ESTIMATING AZIMUTH AND ELEVATION FROM INTERAURAL DIFFERENCES

Keith D. Martin

Perceptual Computing Section
MIT Media Lab, E15-401
Cambridge, MA 02139
kdm@media.mit.edu

ABSTRACT

Modeling of human auditory localization has largely been limited to lateralization, or left-to-right position. This paper describes an attempt to tackle the more complicated problem of position estimation with two degrees of freedom (azimuth and elevation). Differences in interaural intensity and arrival time are extracted from the acoustic signals at the left and right eardrums, and an estimate of position is formed which is optimal for certain classes of source signals. Examples of "spatial likelihood maps" generated by the model are given and the types of errors made by the model are quantified. It is suggested that such a model may work well in conjunction with a spectral cue model like the one suggested by Zakarauskas and Cynader (J. Acoust. Soc. Amer., Vol. 94, 1993, pp. 1323-1331).

1. INTRODUCTION

1.1. HRTFs and Eardrum Recordings

It is generally accepted that the cues used for localization are embodied in the free-field to eardrum, or head-related, transfer function (HRTF). The HRTF includes the high frequency shadowing due to the presence of the head and torso, as well as the directional-dependent spectral variations imparted by the diffraction of sound waves by the pinna.

For free-field sound sources more than a few feet away, the acoustic wave front reaching the head may be approximated by a plane wave. To the degree that this approximation is valid, interaural differences do not vary perceptibly with distance. Therefore, source distance is ignored in this paper, and HRTFs are assumed to be constant with respect to distance.

1.2. Interaural Differences, Localization and the Precedence Effect

Interaural differences are often cited as the most significant cues for lateralization (left-to-right position), and spectral cues based on features of the HRTF are given credit for humans' ability to perform vertical localization tasks [1]. Zakarauskas and Cynader describe a localization model based on spectral features [2]. Interaural differences are largely dismissed as cues for vertical localization, possibly because most studies of vertical localization are conducted with source locations on the median plane, where interaural differences are minimized, although Searle *et al.* have described a localization model based on interaural differences for sources on or near the median plane [3], and Lim and Duda have shown that interaural intensity differences are viable vertical localization cues for sources away from the median plane [4]. These models estimate azimuth and elevation independently rather than in combination. Also, they do not consider the "precedence effect," which is very significant for localization of sources in the presence of reflections [5].

1.3. Goal

In this paper, we will describe a system that addresses these points. The goal is to produce a model that calculates a set of interaural cues and infers a position on the two-dimensional surface of a sphere directly. Further, the model should be able to exhibit human-like robustness in ambient environments; thus, we include a mechanism corresponding to the precedence effect.

2. FORM OF THE MODEL

For purposes of this paper, we define the following notations for interaural differences. We shall refer to the *interaural intensity difference* (IID), which is the difference (in dB) between the signal levels at the two ears, the *interaural phase delay* (IPD), which is the time delay between the fine structure of the signals at the two ears, and the *interaural envelope delay* (IED), which is roughly equivalent to the difference in group delay of the signals at the two ears.

The model described in this paper can be broken into several layers, as shown in figure 2. In describing the model, the layers are grouped into two sections: (1) the model's "front end," which transforms the acoustic signals at the two eardrums into measures of interaural differences, and (2) a statistical estimator which determines the (θ,ϕ) position which is most likely to have given rise to the measured interaural differences.

The structure of the model is conceptually simple. At the input, the eardrum signals are passed through identical filter banks, which are intended to model the time-frequency analysis performed by the cochlea. Envelopes are estimated for each channel by squaring and smoothing the filter outputs. The envelope signals pass through onset detectors, which note the time and relative intensity of energy peaks in each signal. This information is used by the interaural

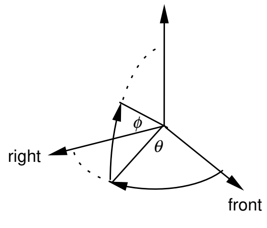

Figure 1: The coordinate system used in this paper: azimuth $(-180^\circ < \theta < 180^\circ)$ and elevation $(-90^\circ < \phi < 90^\circ)$.

