CHAPTER 5

BINAURAL SOUND LOCALIZATION

5.1 INTRODUCTION

We listen to speech (as well as to other sounds) with two ears, and it is quite remarkable how well we can separate and selectively attend to individual sound sources in a cluttered acoustical environment. In fact, the familiar term 'cocktail party processing' was coined in an early study of how the binaural system enables us to selectively attend to individual conversations when many are present, as in, of course, a cocktail party [23]. This phenomenon illustrates the important contribution that binaural hearing makes to auditory scene analysis, by enabling us to localize and separate sound sources. In addition, the binaural system plays a major role in improving speech intelligibility in noisy and reverberant environments.

The primary goal of this chapter is to provide an understanding of the basic mechanisms underlying binaural localization of sound, along with an appreciation of how binaural processing by the auditory system enhances the intelligibility of speech in noisy acoustical environments, and in the presence of competing talkers. Like so many aspects of sensory processing, the binaural system offers an existence proof of the possibility of extraordinary performance in sound localization, and signal separation, but it does not yet provide a very complete picture of how this level of performance can be achieved with the contemporary tools of computational auditory scene analysis (CASA).

We first summarize in Sec. 5.2 the major physical factors that underly binaural perception, and we briefly summarize some of the classical physiological findings that describe mechanisms that could support some aspects of binaural processing. In Sec. 5.3 we briefly review some of the basic psychoacoustical results which have motivated the development

Computational Auditory Scene Analysis. By DeLiang Wang and Guy J. Brown (eds.)
ISBN 0-471-45435-4

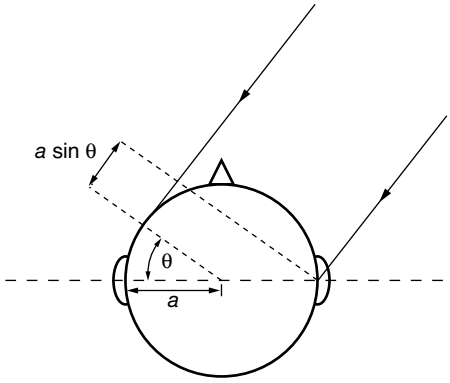

Figure 5.1 Interaural differences of time and intensity impinging on an ideal spherical head from a distant source. An interaural time delay (ITD) is produced because it takes longer for the signal to reach the more distant ear. An interaural intensity difference (IID) is produced because the head blocks some of the energy that would have reached the far ear, especially at higher frequencies.

of most of the popular models of binaural interaction. These studies have typically utilized very simple signals such as pure tones or broadband noise, rather than more interesting and ecologically relevant signals such as speech or music. We extend this discussion to results involving the spatial perception of multiple sound sources in Sec. 5.4. In Sec. 5.5 we introduce and discuss two of the most popular types of models of binaural interaction, cross-correlation based models and the equalization-cancellation (EC) model. Section 5.6 then discusses how the cross-correlation model may be utilized within a CASA system, by providing a means for localizing multiple (and possibly moving) sound sources. Binaural processing is a key component of practical auditory scene analysis, and aspects of binaural processing will be discussed in greater detail later in this volume in Chapters 6 and 7.

5.2 PHYSICAL CUES AND PHYSIOLOGICAL MECHANISMS UNDERLYING AUDITORY LOCALIZATION

5.2.1 Physical cues

A number of factors affect the spatial aspects of how a sound is perceived. Lord Rayleigh's 'duplex theory' [86] was the first comprehensive analysis of the physics of binaural perception, and his theory remains basically valid to this day, with some extensions. As Rayleigh noted, two physical cues dominate the perceived location of an incoming sound source, as illustrated in Fig. 5.1. Unless a sound source is located directly in front of or behind the head, sound arrives slightly earlier in time at the ear that is physically closer to the source, and with somewhat greater intensity. This *interaural time difference* (ITD) is produced because it takes longer for the sound to arrive at the ear that is farther from the source. The *interaural intensity difference* (IID) is produced because the 'shadowing' effect of the head prevents some of the incoming sound energy from reaching the ear that is turned away from the direction of the source. The ITD and IID cues operate in complementary ranges of frequencies at least for simple sources in a free field (such as a location outdoors or in an anechoic chamber). Specifically, IIDs are most pronounced at frequencies above approximately 1.5 kHz because it is at those frequencies that the head is large compared

to the wavelength of the incoming sound, producing substantial reflection (rather than total diffraction) of the incoming sound wave. Interaural timing cues, on the other hand, exist at all frequencies, but for periodic sounds they can be decoded unambiguously only for frequencies for which the maximum physically-possible ITD is less than half the period of the waveform at that frequency. Since the maximum possible ITD is about 660 μs for a human head of typical size, ITDs are generally useful only for stimulus components below about 1.5 kHz. Note that for pure tones, the term *interaural phase delay* (IPD) is often used, since the ITD corresponds to a phase difference.

If the head had a completely spherical and uniform surface, as in Fig. 5.1, the ITD produced by a sound source that arrives from an azimuth of θ radians can be approximately described (e.g., [64, 65, 106]) using diffraction theory by the equation

$$\tau = (a/c)2\sin\theta \tag{5.1a}$$

for frequencies below approximately 500 Hz, and by the equation

$$\tau = (a/c)(\theta + \sin\theta) \tag{5.1b}$$

for frequencies above approximately 2 kHz. In the equations above, a represents the radius of the head (approximately 87.5 mm) and c represents the speed of sound.

The actual values of IIDs are not as well predicted by wave diffraction theory, but they can be measured using probe microphones in the ear and other techniques. They have been found empirically to depend on the angle of arrival of the sound source, frequency, and distance from the sound sources (at least when the source is extremely close to the ear). IIDs produced by distant sound sources can become as large as 25 dB in magnitude at high frequencies, and the IID can become greater still when a sound source is very close to one of the two ears.

The fissures of the outer ears (or *pinnae*) impose further spectral coloration on the signals that arrive at the eardrums. This information is especially useful in a number of aspects of the localization of natural sounds occurring in a free field, including localization in the vertical plane, and the resolution of front-back ambiguities in sound sources. Although measurement of and speculation about the spectral coloration imposed by the pinnae have taken place for decades (e.g., [2, 51, 65, 81, 102, 105]), the 'modern era' of activity in this area began with the systematic and carefully controlled measurements of Wightman and Kistler and others (e.g., [82, 124, 125, 126]), who combined careful instrumentation with comprehensive psychoacoustical testing. Following procedures developed by Mehrgardt and Mellert [81] and others, Wightman and Kistler and others used probe microphones in the ear to measure and describe the transfer function from sound source to eardrum in anechoic environments. This transfer function is commonly referred to as the *head-related transfer function* (HRTF), and its time-domain analog is the *head-related impulse response* (HRIR). Among other attributes, measured HRTFs show systematic variations of frequency response above 4 kHz as a function of azimuth and elevation. There are also substantial differences in frequency response from subject to subject. HRTF measurements and their inverse Fourier transforms have been the key component in a number of laboratory and commercial systems that attempt to stimulate natural three-dimensional acoustical environments using signals presented through headphones (e.g., [7, 43]).

