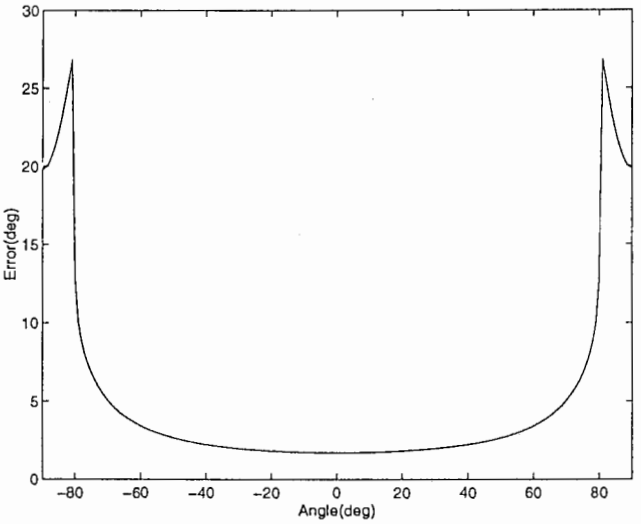


Figure 5: Error in the waveform angle estimation as a function of the angle of incidence θ

two ears, i.e. if x_r and x_l are the received signals then the resulting cross-correlation function R(τ), computed over some finite interval T, is given by the equation

$$R(\tau) = \sum_{t=1}^{T} x_r(t)x_l(t+\tau) \tag{2}$$

In noise free, anechoic conditions the offset of the largest peak in the cross-correlation function corresponds to the ITD. However, depending on the spectral characteristics of the signal, multiple ambiguous peaks can arise at multiples of the dominant spectral component periodicities (a phenomenen known as phase ambiguity [3]). As the frequency increases so does the number of ambiguous peaks. Figure 6 shows this phase ambiguity effect for several different sinusoidal inputs ranging from 500 to 2000Hz. As a consequence it can make sense to ignore high frequencies by low pass filtering the input signals prior to cross-correlation. Another way of reducing the effect of such ambiguities is to replace the input signals with their envelopes. In humans, where the received signals are split into frequency bands prior to processing, envelope extraction is believed to occur only for those bands greater than about 1.4kHz, where phase ambiguity effects are greatest. On the other hand, reducing the signal bandwidth can actually enhance ambiguities. For example, in figure 7, which shows the cross-correlation of the sum of the sinusoids in figure 6, there are no ambiguous peaks. The optimal choice of filter thus depends on both the signal and noise spectral characteristics.

1.2 Interaural Amplitude Difference(IAD)

The IAD is a complex non-linear function of both the location and spectral content of the sound source. The IAD arises partially from that fact that the ears are separated by the head, resulting in a location and frequency dependent head related transfer function(HRTF) between the two auditory signals. Essentially the head produces a shadow, resulting in a relative signal attenuation at the ear farthest from the source. Due to diffraction effects the extent of this attenduation increases with frequency, producing to a first approximation a low pass transfer function between the near and far ear auditory signals. An example of a HRTF [4] is shown in Figure 8. Another factor in the HRTF is the shape of the ear pinna, which also give rise to a frequency and spatially dependent directional gain.

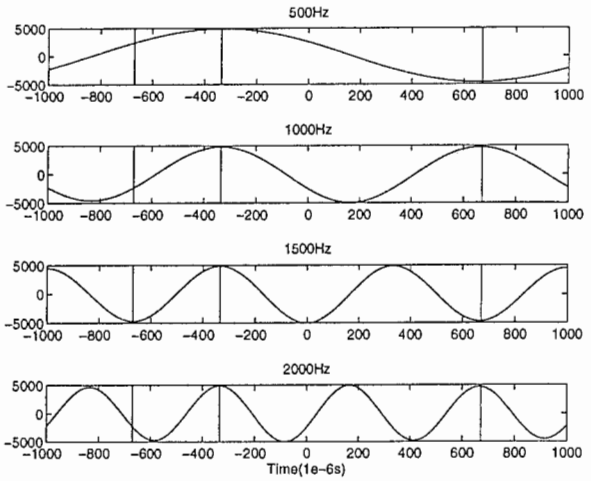


Figure 6: Phase ambiguity in the Cross-correlation function of several different sinusoids.

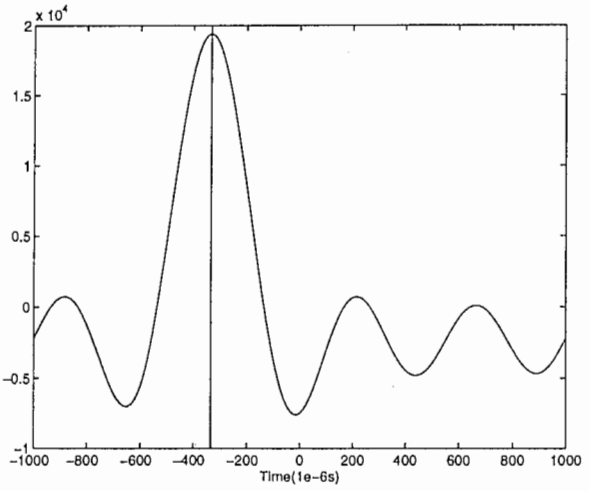


Figure 7: Cross-correlation function of the sum of the sinusoids in Figure 6.

