Psychological Review
1996, Vol. 103, No. 1, 137–142

0033-295X/96/$3.00

An Auditory Illusion Predicted From a Weighted Cross-Correlation Model of Binaural Interaction

Kourosh Saberi
University of Florida

In humans, the lateral movement of an acoustic source produces dynamic changes in the relative sound-pressure level and time of arrival of the acoustic wave at the 2 ears. The dynamic nature of these cues is assumed to play an important role in the perception of lateral motion. A phenomenon of auditory motion is reported whose lateral direction and relative velocity may be specified while interaural differences are kept constant. The stimulus producing this percept is a narrowband waveform whose instantaneous bandwidth is a cosine function of time. This phenomenon is predicted from a model of cross-correlation that estimates the running position of an image from a weighted combination of 2 variables: (a) magnitude of interaural delay, with smaller delays receiving more weight, and (b) consistency of interaural information across frequency.

The movement of an acoustic source along an azimuthal trajectory produces dynamic changes in interaural differences of time (IDT) and level (IDL). When a sound source is in front of the observer, the acoustic waveform reaches both ears simultaneously and has equal sound-pressure level at the two ears. As the source begins its lateral movement, interaural differences progressively increase and favor the ear that is closer to the sound source. Changes in interaural differences are considered important in the perception of lateral auditory motion, illusory (Burtt, 1917; Grantham, 1986; Strybel, Manligas, Chan, & Perrott, 1990; Strybel, Witty, & Perrott, 1992) or real (Perrott & Musicant, 1977; Perrott & Tucker, 1988). This report describes a phenomenon of auditory motion whose lateral direction and velocity may be specified while interaural parameters are kept constant for the duration of the stimulus. The article begins by describing the stimulus presented to each ear; it then describes the features of the perceived motion and discusses a likely physiological mechanism and a theoretical model consistent with these observations.

Observations

The sound presented to each ear was a complex acoustic stimulus. It consisted of a narrowband waveform whose bandwidth was sinusoidally modulated. This waveform was presented to both ears, except that the presentation of the signal to one ear was delayed by 1,500 μs. The delay was constant throughout the presentation of the stimulus, and the interaural level difference was zero. The rise–fall envelopes were not interaurally delayed. The details of waveform generation are provided in the Appendix, but as an example, the stimulus consisted of a 5-tone complex whose components were linearly spaced from 300 to 700 Hz (Δf = 100 Hz). Each component was sinusoidally modulated at f_m = 2 Hz, and the stimulus duration was 500 ms.[1] The depth of modulation was frequency dependent and exponentially increased with deviation from the center component (500 Hz). Figure 1, which interpolates across frequency, shows how the instantaneous bandwidth of such a waveform was a function of time.

When presented with this stimulus, observers reported the movement of an intracranial auditory image from the side of the leading ear toward the side of the lagging ear. The velocity of movement depended on f_m, and the movement trajectory was somewhat similar to that of binaural beats (Licklider, Webster, & Hedlun, 1950; von Bekesy, 1960; Perrott & Nelson, 1969) except that it moved in both left and right directions depending on whether the instantaneous bandwidth was increasing or decreasing. Binaural beats, however, are unidirectional. For small f_m (less than 2 Hz), the bandwidth modulation rate was sufficiently slow to produce the percept of a moving image; at higher rates a spatially stationary but diffused image was perceived whose loudness fluctuated at the rate f_m. When naive listeners were asked to describe the percept of the auditory image in terms of its spatial properties, they usually volunteered the report of lateral motion (the experimenter made no reference to this). Another important observation is that the motion of the image was continuous. This may be illustrated by increasing the stimulus duration (e.g., 5 s) and decreasing the magnitude of f_m, for example, to 0.1 Hz. In this case it takes the auditory image longer (5 s) to move from one side of the head to the other, and observers readily trace its trajectory. To quantify these reports, 3 observers completed a single-interval forced-choice task in which they were instructed to determine if they heard a moving or a stationary image. On each trial, one of two

This research was supported by the National Institutes of Health and the Air Force Office of Scientific Research. I thank Richard M. Stern, David M. Green, David R. Perrott, and Thomas Z. Strybel for helpful comments.

