A BIOLOGICALLY PLAUSIBLE ACOUSTIC MOTION DETECTION NEURAL NETWORK

Sofia Cavaco
Departamento de Informática, Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
2825-114 Monte da Caparica, Portugal
e-mails: sc@di.fct.unl.pt, scavaco@mail.telepac.pt
&
John Hallam
Division of Informatics, University of Edinburgh
5 Forrest Hill, EH1 2QL Edinburgh, U.K.
e-mail: john@dai.ed.ac.uk

In this paper we present an acoustic motion detection system to be used in a small mobile robot. While the first purpose of the system has been to be a reliable computational implementation, cheap enough to be built in hardware, effort has also been taken to construct a biologically plausible solution. The motion detector consists of a neural network composed of motion-direction sensitive neurons with a preferred direction and a preferred region of the azimuth. The system was designed to produce a higher response when stimulated by motion in the preferred direction than in the null direction and that is in fact what the system does, which means that, as desired, the system can detect motion and distinguish its direction.

1 Introduction

Sounds arriving at one ear are slightly different from the same sounds received by the other ear. It is these interaural differences, along with other spectral cues, that are used by the auditory system to find out the sound sources' locations.

In the human species the localisation of sound sources starts at the superior olivary nucleus. The medial superior olive uses the difference in time between the arrival of the sound at each of the two ears, which is called the Interaural Time Difference (ITD), to localise sound sources. The azimuth of the source can be extracted from this difference[a]. If the sound comes from 0° in azimuth or 180° it will reach both ears synchronously (ITD=0); the maximum absolute value of ITDs (max_{ITD}) is obtained whenever the sound comes from 90° or -90°. Other positions will have ITDs between 0 and max_{ITD}. On the other hand, the lateral superior olive uses the difference in intensity between the signals arriving at each ear, which is called the Interaural Intensity Difference (IID), to localise sound sources [5]. If the wavelength is less than or equal to the distance between the ears there is ambiguity when extracting the azimuth of the source from ITDs. For this reason and also from results of experiments it is thought that ITDs are only used for frequencies below a certain value which depends on the size of the head, while IIDs are used for frequencies above that value.

The cells in the medial superior olive are binaural neurons with two sets of dendrites, which receive their inputs from each cochlear nucleus. These neurons are sensitive to specific temporal delays between the arrival of the signal at the two ears [5]. The neurons are arranged in a pattern such that neurons close to one extreme of this nucleus are maximally activated by short delays between the arrival of the signal at each ear, while neurons close to the opposite end of the nucleus are maximally activated by long delays. The cells which lie between the two ends of the nucleus respond maximally to intermediary delays [5]. In this

[a] Strictly speaking, it cannot. The sound can be located on a hyperboloid of revolution, the surface for which the difference in path lengths to the ears is fixed. However, assuming that the sound comes from a certain elevation we can extract the azimuth from the ITD.

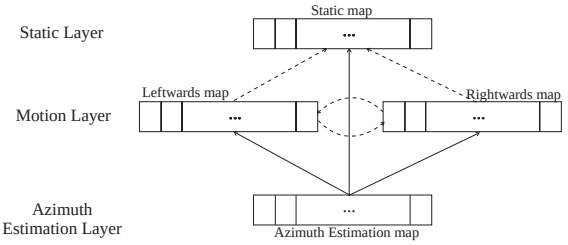

Figure 1. The motion detector. The lines '—' represent excitatory connections while the lines '_ _' represent inhibitory connection.

way, this nucleus has a spatial pattern of neuronal stimulation, or in other words, it has a topographic map of ITDs.

This spatial organisation of ITDs is spread through the auditory pathways up to the auditory cortex, where the region which is most stimulated determines the direction from where the sound comes from [5].

Since it is well understood how ITDs vary with azimuth and also a system based on these cues to localise sound can be easily implemented in hardware and computationally cheap software, the motion detector we have built works on ITDs to localise and detect motion of sound sources. Also, it is composed of four auditory topographic maps, which are divided into three layers (Fig. 1). The neurons in each map are organised in the same way as the neurons in the medial superior olive nuclei, apart from the fact that instead of having two maps (one to localise sounds coming from each side of the brain) we have joined each pair of maps into a larger vector whose middle position corresponds to the frontal azimuth. All four maps have the same spatial organisation of ITDs.

