A computational model of sound localization in the barn owl

Brian J. Fischer and Charles H. Anderson
Washington University in St. Louis

February 21, 2003

Abstract

We present a computational model of sound localization in the barn owl that is based on the extraction of temporal and intensity information in parallel pathways. Localization is accomplished by performing statistical inference on cues extracted from the auditory input signals. Our formulation of the problem allows us to predict the form of the nonlinear operation that is reportedly performed on the outputs of the time and intensity pathways. We use a simple set of stimuli to show that our model performs in a manner consistent with experimental results in both single and multiple source environments.

1 Introduction

The owl is a nocturnal hunter that is able to localize sound sources in its environment using auditory cues alone ([2]). The neural basis for this localization behavior is the existence of neurons with spatially restricted receptive fields arranged in a topographic map in the external nucleus of the inferior colliculus ([1]). It is thought that the space specific neurons encode a product of signals computed in parallel by the time and intensity pathways ([3]). Here we propose a computational model of sound localization in the barn owl that is consistent with the extraction of temporal and intensity information in parallel pathways. Our probabilistic formulation of the problem allows us to predict the form of nonlinear operation that is performed on the outputs of the time and intensity pathways to be the multiplication of two likelihood functions, one computed from temporal information and one computed from intensity information. We use a simple set of stimuli to show that our model performs in a manner consistent with experimental results in both single and multiple source environments.

2 Problem Description

The sound localization problem is to estimate the location of sound sources in the environment using the auditory signals that arrive at the two ears. Here we consider a simple example where locations are defined by the azimuth, θ, and elevation, ϕ, with $\theta, \phi \in [-90^o, 90^o]$. Sound signals are taken to be of the form

$$s(t) = \sum_{i=1}^{n} A_i \cos(\omega_i t)$$

where $n \leq 4$ and $\omega_i \in [1, 10]$ kHz.
To produce the input signals to each ear, source signals are scaled and delayed to model the position dependent intensity and temporal disparities that exist between the input signals at each ear. Specifically, the input signal at the left ear, $s_L(t)$, and the input signal at the right ear, $s_R(t)$, are modelled as

$$s_L(t) = \sum_{i=1}^{n} a_L(\phi) A_i \cos(\omega_i (t - d_L(\theta)))$$

and

$$s_R(t) = \sum_{i=1}^{n} a_R(\phi) A_i \cos(\omega_i (t - d_R(\theta)))$$

The elevation dependent gains are given by $a_L(\phi) = 1 - \phi/90^o$ and $a_R(\phi) = 1 + \phi/90^o$. The azimuth dependent delays are given by $d_L(\theta) = [-ITD]_+$ and $d_R(\theta) = [ITD]_+$ where $ITD = 150 \sin(\theta)\ \mu s$.
We formulate the resulting localization problem as a statistical estimation problem. We suppose that θ and ϕ are random variables with prior density $p(\theta, \phi)$ and that signals $s(t)$ are stochastic processes characterized by the coefficients A_i, so that $p(s(t)) = p(\mathbf{A})$. The solution to the localization problem is obtained by computing the posterior probability density

$$p(\mathbf{A}, \theta, \phi | s_L(t), s_R(t))$$

From this density we estimate the location using a maximum likelihood estimate. We hypothesize that this conditional density is computed by performing inference on information extracted from the input signals within the time and intensity pathways.

3 Cue Extraction

The localization process begins by extracting information from the input signals. It is well known that operations performed in the time pathway allow for the estimation of the interaural temporal disparity (ITD), while operations performed in the intensity pathway allow for the estimation of the interaural intensity disparity (IID) ([1]). Our model of cue extraction is based on operations believed to be performed within these two pathways. In addition to computing ITD and IID we propose that computations must be performed in the intensity pathway that will allow for estimation of the input signal spectrum. Estimating the input signal spectrum is important for resolving single sources in multiple source environments and for exposing position dependent filtering of the left and right input signals.

3.1 Cochlear Filter Bank

Again, the input signals are given by $s_L(t)$ and $s_R(t)$. To model the filtering of auditory signals by the cochlea, these signals are each filtered by an identical filter bank of band-pass filters, $h_k(t)$, where the index k indicates that this filter has characteristic frequency ω_k. The filtered signals are given by,

$$u_L(t, \omega_k) = h_k \star s_L(t)$$

and

$$u_R(t, \omega_k) = h_k \star s_R(t)$$

where

$$h_k(t) = te^{-t/\tau_k} \cos(\omega_k t) H(t)$$

and $H(t)$ is the unit step function.