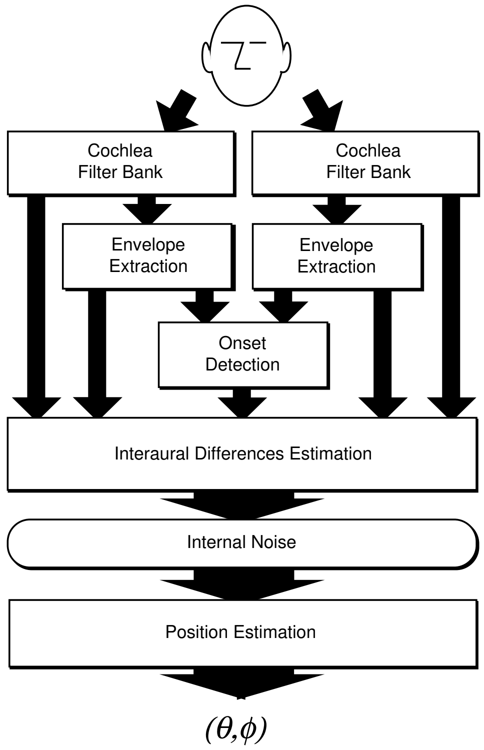

Figure 2: Block diagram of the proposed position estimator.

difference estimator to model the "precedence effect" [5]. The interaural difference estimators use the outputs of the filter bank, the envelope estimators, and the onset detectors to extract the IID, IPD, and IED at onsets in each frequency band. A "spatial likelihood" map is generated from the interaural differences, based on probability distributions for interaural differences derived from HRTF data. The global maximum of the likelihood map, which corresponds to the maximum likelihood position estimate, is interpreted as the "perceived sound source location."

2.1. The Front End

Eardrum Signals. The test signals used for this research were derived from HRTF measurements of a KEMAR dummy-head microphone [6]. The KEMAR data set is densely sampled in both azimuth and elevation and stored in the form of impulse responses. It is therefore straightforward to synthesize eardrum signals from arbitrary input signals by convolution.

Cochlea Filter Bank. At present, a 24 channel constant Q filter bank is employed to model the frequency analysis performed by the cochlea. All filters are fourth order IIR, with repeated conjugate-symmetric poles at the center frequency (CF) and zeros at D.C. and at the Nyquist frequency. The pole moduli are set such that the Q of each filter ($\mathrm{Q} = \mathrm{CF}/(-3\mathrm{dB\ BW})$) is 8.0, and each filter is scaled such that it has unity gain at its CF. The center frequencies are spaced evenly on a logarithmic scale, with three filters per octave, starting at a CF of 80 Hz in the first filter, and reaching a CF of approximately 18 kHz in the last filter.

The basic results of the model are not sensitive to the specific form of the filter bank, so no effort has been made to model the cochlea more closely.

Envelope Extraction. Amplitude-modulation envelopes are extracted from each filter channel by squaring and smoothing the outputs of the filter bank. The cutoff frequency for the smoothing filter is set to the CF of each channel, but limited to 800 Hz, which has been reported to be a good match to the envelope-following ability of the auditory periphery [1]. This limit allows reasonable interaural level comparisons to be made, while still retaining enough modulation energy for IED estimation. The envelopes are equivalent to a running estimate of rms energy in each channel.

Onset Detection. There are two major reasons for detecting onsets and evaluating interaural differences in small time windows around them. Pragmatically, in a normal listening environment, with multiple sources and reverberant energy, onsets (i.e. energy peaks) generally provide a small time window with a locally high signal-to-noise ratio for evaluating interaural differences. Psychologically, there is evidence that interaural differences are strongly weighted by human listeners at onsets. This has been demonstrated in experiments related to the precedence effect [7].

At present, an extremely simple method of onset detection has been implemented. In each frequency band, the two envelope signals are summed, and local maxima are identified. A simple suppression mechanism has been used to model the precedence effect, whereby selecting an onset causes envelope peaks occuring in the next 2–10 ms to be ignored. The selected onsets therefore occur no more frequently than once every 10 ms in a particular frequency channel. A small backward-masking effect has also been introduced in order to suppress small peaks followed immediately by a much larger peak.