2 The Acoustic Motion Detection Neural Network

The Acoustic Motion Detection (AMD) neural network is composed of three layers, the azimuth estimation layer, the motion layer and the static layer, (see Fig. 1). The azimuth estimation layer includes the azimuth estimation map that is built using ITDs. The motion layer is composed of two maps which are sensitive to motion. One of the maps is sensitive to leftwards motion in the azimuth while the other is sensitive to rightwards motion. Finally, the static layer contains the static map, whose function is to detect static sound sources.

All four maps have the same number of neurons, which are related to regions in the azimuth. The leftmost position of each map is related to $-90°$ in azimuth, the middle position corresponds to the frontal azimuth ($0°$) and the rightmost position corresponds to $90°$ in azimuth. Also, the maps have the same topographic organisation, that is, neuron i in the azimuth estimation map is related to the same region of the azimuth as neuron i in each of the other maps.

Using ITDs to estimate the position of sound sources does not allow us to distinguish whether a sound comes from ahead or behind. Nevertheless, with a head rotation the ambiguity can be overcome. In fact, head rotations are used in biological systems to help localising sound sources.

The neighbourhood preserving characteristics of the auditory system are used in this artificial system. Not only are the cells of each map arranged by regions of the azimuth they are related to (nearby cells correspond to nearby regions of azimuth) but also that organisation is similar in all the maps. In other words, the topographic organisation of the azimuth estimation map is preserved in the motion and static layers. Additionally, according to what is said in [8, 18], the axons that come from the same neighbourhood in the source map reach the same neighbourhood in the target map and different cells in the same map have similar connections.

In the following sections the three layers that compose the AMD network will be seen in detail.

2.1 The Azimuth Estimation layer

The Azimuth Estimation layer is used to estimate the azimuth of sound sources with or without motion. This layer is composed of a map, which resembles the organization of the medial superior olive and is produced by the Sound Source's Azimuth Estimation (SSAE) system [3].

The SSAE system estimates the azimuth of sound sources using ITDs, which are found using Zero Crossings. This system divides the signal into several frequency bands and produces a map for each band. Each of these maps is used as the input layer of a different AMD neural network. In the remaining of this paper we will refer to just one of these maps.

The map produced by the SSAE system works as a voting system and every ITD contributes with a vote to its construction. For instance, if the arriving ITD says that sound comes from $45°$, a vote will be added to the position of the map that corresponds to $45°$ in azimuth. Each cell contains the number of votes to a given region of the azimuth. For that reason this map can also be called the vector of votes. The value of each cell in the map is a measure of how certain the

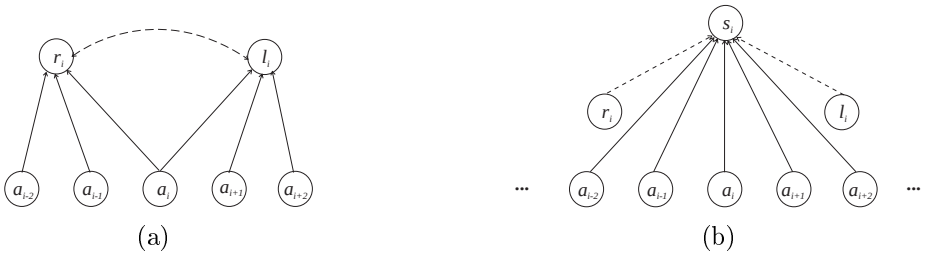

Figure 2. The AMD neural network connections. The neurons on the bottom belong to the azimuth estimation map. Neuron l_i belongs to the leftwards map, neuron r_i belongs to the rightwards map and neuron s_i belongs to the static map. The neurons' subscript tells the regions of the azimuth they are related to, for instance, neuron a_i is related to region i, just like neuron l_i, r_i and s_i. The lines '—' represent excitatory connections while the lines '_ _' represent inhibitory connections. (a) The motion layer connections. (b) The static layer connections.

system is that sound comes from the region related to that cell. If there is a cell with a value much higher than the others, it means that the system thinks that sound comes from the region related to that cell. (For more details on the SSAE system see [3].)

2.2 The Motion layer

The Motion layer is responsible for detecting acoustic motion. This layer is composed of two maps, the leftwards and the rightwards map, which have the azimuth estimation map's topographic organisation. The value of each cell in these maps indicates if there is a sound source moving in the region the cell is related to. The leftwards map indicates if the sound sources are moving leftwards while the rightwards map indicates if the sound sources are moving rightwards.