As Wightman and Kistler note [126], the ITD as measured at the eardrum for broadband stimuli is approximately constant over frequency and it depends on azimuth and elevation in approximately the same way from subject to subject. Nevertheless, a number of different azimuths and elevations will produce the same ITD, so the ITD does not unambiguously

signal source position. IIDs measured at the eardrum exhibit much more subject-to-subject variability, and for a given subject the IID is a much more complicated function of frequency, even for a given source position. Wightman and Kistler suggest that because of this, IID information in individual frequency bands is likely to be more useful than overall IID.

5.2.2 Physiological estimation of ITD and IID

There have been a number of physiological studies that have described cells that are likely to be useful in extracting the basic cues used in auditory spatial perception, as have been described in a number of comprehensive reviews (e.g., [27, 29, 87, 66, 128]). Any consideration of neurophysiological mechanisms that potentially mediate binaural processing must begin with a brief discussion of the effects of processing of sound by the auditory periphery. As was first demonstrated by von Békésy [122] and by many others since, the mechanical action of the cochlea produces a frequency-to-place transformation. Each of the tens of thousands of fibers of the auditory nerve for each ear responds to mechanical stimulation along only a small region of the cochlea, so the neural response of each fiber is highly frequency specific, and stimulus frequency to which each fiber is most sensitive is referred to as the *characteristic frequency* (CF) for that fiber. This frequency-specific 'tonotopic' representation of sound in each channel of parallel processing is preserved at virtually all levels of auditory processing. In addition, the response of a fiber with a low CF is 'synchronized' to the detailed time structure of a low-frequency sound, in that neural spikes are far more likely to occur during the negative portion of the pressure waveform than during the positive portion.

The auditory-nerve response to an incoming sound is frequently modeled by a bank of linear bandpass filters (that represent the frequency selectivity of cochlear processing), followed by a series of nonlinear operations at the outputs of each filter that include half-wave rectification, nonlinear compression and saturation, and 'lateral' suppression of the outputs of adjacent frequency channels (that represent the subsequent transduction to a neural response). A number of computational models incorporating varying degrees of physiological detail or abstraction have been developed that describe these processes (e.g., [75, 80, 89, 133]; see also Sec. 1.3.2). One reason for the multiplicity of models is that it is presently unclear which aspects of nonlinear auditory processing are the most crucial for the development of improved features for robust automatic speech recognition, or for the separation of simultaneously-presented signals for CASA. Any physiological mechanism that extracts the ITD or IID of a sound (as well as any other type of information about it) must operate on the ensemble of narrowband signals that emerge from the parallel channels of the auditory processing mechanism, rather than on the original sounds that are presented to the ear.

The estimation of ITD is probably the most critical aspect of binaural processing. As will be discussed below in Sec. 5.5.2, many models of binaural processing are based on the cross-correlation of the signals to the two ears after processing by the auditory periphery, or based on other functions that are closely related to cross-correlation. The physiological plausibility of this type of model is supported by the existence of cells first reported by Rose *et al.* [95] in the inferior colliculus in the brainstem. Such cells appear to be maximally sensitive to signals presented with a specific ITD, independent of frequency. This delay is referred to as a *characteristic delay* (CD). Cells exhibiting similar response have been reported by many others in other parts of the brainstem, including the medial superior olive and the dorsal nucleus of the lateral lemniscus.

Several series of measurements have been performed that characterize the distribution of the CDs of ITD-sensitive cells in the inferior colliculus (e.g., [129, 66]), and the medial geniculate body (e.g., [109]). The results of most of these studies indicate that ITD-sensitive cells tend to exhibit CDs that lie in a broad range of ITDs, with the density of CDs decreasing as the absolute value of the ITD increases. While most of the CDs appear to occur within the maximum ITD that is physically possible for a point source in a free field for a particular animal at a given frequency, there is also a substantial number of ITD-sensitive cells with CDs that fall outside this 'physically-plausible' range. In a recent series of studies, McAlpine and his collaborators have argued that most ITD-sensitive units exhibit characteristic delays that occur in a narrow range that is close to approximately one-eighth of the period of a cell's characteristic frequency (e.g., [78]), at least for some animals.

The anatomical origin of the characteristic delays has been the source of some speculation. While many physiologists believe that the delays are of neural origin, caused either by slowed conduction delays or by synaptic delays (e.g., [22, 132]), other researchers (e.g., [101, 104]) have suggested that the characteristic delays could also be obtained if higher processing centers compare timing information derived from auditory-nerve fibers with different CFs. In general, the predictions of binaural models are unaffected by whether the internal delays are assumed to be caused by neural or mechanical phenomena.

A number of researchers have also reported cells that appear to respond to IIDs at several levels of the brainstem (e.g., [11, 19]). Since the rate of neural response to a sound increases with increasing intensity (at least over a limited range of intensities), IIDs could be detected by a unit which has an excitatory input from one ear and an inhibitory input from the other.

5.3 SPATIAL PERCEPTION OF SINGLE SOURCES

5.3.1 Sensitivity to differences in interaural time and intensity

Humans are remarkably sensitive to small differences in interaural time and intensity. For low-frequency pure tones, for example, the *just-noticeable difference* (JND) for ITDs is on the order of 10 μs, and the corresponding JND for IIDs is on the order of 1 dB (e.g., [34, 53]). The JND for ITD depends on the ITD, IID, and frequency with which a signal is presented. The binaural system is completely insensitive to ITD for narrowband stimuli above about 1.5 kHz, although it does respond to low-frequency envelopes of high-frequency stimuli, as will be noted below. JNDs for IID are a small number of decibels over a broad range of frequencies. Sensitivity to small differences in interaural correlation of broad-band noise sources is also quite acute, as a decrease in interaural correlation from 1 to 0.96 is readily discernable (e.g., [40, 45, 90]).

5.3.2 Lateralization of single sources

Until fairly recently, most studies of binaural hearing have involved the presentation of single sources through headphones, such as clicks (e.g., [32, 49]), bandpass noise (e.g., [120]), pure tones (e.g., [34, 96, 130]), and amplitude-modulated tones (e.g., [3, 52]). Such experiments measure the *lateralization* of the source (i.e., its apparent lateral position within the head) and are therefore distinct from *localization* experiments (in which the task is to judge the apparent direction and distance of the source outside the head).

Unsurprisingly, the perceived lateral position of a narrowband binaural signal is a periodic (but not sinusoidal) function of the ITD with a period equal to the reciprocal of the center frequency. It was found that the perceived lateralization of a narrowband signal such as

a pure tone was affected by its ITD only at frequencies below approximately 1.5 kHz, consistent with the comments on the utility of temporal cues at higher frequencies stated above. Nevertheless, the perceived laterality of broader-band signals, such as amplitude-modulated tones and bandpass-filtered clicks, can be affected by the ITD with which they are presented, even if all components are above 1.5 kHz, provided that the stimuli produce low-frequency envelopes (e.g., [4, 5, 52, 79]).