Interaural Differences. In the current model, the IID is simply formed by the weighted log ratio of the envelope signals. The weighting is accomplished by a window function, which takes the form $w(t) = At\exp(-t/\tau)$. The effective "width" of the window used is 2–3 ms, as suggested by the results in [7], and the peak of the window function is centered at each detected onset. This weighting function is an important part of the precedence effect model, allowing suppressed onsets to have some small influence on the IID estimated by the model.

The IPD is estimated by a running cross-correlator similar to those proposed by Blauert [8] and Lindemann [9]:

$$\mathrm{IPD} = \operatorname*{argmax}_{\tau} \int_{-\infty}^{\infty} L_k(t-\frac{\tau}{2})R_k(t+\frac{\tau}{2})w(t)dt \qquad (1)$$

where $L_k(t)$ and $R_k(t)$ are the signals in the kth channel of the the left and right ear filter banks respectively, $w(t)$ is the window function previously described, and τ is limited to the range $-1 < \tau < 1$ ms, which includes the range of IPDs encountered in natural listening conditions. The IED estimation is identical to the IPD estimation except that the envelope signals are substituted for the filter bank signals in equation (1). In general, the window function $w(t)$ used for the IPD and IED estimators can be different from the one used for IID estimation, but the same window is used presently for simplicity.

Internal Noise. Currently, the output of the model is deterministic, but a model for internal noise in the system will clearly be required in order for model results to be meaningfully compared with human psychoacoustic data.

2.2. Position Estimation

Development of the ML estimator. If we make the restriction that the model has no *a priori* knowledge of the input spectrum, we might assume, for purposes of constructing a model, that the "average" spectrum is locally "white" (i.e. that the probability distribution of energy is constant over frequency). We can then argue that the interaural differences for a source at a given position will vary in a noise-like manner about some mean as changes in the input spectrum interact with the fine structure of the HRTFs within any given bandpass channel. We further make the doubtful, but nonetheless common, assumption that the noise is Gaussian and zero-mean in nature.

If we also assume that the noise in the various interaural differences is independent of the noise in other channels (an assumption that has been empirically verified for the front-end described in this paper and white-noise stimuli), then the "likelihood" that a source is located at a particular position is given by:

$$\mathcal{L}_{\theta,\phi} = K \exp\left[-\frac{1}{2}\sum_k \sum_i \frac{(P_{k,i} - \overline{P}_{k,\theta,\phi})^2}{\sigma_k^2}\right], \qquad (2)$$

where $P_{k,i}$ is the measured value of the ith interaural difference under consideration in channel k, $P_{k,\theta,\phi}$ is the the mean difference expected for a source located at (θ,ϕ), σ_k^2 is the variance of the interaural difference under consideration (the variance may be inferred from psychoacoustic data such as the JND data reported in [10]), and K is a normalizing constant.

It should be noted that σ_k^2 does not vary with θ and ϕ in the current model, though it might be allowed to vary with position in a more general model.

The independence assumption makes the likelihood function clearly separable into terms corresponding to the various interaural differences measured at the various onsets. Onsets arising from different sources in a multiple-source signal might therefore be separated, and "unnatural" interaural differences (such as extremely large IIDs) might be discounted, as suggested by Rakerd and Hartmann [11]. Further, onsets might be weighted by a measure of their energy, giving rise to a more intuitively satisfying model.

Template Extraction from HRTF Data. Reapplying the input signal assumptions that we used to derive the probabilistic model, we may calculate the expected mean values of the interaural differences in the various frequency bands at the measured positions. Some reasonable expressions for the interaural "templates" are given in equations (3), (4) and (5). These templates are based on a flat input spectrum in general, and a Dirac δ-function input signal (the most simple flat input) when the time structure is required specifically.

$$\mathrm{IID}_{k,\theta,\phi} = 10\log\left(\frac{\int_{-\infty}^{\infty} |([H_k R_{\theta,\phi}] * [H_k R_{\theta,\phi}])H_E|\, d\omega}{\int_{-\infty}^{\infty} |([H_k L_{\theta,\phi}] * [H_k L_{\theta,\phi}])H_E|\, d\omega}\right) \qquad (3)$$

$$\mathrm{IPD}_{k,\theta,\phi} = \underset{\tau}{\operatorname{argmax}}\ \mathcal{F}^{-1}(|H_k|^2 L_{\theta,\phi} R^*_{\theta,\phi}) \qquad (4)$$

$$\mathrm{IED}_{k,\theta,\phi} = \underset{\tau}{\operatorname{argmax}}\ \mathcal{F}^{-1}(|H_E|^2 \left[(H_k L_{\theta,\phi}) * (H_k L_{\theta,\phi})\right] \times \left[(H_k R_{\theta,\phi}) * (H_k R_{\theta,\phi})\right]^*) \qquad (5)$$