Motion-direction sensitive neurons

Several sources [1, 4, 10, 15, 16] not only show that there are acoustic motion sensitive neurons but also agree that these neurons are motion-direction sensitive, that is, the neurons not only respond to moving sounds but also they have a higher activation when responding to sounds moving in a specific direction, which is called their preferred direction. In [16] the authors go even further and say that these neurons have a preferred region in the azimuth, which suggests the existence of auditory motion-direction maps in the auditory system. Also, Mäkelä and McEvoy say that the cells with the same preferred direction are grouped together [10].

Many of the neurons sensitive to motion do not distinguish between motion in the preferred direction and static sounds (that is, they have a *direction-independent excitation*). Nevertheless, some neurons do in fact distinguish between moving and stationary sounds. These cells can have a higher response to motion in the preferred direction than in the null direction, which is the direction opposite to the preferred direction, or stationary sounds, which suggests a *direction-dependent excitation*. There are also neurons that do not distinguish between motion in the preferred direction and stationary sounds but which can have a lower activation in response to the null direction, which suggests a null-direction inhibition (that is, a *direction-dependent inhibition*). Lastly, some neurons can also have a *direction-independent inhibition* [1, 15].

Wagner and Takahashi have built an acoustic motion detector composed of neurons which are both direction-independent excitatory and direction-dependent inhibitory [15]. In other words, the neurons' activation is stronger when responding to static sounds or motion in the preferred direction than motion in the null direction.

On the other hand, we have tried to build the leftwards and the rightwards motion neurons with direction-dependent excitation and direction-dependent inhibition. Therefore, the activation level of these neurons is higher when stimulated by motion in the preferred direction than when stimulated either by stationary sounds or motion in the null direction.

Additionally, the neurons with the same preferred direction were grouped in the same map, giving rise to the leftwards map and the rightwards map (which agrees with Mäkelä and McEvoy's statement about having the neurons with the same preferred direction grouped together [10]).

The activation levels

Each neuron in the leftwards and in the rightwards map is supposed to tell if there is leftwards or rightwards motion in the region it is related to. For that purpose the neurons must be sensitive to temporal and spatial changes of the sound source's azimuth.

That sensitivity can be obtained by looking to the history of the winning cells in the azimuth estimation map. For instance, if at some instant of time, $t-n$, cell a_{i-1} in the azimuth estimation map wins a vote and the next cell to win a vote is cell a_i at time t, it looks like the sound source is moving rightwards (from region $i-1$ to region i). Thus, cell r_i in the rightwards map should notice it at time t and fire (see Fig. 2 (a)).

If the neurons in the motion layer are able to memorise past activation levels of some of the cells in the azimuth estimation map they will be able to check whether there is movement in their preferred directions. Lets suppose that cell i in the azimuth estimation map wins a vote when sample s has been processed. Then the value of cell i at sample s is higher than its value at sample $s-1$ times the decay (see [3]), that is,

$$votes(s,i) > votes(s-1,i) \times e^{\frac{\Delta t_s}{\tau \times 2^{band-1}}},$$

with Δt_s the time since the previous vote, that is, the time between sample s and the last sample where a vote occurred (Δt_s is always increasing until a new vote appears). τ is a constant smaller than zero and *band* is the highest cutoff frequency of the band being considered.

However, considering consecutive samples makes the system respond to velocities of one azimuth region per inter-sample time, i.e.,

$$\omega = \frac{region}{\Delta t} = region \times \frac{f_s}{2} = 1.47 \times 10^5 (^\circ s^{-1}),$$

where *region* stands for the degrees of the azimuth that each vector cell corresponds to and Δt is the time between two samples, which is $2/f_s$, with f_s the sample frequency[b]. This is unreasonably fast.

Therefore (considering intervals of $n+1$ samples) to be certain that there was at least one vote between sample $s-n$ and sample s one has to make sure that[c] (see [3]):

$$votes(s,i) > votes(s-n,i) \times e^{\sum_{j=s-n}^{s} \frac{\Delta t_j}{\tau \times 2^{band-1}}}.$$

[b] In this system $region = 6.67^\circ$ and $f_s = 44100Hz$. Also, note that Δt and Δt_s have different meanings.

[c] Note that the decay is applied to each cell in the azimuth estimation map every time a new sample is processed.