Correspondence concerning this article should be addressed to Kourosh Saberi, who is now at 36-767, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

[1] Other typical waveforms that produced a sense of motion were 3-tone complexes of 300, 500, 700; 500, 700, 900 Hz; or 300, 400, 500 Hz.

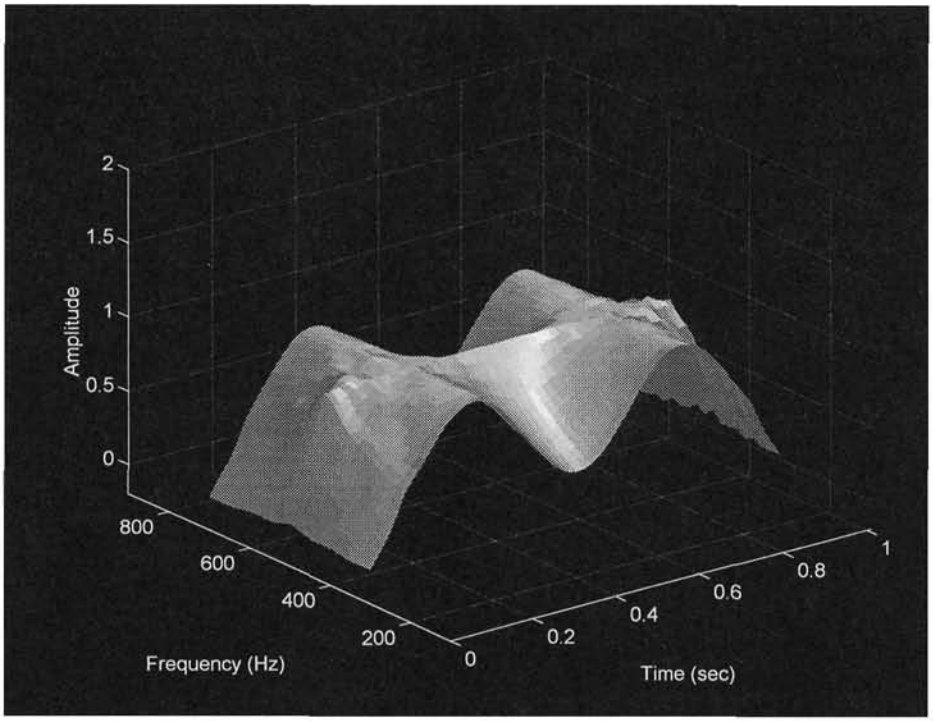

Figure 1. Envelope, $E(t)$, of waveform used in the current study, obtained from the Hilbert Transform, $H[X(t)]$, of Equation A1 (see Appendix); $E(t) = [H[X(t)]^2 + X(t)^2]^{0.5}$. The surface is interpolated across frequencies to show bandwidth modulation. The stimulus is unmodulated at the center frequency of 500 Hz and is increasingly modulated at a depth that is exponentially related to deviation from center frequency (see Equation A2 in Appendix). This produces a waveform whose instantaneous bandwidth is sinusoidally modulated. The envelope of each tone of the 5-tone complex described in text would, for example, trace the envelope changes of this figure at exactly the five frequencies of the complex. This figure was generated with a 21-tone complex at 20-Hz spacing to allow surface interpolation ($\lambda = 20$, $\beta = 400$, $f_c = 500$ Hz, $k = 0.07$, in Equations A1 and A2).

waveforms was presented with equal prior probabilities. One waveform was the one described above (also see the Appendix), and the other was again the same stimulus but without an interaural delay. No feedback was provided. Each observer received 100 presentations of the stimulus. Table 1 shows the results of this experiment. These data support the reports of perceived movement.