IIDs, in contrast, generally affect the lateral position of binaural signals of all frequencies. Under normal circumstances the perceived laterality of stimuli presented through headphones is a monotonic function of IID, although the exact form of the function relating lateral position to IID depends upon the nature of the stimulus including its frequency content, as well as the ITD with which the signal is presented. Under normal circumstances, IIDs of magnitude greater than approximately 20 dB are perceived close to one side of the head, although discrimination of small changes in IID based on lateral position cues can be made at IIDs of these and larger magnitudes (e.g., [34, 53]).

Many studies have been concerned with the ways in which information related to the ITD and IID of simple and complex sources interact with each other. If a binaural signal is presented with an ITD of less than approximately 200 μs and an IID of 5 dB in magnitude, its perceived lateral position can approximately be described by a linear combination of the two cues. Under such conditions, the relative salience of ITD and IID was frequently characterized by the time-intensity trading ratio, which can range from approximately 20 to 200 μs/dB, depending on the type of stimulus, its loudness, the magnitude of the ITDs and IIDs presented, and other factors (*cf.* [39]). While it had been suggested by Jeffress [59] and others that this time-intensity conversion might be a consequence of the observed decrease in physiological latency in response to signals of greater intensity, lateralization studies involving ITDs and IIDs of greater magnitude (e.g., [34, 96]) indicate that response latency alone cannot account for the form of the data. Most contemporary models of binaural interaction (e.g., [8, 70, 110]) assume a more central form of time-intensity interaction.

5.3.3 Localization of single sources

As noted above, Wightman and Kistler developed a systematic and practical methodology for measuring the HRTFs that describe the transformation of sounds in the free field to the ears. They used the measured HRTFs both to analyze the physical attributes of the sound pressure impinging on the eardrums, and to synthesize 'virtual stimuli' that could be used to present through headphones a simulation of a particular free-field stimulus that was reasonably accurate (at least for the listener used to develop the HRTFs) [124, 125, 126]. These procedures have been adopted by many other researchers.

Wightman and Kistler and others have noted that listeners are able to describe the azimuth and elevation of free-field stimuli consistently and accurately. Localization judgments obtained using 'virtual' headphone simulations of the free-field stimuli are generally consistent with the corresponding judgments for the actual free-field signals, although the effect of elevation change is less pronounced and a greater number of front-to-back confusions in location is observed [125]. On the basis of various manipulations of the virtual stimuli, they also conclude that under normal circumstances the localization of free-field stimuli is dominated by ITD information, especially at the lower frequencies, that ITD information must be consistent over frequency for it to play a role in sound localization, and that IID information appears to play a role in diminishing the ambiguities that give rise to front-back confusions of position [126]. While the interaural cues for lateralization are relatively robust across subjects, fewer front-to-back and other confusions are experienced in the localization of

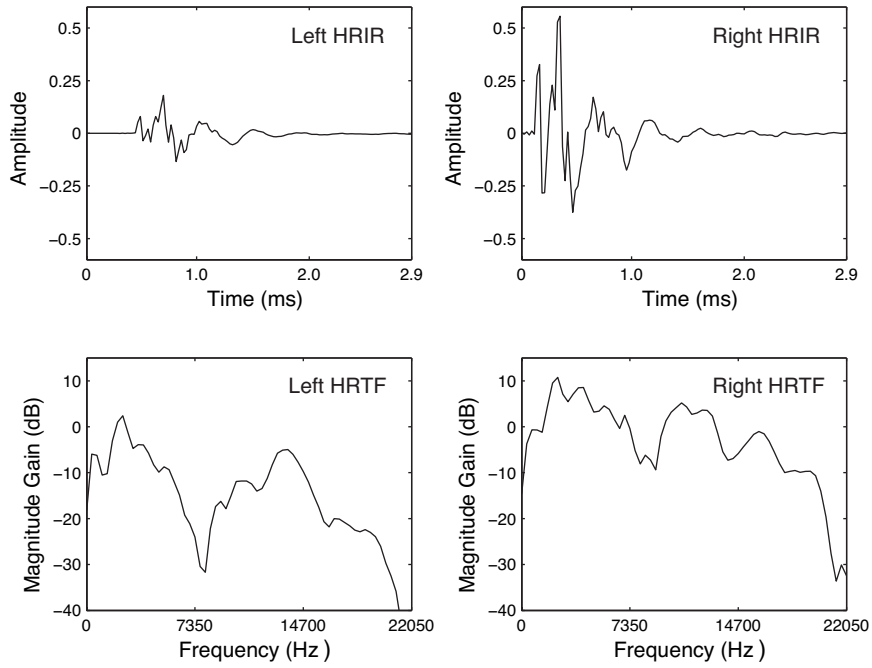

Figure 5.2 Head related impulse responses (HRIRs) (top row) and the corresponding head-related transfer functions (HRTFs) (bottom row) recorded at the left and right ears of a KEMAR manikin, in response to a source placed at an azimuth of 40° to the right of the head with 0° elevation. Note that the stimulus is more intense in the right ear, and arrives first in the right ear before reaching the left ear. Plotted from data recorded by Gardner and Martin [47].

simulated free-field stimuli if the HRTFs used for the virtual sound synthesis are based on a subject's own pinnae [123, 84]. Useful physical measurements of ITD, IID, and other stimulus attributes can also be obtained using an anatomically realistic manikin such as the popular Knowles Electronics Manikin for Acoustic Research (KEMAR) [18]. Examples of HRTFs (and the corresponding HRIRs) recorded from a KEMAR manikin are shown in Fig. 5.2. Experimental measurements indicate that localization judgements are more accurate when the virtual source is spatialized using HRTFs obtained by direct measurement in human ears, rather than from an artificial head such as the KEMAR. However, there is a strong learning effect, in which listeners adapt to unfamiliar HRTFs over the course of several experimental sessions [83].

5.3.4 The precedence effect

A further complication associated with judging the location of a sound source in a natural acoustic environment, such as an enclosed room, is that sound is reflected from various surfaces before reaching the ears of the listener. However, despite the fact that reflections arise from many directions, listeners are able to determine the location of the direct sound quite accurately. Apparently, directional cues that are due to the direct sound (the 'first wave front') are given a higher perceptual weighting than those due to reflected sound. The term *precedence effect* is used to describe this phenomenon (see [72] for a review). Further

discussion of the precedence effect can be found in Chapter 7, which deals with the effect of reverberation on human and machine hearing.

5.4 SPATIAL PERCEPTION OF MULTIPLE SOURCES

5.4.1 Localization of multiple sources

Human listeners can localize a single sound source rather accurately. One accuracy measure for azimuth localization is the minimum audible angle (MAA), which refers to the smallest detectable change in angular position. For sinusoidal signals presented on the horizontal plane, spatial resolution is highest for sounds coming from the median plane (directly in front of the listener) with about 1° MAA, and it deteriorates markedly when stimuli are moved to the side – e.g., the MAA is about 7° for sounds originating at 75° to the side [8, 85]. In terms of making an absolute judgment of the spatial location of a sound, the average error is about 5° for broadband stimuli presented on the median plane, and increases up to 20° for sounds from the side [56, 67].