3.2 Energy

In order to determine the energy in each frequency band of the input signals, zero lag autocorrelations of the filter outputs $u_L(t, \omega_k)$ and $u_R(t, \omega_k)$ are computed. To detect time varying changes in the signal spectrum the energy calculations are performed within a window. We compute

$$y_L(t, \omega_k) = \int |u_L(\sigma, \omega_k)|^2 w(t - \sigma) d\sigma$$

and

$$y_R(t, \omega_k) = \int |u_R(\sigma, \omega_k)|^2 w(t - \sigma) d\sigma$$

where the window is $w(t) = e^{-t/\tau} H(t)$.

3.3 Energy Differences

Once intensity calculations are performed on the input signals through the windowed energy estimates, binaural intensity differences (IIDs) can be calculated as

$$z(t, \omega_k) = y_R(t, \omega_k) - y_L(t, \omega_k)$$

3.4 Intensity Sums

Sums of the energy estimates are computed within each frequency channel. These measures can be used for estimation of the spectrum of the sound source.

$$\psi(t, \omega_k) = y_R(t, \omega_k) + y_L(t, \omega_k)$$

3.5 Cross Correlation

Cross correlations are computed between the outputs of each filter bank with a range of delays imposed on the filter outputs. This type of operation is believed to be performed in the nucleus laminaris of the barn owl. These measures are used to extract temporal differences between the input signals that are related to the azimuthal location of the source. We compute a windowed cross correlation

$$x_L(t, \omega_k, m) = \int u_L(\sigma - m\Delta, \omega_k) u_R(\sigma - (N-m)\Delta, \omega_k) w(t-\sigma) d\sigma$$

and

$$x_R(t, \omega_k, m) = \int u_R(\sigma - m\Delta, \omega_k) u_L(\sigma - (N-m)\Delta, \omega_k) w(t-\sigma) d\sigma$$

where Δ is a delay increment, N is the total number of delay steps between the left and right extremes, $m \in \{0, \ldots, N/2\}$, and the window is the same as for the energy calculation.

3.6 Cue summary

For mathematical convenience we summarize these operations as the computation of a cue vector

$$\xi(t) = [\xi_T(t)\ \xi_I(t)]^T$$

where

$$\xi_T(t) = [x_L(t, \omega_1, 1), \ldots, x_L(t, \omega_n, N/2), x_R(t, \omega_1, 1), \ldots, x_R(t, \omega_n, N/2)]$$

is the vector of cues computed in the temporal pathway and

$$\xi_I(t) = [y_L(t, \omega_1), \ldots, y_L(t, \omega_n), y_R(t, \omega_1), \ldots, y_R(t, \omega_n), z(t, \omega_1), \ldots, z(t, \omega_n), \psi(t, \omega_1), \ldots, \psi(t, \omega_n)]$$

is the vector of cues computed in the intensity pathway. The cue vector $\xi(t)$ represents all of the input derived information that is available for the estimation of the location of a sound source.

3.7 Inference

In the following we drop the explicit dependence on time by writing $s_L = s_L(t)$, $s_R = s_R(t)$, and $\xi = \xi(t)$. We compute the posterior distribution $p(\mathbf{A}, \theta, \phi | s_L, s_R)$ by performing inference on the cue vector ξ in the following way

$$p(\mathbf{A}, \theta, \phi | s_L, s_R) = \int p(\mathbf{A}, \theta, \phi | \xi) p(\xi | s_L, s_R) d\xi$$

Assuming that the cue vector is computed in a deterministic manner from the input signals, ie.,
$p(\xi|s_L, s_R) = \delta(\xi - \xi(s_L, s_R))$ we find

$$\begin{aligned} p(\mathbf{A}, \theta, \phi|s_L, s_R) &= \int p(\mathbf{A}, \theta, \phi|\xi)\delta(\xi - \xi(s_L, s_R))d\xi \\ &= p(\mathbf{A}, \theta, \phi|\xi(s_L, s_R)) \\ &= \frac{1}{Z} p(\xi_T(s_L, s_R)|\mathbf{A}, \theta)p(\xi_I(s_L, s_R)|\mathbf{A}, \phi)p(\mathbf{A}, \theta, \phi) \end{aligned}$$

where we have used the fact that the cue vector computed in the temporal pathway is independent of the cue vector computed in the intensity pathway. So, our model suggests the computations performed in the parallel time and intensity pathways to be the calculation of two likelihood functions, one computed from temporal information and one computed from intensity information. It also requires their combination in a multiplicative manner to compute the posterior density $p(\mathbf{A}, \theta, \phi|s_L, s_R)$.