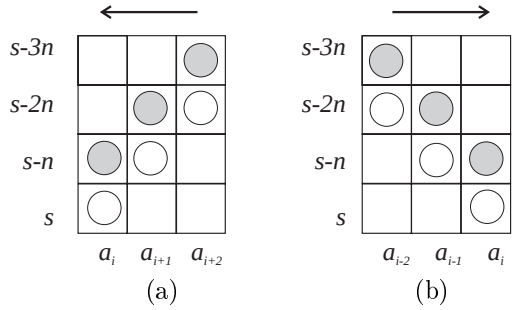

Figure 3. Each line in the figure represents part of an azimuth estimation map. Two consecutive lines are two configurations of that map displaced by $n+1$ samples (line s is the most recent). Therefore, grey circles represent $votes(sample-n, region)$ while white circles represent $votes(sample, region)$. In order to sense motion the neurons in the motion layer must detect votes between black and white circles in the same column. (a) The comparisons to be done to detect leftwards motion. (b) The comparisons to be done to detect rightwards motion.

Having all that in mind, the activation of the neurons in the motion layer can be done by memorising and comparing past activations of the neurons in the azimuth estimation map. Comparing the values between the white and grey circles in each column of the masks in Fig. 3 a neuron can detect motion in its preferred direction approaching its preferred region. For example, neuron l_i can detect leftwards motion in region i comparing the activation values marked in Fig. 3 (a) while neuron r_i can detect rightwards motion in the same region comparing the activation values marked in Fig. 3 (b).

Therefore, the activation functions of leftwards and rightwards neurons are:

$$left(s,i) = \begin{cases} votes(s,i) & \text{if } d(s,i) > t_1 \\ & \text{and } d(s-n,i+1) > t_1 \\ & \text{and } d(s-2n,i+2) > t_1 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$right(s,i) = \begin{cases} votes(s,i) & \text{if } d(s-2n,i-2) > t_1 \\ & \text{and } d(s-n,i-1) > t_1 \\ & \text{and } d(s,i) > t_1 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where:

$$d(s,i) = votes(s,i) - votes(s-n,i) \times e^{\sum_{j=s-n}^{s} \frac{\Delta t_j}{\tau \times 2^{band-1}}}. \quad (3)$$

A threshold function was introduced because computers have not a infinite precision and also to make

the system not only more reliable but more biologically plausible as well.

Apart from these excitatory connections, each neuron in the motion layer has also an inhibitory connection from its corresponding neuron in the map of its null direction. In other words, there is an inhibitory connection both ways between neurons i of the leftwards and rightwards maps. The weakest neuron's activation is inhibited by this connection whereas the strongest neuron's activation is kept unchanged. The reason why the strongest neuron wins is that this neuron was the last to detect motion (i.e. this neuron detected a more recent vote than the vote detected by its partner).

It turned out that even for static sounds the vector of votes' maximum can have some little changes of position over time. The more positions the vote has to travel leftwards or rightwards in order to prove that there is leftwards or rightwards motion the better the system performance is. In order to have a more reliable output the masks used by the motion layer can have more columns (and consequently more lines). Thus, there is a tradeoff between computational expense and how reliable the output is. The AMD neural network is using three columns masks. Its masks look like those in Fig. 3.

2.3 The Static layer

The job of the static layer is to detect static sound sources. This layer contains the static map, which is composed of neurons sensitive to static sounds. Each neuron in the static layer is activated by its corresponding neuron in the vector of votes. The neuron can only be activated if the same position in the vector of votes graph is a peak high enough to pass a threshold which depends on the total activation of the azimuth estimation map. In addition, the neuron's output can be inhibited by the maps of the motion layer (Fig. 2 (b)). If there is motion detected in either way in region i, the output of static neuron i is inhibited. These neurons are thus sensitive to sound but not to motion. The activation function of this layer is:

$$f(i) = \begin{cases} votes(i) & \text{if } left(i) = 0 \text{ and } right(i) = 0 \\ & \text{and } t(votes(i)) > 0 \\ & \text{and } votes(i-1) < votes(i) \\ & \text{and } votes(i+1) < votes(i) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Table 1. Average of the activation of all the units in the leftwards and rightwards maps over time. The same sound, crumpling paper, was used moving in the two possible directions, static and combined with another source of the same kind.

	leftwards map	rightwards map
moving rightwards	12	16
moving leftwards	19	11
a static source	8	6
two static sources	2	1

where:

$$t(x) = \begin{cases} x & \text{if } x/\sum votes > t_2 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

3 Results

To test the motion detector some recordings of voices, coughs, whistles, clapping hands, crumpling and tearing paper, feet beating on the floor, rolling chairs, among others, were made using two microphones displaced by $9.5cm$, into PCW files which can be directly used by MATLAB (the language used to build both the SSAE system and the AMD neural network). The recordings were performed in a real work environment. There were background noises like the sound of keyboards and voices. Also we used no pinnae. The same sounds were recorded moving into different directions, in different regions of the azimuth and with different velocities.