In the context of this book, a very relevant issue is the ability of human listeners to localize a sound in multisource scenarios. An early study was conducted by Jacobsen [58] who observed that, for pure tones presented with a white-noise masker, the MAA is comparable to that when no masker is presented so long as the signal-to-noise ratio (SNR) is relatively high (10 to 20 dB). In a comprehensive investigation, Good and Gilkey [48] examined the accuracy of localization judgments by systematically varying SNR levels and sound azimuths. In their study, the signal is a broadband click train that may originate from any of a large number of spatial locations, and a broadband masker is always located on the median plane. As expected, the accuracy of localization decreases almost monotonically when the SNR is lowered. However, Good and Gilkey found that azimuth localization judgments are not strongly influenced by the interference and remain accurate even at low SNRs. This holds until the SNR is in the negative range, near the detection threshold for the target, beyond which the target will be inaudible due to masking by the interference. The same study also reported that the effect of the masker on localization accuracy is a little stronger for judgments of elevation, and is strongest in the front-back dimension. Using multisource presentation of spoken words, letters and digits, Yost *et al.* [131] found that utterances that can correctly be identified tend to be correctly localized.

Similar conclusions have been drawn in a number of subsequent studies. For example, Lorenzi *et al.* [74] observed that localization accuracy is unaffected by the presence of a white-noise masker until the SNR is reduced to the 0-6 dB range. They also reported that the effect of noise is stronger when it is presented at the side than from the median plane. Hawley *et al.* [50] studied the localization of speech utterances in multisource configurations. Their results show that localization performance is very good even when three competing sentences, each at the same level as the target sentence, are presented at various azimuths. Moreover, the presence of the interfering utterances has little adverse impact on target localization accuracy so long as the target is clearly audible, consistent with the earlier observation by Yost *et al.* [131]. On the other hand, Drullman and Bronkhorst [35] reported rather poor performance in speech localization in the presence of one to four competing talkers, which may have been caused by the complexity of the task that required the subjects to first detect the target talker and then determine the talker location. They further observed that localization accuracy decreases when the number of interfering talkers increases from 2 to 4.

Langendijk *et al.* [67] studied listeners' performance in localizing a train of noise bursts presented together with one or two complex tones. Their findings confirm that for positive SNR levels, azimuth localization performance is not much affected by the interference. In addition, they found that the impact of maskers on target localization is greater when azimuth separation between multiple sources is reduced. Using noise bursts as stimuli for both target and masker, Braasch [12] showed that localization performance is enhanced by introducing a difference in onset times between the target and the masker when the SNR is 0 dB.

In summary, the body of psychoacoustical evidence on sound localization in multisource situations suggests that localization and identification of a sound source are closely related. Furthermore, auditory cues that promote auditory scene analysis also appear to enhance localization performance.

5.4.2 Binaural signal detection

While the role that the binaural system plays in sound localization is well known, binaural processing also plays a vital role in detecting sounds in complex acoustical environments. In the following, we review some classical studies on binaural signal detection, and then discuss the ability of binaural mechanisms to enhance the intelligibility of speech in noise.

Classical binaural detection Most of the classical psychoacoustical studies of binaural detection have been performed with simple stimuli such as pure tones, clicks, or broadband noise. Generally, it has been found that the use of binaural processing significantly improves performance for signal detection tasks if the overall ITD, IID, or interaural correlation changes as a target is added to a masker. For example, if a low-frequency tone and a broadband masker are presented monaurally, the threshold SNR that is obtained will generally depend on various stimulus parameters such as the target duration and frequency, and the masker bandwidth. Diotic presentation of the same target and masker (i.e., with identical signals presented to the two ears) will produce virtually the same detection threshold. On the other hand, if either the target or masker are presented with a nonzero ITD, IID, or IPD, the target will become much easier to hear. For example, using 500 Hz tones as targets and broadband maskers, presentation of the stimuli with the masker interaurally in phase and the target interaurally out of phase (the 'N_0S_π' configuration) produces a detection threshold that is about 12 to 15 dB lower than the detection threshold observed when the same target and masker are presented monaurally or diotically (the 'N_0S_0' configuration) [54]. This 'release' from masking is referred to as binaural 'unmasking' or the binaural masking level difference (MLD or BMLD).

Detection performance improves for stimuli which produce an MLD because the change in net ITD or IID that occurs when the target is added to the masker is detectable by the binaural processing system at lower SNRs than those needed for monaural detection. The MLD is one of the most robust and extensively studied of all binaural phenomena, particularly for tonal targets and noise maskers. Durlach and Colburn's comprehensive review [39] describes how the MLD depends on stimulus parameters such as SNR, target frequency and duration, masker frequency and bandwidth, and the ITD, IID, and IPD of the target masker. These dependencies are for the most part well described by modern theories of binaural hearing (e.g., [15, 25, 26]).

Binaural detection of speech signals While most of the MLD literature was concerned with the detection of tonal targets in noise maskers, the importance of interaural

differences for improving speech intelligibility in a noisy environment has been known since at least the 1950s, when Hirsch, Kock, and Koenig first demonstrated that binaural processing can improve the intelligibility of speech in a noisy environment [55, 60, 61]. As reviewed by Zurek [134], subsequent studies (e.g., [20, 21, 33, 77]) identified two reasons why the use of two ears can improve the intelligibility of speech in noise. First, there is a *head-shadow advantage* that can be obtained by turning one ear toward the direction of the target sound, at the same time increasing the likelihood that masking sources will be attenuated by the shadowing effect of the head if they originate from directions that are away from that ear. The second potential advantage is a *binaural-interaction advantage* that results from the fact that the ITDs and IIDs of the target and masker are different when the sources originate from different azimuths. Many studies confirm that speech intelligibility improves when the spatial separation of a target speech source and competing maskers increases, both for reasons of head shadowing and binaural interaction.

Levitt and Rabiner [68] developed a simple model that predicts the binaural-interaction advantage that will be obtained when speech is presented with an ITD or IPD in the presence of noise. They predicted speech intelligibility by first considering the ITD or IPD and SNR of the target and masker at each frequency of the speech sound, and assuming that the binaural-interaction advantage for each frequency component could be predicted by the MLD for a pure tone of the same frequency in noise, using the same interaural parameters for target and masker. The effects of the MLD are then combined across frequency using standard articulation index theory (e.g., [42, 44, 63]).

Zurek [134] quantified the relative effects of the head-shadow advantage and the binaural advantage for speech in the presence of a single masking source. He predicted the head-shadow of speech from the SNR at the 'better' ear for a particular stimulus configuration, and the binaural-interaction advantage from the MLD expected from the stimulus components at a particular frequency, again combining this information across frequency by using articulation index theory. Zurek found the predictions of this model to be generally consistent with contemporary data that described the average dependence of intelligibility on source direction and listening mode, although it did not model variations in data from subject to subject. Hawley *et al.* [50] extended Zurek's approach in a series of experiments that employed multiple streams of competing speech maskers. The predictions of both models indicate that many phenomena can be accounted for simply by considering the monaural characteristics of the signals that arrive at the 'better' ear, but that processing based on binaural interaction plays a significant role for certain configurations of targets and competing maskers.