4 Results

We tested the model with single sound sources of the type described above at locations ranging from -90^o to 90^o in both azimuth and elevation. When the input signal consisted of more than one sinusoid the posterior density was concentrated around the true azimuth and elevation. A typical example is shown in figure 1 where $\theta = -50^o$, $\phi = 50^o$, and $\mathbf{A} = [.25\ .25\ .75\ .75]^T$. Here we plot the

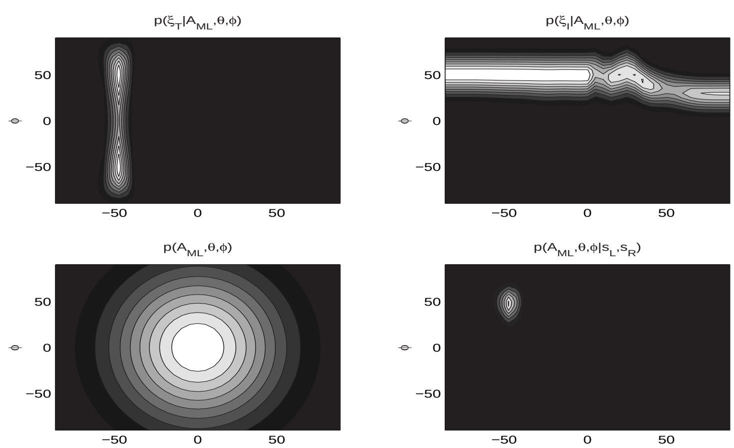

Figure 1: Elements of the Posterior Distribution for $\theta = -50^o$, $\phi = 50^o$, $t = 50$ $msec$, and $\mathbf{A}$ fixed at the maximum likelihood estimate.

prior, the two likelihood functions, and the posterior distribution. As expected, the temporal likelihood is concentrated around the true azimuthal position, but

provides no information about elevation and the opposite is true for the intensity likelihood.
We tested our model with multiple sound sources in a manner similar to the experimental procedure used by Takahashi et. al. ([4]). In our model, two sound sources were presented on the horizontal plane, $\phi = 0$, at a fixed angle apart. Takahashi et. al. report that when two sources on the horizontal plane produce the same signal, neurons responded to a phantom source located between the two sources and when the two sources produced sounds consisting of distinct frequency components the sounds were individually resolvable. As shown in figure 2, when we used two sources located on opposite sides of the vertical plane that produced the same signal a phantom bump was seen in the posterior distribution. We also see that when the signals consist of sinusoids of different frequencies the individual sources are resolvable in the posterior distribution (figure 2). Thus, our model is consistent with the experimentally reported behavior of neurons in the external nucleus of the inferior colliculus.
We note that these results were obtained while using a broad prior distribution. By adjusting the prior it is possible to resolve a single peak in the posterior distribution when multiple sources are present.

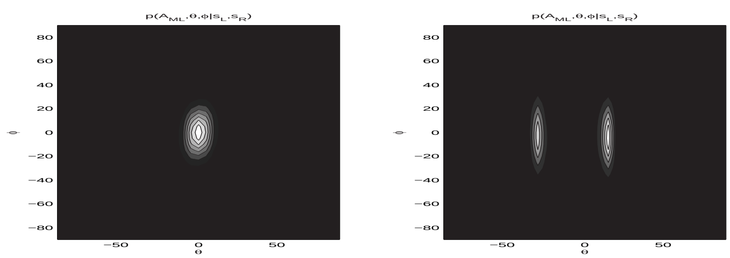

Figure 2: Left: Posterior Distribution for $\theta_1 = -50^o$, $\theta_2 = 50^o$ $\phi_{1,2} = 0^o$, $t = 50$ $msec$, and $\mathbf{A}$ fixed at the maximum likelihood estimate. Here the signals are identical, $\mathbf{A} = [.5\ .5\ .5\ .5]$. Right: Posterior Distribution for $\theta_1 = -30^o$, $\theta_2 = 30^o$ $\phi_{1,2} = 0^o$, $t = 50$ $msec$, and $\mathbf{A}$ fixed at the maximum likelihood estimate. Here $\mathbf{A}_1 = [0\ .25\ 0\ .75]$ and $\mathbf{A}_2 = [.75\ 0\ .25\ 0]$

5 References

1. Konishi M. Listening with Two Ears. Sci. Am. 268:66-73 (1993).

2. Payne R.S. Acoustic Localization of Prey by Barn Owls. J. Exp. Biol. 54:535-573 (1971).

3. Peña J.L., M. Konishi, Auditory Spatial Receptive Fields Created by Multiplication. Science 292:249-252 (2001).

4. Takahashi T.T., C.H. Keller, Representation of Multiple Sound Sources in the Owl's Auditory Space Map. J. Neurosci. 14:4780-4793 (1994).