Estimating the Azimuth and Elevation of a Sound Source from the Output of a Cochlear Model

Chuck Lim

Advanced Linear Devices
415 Tasman Drive
Sunnyvale, CA 94089

Richard O. Duda

Dept. of Electrical Engineering
San Jose State University
San Jose, CA 95192

Abstract

Interaural intensity differences (IID's) and interaural time differences (ITD's) give important information for estimating the elevation as well as the azimuth of a sound source. We extract this information from a filter-bank model of the cochlea using straightforward short-time autocorrelation and crosscorrelation operations. With the appropriate coordinate system, the resulting ITD varies primarily with azimuth, while the IID varies with azimuth and elevation. For a single sound source in an anechoic environment, a maximum-likelihood estimation procedure is shown capable of recovering azimuth to within 1 degree and elevation to within 16 degrees.

1 Introduction

This paper concerns the estimation of both the azimuth and the elevation of a sound source in an echo-free environment. We present the results of computer simulations showing that it is possible to localize a broad-band sound source with an accuracy comparable to that of humans from the information available on the left and right auditory nerves of human listeners. Our basic approach employs (a) experimentally measured transfer functions that provide the signals at the left and right ear drums, (b) a model of the cochlea that converts these signals into average neural firing rates, (c) crosscorrelation and energy measurements that extract the interaural time and intensity differences, and (d) a nearest-neighbor procedure that yields the azimuth and elevation estimates. We begin by describing the motivation for this work. We then describe the cochlear model, the localization system, and the experimental results.

2 Background

It is well known that humans use the differences in the acoustic signals reaching the two ears to localize sound sources. Many years ago, Lord Rayleigh determined how acoustic waves were diffracted by the head, and identified the interaural time difference (ITD) and the interaural intensity difference (IID) as the two major cues for determining azimuth [9]. Rayleigh's *duplex theory* held that the ITD is used at low frequencies, where phase shift is unambiguous, and the IID is used at high frequencies, where diffraction or "head shadow" produces large amplitude differences. Rayleigh was less certain about how one could determine the elevation of the source, or how one could distinguish front from rear, which are more complicated problems. However, since his pioneering work, other researchers have identified a number of monaural and binaural cues for azimuth, elevation and range [1, 10].

The physical basis for many of these cues comes from the orientation-dependent way in which sound waves are diffracted by the torso, shoulders, head, and outer ears or pinnae. These effects are captured in the so-called head-related transfer function (HRTF) that relates the spectrum of the sound source to the spectrum of the signals reaching the ear drums. To be specific, if $X(\omega)$ is the Fourier transform of the source signal, then

$$X_L(\omega) = H_L(\omega)X(\omega) \quad (1)$$

and

$$X_R(\omega) = H_R(\omega)X(\omega) \quad , \quad (2)$$

where X_L and X_R are the Fourier transforms of the left-ear and right-ear signals, and H_L and H_R are the left-ear and right-ear HRTF's, respectively.[1]

[1] In defining the HRTF, it is conventional to assume that the acoustic environment is anechoic. The effects of echoes and

1058-6393/95 $4.00

In addition to varying with frequency, both H_L and H_R depend on the position of the source relative to the head. In our work, we locate the source by its azimuth θ, elevation ϕ, and range r in a head-centered, interaural-polar, spherical coordinate system (see Fig. 1). Thus, the HRTF's can be thought of as functions of four variables — ω, θ, ϕ, and r. For simplicity, we shall ignore the range dependence in this paper, and reduce the number of variables to three.

The ratio $H(\omega,\theta,\phi) = H_R(\omega,\theta,\phi)/H_L(\omega,\theta,\phi)$, which is called the interaural transfer function or ITF, captures the binaural "difference" cues. To be specific, the derivative of the phase of H gives the spectral ITD, and the amplitude of H (measured in dB) gives the spectral IID. When only one broad-band sound source is active and the environment is anechoic, one can estimate the ITF merely by forming the ratio of the Fourier transforms of the right-ear and left-ear signals, which is independent of the spectrum of the source:

$$\frac{X_R(\omega)}{X_L(\omega)} = \frac{H_R X}{H_L X} = \frac{H_R}{H_L} = H_m(\omega) \quad . \tag{3}$$

Here we have used the subscript m to indicate that in practice this calculation yields a *measured* value of H for some unknown azimuth and elevation, and will be subject to the inevitable measurement errors.