5.5 MODELS OF BINAURAL PERCEPTION

In this section we review some of the classical and more recent models of binaural interaction that have been applied to simple and more complex binaural phenomena. We begin with a discussion of two seminal theories of binaural interaction, the *coincidence-based model* of Jeffress [59] and the *equalization-cancellation model* (EC model) of Durlach [37, 38]. Most current binaural models trace their lineage to one (and in some cases both) of these theories. We continue with a discussion of modern realizations of the Jeffress model that typically include a model for auditory-nerve activity, a mechanism that extracts a frequency-dependent measurement of the interaural cross-correlation of the signals, along with some additional processing to incorporate the effects of IIDs and to develop the perceptual representation that is needed to perform the particular psychoacoustical task at hand. In addition to this

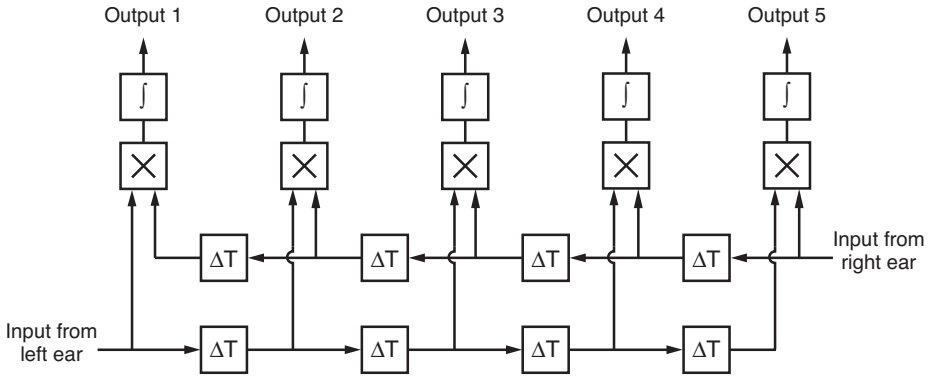

Figure 5.3 Schematic representation of the Jeffress place mechanism. Boxes containing crosses are correlators (multipliers) that record coincidences of neural activity from the two ears after the internal delays (ΔT).

discussion, the reader is referred to a number of excellent reviews of models of binaural interaction that have appeared over the years [13, 27, 28, 29, 115, 116, 119].

5.5.1 Classical models of binaural hearing

The Jeffress hypothesis. Most modern computational models of binaural perception are based on Jeffress's description of a neural 'place' mechanism that would enable the extraction of interaural timing information [59]. Jeffress postulated a mechanism that consisted of a number of central neural units that recorded coincidences in neural firings from two peripheral auditory-nerve fibers, one from each ear, with the same CF. He further postulated that the neural signal coming from one of the two fibers is delayed by a small amount that is fixed for a given fiber pair, as in the block diagram of Fig. 5.3. Because of the synchrony in the response of low-frequency fibers to low-frequency stimuli, a given binaural coincidence-counting unit at a particular frequency will produce maximal output when the external stimulus ITD at that frequency is exactly compensated for by the internal delay of the fiber pair. Hence, the external ITD of a simple stimulus could be inferred by determining the internal delay that has the greatest response over a range of frequencies. While the delay mechanism was conceptualized by Jeffress and others in the form of the ladder-type delay shown in Fig. 5.3, such a structure is only one of several possible realizations. These coincidence-counting units can be thought of as mathematical abstractions of the ITD-sensitive units which were first described by Rose *et al.* decades later, as discussed in Sec. 5.2.2 above. The important characteristic-delay parameter of the ITD-sensitive units is represented by the difference in total delay incurred by the neural signals from the left and right ears that are input to a particular coincidence-counting unit. This parameter will be referred to as the net internal delay τ for that particular unit. As will be discussed below, the short-term average of a set of such coincidence outputs at a particular CF plotted as a function of their internal delay τ is an approximation to the short-term cross-correlation functions of the neural signals arriving at the coincidence detectors. Licklider [69] proposed that such a mechanism could also be used to achieve an auto-correlation of neural signals for use in models of pitch perception (see Sec. 2.3.2).

We note that the interaural coincidence operation cannot by itself account for effects related to the IID of the stimulus. Jeffress proposed the *latency hypothesis* which is based on the common observation that the neural response to more intense sounds tends to be initiated more rapidly than the response to less intense sounds. Jeffress postulated that this effect enabled IIDs to be converted to ITDs at the level of the peripheral auditory response to a sound. It is now widely believed that the interaction between the effects of IIDs and ITDs is mediated at a more central level.

The equalization-cancellation model. The equalization-cancellation model (or EC model) has also been extremely influential for decades, no doubt because of its conceptual simplicity and because of its ability to describe a number of interesting binaural phenomena. The model was first suggested by Kock [60] and was subsequently developed extensively by Durlach (e.g., [37, 38]). While the EC model was primarily developed to describe and predict the binaural masking-level differences described in Sec. 5.4.2, it has been applied with some success to other phenomena as well. In the classic application to binaural detection, the EC model assumes that the auditory system transforms the signals arriving at the two ears so that the masker components are 'equalized', or made equal to one another to the extent possible. Detection of the target is achieved by 'cancelling', or subtracting the signals to the two ears after the equalization operation. Considering the two classical binaural signal configurations described in Sec. 5.4.2 for MLD experiments, it is clear that signals in the N_0S_0 configuration will not be detectable by binaural processing because if the equalization and cancellation operations are accurate and successful the target will be cancelled along with the masker. Binaural detection stimuli presented in the N_0S_π configuration should be easily detected, however, because the target is reinforced as the masker is cancelled. Quantitative predictions for the EC model are obtained by specifying limits to the operations used to achieve the cancellation process, as well as sources of internal noise. The performance and limitations of the EC model are discussed in detail in the excellent early review of binaural models by Colburn and Durlach [28].

5.5.2 Cross-correlation-based models of binaural interaction

Stimulus-based cross-correlation models. While both Jeffress and Licklider had introduced the concept of correlation in their models of binaural interaction and processing of complex signals, the work of Sayers and Cherry (e.g., [24, 96, 97, 98, 99]) represented the first comprehensive attempt to relate the fusion and lateralization of binaural stimuli to their interaural cross-correlation. Sayers and Cherry considered the *short-time cross-correlation function* (or the 'running' cross-correlation function) of the stimuli:

$$R(t,\tau) = \int_{-\infty}^{t} x_L(\alpha)x_R(\alpha-\tau)w(t-\alpha)p(\tau)d\alpha \tag{5.2}$$

where $x_L(t)$ and $x_R(t)$ are the signals to the left and right ears. The function $w(t)$ represents the temporal weighting of the short-time cross-correlation operation, and is exponential in form in most of Sayers and Cherry's calculations. The function $p(\tau)$ is typically a double-sided decaying exponential that serves to emphasize the contributions of internal delays τ that were small in magnitude. We refer to this type of emphasis as 'centrality'. The reader should note that Sayer's and Cherry's function does not take into account any signal processing by the peripheral auditory system. As is true for all cross-correlation-based models, an additional mechanism is needed to account for the effects of IID.