3.1 The Motion layer

In order to observe how the system reacts to the null direction we tested the system with and without the cross inhibition between the leftwards and rightwards maps.

As explained in the previous sections, the neurons in the motion layer have a stronger activation when responding to motion in the preferred direction than when responding to motion in the null direction or to static sounds. Just as with cortical motion-sensitive neurons that have a higher response to motion in the preferred direction than in the null direction, some activation is still expected in response to motion in the null direction [1]. Table 1 is a typical example of the results obtained. This table compares the mean activation of the motion maps with moving and static sounds. The same sound was used with and without motion, also the results using a two sources sample are included. The first two lines of this table show the relation between the activation of the two motion maps

over time in response to the same sound moving into different directions. The mean activation as response to the preferred direction is higher than as response to the null direction, for instance the leftwards map's mean response to the preferred direction is 19, while to the null direction is 12. In addition, the highest mean activation is found in the map whose preferred direction is the same as the sound source's movement, for example when the source is moving rightwards the rightwards map's activation is 16, while the leftwards map has a lower activation, 12.

The reason why there is activation in response to the null direction is that though the highest peaks in the vector of votes are moving in the preferred direction, there are still some fluctuations to the other direction due to wrong guesses of the azimuth estimation map (see [3]). The response to the null direction depends on the source's velocity as well as the number of samples considered in one interval ($n+1$ in Eq. (3)). Even with fluctuations from left to right and vice-versa if the masks have more columns the sensitivity to the null direction decreases. Note that when introduced, the cell-to-cell cross inhibition between the leftwards and rightwards maps removes much sensitivity to the null direction. With a neighbourhood-to-cell cross inhibition between the two maps the sensitivity to the null direction decreases even more.

The response to static sounds is much lower than the activation in the preferred and in the null direction. Just a few cells are active in the graphs of the leftwards and rightwards maps when a static sound is heard. Table 1 shows that the motion maps activation is much lower when responding to static sounds. Take as an example, the leftwards map: its response to the preferred direction is 19, to the null direction is 12, while to a static sound is much lower, 8. Also, the system does not confuse sounds from two static sources with movement. Notice the last line in the table, which has much lower values than the first two lines.

3.2 The Static layer

While the azimuth estimation map shows some output outside the region with the strongest value [3], the static map filters that activation and just the regions with the highest values, that is, the local maxima, are active in this map. The accuracy of the static map is therefore the same as discussed in [3] for the azimuth estimation map. Comparing the azimuth estimation of different sounds at the same location, it was observed that the estimation accuracy depends on the kind of sound. Additionally, comparing the azimuth estimation of the same sound at different locations, it was concluded that the accuracy also depends on the region of the azimuth. The humans' estimation accuracy also depends on the frequency of the sound and the region of the azimuth.

It turned out that the system has an estimation error between 0° and 13.34°, depending on the sound source and the source's location. Though this estimation error may seem a lot, it is in fact better than the human estimation error, which can be as high as 16.3° [7].

Additionally, the static map and the azimuth estimation map share some other properties. To begin with, different sound sources emitting sound at the same time can be localised with the same estimation errors as the localisation of a single sound source. On the other hand, even though the duplex theory states that ITDs are used just for low frequencies (such that the wavelength is bigger than the distance between the ears) it turned out that the system is able to use ITDs to correctly localise sound sources with frequencies higher than those, provided that the interaural phase difference is less than 2π ($IPD < 2\pi$). (See [3] for more details on these results.)

4 Conclusion

In this paper we presented a biologically plausible neural network that localises sound sources and detects acoustic motion. The neural network builds three topographic maps as output, one sensitive to leftwards motion, other to rightwards motion and the last sensitive to static sounds. As input the network uses an azimuth estimation map built by the SSAE system [3], which uses ITDs to estimate the azimuth of sound sources.