If the exact ITF is known, this suggests a simple way to estimate the azimuth and elevation from signals reaching the two ears: compare the measured value $H_m(\omega)$ to the known ITF $H(\omega,\theta,\phi)$ and find values $\hat{\theta}$ and $\hat{\phi}$ for which $H(\omega,\hat{\theta},\hat{\phi})$ best approximates $H_m(\omega)$.

In prior work, the second author employed a special case of this procedure, using FFT's to measure the signal spectra, and matching only the amplitudes of the ITF's [4]. Away from the median plane, the resulting localization performance was remarkably good, with angular accuracies close to measured human abilities. However, in the important high-frequency region, much greater frequency resolution was available than people actually possess [2]. Thus, although we were able to show that the ITF captured enough information to allow accurate localization in elevation as well as azimuth, we might have been exploiting spectral information that lies within the pass bands of critical-band filters, and thus not available to human listeners.

In this paper, we resolve this question by extracting the ITD and IID information from a model of the cochlea. While this approach has the disadvantage that the resulting estimates are no longer guaranteed to be independent of the source spectrum, it demonstrates that spectral fine structure is not necessary for localization accuracy.

room reverberation can be very complex, but can often be approximated by introducing additional, "mirror image" sources. We ignore this important complication in this paper.

3 The Response of the Cochlear Model

The cochlear model that we used was developed by Lyon [5, 6, 7], and implemented in C by Slaney [11]. It consists of a set of roughly constant-Q bandpass filters that simulate the basilar membrane, half-wave rectifiers that simulate the inner hair cells, and a four-stage, laterally-interacting automatic gain control (AGC) system that accounts for the compressive nonlinearities in the cochlea (see Fig. 2). The output of the cochlear model is a set of N signals that represent the average neural firing rates at N points along the basilar membrane. (In our work, we sampled at $N = 45$ points corresponding to center frequencies from about 4.2 kHz to 18.5 kHz.)

In our experiments, the source signal was a unit impulse. To simulate the effects of the head and the outer ears or pinnae, we used experimentally measured HRTF's for the KEMAR manikin [3]. Thus, the time-domain inputs to the cochlear model were the impulse responses $h_L(t,\theta,\phi)$ for the left ear and $h_r(r,\theta,\phi)$ for the right ear, sampled at a 44.1-kHz rate.

To compute the ITD, we crosscorrelated corresponding left-ear and right-ear channel outputs and found the time shift for maximum crosscorrelation. To a first approximation, the resulting 45 time shifts turned out to be essentially functions of azimuth only (see Fig. 3a). As expected, the ITD increased systematically (and roughly sinusoidally) with azimuth. Since the frequency variation of the ITD was small, we decided to average the time lags to obtain a single ITD measurement.

To compute the IID, we disabled the AGC system to prevent it from reducing the interaural differences. We used the logarithm of the zero-lag autocorrelation of each channel to approximate the amplitude spectrum, using channel-by-channel differences to obtain a measure of the IID spectrum in decibels. The resulting 45 IID's varied with everything — frequency, azimuth and elevation. Fig. 3b shows that the IID spectrum for $\phi = 0°$ increases systematically with azimuth. Because the curves start at 2.8 kHz, the characteristic rising response due to "head-shadow" is not evident. However, the curves display a prominent peak around 8 kHz and a subsequent dip in the general vicinity of 12 kHz. These results are consistent with results we had obtained earlier with FFT analysis [4].

More important, when the azimuth was held constant, the resulting function IID(ω, ϕ) provided a spectral profile that seemed to be characteristic of the elevation angle, ϕ. (The elevation dependence for $\theta = 30°$ is illustrated in Fig. 4). This suggests using the single ITD number to determine azimuth, and the IID profile to determine elevation.

4 The Localization System

The ITD value and the IID profile were used to form the components of a 46-dimensional vector $\mathbf{x}(\theta, \phi)$. Because the ITD was measured in seconds and the 45 IID values were in dB, and because we wanted to use simple Euclidean distance to compare vectors, we multiplied the ITD value by a scale-conversion factor $w = 45f$. We found empirically that the value $w = 25$ dB/sample (f=24.5 dB/ms) yielded good results.