Sayers and Cherry added a constant proportional to the intensity of the left-ear signal to values of the internal delay τ that were less than zero and a (generally different) constant proportional to the intensity of the right-ear signal to values of τ that were greater than zero. A judgment mechanism then extracted subjective lateral position using the statistic

$$\hat{P} = \frac{I_L - I_R}{I_L + I_R} \tag{5.3}$$

where I_L and I_R are the integrals of the intensity-weighted short-time cross-correlation function over negative and positive values of τ, respectively. Sayers and Cherry considered the lateral position of a variety of stimuli including speech, pure tones of various frequencies, and click trains, and found that they were predicted at least qualitatively by the above lateralization function or variants of it. Furthermore, their data indicated that the latency hypothesis could not adequately describe all of the complexities of the dependence of perceived laterality on the IID of the stimulus.

Models incorporating the auditory-nerve response to the stimuli. While the models of Sayers and Cherry predicted binaural phenomena from the cross-correlation of the auditory stimulus itself, more recent theories of binaural interaction have been based on the interaural correlation of the *neural response* to the stimulus, rather than to the auditory stimulus itself. Early physiologically-based models focussed more on the development of closed-form mathematical analytical functions to describe and predict the data under consideration. Nevertheless, the trend over the last several decades has been to use computational models to calculate both the auditory-nerve response to sounds and the subsequent processing of the signal needed to obtain both the representation of the interaural timing information that is associated with the signal, as well as subjective variables such as lateral position that are developed from that representation. The use of computational simulation (rather than analytical description) has the advantage that models can make use of more complex and accurate characterization of the auditory-nerve response to the stimuli as well as the advantage that far more complex stimuli can be considered. On the other hand, the ever-increasing complexity of the computational models can make it more difficult to understand exactly which aspect of a model is most important in describing the data to which it is applied.

We first consider in some detail the influential quantification of Jeffress's hypothesis by Colburn [25, 26]. We then review several extensions to the Jeffress-Colburn model in the 1970s and 1980s by Stern and Trahiotis [110, 111, 113, 114, 115, 117, 118], by Blauert and his colleagues Lindemann and Gaik [6, 9, 46, 70, 71], and finally by Breebaart and colleagues [14, 15, 16].

Colburn's quantification of the Jeffress hypothesis. Colburn's model of binaural interaction based on auditory-nerve activity consisted of two components: a model of the auditory-nerve response to sound, and a 'binaural displayer' that can be used to compare the auditory-nerve response to the signals of the two ears. The model of auditory-nerve activity used in the original Colburn model was simple in form to facilitate the development of analytic expressions for the time-varying means and variances of the neural spikes produced by the putative auditory-nerve fibers. Based on an earlier formulation of Siebert [107], Colburn's peripheral model consisted of a bandpass filter (to depict the frequency selectivity of individual fibers), an automatic gain control (which limits the average rate of response to stimuli), a lowpass filter (which serves to limit phase-locking to stimulus fine structure at higher frequencies), and an exponential rectifier (which roughly characterizes

peripheral nonlinearities). The result of this series of operations is a function $r(t)$ that describes the putative instantaneous rate of firing of the auditory-nerve fiber in question. The responses themselves were modeled as nonhomogeneous Poisson processes. Similar functional models have been used by others (e.g., [9, 46, 70, 103]). In recent years, models of the peripheral auditory response to sound have become more computationally oriented and physiologically accurate (e.g., [80, 133]).

The heart of Colburn's binaural model [25, 26] is an ensemble of units describing the interaction of neural activity from the left and right ears generated by auditory-nerve fibers with the same CF, with input from one side delayed by an amount that is fixed for each fiber pair, as in Jeffress's model depicted in Fig. 5.3. Following Jeffress's hypothesis, each coincidence-counting mechanism emits a pulse if it receives two incoming pulses (after the internal delay) within a sufficiently short time of each other. If the duration of the coincidence window is sufficiently brief, the Poisson assumption enables us to compute statistics for the coincidence counts as a function of running time, CF, and internal delay. For example, it can be shown [25] that the average number of coincidences observed at time t from all fiber pairs with CF f and internal delay τ, $E[L(t,\tau,f)]$, is approximately

$$E[L(t,\tau,f)] = \int_{-\infty}^{t} r_L(\alpha) r_R(\alpha-\tau) w_c(t-\alpha) p(\tau,f) d\alpha \tag{5.4}$$

where $r_L(t)$ and $r_R(t)$ are now the functions that describe the instantaneous rates of the Poisson processes that describe the activity of the two auditory-nerve fibers that are the inputs to the coincidence-counting unit. Here, $L(t,\tau,f)$ is a binaural decision variable and $E[\cdot]$ denotes the expectation. The function $w_c(t)$ represents the temporal weighting function as before, and $p(\tau,f)$ now represents the relative number of fiber pairs with a particular internal delay τ and CF f. Comparing Eq. 5.4 to Eq. 5.2 it can readily be seen that the relative number of coincidence counts of the Jeffress-Colburn model, considered as a function of the internal-delay τ at a particular CF f, is an estimate of the short-time interaural cross-correlation of the *auditory-nerve responses to the stimuli* at each CF.

Colburn and Durlach [28] have noted that the cross-correlation mechanism shown in Fig. 5.3 can also be regarded as a generalization of the EC model of Durlach [37]. As described above, the EC model yields predictions concerning binaural detection thresholds by applying a combination of ITD and IID that produces the best equalization of the masker components of the stimuli presented to each of the two ears. Cancellation of the masker is then achieved by subtracting one of the resulting signals from the other. Predictions provided by the EC model are generally dominated by the effects of the ITD-equalization component rather than the IID-equalization component. Because the interaural delays of the fiber pairs of the Jeffress-Colburn model perform the same function as the ITD-equalizing operation of the EC model, most predictions of MLDs for the two models are similar. The Jeffress-Colburn model can explain most (but not all) binaural detection phenomena by the decrease in correlation of the stimuli (and the corresponding decrease in response by the coincidence-counting units) that is observed at the ITD of the masker and the frequency of the target when the target is added to the masker in a configuration that produces binaural unmasking.