As desired, the response of the motion maps to a static sound is much lower than their response to moving sounds (either in the preferred direction or the null direction). Also, as expected, it turned out that when a sound source is moving, the map whose preferred direction is the sound source's direction, acquires a higher activation level than the other map. However some solutions to obtain a lower level of activation in response to the null direction were considered. The cell-to-cell cross inhibition between the leftwards and rightwards maps could be changed to a neighbourhood-to-cell cross inhibition. In this way, if leftwards motion were detected in a certain region, no rightwards motion could be detected in the region's neighbourhood.

Although the static map has a good performance

concerning the detection of static sounds, its overall response to moving sounds could be improved changing its cell-to-cell inhibition into a neighbourhood-to-cell inhibition. Thus, if some motion were detected in a region's neighbourhood, the static cell related to that region would not be allowed to sense static sound sources.

Apart from testing the system's sensitivity to motion, its ability to localise sound sources was also tested. It was observed that, just like what happens with humans, the azimuth estimation accuracy depends on the kind of sound as well as on the region of the azimuth. The system has an azimuth estimation error between 0° and 13.34°, which, as seen above is better than the humans' estimation error.

Additionaly, the system identifies and localises different sound sources emitting sound at the same time with the same range of errors as mentioned in the previous paragraph.

Finally, it turned out that even though the duplex theory states that ITDs are used to localise sounds with a wavelength bigger than the distance between the ears, the system is able to localise sounds with lower wavelengths, provided that the IPD is less than 2π.

Acknowledgments

This research was performed in The University of Edinburgh and was supported by a grant from the Junta Nacional de Investigação Científica e Tecnológica.

References

[1] L. Aitkin. *The Auditory Cortex, structural and functional bases of auditory perception.* Chapman and Hall, 1990.

[2] A. Babeanu. Steerable ears for robotic cat. Master's thesis, Dept. of Artificial Intelligence, University of Edinburgh, 1994.

[3] S. Cavaco and J. Hallam. A biologically plausible acoustic azimuth estimation system. In H.G. Okuno and F. Klassner, editors, *Working Notes of IJCAI'99 Workshop on Computational Auditory Scene Analysis (CASA'99)*, Aug 1999.

[4] T.D. Griffiths, A. Rees, C. Witton, R.A. Shakir, G.B. Henning, and G.G.R. Green. Evidence for a sound movement area in the human cerebral cortex. *Nature*, 383:425–427, October 1996.

[5] A.C. Guyton. *Fisiologia Humana.* Guanabara, 1984.

[6] S. Handel. *Listening, An Introduction to the Perception of Auditory Events.* A Bradford Book, the MIT Press, 1989.

[7] L.A. Jeffress. Localization of sound. In W.D. Keidel and W.D. Neff, editors, *Auditory System Physiology (CNS). Behavioural Studies Psychoacoustics*, chapter 10, pages 449–459. Springer-Verlag, 1975.

[8] E.R. Kandel, J.H. Schwartz, and T.M. Jessell. *Principles of Neural Science.* Appleton & Lange, 1991.

[9] D. MacFarland. *Animal Behaviour.* Addisson Wesley, 1993.

[10] J.P. Mäkela and L. McEvoy. Auditory evoked fields to illusory sound source movements. *Experimental Brain Research*, 110(3):446–454, 1996.

[11] D. Marr. *Vision.* Freeman, 1982.

[12] W. Mills. Auditory localization. In J.N. Tobias, editor, *Foundations of Modern Auditory Theory*, volume 3. Academic Press, 1972.

[13] R.S. Payne. How the barn owl locates prey by hearing. In *Living Bird*, volume 1, pages 151–159. 1962.

[14] W.J. Tan. The cat ears project 1996. Master's thesis, Dept. of Artificial Intelligence, University of Edinburgh, 1996.

[15] H. Wagner and T. Takahashi. Influence of temporal cues on acoustic motion-direction sensitivity of auditory neurons in the owl. *Journal of Neurophysiology*, 68(6):2063–2076, December 1992.

[16] H. Wagner, T. Trinath, and D. Kautz. Influence of stimulus level on acoustic motion-direction sensitivity on barn owl midbrain neurons. *Journal of Neurophysiology*, 71(5):1907–1916, May 1994.

[17] H.T. Wang, B. Mathur, and C. Koch. I though i saw it move: Computing optical flow in the primate visual system. In M.A. Gluck and D.E. Rumelhart, editors, *Neuroscience and Connectionist Theory*, chapter 6, pages 237–265. Lawrence Erlbaum Associates, Publishers, 1990.

[18] W.R. Zemlin. *Speech and Hearing Science, Anatomy and Physiology.* Prentice Hall, 1988.