Weighted Cross-Correlation

In 1948, Jeffress proposed a theoretic model of binaural localization that translates the acoustic temporal code (i.e., IDT) into a physiologically spatial code. In his model, he used two-dimensional arrays of tonotopically organized coincidence-detector neurons that perform the mathematical equivalent of cross-correlating the temporal firing patterns that result from the acoustic inputs to the left and right ears. Recent physiological studies have associated the medial superior olive with this mechanism (Carr & Konishi, 1990; Yin & Chan, 1990). Several mathematical models that incorporate findings from psychoacoustics and physiology also have been developed.

Table 1
Detection of Motion for IDT = 1,500 μs

Observer	Hits	False alarms
1	99	1
2	98	2
3	100	0

Note. IDT = interaural differences of time.

One recent model is the weighted-image model of Stern, Zeiberg, and Trahiotis (1988), again driven by insights from Jeffress (1972; p. 358), which postulates combining two weighted cues in determining the lateral position of an auditory image. One cue is referred to as *centrality.* Psychophysical data show that smaller interaural delays in narrowband periodic waveforms are more accurately detected than larger delays (Klumpp & Eady, 1956; Stern et al., 1988; Stern & Trahiotis, 1991; Yost, 1974). This is usually considered as evidence that a larger population of neurons is associated with smaller interaural delays; these neurons code the location of sources near the front of the observer (i.e., near the mid-sagittal plane). Smaller delays are usually plotted at the center of the cross-correlation plane, hence the term *centrality.*

The second cue represents consistency of interaural information across *frequency.* This cue, which represents the true interaural delay in the waveform, has been referred to as *straightness* by Stern et al. (1988) in reference to the straightness of the trajectory of the cross-correlation peaks across frequency (see Figure 2). One may use any of several measures to quantify consistency of interaural information across frequency. Stern et al. (1988) used the frequency variance about each trajectory; the straighter trajectories have a smaller variance. Others (Shackleton, Meddis, & Hewitt, 1992) have used a simple summation scheme across frequency. Although the precise measure is not critical to the results reported (both measures produce similar results), a modified method of summation is preferred here because it is physiologically realistic, simple, and allows use of information from the entire cross-correlation plane (not just the peaks).

Figure 2 shows the output of an auditory model that consists of peripheral bandpass filtering, a running cross-correlation mechanism, and an estimator of running position based on a weighted combination of centrality and consistency across frequency. The plot shows two instants in time: one when the stimulus is near its maximum instantaneous bandwidth (t = 0.255 s; top panels), and the other near minimum bandwidth (t = 0.495 s; bottom panels). The left panels show the results of cross-correlation without the centrality weighting, and the right panels show this output after centrality weighting. Note that the ridge near a delay of −1.5 ms (true delay) is consistent across frequency. Thus, when the activity within this plane is integrated across frequency, it will produce a larger peak at $\tau = -1.5$ ms, compared with a ridge at the same position that is curved (not shown). This consistency across frequency is, of course, more evident when the bandwidth is wider (top panels). On the other hand, it is also clear that the curved ridge positioned near τ = 0–1 ms is more heavily weighted because it is associated with smaller delays (top right). At Time 0.495 s, the bandwidth of the waveform is near its minimum (lower panels); therefore, the weight given to consistency is very small, and the centrality cue dominates (i.e., it is difficult to determine which peak may be associated with consistent interaural information across frequency, because the bandwidth is so narrow). During the course of stimulus presentation, the activity within this putative cross-correlation mechanism oscillates between the two extremes. A weighted combination of centrality and consistency determines the instantaneous position of the image. This model and its output are described in the Appendix. Figure A2 shows the running-position estimate from the model for the stimulus used to collect the data of Table 1.