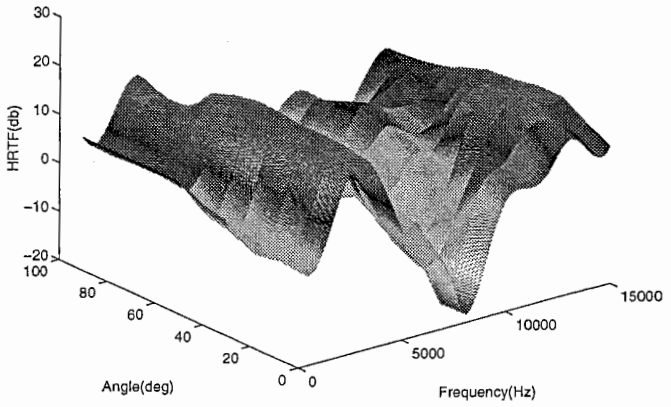
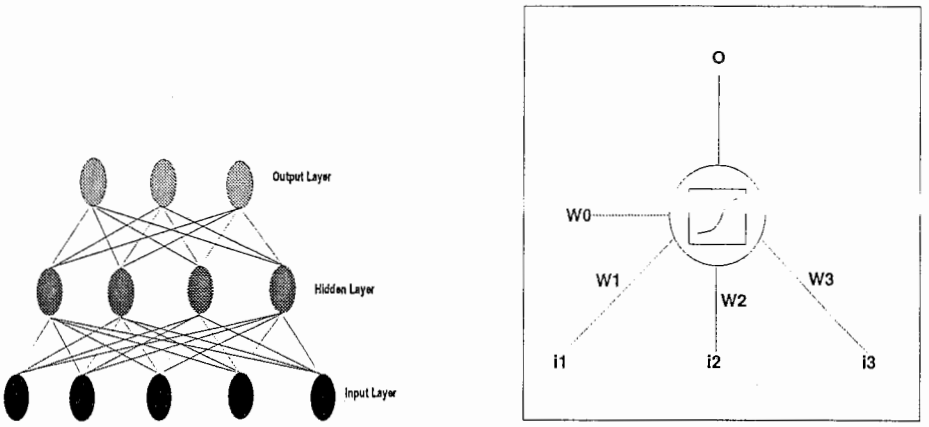


Figure 8: The measured HRTF of a human subject plotted as a function of angle and frequency.

1.3 Learning the mapping between ITD, IAD and angle of incidence θ

The aim of this work was to study the use of ITD and IAD for angle estimation, both separately and together. In particular the question as to how to combine these two cues is an important subject of current research. It is known that humans solve this problem [6], but it is not yet known how.

Neural networks were chosen to learn the mapping between both ITD, IAD and angle of incidence. Given the highly nonlinear nature of the mapping from IAD to angle, the use of a neural network seemed a logical choice. For ITD on the other hand it could be argued that a simple peak picking algorithm would suffice. In practice however it was found that noise and reverberation effects led to significant errors with such a simple peak picking approach. Hence it was decided to employ a neural network for both mappings. An advantage of this is that ITD and IAD information can be combined by simply inputing the joint feature vector to a single network.



a)The Neural Network Architecture. b)The Description of a Single Unit.

Figure 9: The Neural Network Description.

1.4 The Neural Network

The neural network chosen was a simple backpropagation network of the form shown in Figure 9a). The network had 3 layers, an input layer, a hidden layer and an output layer. Each unit of the hidden layer was fully connected to all the inputs and outputs units. Figure 9b) shows a unit which takes the weighted sum of its inputs (i1, i2, i3 are the inputs, w1, w2 and w3 are the unit weights) plus a bias weight(w0) and passes it to a sigmoid transfer function to calculate the output O:

$$F(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

Gradient descent was used for network learning (for more details about the algorithm and the way the network was built, see Appendix A and [12]).

1.5 Extracting ITD and IAD cues from the acoustic waveform

The first problem in extracting ITD and IAD information from the waveform is that of signal detection; i.e. deciding when the signal is present. The

approach taken here was to use a simple threshold on the resulting cross-correlation peak height.

The second problem is that of deciding which parts of the acoustic waveform to use for localization purposes. There is evidence from the studies of both humans[REF] and animals[REF] that the initial transient portions of the waveform are the most important for localization purposes. This is known as the Franssen effect [9]and was demonstrated most clearly by an experiment performed at AT&T Laboratories, Murray Hill. In that experiment a tone signal was fed into two loudspeakers, the first speaker radiating only a short transient and then falling silent. The second speaker radiated a slightly delayed and softly turned-on steady tone. In a reverberant environment, listeners invariably perceived the tone as coming from the silent speaker and were amazed when the demonstrator pulled the plug on this loudspeaker. The idea that humans can use just the initial transient portion of a sound to evaluate the incident direction make sense, since only this portion of the sound is guaranteed to be free of reflected energy.

The above implies that ideally the windows for computing the ITD and IAD cues should be centered upon sound initial transients occuring in the recorded binaural waveforms. However, in practice accurate location of such transients can be problematic. Mistakes by the window positioning algorithm can have adverse effects on localization performance. In order to make our system as robust as possible a very simple window positioning algorithm was employed. First, each waveform file was split into 10 frames. For the clapping files each frame was positioned to contain a single clapping, while the speech and music files were simply split into 10 equal sized frames. For each corresponding binaural frame pair the position of the left and right ear waveform maxima were found. The window for computing the IAD and ITD cues was then centered at the time corresponding to the biggest of the left and right maxima. Thus 10 ITD and IAD cues were obtained from each binaural waveform pair.

2 Experiments

Recording a binaural sound database

A head and torso simulator (Type 4128 from Bruel & Kjaer) was used to simulate human binaural hearing(for more details about the simulator, see [11]). Sound recorded by microphones in the simulator, or manikin, were pre-amplified and then digitized using a 48 kHz, 16 bit A/D converter. The resulting digitized signals were then fed via a DAT-link to a workstation filesystem.

Binaural waveforms were recorded in the variable reverberation chamber at ATR. The recorded sounds comprised of human speech, spoken by the author, human clapping and music played from a cassette player. Sounds were generated from 10 different positions at approximatley equally spaced angles around the front of the manikin and at approximately equal distance from the manakin. For each of the 10 positions, 5 sentences, 3 different pieces of 30s music and 10 clapping were recorded. The reason for recording such a wide range of different sounds was to attempt to prevent the neural networks from learning the sound spectral characteristics themselves, as opposed to the ITD or IAD cues. Half the recorded waveforms were set aside for training the neural networks and the other half set aside for testing.

Three different neural networks were built. The first received just IAD cues, the second just ITD cues and the third both ITD and IAD cues.

2.1 Localization with just the IAD Cues

2.1.1 Computation of the IAD cue

The goal of the experiment was to see how the recognition rate for the IAD network varied according to the length of window used to compute the IAD cues.