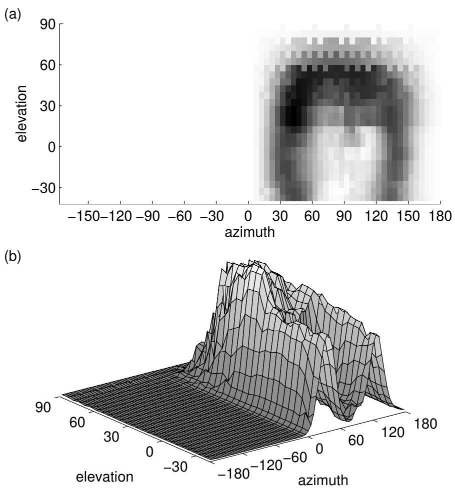

Figure 3: A sample "spatial likelihood map" generated by the current model. This example is for a white noise source located at $(\theta = 40°, \phi = 20°)$. (a) This map is essentially a Mercator projection of the sphere. The main feature is a circular band, corresponding to a cone of confusion. (b) Surface plot of the map in (a). The surface height increases with increasing likelihood.

In equations (3), (4) and (5), H_k refers to the transfer function of the kth bandpass filter, $L_{\theta,\phi}$ and $R_{\theta,\phi}$ refer to the left and right eardrum signals, H_E refers to the transfer function of the envelope smoothing filter, τ is the time parameter resulting from the inverse Fourier transform ($\mathcal{F}^{-1}$), and $*$ and * denote the convolution and complex-conjugation operators respectively.

Equation (3) represents the IID as the ratio of weighted average intensities in the specified channel. In equation (4), the IPD is given by the time corresponding to the maximum of the cross-correlation of the impulse responses at two ears in filter bank channel channel k for a sound source at position (θ,ϕ). Similarly, the IED is given by the time corresponding to the maximum of the cross-correlation of the impulse responses of the envelope filters in channel k. The input signal is assumed to be stationary over the duration of the measurement window, so $w(t)$ does not appear in the equations.

3. METHOD

The KEMAR HRTF data set is sampled somewhat regularly on the surface of a 1.4 m radius sphere around the dummy-head. Sample points are located on circles of equal elevation (in 10° elevation increments). On each circle, samples are spaced by 5° great-circle increments. To train the model, *head-related impulse responses* (HRIRs) were used from KEMAR's left ear, and reflected across the median plane, resulting in a symmetric set. With such a data set, interaural differences are uniformly zero on the median plane. 177 positions in the right hemisphere were selected as the training data, leaving 165 test positions interlaced between the training positions. At each training position, interaural difference templates were calculated as described in section 2.2. Interaural differences for training points on the left hemisphere were derived from the right hemisphere points by symmetry.

Interaural differences for other positions on the sphere were inter-

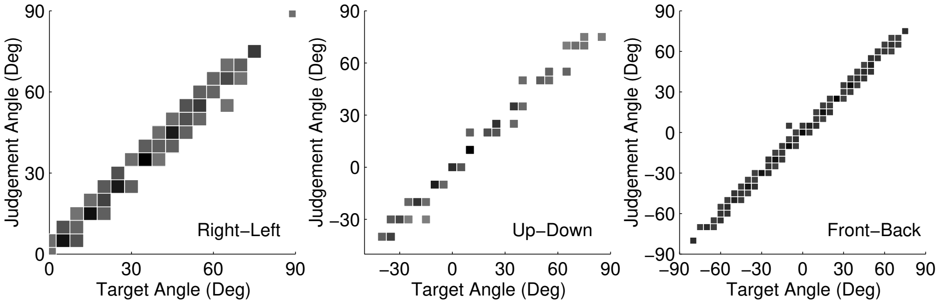

Figure 4: Apparent direction judgments for measured test positions. (a) Angle with the median plane (i.e. left-right judgment) (b) Angle with the horizontal plane (i.e. up-down judgment) (c) Angle with the vertical plane containing the interaural axis (i.e. front-back judgment).

polated by assuming that the surface is piece-wise planar. This is obviously not the case, but "spherical" interpolation with irregularly spaced points is a very difficult problem. Positions were sampled at 5° increments in both azimuth and elevation.