Our HRTF data provided impulse responses for 144 points essentially uniformly sampled over the right hemisphere.[2] We traced a spiral trajectory through these points, and used every other point to define a set of 72 reference vectors $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_{72}\}$ for which the azimuth and elevation were known. To provide better angular resolution, we added 71 additional reference vectors at intermediate sample points by averaging successive vectors. This produced an expanded set of 143 vectors $X' = \{\mathbf{x}_1, 0.5(\mathbf{x}_1 + \mathbf{x}_2), \mathbf{x}_2, \cdots, 0.5(\mathbf{x}_{71} + \mathbf{x}_{72}), \mathbf{x}_{72}\}$ having known azimuths and elevations. In neural network terms, these 143 vectors represented the training data. The localization system, diagrammed in Fig. 5, used these reference vectors to estimate the azimuth and elevation for an unknown input.

When a sound source is located at a position (θ, ϕ), not included in X', the resulting vector $\mathbf{x}$ will be different from the vectors in X', partly because of the spatial variation of the response and partly because of inevitable random noise. With standard assumptions about the noise being additive and independent and the training set being arbitrarily large, it is easy to show that a maximum likelihood estimate $(\hat{\theta}, \hat{\phi})$ for the azimuth and elevation for $\mathbf{x}$ can be obtained by finding the vector in X' that is nearest to $\mathbf{x}$. However, it is not obvious that such a nearest-neighbor approach is optimal with a small training set, or that information extracted by the cochlear model, even if used optimally, can yield accurate estimates.

[2] There were an additional 69 points on the median plane, but since it is essentially impossible to localize in the median-plane using interaural differences, these points were excluded.

5 Experimental Results

The nearest-neighbor localization system was tested with the 72 samples not used to form the training data. The results were remarkably good — an average absolute error of 0.8° in azimuth and 16° in elevation. As the following table shows, accuracy in elevation estimation was substantially degraded when only ITD or IID information was used.

Conditions	Azimuth error	Elevation error
ITD only	0.6°	73°
IID only	2.0°	21°
ITD and IID	0.8°	16°

These results show that very good localization in *both* azimuth and elevation can be extracted by simple autocorrelation and crosscorrelation of the outputs of a cochlear model.

It must be acknowledged that these results were obtained under the rather ideal conditions of a single impulse source in an anechoic environment. With a narrow-band source, one would have to discount or omit channels for which the signal levels at both ears were low, and with fewer effective channels there could be a serious drop in performance. With multiple sources, one would have to introduce new mechanisms to prevent the system from localizing on a meaningless "center of gravity" of the sources. Echoes and reverberation can be viewed as a particularly troublesome case of strongly correlated multiple sources. However, since the localization estimates are obtained in a few milliseconds, there is hope that an onset-triggered analysis could cope with all of these problems. In any case, our results show that excellent localization can be achieved using the kind of information that is extracted by the human cochlea.

Acknowledgements

This work was supported by the National Science Foundation under NSF Grant No. IRI-9214233, and is based on the M.S. project by the first author [8]. We also appreciate the assistance and encouragement we have received from Mr. Richard F. Lyon at Apple Computer, Inc., and Dr. Malcolm Slaney at Interval Research Corp.

References

[1] Blauert, J., *Spatial Hearing* (MIT Press, Cambridge, MA, 1983).

[2] Carlile, S., and D. Pralong, "The location-dependent nature of perceptually salient features of the human head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 95, pp. 3445-3459 (June, 1994).

[3] Duda, R. O.,"Short-time measurement of the KEMAR head-related transfer function," a report submitted to Richard F. Lyon, Apple Computer, Inc. (August 1991).

[4] Duda, R. O., "Elevation dependence of the interaural transfer function," in T. R. Anderson and R. H. Gilkey, Eds., *Binaural and Spatial Hearing* (Lawrence Erlbaum Associates, Hillsdale, NJ, in press). (This paper was originally presented at the *Conference on Binaural and Spatial Hearing* (Dayton, OH), Sept. 9-12, 1993).

[5] Lyon, R. F., "A computational model of filtering, detection, and compression in the cochlea," *Proc. ICASSP '82* (Paris), pp. 1282-1285 (1982).