Figure 5.4 is an example of the representation of simple stimuli by a contemporary implementation of the Jeffress-Colburn model. Responses are shown to a binaural bandpass noise with nominal corner frequencies of 100 and 1000 Hz, presented with an ITD such that the left ear is leading by 0.5 ms. The upper panel shows the relative rate of coincidences that would be produced by the coincidence-counting units as a function of their internal delay τ and CF f. The calculations in this figure were implemented by passing the incoming

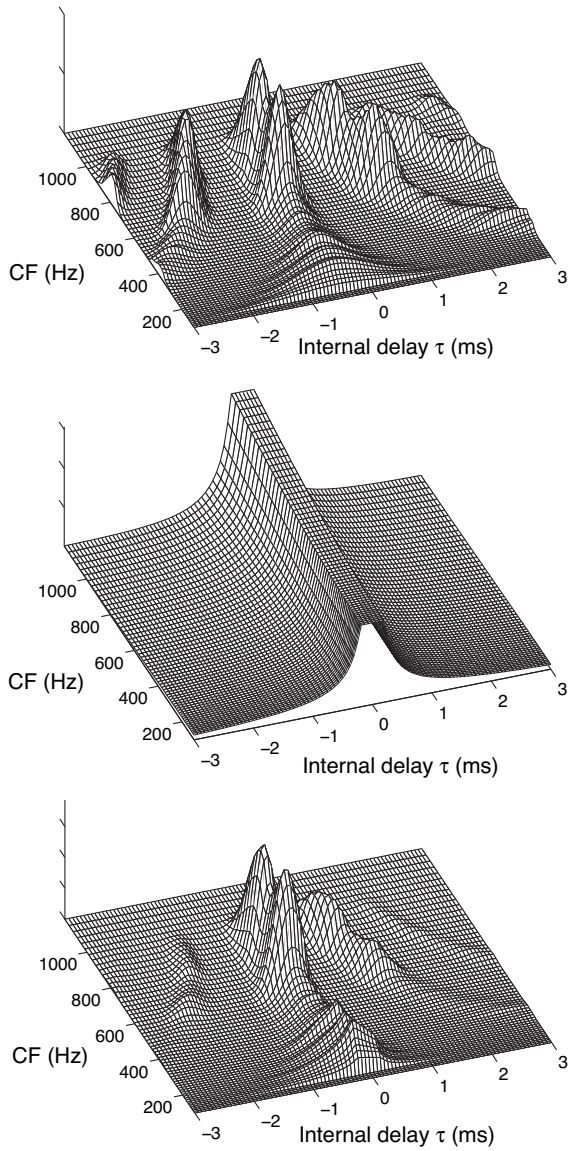

Figure 5.4 Representation of bandpass noise by the Jeffress-Colburn model. The upper panel of the figure shows the rate of response of individual Jeffress-Colburn coincidence detectors to bandpass noise with nominal corner frequencies of 100 and 1000 Hz, presented with an ITD of -0.5 ms. Shown are the relative rate of coincidences of a fiber pair, reported as a joint function of CF and internal delay. The central panel shows the putative distribution of the fiber pairs, again in terms of CF and internal delay [113]. The lowest panel shows the product of the two functions depicted in the upper two panels, which represents the total number of coincidence counts recorded in response to the signal.

signals through a bank of gammatone filters as realized in Slaney's auditory toolbox [108], followed by a simple half-wave square-law rectifier that squares positive input values and sets negative input values to zero. This model lacks many of the known complexities of the auditory-nerve response to sound. We note that maximum response is noted at internal delays equal to -0.5 ms over a broad range of frequencies, and that secondary maxima are observed at internal delays that are separated by integer multiples of the reciprocal of the CF for a particular fiber pair. Although these secondary ridges provide some ambiguity, it is clear that for natural stimuli the ITD can be inferred from the location of the 'straight' ridge that is consistent over frequency [114, 117]. The central panel of the figure depicts the function $p(\tau)$, which specifies the distribution of the coincidence counting units as a function of internal delay and CF. The specific function shown here was developed by Stern and Shear [113] and postulates a greater number of units with internal delays that are smaller in magnitude. The lower panel shows the total response of all the fiber pairs as a function of τ and f which reflects both the average response per fiber pair and the distribution per fiber pair; it is the product of the functions shown in the upper two panels.

5.5.3 Some extensions to cross-correlation-based binaural models

Extensions by Stern and Trahiotis. Stern and his colleagues (e.g., [110, 111, 114, 113, 117]) extended Colburn's model to describe the subjective lateral position of simple stimuli, and evaluated the extent to which the position cue could be used to account for performance in the largest possible set of psychoacoustical results in subjective lateralization, interaural discrimination, and binaural detection. Stern's extensions to Colburn's coincidence-counting mechanism include explicit assumptions concerning time-intensity interaction, and a mechanism for extracting subjective lateral position. These extensions of the Jeffress/Colburn model are referred to as the 'position-variable model' by Stern and his colleagues. In addition a second coincidence-based mechanism was proposed that serves to emphasize the impact of ITDs that are consistent over a range of frequencies [114].

One aspect of the model that received particular attention was the form of the function $p(\tau, f)$ that specifies the distribution of internal delays at each CF. The distribution function shown in the central panel of Fig. 5.4 was developed based on careful consideration of several key lateralization and detection results [113]. This function specifies a greater number of coincidence-counting units with internal interaural delays of smaller magnitude, which has been confirmed by physiological measurements (e.g., [66]). Nevertheless, a substantial fraction of the coincidence counters is assumed to have internal delays that are much greater in magnitude than the largest delays that are physically attainable with free-field stimuli. The existence of these very long internal delays is in accord with psychoacoustical as well as physiological data (see Sec. 5.2.2).

To account for the effects of IID, Stern and Colburn proposed that the representation of timing information shown in Fig. 5.4 be multiplied by a pulse-shaped weighting function with a location along the internal-delay axis that varied according to the IID of the stimulus. They further proposed that the subjective lateral position of a simple binaural stimulus could be predicted by the 'center of mass' along the internal-delay axis of the combined function that reflects both ITD and IID. More recently, Stern and Trahiotis [114] incorporated an additional modification to the model called 'straightness weighting' that is designed to emphasize the modes of the function that appear at the same internal delay over a range of CFs. This second-level mechanism which emphasizes ITDs that are consistent over a range of frequencies has the additional advantage of sharpening the ridges of the cross-correlation patterns in the Jeffress-Colburn model along the internal-delay axis. It should be noted that

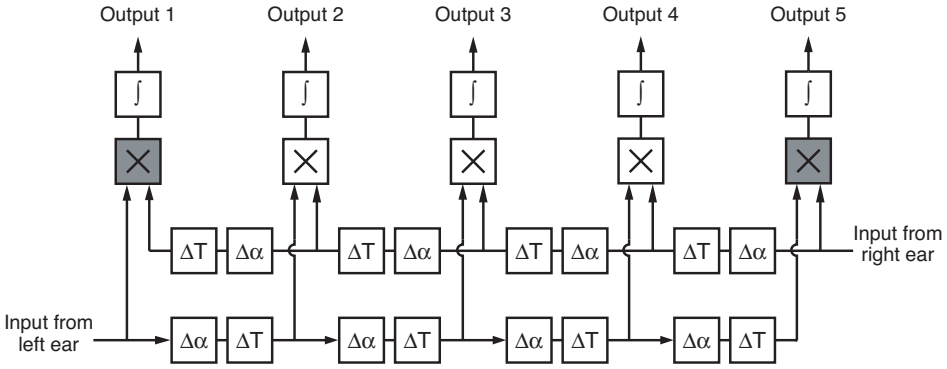

Figure 5.5 Schematic diagram of Lindemann's model. ΔT denotes a time delay and $\Delta\alpha$ denotes an attenuator. Boxes containing crosses are correlators (multipliers). At the two ends of the delay lines, the shaded boxes indicate correlators that are modified to function as monaural detectors. Adapted from Fig. 3(a) of [70] with the permission of the publisher.