A single IAD cue (i.e. a single network input vector) was computed by extracting two vectors x_l^t and x_r^t of length N, from the binaural waveforms x_l and x_r at a selected time t(see section 1.5), i.e.

$$\begin{aligned} x_l^t &= [x_l(t), x_l(t+1), \ldots, x_l(t+N)] \\ x_r^t &= [x_r(t), x_r(t+1), \ldots, x_r(t+N)] \end{aligned} \tag{4}$$

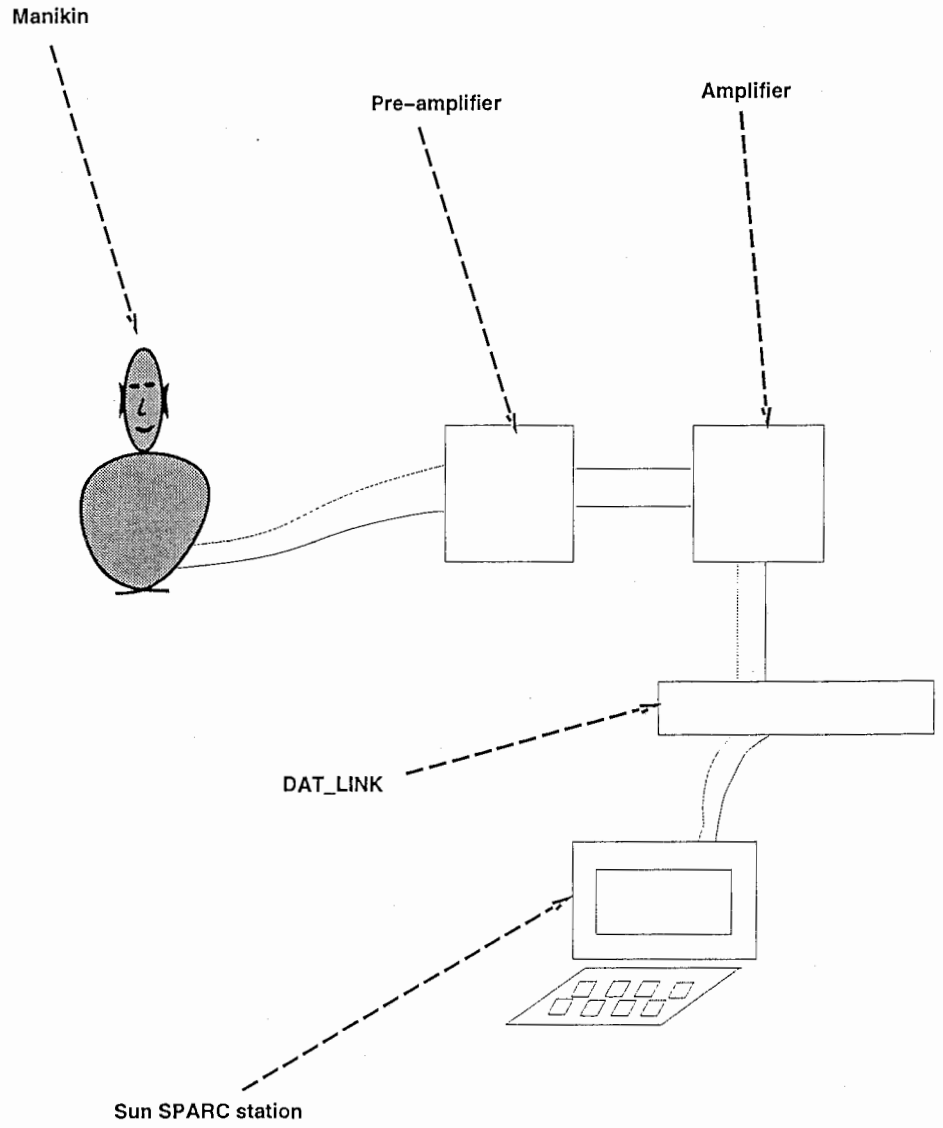


Figure 10: The database recording set up.

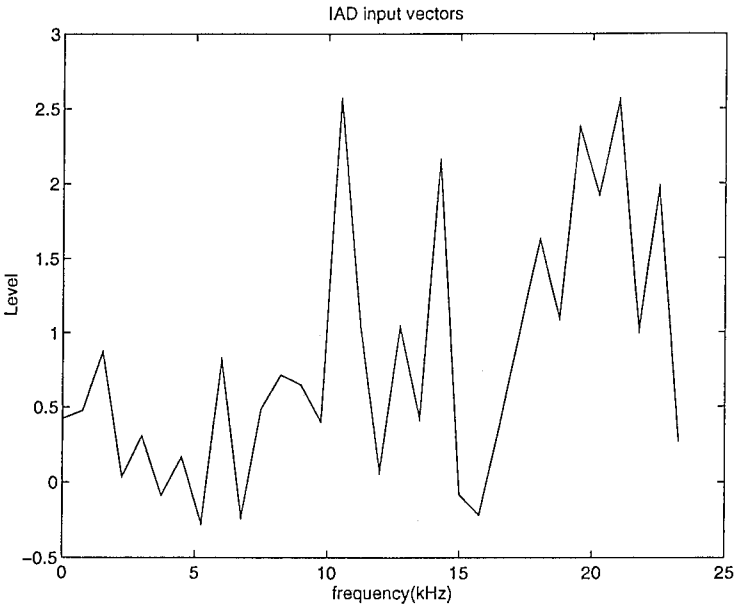


Figure 11: A typical IAD network input vector

where $x_r(\cdot)$ is the right binaural waveform and $x_l(\cdot)$ the left binaural waveform. The corresponding network input vector $\mathrm{IAD}^t(\omega)$ was then computed as the log of the ratio of the power spectra of x_l^t and x_r^t

$$\mathrm{IAD}^t(\omega) = 2\log\left(\frac{|X_l^t(\omega)|}{|X_r^t(\omega)|}\right). \tag{5}$$

$X_l^t(\omega)$ and $X_l^t(\omega)$ both had 32 spectral bins, with ω ranging from 0 to the Nyquist frequency.

Figure 11 shows a typical IAD network input vector.