Although continuous motion suggests that at least one image generated from the two cues is their weighted combination, it is also important to consider circumstances that produce multiple

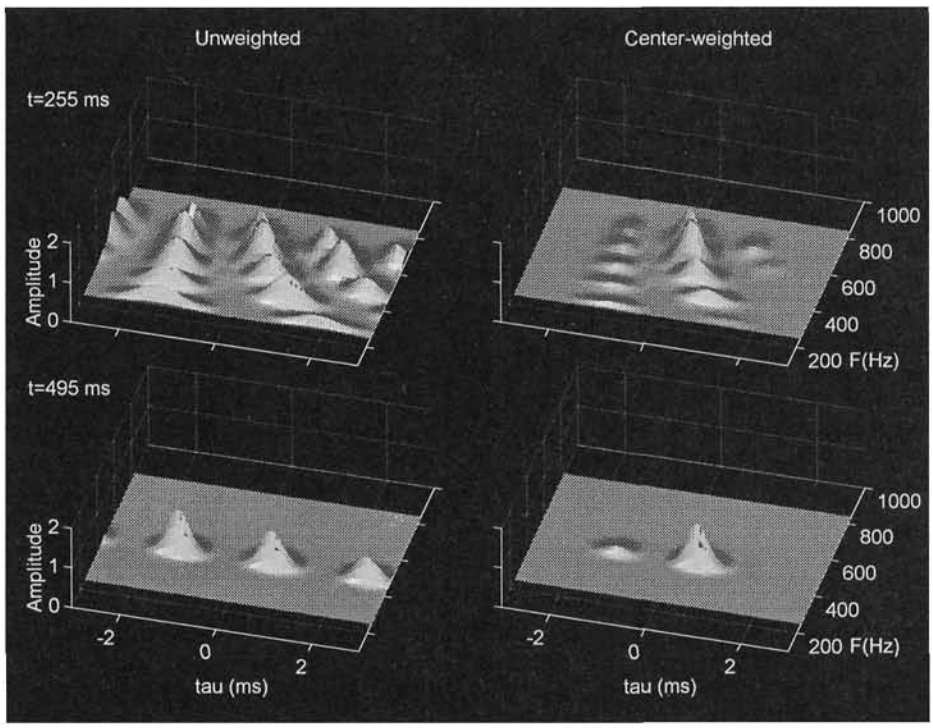

Figure 2. Output of the running cross-correlation described in the text and Appendix for the 5-tone complex used in collecting the data from human observers (see Table 1). The top panels show this output at t = 0.255 s when the waveform is near its maximum instantaneous bandwidth, and the lower panels show this output at t = 0.495 when it is near its minimum instantaneous bandwidth. Left panels are this output without the centrality weighting; consistency of interaural information across frequency is observed for the true delay of −1.5 ms, where the negative sign represents a waveform-lead to the left ear. The right panels show the same output as the left panels, with centrality weighting $p(\tau)$ in Equation A4 (see Appendix).

or diffused images. Although modeling the many factors that contribute to the perception of multiple images is beyond the message of this article, one interesting example observed during these experiments should be noted. The 3- and 5-tone complexes always produced motion. However, when observers listened to 21-component stimuli (at 20-Hz spacing between components, this produced a 400-Hz wide signal), there was a clear change in sound quality from a noiselike pitch to a tonal one during the 0.5-s duration of the stimulus. In this case, one image (noisy) seemed to disappear as another (tonal) emerged, and motion was rarely perceived. In such cases, it may be preferable to consider a measure of position estimate related to individual trajectories (e.g., the largest peak after integration across frequency) instead of the centroid about the τ axis.

Finally, some other interesting implications of these observations also are worth noting. First, the use of a 5-tone or 3-tone complex with a wide component spacing (100–200 Hz) suggests that consistency of interaural information may apply to stimuli that are spectrally discontinuous (e.g., stimuli whose components fall within different critical bands). Second, the hypothesized weighting mechanism seems to have a fairly short time constant; perceived movement would mean a running combination of weighted centrality and consistency information that must be updated at a fairly rapid pace.

References

Bekesy, G. von (1960). *Experiments on hearing.* New York: McGraw-Hill.

Blauert, J., & Cobben, W. (1978). Some consideration of binaural cross-correlation analysis. *Acustica, 39,* 96–103.

Burtt, H. E. (1917). Auditory illusions of movement—A preliminary study. *Journal of Experimental Psychology, 2,* 63–75.