[6] Lyon, R. F., "A computational model of binaural localization and separation," *Proc. ICASSP '83* (Boston, MA), pp. 1148-1151 (1983).

[7] Lyon, R. F., and C. Mead, "An analog electronic cochlea," *IEEE Trans. ASSP*, vol. 36, pp. 1119-1134 (1988).

[8] Lim, C., "A Sound Localization System Using Correlograms," Technical Report No. 9, NSF Grant No. IRI-9214233, Department of Electrical Engineering, San Jose State University, San Jose, CA (September, 1994).

[9] Lord Rayleigh (J. W. Stutt), "On our perception of sound direction," *Phil. Mag.*, vol. 13, pp. 214-232 (1907).

[10] Middlebrooks, J. C., and D. M. Green, "Sound localization by human listeners," *Annu. Rev. Psychol.*, vol. 42, pp. 135-159 (1991).

[11] M. Slaney, "Lyon's cochlear model," Apple Technical Report No. 13, Advanced Technology Group, Apple Computer Inc., Cupertino, CA (1988).

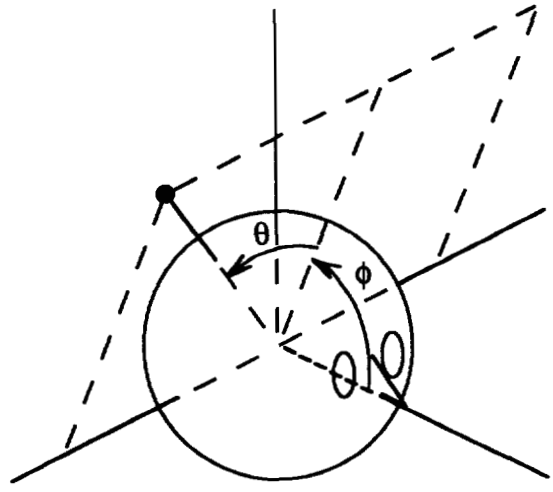

Fig. 1 The interaural-polar coordinate system

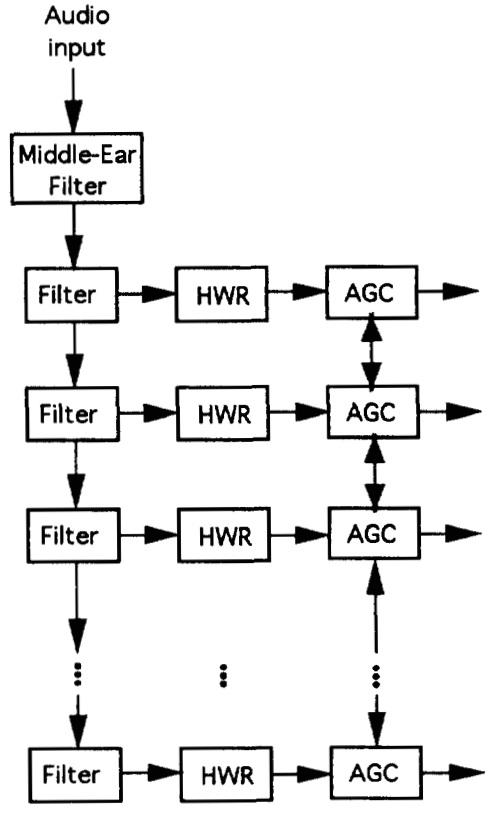

Fig. 2 Lyon's cochlear model

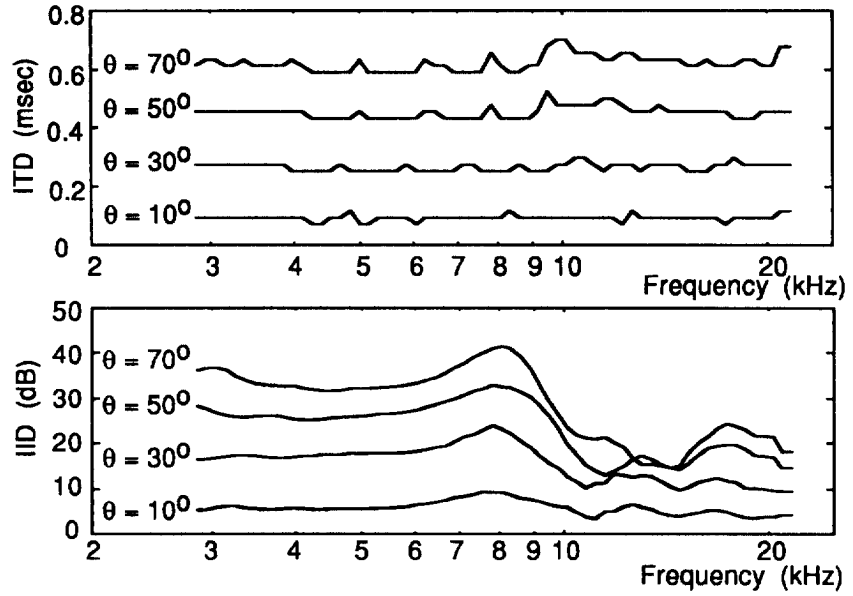

Fig. 3 Azimuth dependence of the ITD and IID spectrum, elevation = 0^o

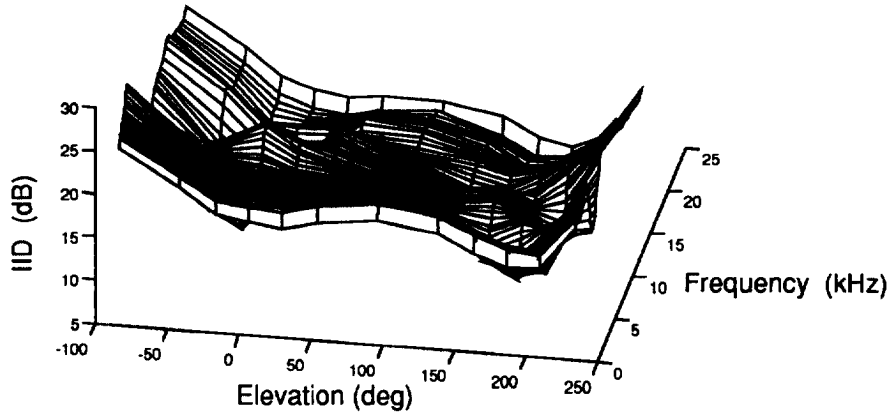

Fig. 4 Elevation dependence of the IID spectrum, azimuth = 30^o

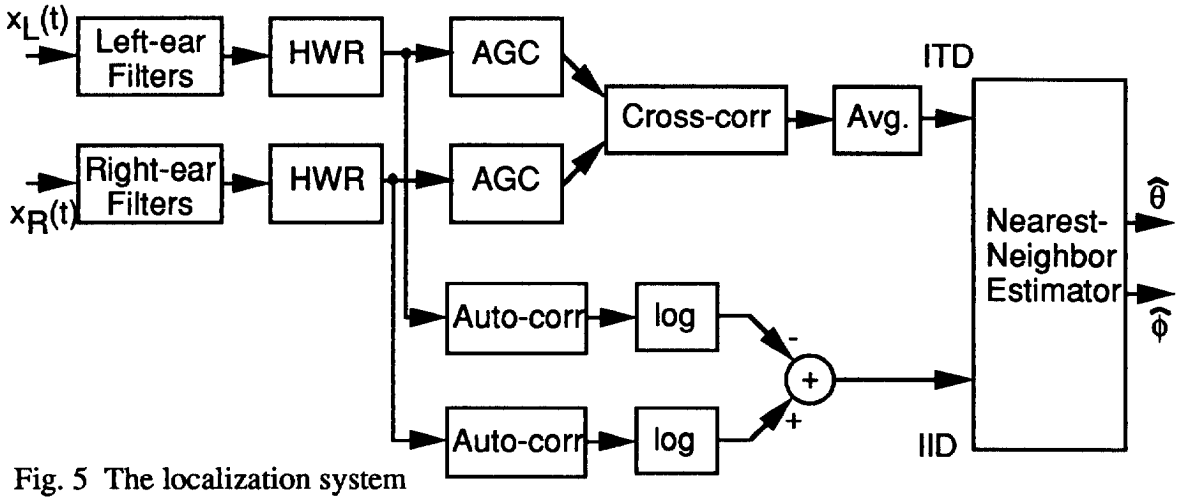

Fig. 5 The localization system