2.1.2 Experimental Results

The graphs in Figure 12 shows how the recognition rate varied for training window lengths varying geometrically from 1 to 680ms. Each curve is made from a fixed training window length: Six identical networks were trained with six different window lengths (20, 40, 85, 170, 340, 680ms). Each network was

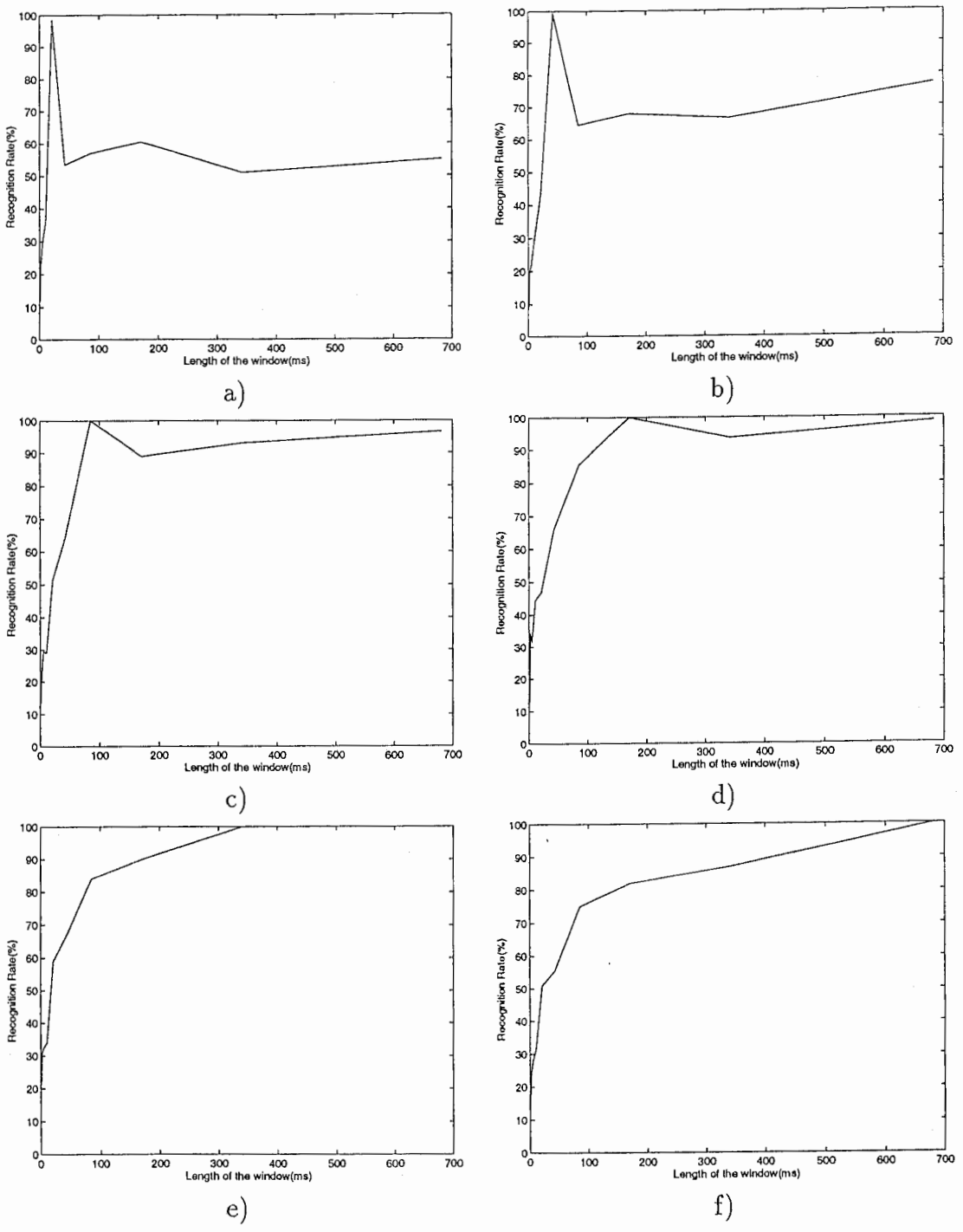


Figure 12: Recognition rate for IAD network, as a function of the testing window length with training window lengths of a) 20ms, b) 40ms, c) 85ms, d) 170ms, e) 340 ms, f) 680ms

then tested using all the different window lengths.

The results show that networks trained using short window lengths(i.e. 20 and 40ms) gave good localization results when tested using the same window length but were unable to generalize to longer window lengths, despite the fact that the S/N ratio of the IAD feature increases with increasing window length. Networks trained using intermediate window lengths (i.e. 85 and 170ms) showed much better stability across the range of different window lengths. Of particular interest is the fact that the network trained using the 85 ms window gave consistently better performance over the 100 to 680 ms window length range than the networks trained using the 680 ms window. The reason for this may well be that IAD cues computed using the 85 ms window were more noisy than those computed using the 680 ms window. Consequently the 85ms network learnt a more robust set of boundaries than the in the 680 ms network case. From these results it would seem that the optimum window length is around 100 to 200ms .

2.2 Localization with just the ITD cues

2.2.1 Computation of the ITD cue

Just as in the previous experiment the goal here was to find the optimal window length, but this time for computing the cross-correlation function between the input signals. The ITD cue extracted from the binaural waveform at time $t(\mathrm{ITD}^t)$ was defined as

$$\mathrm{ITD}^t(\tau) = \sum_{n=t}^{t+N} x_l(n)x_r(n+\tau) \qquad -\tau_{\max} \leq \tau \leq \tau_{\max} \tag{6}$$

where N is the window length and $x_l(\cdot), x_r(\cdot)$ the left and right binaural waveforms respectively. The maximum delay $\tau_{\max}$ was fixed at 55 sample periods for all window lengths. The value of $\tau_{\max}$ was chosen to cover approximately twice the range of physically possible delays(the factor of two being an arbitrary "safety factor"). Figure 13 shows a typical input vector.

2.2.2 Results of the Experiment

With the ITD network it was found that the localization results improved with window length. Figure 14 shows the average localization performance

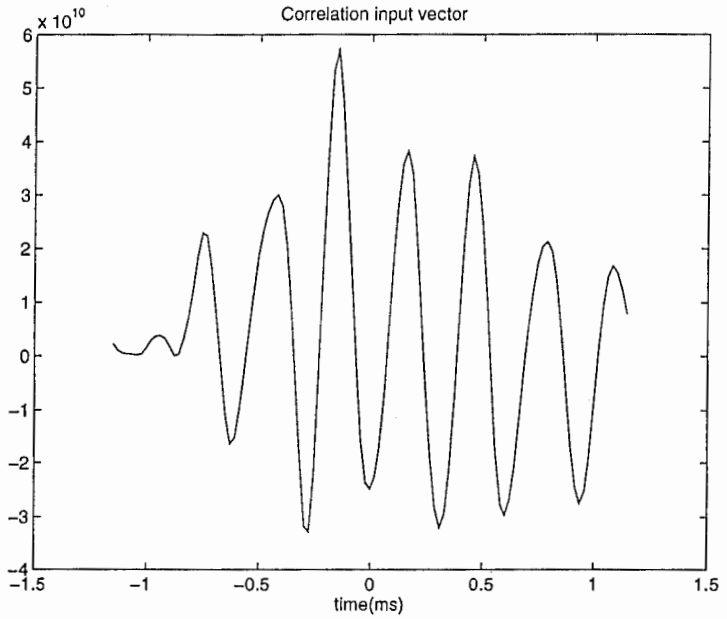


Figure 13: A typical ITD network input vector.