Carr, C. E., & Konishi, M. (1990). A circuit for detection of interaural time differences in the brain stem of the barn owl. *Journal of Neuroscience, 10,* 3227–3246.

Grantham, D. W. (1986). Detection and discrimination of simulated motion of auditory targets in the horizontal plane. *Journal of the Acoustical Society of America, 79,* 1939–1949.

Holdsworth, J., Nimmo-Smith, I., Patterson, R., & Rice, P. (1988). Implementing a gammatone filter bank. *Annex C or the SVOS Final Report* (Part A: The auditory filter bank).

Jeffress, L. A. (1948). A place mechanism of sound localization. *Journal of Comparative Physiological Psychology, 41,* 35–39.

Jeffress, L. A. (1972). Binaural signal detection: Vector theory. In J. V. Tobias (Ed.), *Foundations of modern auditory theory* (Vol. 2, pp. 349–368). New York: Academic Press.

Klumpp, R. G., & Eady, H. R. (1956). Some measurements of interaural time difference thresholds. *Journal of the Acoustical Society of America, 28,* 859–860.

Licklider, J. C. R., Webster, J. C., & Hedlun, J. M. (1950). On the frequency limits of binaural beats. *Journal of the Acoustical Society of America, 22,* 468–473.

Perrott, D. R., & Musicant, A. D. (1977). Minimum audible movement angle: Binaural localization of moving sound sources. *Journal of the Acoustical Society of America, 62,* 1463–1466.

Perrott, D. R., & Nelson, M. A. (1969). Limits for the detection of binaural beats. *Journal of the Acoustical Society of America, 46,* 1477–1481.

Perrott, D. R., & Tucker, J. (1988). Minimum audible movement angles as a function of signal frequency and the velocity of the source. *Journal of the Acoustical Society of America, 83,* 1522–1527.

Raatgever, J. (1980). *On the binaural processing of stimuli with different interaural phase relations.* Unpublished doctoral dissertation, Technische Hogeschool Delft, Delft, The Netherlands.

Sayers, B. M., & Cherry, E. C. (1957). Mechanism of binaural fusion in the hearing of speech. *Journal of the Acoustical Society of America, 29,* 973–987.

Shackleton, T. M., Meddis, R., & Hewitt, M. J. (1992). Across frequency integration in a model of lateralization. *Journal of the Acoustical Society of America, 91,* 2276–2279.

Sheer, G. D. (1987). *Modeling the dependence of auditory lateralization on frequency and bandwidth.* Unpublished master's thesis, Carnegie Mellon University.

Stern, R. M., & Trahiotis, C. (1991). Consistency in interaural timing over frequency. In Y. Cazals, L. Demany, & K. Horner (Eds.), *Advances in the biosciences: Auditory physiology and perception* (pp. 547–554). New York: Pergamon.

Stern, R. M., Zeiberg, A. S., & Trahiotis, C. (1988). Lateralization of complex binaural stimuli: A weighted-image model. *Journal of the Acoustical Society of America, 84,* 156–165.

Strybel, T. Z., Manligas, C. L., Chan, O., & Perrott, D. R. (1990). A comparison of the effects of spatial separation on apparent motion in the auditory and visual modalities. *Perception & Psychophysics, 47,* 439–448.

Strybel, T. Z., Witty, A. M., & Perrott, D. R. (1992). Auditory apparent motion in the free field: The effects of stimulus duration and separation. *Perception & Psychophysics, 52,* 139–143.

Yin, T. C. T., & Chan, C. K. (1990). Interaural time sensitivity in medial superior olive of cat. *Journal of Neurophysiology, 64,* 465–487.

Yost, W. A. (1974). Discrimination of interaural phase differences. *Journal of the Acoustical Society of America, 55,* 1299–1303.