A test signal was generated for each of the test positions. A single 0.5 s burst of monophonic Gaussian white noise sampled at 44.1 kHz was used as a template, and the test signals were generated by convolution with the left and right ear HRIRs for each test position, resulting in signals that are equivalent to the KEMAR mannequin's eardrum signals for a source at each position.

Each test signal was passed through the model front-end, and the resulting output was divided into five 100 ms segments. Position estimates were formed by evaluating a spatial likelihood map for each subset of onsets and choosing the global maximum as the "perceived position."

4. RESULTS

Localization errors are quantified in figure 4, using a three-pole, head-centered spherical coordinate system. The model's performance is surprisingly good, rarely making a judgment error of more than 5° (many of the error judgments were in fact due to the 5° quantization of the spatial likelihood map). This excellent performance is probably due to the absence of internal noise in the model, which, when added, will smooth out the features of the likelihood map, making the peaks more broad and increasing the chance of error.

5. CONCLUSIONS

The model described in this paper is similar to the one described by Lim and Duda, but it has important differences. The current model estimates azimuth and elevation simultaneously, whereas the model of Lim and Duda estimates them separately. The biggest strength of the current model is that it "knows when it's doing a good job." The model is capable of rating its own certainty by examining the maximum value of the likelihood function and the "broadness" of the peak.

There are several necessary changes to the current model. Internal noise must be added so that localization errors made by the model more closely match those made by humans. Currently, the model can not distinguish two sources at different positions on the median plane. To rectify this problem, it is suggested that a spectral cue model similar to the one proposed in [2] be integrated with the current model.

6. ACKNOWLEDGMENTS

Financial support was provided by the National Science Foundation and the MIT Media Lab. Special thanks are extended to Dan Ellis and Eric Scheirer for their comments on a draft of this paper. Thanks also to Pat Zurek for his comments on the model itself.

References

1. Middlebrooks, J.C. and Green, D.M., "Sound Localization By Human Listeners," Annu. Rev. Psychol., Vol. 42, 1991, pp. 135–159.
2. Zakarauskas, P. and Cynader, M.S., "A computational theory of spectral cue localization," J. Acoust. Soc. Amer., Vol. 94, 1993, pp. 1323–1331.
3. Searle, C.L., Braida, L.D., Cuddy, D.R., and Davis, M.F., "Binaural pinna disparity: another auditory localization cue," J. Acoust. Soc. Amer., Vol. 57, 1975, pp. 448–455.
4. Lim, C. and Duda, R.O., "Estimating the Azimuth and Elevation of a Sound Source from the Output of a Cochlear Model," presented at the 28th Asilomar Conference on Signals, Systems, and Computers, 1994.
5. Zurek, P.M., "The Precedence Effect" in Yost, W.A. and Gourevitch, G., editors, *Directional Hearing*, Springer-Verlag, New York, 1987, pp. 85–105.
6. Gardner, W.G. and Martin, K.D., "HRTF measurements of a KEMAR," J. Acoust. Soc. Amer., Vol. 97 (6), 1995.
7. Zurek, P.M., "The precedence effect and its possible role in the avoidance of interaural ambiguities," J. Acoust. Soc. Am., Vol. 67, 1980, pp. 952–964.
8. Blauert J. and Cobben, W. "Some Consideration of Binaural Cross Correlation Analysis," *Acustica*, Vol. 39, 1978, pp. 96–104.
9. Lindemann, W., "Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals," J. Acoust. Soc. Am., Vol. 80, 1986, pp. 1608–1622.
10. Hershkowitz, R.M. and Durlach, N.I., "Interaural Time and Amplitude jnds for a 500-Hz Tone," J. Acoust. Soc. Am., Vol. 46, 1969, pp. 1464–1467.
11. Rakerd, B. and Hartmann, W.M., "Localization of sound in rooms, II: The effects of a single reflecting surface," J. Acoust. Soc. Am., 78, 1985, pp. 524–533.