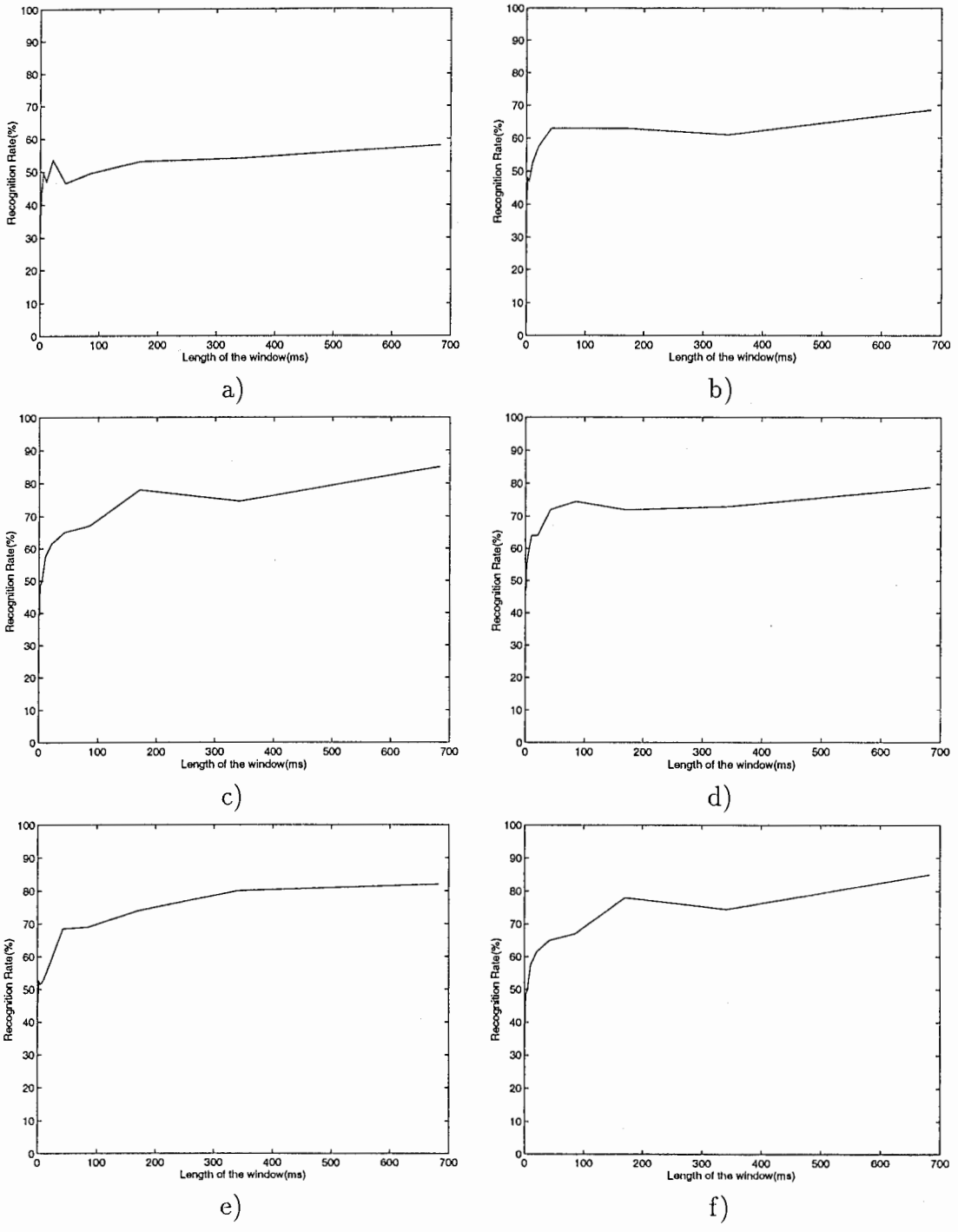


Figure 14: Recognition Rate for ITD network, as a function of the testing window length, with training window length of a) 20ms, b) 40ms, c) 85ms, d) 170ms, e) 340 ms, f) 680ms

over all three sound types (i.e. speech, music and clapping) for varying window lengths. As in the IAD case, 6 identical networks were trained with different window lengths(20, 40, 85, 170, 340 and 680ms) and then tested using all 6 window lengths.

The most likely explanation for the results is that the cross correlation estimates become less noisy when computed using longer windows. The reason being that the noise is less correlated than the sounds being localized. Note also that none of the results got better than 85%. A reason for this may be insufficient training of the networks(both the ITD and IAD networks were trained using 32 iterations, but obviously the IAD network was much smaller having only 32 inputs, while the ITD network had 111).

From the point of view of the performance /computation tradeoff a value of around 200ms for the window length seemed optimal. This agrees nicely with similar results in the literature using cross correlation vectors and peak picking algorithms for localization.

Although only the average localization performance over all 3 sound types is plotted in Figure 14, when analysed individually it was found that the music was far more difficult to localize than either the clapping or the speech. Localization performance for the music never got above 51%. This may well have been due to the large high frequency content of the music sounds (see section 1.1).

2.3 Experiment 3: Localization using both ITD and IAD cues

2.3.1 Goal of the Experiment

The goal of this experiment was to determine if both ITD and IAD cues can be combined to obtain a more accurate determination of sound direction. The input vector to this joint ITD/IAD network comprised simply of the concatendated ITD and IAD feature vectors. No additional information was provided to the network in the form of the joint feature partition boundary or relative "importance" weightings to attach to either of the two sub features. The network was thus free to interpret the joint ITD/IAD vector as it felt fit.