Appendix

Stimuli and Cross-Correlation Model

The stimulus was a narrowband waveform, $X(t)$, whose effective instantaneous bandwidth modulated at the rate f_m,

$$X_{l,r}(t) = \sum_{i=1}^{\beta/\lambda+1} \frac{[1 + m(f_i)cos(2\pi f_m t)]sin(2\pi f_i t + \phi_i)}{1 + m(f_i)};$$

$$f_i = f_c - \frac{\beta}{2} + \lambda(i-1) \quad [A1]$$

where f_c is the frequency of the center component, f_i is the frequency of the ith component, f_m is the amplitude–modulation rate, β is bandwidth, λ is component spacing (i.e., if $f_c = 500$, $\lambda = 100$, $\beta = 400$; then $f = [300, 400, \ldots, 700]$), and

$$m(f_i) = 1 - exp(-k|f_c - f_i|) \quad [A2]$$

is the frequency-dependent modulation depth. The parameter k determines the rate at which m increases with deviation from center frequency.

The waveform was delayed to one ear using the linear phase shift $\phi_i = 2\pi f_i$ IDT for the delayed signal, and $\phi_i = 0$ otherwise. Typical parameter values that worked well were $f_c = 500$, $f_m = (0.1, \ldots, 3$ Hz), $\beta = 400$, $\lambda = 100$ or 200 Hz, IDT = 1,500 μs, $k = (0.01, \ldots, 6.0)$ and signal durations of 0.1 to 5 s. The spectrally dynamic nature of this waveform is shown in Figure 1.

The waveform generated by Equation A1 was passed through the auditory model shown in Figure A1. The GammaTone filter bank

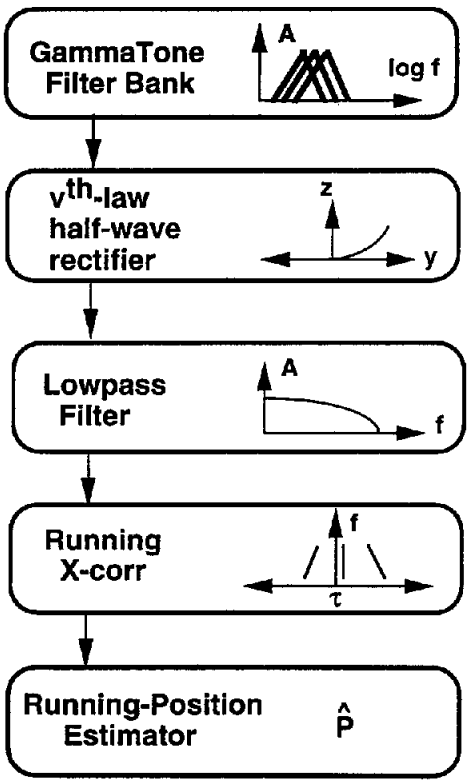

Figure A1. A schematic of the stages of the cross-correlation model. Corr = correlation.

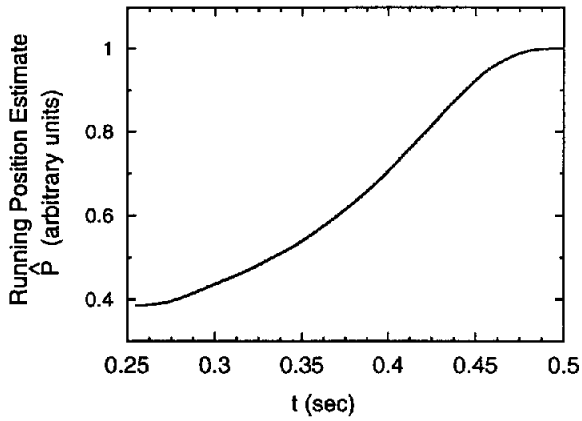

Figure A2. Running-position estimate, $\hat{P}$, calculated from Equation A7 (see Appendix) for the stimuli used to collect the data of Table 1. These estimates were made at 10-ms intervals and are interpolated between these values. The smaller values favor the side of the leading ear (the ear at which the waveform is advanced by 1,500 μs), and the larger values favor the side of the lagging ear.