2.3.2 Results of the Experiment

As in the previous experiments six different networks were trained using six different window lengths. All the networks were then tested using all the window lengths. As before all the networks learned to localize the sounds with varying degrees of accuracy. As a general rule performance increased with window length. For the longer windows the performance increased only very slightly when compared to that of the best ITD and IAD results at the same window length. It was only at the shorter window lengths that the use of both cues showed significant performance increases over the single cue case(see 12b), 14b), and 15b)). Since the use of shorter window lengths is important for real time localization these results are significant in suggesting a definite advantage in using both ITD and IAD cues in combination.

Conclusion

The results show that both IAD and ITD are important cues for sound source localization. In addition it was found that networks using a combination of both IAD and ITD cues outperformed those using either ITD or IAD cues alone. One important aspect of using both cues is that it allows good localization performance with shorter window lengths. Thus for the construction of real time localization systems it would appear that the use of both cues would be advantageous. The final section of this report describes the construction of one such real time localization device designed to exploit the above findings.

3 The Construction of a Real Time Localization Device

The aim was to construct a localization device capable of exploiting both ITD and IAD cues in near real time. The system architecture was inspired by Konishi's description of the parallel pathways in the owl's brain, along which ITD and IAD cues are separately processed [1], prior to combination at a higher level brain center.

In [1], Konishi describes how in the owl's brain acoustic signals from both ears are fused to produce a single spatial perception. Since different combi-

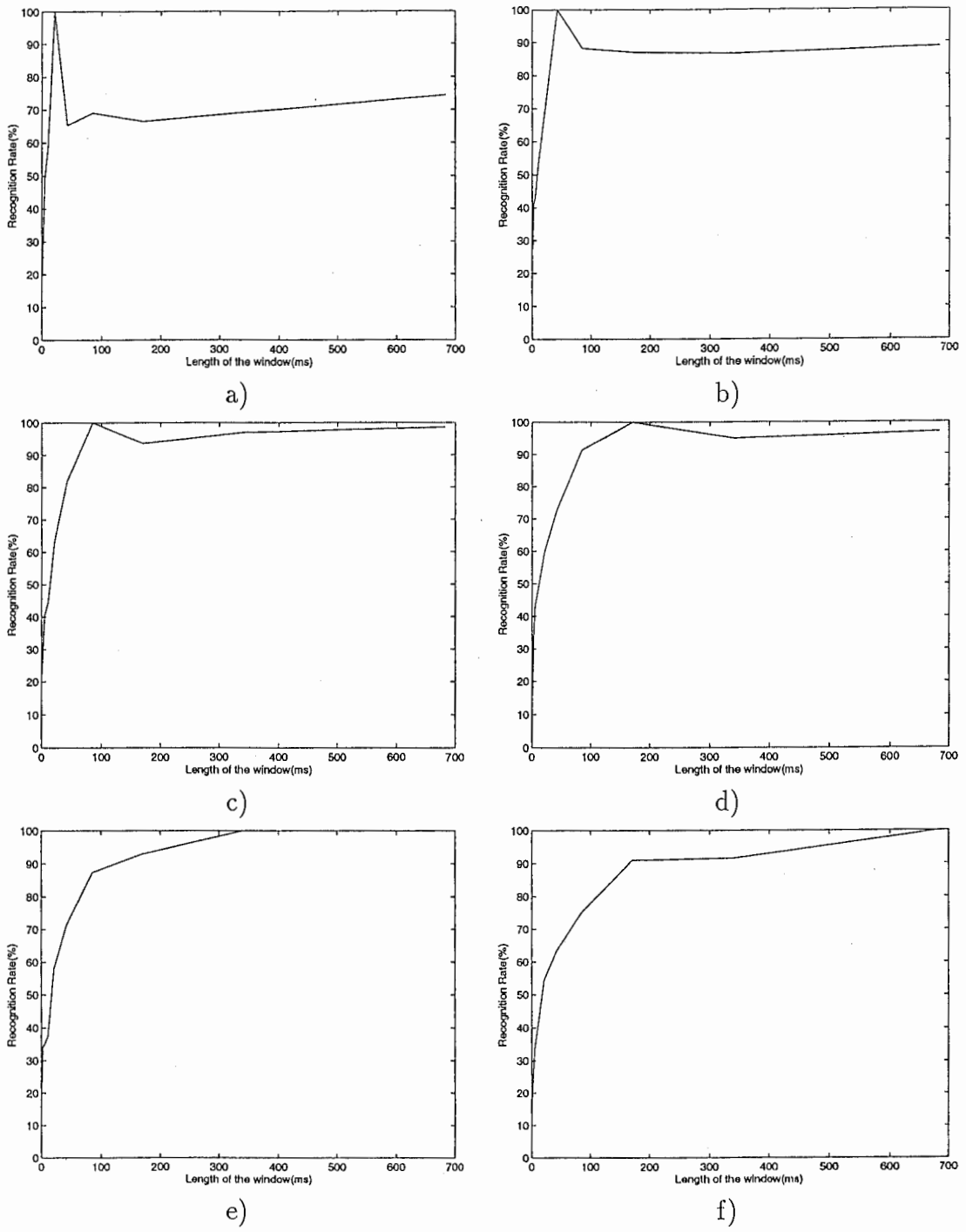


Figure 15: Recognition Rate for ITD&IAD network as a function of the testing window length, with training window length of a) 20ms, b) 40ms, c) 85ms, d) 170ms, e) 340 ms, f) 680ms

nations of signal timing and intensity differences cause the owl to turn its head in predictable directions, Konishi argued that both ITD and IAD cues must somehow be combined to produce a single spatial percept. On physical examination of the brains from several owls he found in the lower regions of the brain neurons sensitive only to interaural time differences (in the *magnocellular nucleus*) and other neurons sensitive only to intensity differences (in the *angular nucleus*). Tracing the pathways he found that time and intensity difference cues are processed separately and then converge at a higher brain level (the *lateral shell of the midbrain auditory area*).

This model of processing inspired us to build a parallel modular system with different modules corresponding loosely to different brain regions, in the sense that each module was assigned a specific task or cue. The pathways joining the brain regions were crudely modelled using a message passing network. The higher region of the brain was implemented by a center controller, receiving the outputs of the lower level ITD and IDA pathways. The advantages of this modular parallel approach were speed, flexibility and extensibility. The next subsection provides a brief description of the various modules currently implemented.

3.1 Brief Description of the Different Modules

This subsection describes the various modules currently implemented in the real time acoustic localization device. Each module runs on a separate workstation, all the modules communicating via RPC based messages.

3.1.1 Signal Acquisition Module

The task of this module is simply to establish a connection from the computer to the external A/D converter and acquire the binaural acoustic signal from the head and torso simulator. At each cycle a single binaural signal frame is output.

3.1.2 Binaural Signal Separation Module

At each cycle this module receives a binaural signal frame from the signal acquisition module, separates the left and right components and then outputs two separate monoaural signal frames.

3.1.3 Sound Level and Cue Separation Module

This module receives the left and right monoaural signals and performs sound detection using a simple level threshold. If a sound is detected the left and right signals are duplicated, one left-right pair being sent to the ITD pathway, the other left-right pair being sent to the IAD pathway.

3.1.4 Spectrum Analysis Modules (IAD pathway)

There are two separate spectral analysis modules, one for the left signal and one for the right. Each module computes its input signal power spectrum using a standard FFT based approach.

3.1.5 Spectral Averaging and Dimensionality Reduction Modules (IAD pathway)

There are two separate modules, one for left power spectrum and one for the right power spectrum. The input power spectra dimensionality is reduced to the dimensionality of the IAD network input vector by smoothing and downsampling in frequency.

3.1.6 Spectral Log Difference Module (IAD pathway)

This module computes the spectal log difference between the dimensionality reduced left and right power spectra. The output of this module is the IAD network input vector.

3.1.7 IAD Neural Network Module (IAD pathway)

The IAD network is as described in subsection 2.1. It evaluates the direction of a sound using IAD cues.

3.1.8 Peak Holding Module (ITD pathway)

There are two of these modules, one each for the left and right monoaural signals. These two modules perform the peak holding technique described by Kaneda in [10]. The power of the waveform is determined and passed to a peak holder, the level of which attenuates over time. Finally the output is differentiated and passed on to the cross correlator module.

3.1.9 Cross Correlator (ITD pathway)

This module cross correlates the outputs from the two peak holding modules. The resulting cross correlation vector becomes the ITD network input vector.

3.1.10 ITD Neural Network (ITD pathway)

The ITD network is as described in subsection 2.2. It evaluates the direction of a sound using ITD cues.

3.1.11 ITD and IAD cue fusion (High Level Center)

This module combines the outputs of the ITD and IAD networks to produce a single location percept.

3.1.12 Graphical Interface (System Utility)

This module can interactively display system parameters on the screen or log them to a file for later analysis.

3.2 Results

Over 20 workstations are employed to run both the modules described above and some additional experimental modules not described here. A typical screen display is shown in figure 17. The graphs represent the input and output of the various system modules. The image of the head displays a recognized sound source direction by moving to face in that direction. Ideally the head would be able to track a moving sound source in real time as it moves about the head and torso simulator. Although the system does make mistakes it is able to track suitably loud sounds with a delay of about 5 seconds between the utterence of a sound and motion of the head on the screen.