(Holdsworth, Nimmo-Smith, Patterson, & Rice, 1988) consisted of 40 logarithmically spaced filters from 250 to 840 Hz ($CF_{i+1} = 1.0315633\ CF_i$). The waveforms were filtered directly through the frequency-domain representation

$$G(t) \propto \left[1 + \frac{j(f - f_0)}{b}\right]^{-4} + \left[1 + \frac{j(f + f_0)}{b}\right]^{-4};$$

$$(-\infty < f < \infty), \quad [A3]$$

where f_0 is the filter's center frequency, b controls the width of the filter and is a function of f_0 (see Holdsworth et al., 1988), and j is the complex number $\sqrt{-1}$. The output of this filter bank was followed by vth law, half-wave rectification ($v = 3$; $z[t] = y[t]^v$ if $y[t] > 0$; $z[t] = 0$ otherwise; Sheer, 1987) and a lowpass filter with a 1,200-Hz cutoff; these two stages, respectively, model the mechanical-response properties of the basilar membrane and the frequency limit for detection of interaural delays (Klumpp & Eady, 1956). The lowpass filter was followed by running cross-correlation

$$\Gamma_{l,r}(t, \tau, f) = q(f)p(\tau)\int_{-\infty}^{T} Y_r(f, t)Y_l(f, t - \tau)e^{-(T-t)/0.005}dt, \quad [A4]$$

where the exponentiation heavily weights the more recent cross-correlation activity (Blauert & Cobben, 1978; Sayers & Cherry, 1957), and the Gaussian center-weighting function ($p[\tau] = \phi[\tau/\sigma]$; $\sigma = 0.6$ ms; Shackleton et al., 1992) scales the output of the cross-correlator such that values near $\tau = 0$ are also more heavily weighted (i.e., centrality function). There is in addition the frequency-weighting function $q(f)$, which is used to reflect psychophysical findings on the dominance region of lateralization (Raatgever, 1980; Stern et al., 1988). This function was of the form

$$q(f) = 10^{b_1 f + b_2 f^2 + b_3 f^3}, \quad [A5]$$

where $b1 = 9.383\,(10^{-3})$, $b2 = -1.126\,(10^{-5})$, and $b3 = 3.992\,(10^{-9})$;

f_0 in Hz is the filter center frequency. The consistency of interaural delays over frequency was modeled by integrating the cross-correlation output across frequency; that is, after centrality weighting. The running-position estimate, $\hat{p}$, was then simply the centroid of the area under the resultant function. Because by definition the centroid about the τ axis is

$$\bar{\tau} = \left[\int f(\tau)d\tau\right]^{-1} \int \tau f(\tau)d\tau, \qquad [A6]$$

then

$$\hat{p}(t,\tau,f) = \frac{\iiint_Z \tau q(f)p(\tau)Y_r(f,t)Y_l(f,t-\tau)e^{-(T-t)/0.005}dtd\tau df}{\iiint_Z q(f)p(\tau)Y_r(f,t)Y_l(f,t-\tau)e^{-(T-t)/0.005}dtd\tau df}$$

$$Z = \{(t,\tau,f)|-\infty < t \le T, \quad -0.003 \le \tau \le 0.003, \quad 250 \le f \le 840\} \qquad [A7]$$

where τ is in seconds, and f is in Hz. It is likely that this measure will produce results qualitatively similar to those from Stern et al.'s (1988) variance measure, or Shackleton et al.'s (1992) average-peak model; however, this centroid estimate was used for its comprehensive nature. The centroid measure encodes consistency by attenuating the weight given to the low-frequency segment of the curved trajectories (due to centrality weighting) and amplifying the higher frequencies, which tend to curve toward the true delay. Although simulations show that this measure produces reasonable results, one measure of consistency is not considered superior to others in the absence of further empirical evidence. Figure A2 shows the output of this model (Equation A7) for the stimulus used in collecting data from human observers (see Table 1). The units of laterality are arbitrary and normalized to unity; the smaller values favor the side of the leading waveform, and the larger values favor the lagging waveform.

Received October 3, 1994
Revision received April 3, 1995
Accepted April 3, 1995 ■