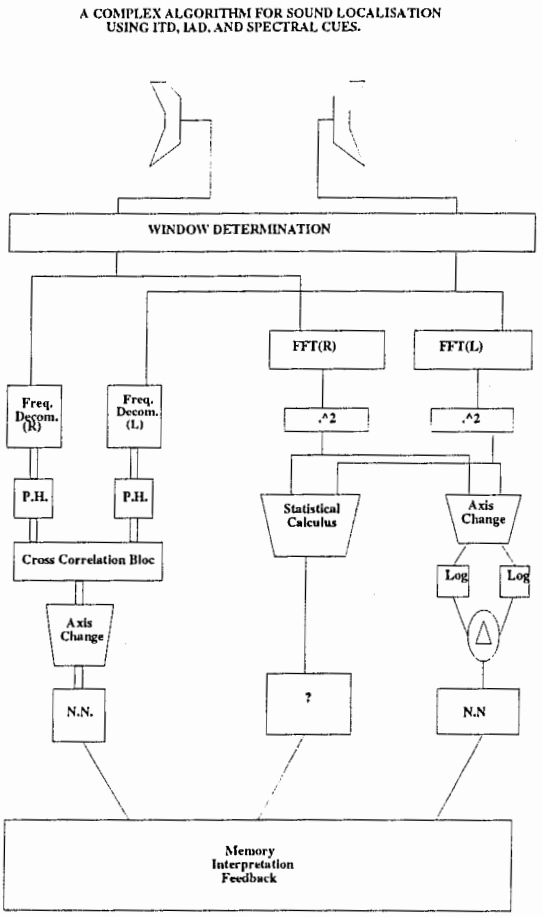


Figure 16: A Description of the Algorithm

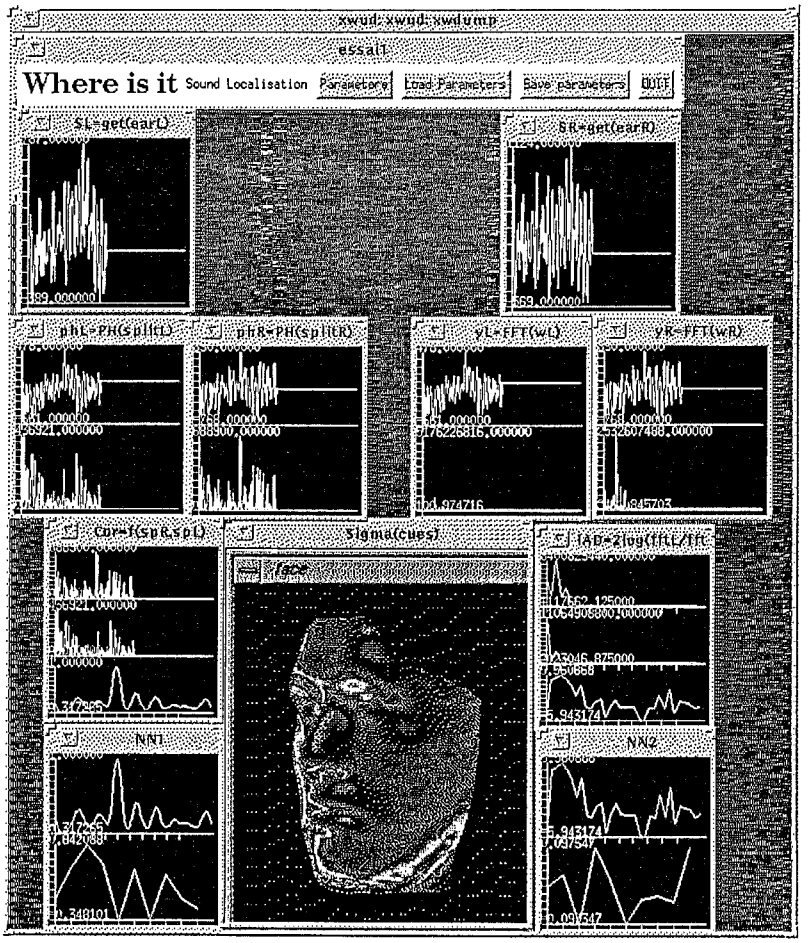


Figure 17: The screen display seen when the localization device is running

3.3 Further study

Much remains still to be done. In particular future work should employ a front end auditory model to simulate the kind of signal processing performed within the human auditory system. Also different ways of computing the ITD and IAD features should be investigated. Most important however is the study of how ITD and IAD cues are best combined. For example, one possible alternative to simply concatenating the cues prior to network classification would be to use the output of the IAD network to somehow constrain the search space of the ITD network. Many other possible solutions exist for what is clearly an important problem for future research.

Conclusion

This study has demonstrated the importance of both IAD and ITD cues for sound source localization. Localization using a combination of both cues appears particularly advantageous when using short analysis window lengths. The construction of a real time localization device went part way to demonstrating the feasability of the proposed algorithms. However, much remains to be done before anything near human like performance levels are attained. Although other approaches such as microphone arrays also exist, which may provide better localization results, the approach presented in this report has the advantage of perhaps providing some insight into the human localization process.

A The Neural Network

This section describes how a network is built and the different network parameters. For more details about Aspirin or about the gradient descent algorithm, see [12].

A.1 Description of a Network

To build a network, first a description file is created, where each network layer is completely described. Below is a sample of such a file :

```
#define NInp 111
#define NHid 30
#define NOut 10

DefineBlackBox W
   {
        OutputLayer-> Angle
        InputSize-> NInp
        Components->
           {
              PdpNode3 HidenLayer [NHid]
                 {
                    InputsFrom-> $INPUTS
                 }
              PdpNode3 Angle [NOut]
                 {
                    InputsFrom-> HidenLayer
                 }
           }
   }
```

With this file, a C-program and an executable file can then be automatically generated.

A.2 Learning parameters

Gradient descent is used to make the network learn. The weights are updated according to the formula:

$$\Delta w_{ji}(t) = -\alpha \frac{\delta E}{\delta w_{ji}} + \gamma \Delta w_{ji}(t-1) \tag{7}$$

where w_{ji} is the weight of the connection from node i to node j, E the total error, α the learning rate, and γ the momentum. Hence, the weight change of a particular weight is not simply proportional to the contribution of that weight to the total error, there is an additional inertia term in the equation. The value of α was fixed to 0.01 and γ to 0.9 . The number of iterations for the learning was kept constant to try and ensure consistency across experiments.

B The parallel implementation

For the parallel implementation of the acoustic localization algorithm the *p4 parallel programming system* by Ralph Butler and Edwing Lusk (ARGONNE NATIONAL LABORATORY) was used. Although this package supports process management, cluster management, global operations, timing functions, debugging functions, memory management and monitor building the message passing facilities were the most heavily used part of the package. The localization algorithm was divided into a number of elementary processes (see section 3) each process being executed on a single machine. In total, twenty workstations in the HIP Sun Spark Station network were used to execute the program.

A single executable file was created. When executed the program first examines a host machine table and then copies and starts itself running on each host in the table. Thus multiple copies of this executable are spawned across the network. Since each host has a unique id, each individual executable performs different functions using conditional branches based on its host id. Thus although a single program containing all the procedures and functions for all the various modules is distributed over the network, each individual process only executes an appropriate portion of that program. Process syncronization is achieved using message passing.

Acknowledements

I would like to express my gratitude to all the people in the ATR Human Information Processing Research Laboratory, who have contributed to a very friendly work atmosphere, who have helped me during my internship, and made this work possible. I also wish to express sincere thanks to Dr. David Rainton for his continual support and very helpful advice during the length of my stay in Japan, to Alain Biem for his encouragements and his help with the neural networks, to Jari Vaario for his help with the p4 package and to Tsuzaki-san for all the help he gave me with the audio equipment. Thank you all.

References

[1] M. Konishi, *Listening with 2 ears*, Scientific American, April 1993.

[2] J. Blauert, *Spatial Hearing*, MIT Cambridge MA.

[3] P.M. Zurek, *The Precedence effect and its possible role in the avoidance of interaural ambiguities*, J. Acoust. Soc. Am. Mars 1993.

[4] S.Carlile and D. Pralong *The location dependent nature of perceptually salient features of the human head-related transfer function*, J. Acoust. Soc. Am. June 1994.

[5] C. Neti, E.D. Young and M.H. Schneider *Neural network models of sound localisation based on directional filtering by the pinna.* J. Acoust Soc. Am. December 1992.

[6] D. Algom, R. Adam and Lior Cohen-Raz *Binaural summation and lateralization of transients: A combined analysis*, J. Acoust. Soc. Am October 1988.

[7] R. Meddis *The Conceptual Basis of Modelling auditory Processing in the Brainstem* The ATR Workshop September 16 and 17, 1994.

[8] J. Backman and M. Karjalainen *Modelling of human directionnal and spatial hearing using neural networks* 1993 IEEE.

[9] M.R. Schroeder *Listening with two ears*, Music Perception, Spring 1993.

[10] Yutaka Kaneda *Sound source localization for wide-band signals under a reverberant condition* J. Acoust. Soc. Jpn 1993.

[11] Bruel & Kjaer, *Head and Torso Simulator Type 4128*, Instruction Manual.

[12] Russell Leighton and the MITRE Corporation, *Aspirin/MIGRAINES*, User's Manual.

[13] R. Butler and Edwing Lusk, *The p4 Parallel Programming System*, User's Guide.