Stern and his colleagues considered only the lateral position of simple stimuli such as pure tones, clicks, or broadband noise, and never carefully addressed the problem of the separate lateralization of individual sources when they are presented simultaneously.

Extensions by Blauert and his colleagues. Blauert and his colleagues have made important contributions to correlation-based models of binaural hearing over an extended period of time. Their efforts have been primarily directed toward understanding how the binaural system processes more complex sounds in real rooms and have tended to be computationally oriented. This approach is complementary to that of Colburn and his colleagues, who (at least in the early years) focussed on explaining 'classical' psychoacoustical phenomena using stimuli presented through earphones. More recently, Blauert's group worked to apply knowledge gleaned from fundamental research in binaural hearing toward the development of a 'cocktail party processor' which can identify, separate, and enhance individual sources of sound in the presence of other interfering sounds (e.g., [10]).

In a relatively early study, Blauert and Cobben [9] combined the running cross-correlator of Sayers and Cherry [97] with a model of the auditory periphery suggested by Duifhuis [36] that was similar to the model proposed by Siebert and adopted by Colburn. They subsequently developed a series of mechanisms that explicitly introduced the effects of stimulus IIDs into the modelling process. One of the most interesting and best known of these mechanisms was proposed by Lindemann [70], which may be regarded as an extension and elaboration of an earlier hypothesis of Blauert [6]. Lindemann extended the original Jeffress coincidence-counter model in two ways (see Fig. 5.5). Firstly, he added a mechanism that inhibits outputs of the coincidence counters when there is activity produced by coincidence counters at adjacent internal delays. Secondly, he introduced monaural-processing mechanisms at the 'edges' of the display of coincidence-counter outputs that become active when the intensity of the signal to one of the two ears is extremely small.

One of the properties of the Lindemann model is that the interaction of the inhibition mechanism and the monaural processing mechanism causes the locations of peaks of the coincidence-counter outputs along the internal-delay axis to shift with changes in IID. In other words, this model produces a time-intensity trading mechanism at the level of the coincidence-counter outputs. While the net effect of IIDs on the patterns of coincidence-counter outputs in the Lindemann model is not unlike the effect of the intensity-weighting

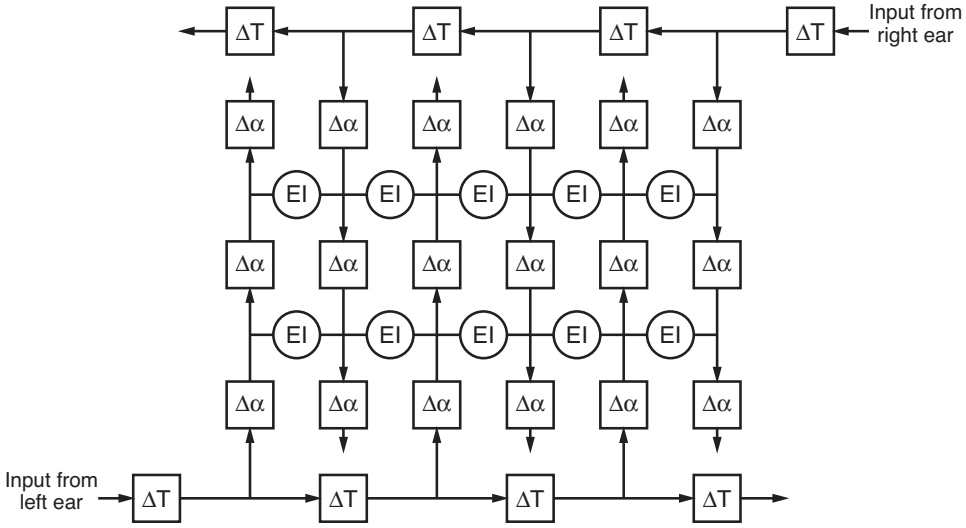

Figure 5.6 Schematic diagram of Breebaart's model. ΔT denotes a time delay and $\Delta\alpha$ denotes an attenuator. The circles denote excitation-inhibition (EI) elements. Redrawn from Fig. 3 of [14] with the permission of the publisher.

function in the model of Stern and Colburn [110], the time-intensity interaction of the Lindemann model arises more naturally from the fundamental assumptions of the model rather than as the result of the imposition of an arbitrary weighting function. In addition to the time-intensity trading properties, Lindemann also demonstrated that the contralateral-inhibition mechanism could also describe several interesting phenomena related to the precedence effect [71]. Finally, the inhibitory mechanisms of Lindemann's model produce a 'sharpening' of the peaks of the coincidence-counter outputs along the internal-delay axis, similar to that which is achieved by the second-level coincidence layer of Stern's model that was designed to emphasize ITDs that are consistent over frequency.

Gaik [46] extended the Lindemann mechanism further by adding a second weighting to the coincidence-counter outputs that reinforces naturally-occurring combinations of ITD and IID. This has the effect of causing physically-plausible stimuli to produce coincidence outputs with a single prominent peak that is compact along the internal-delay axis and that is consistent over frequency. Conversely, very unnatural combinations of ITD and IID (which tend to give rise to multiple spatial images) produce response patterns with more than one prominent peak along the internal-delay axis. The Blauert-Lindemann-Gaik model has been used as the basis for several computational models for systems that localize and separate simultaneously-presented sound sources (e.g., [10]).

Models that incorporate interaural signal cancellation. Although the coincidence-counting mechanism proposed by Jeffress and quantified by Colburn has dominated models of binaural interaction, the EC model developed by Durlach has retained its appeal for many years, and figures strongly in the conceptualization of data by Culling and Summerfield [30] and others (see Sec. 5.7). As noted above in Sec. 5.5.2, the Jeffress-Colburn model is loosely based on neural cells that are excited by inputs from both ears (called EE cells). Breebaart has recently proposed an elaboration of the Jeffress-Colburn model that includes an ab-

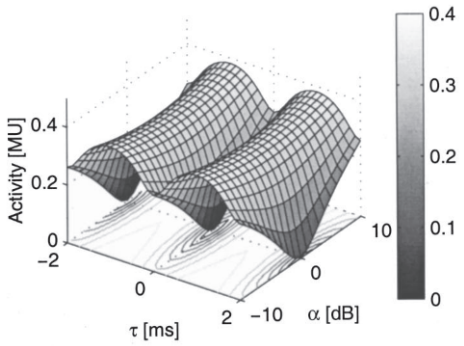

Figure 5.7 Activity of EI units in Breebaart's model, as a function of the characteristic IID (α) and ITD (τ) of each unit. The input was a 500 Hz tone presented diotically (i.e., with zero IID and ITD) and there is no internal noise. Reproduced from Fig. 5 of [14] with the permission of the publisher.

straction of cells that are excited by input from one ear but inhibited by input from the other ear (EI cells) [14, 15, 16]. The incorporation of EI units into binaural models provides an explicit mechanism that can in principle estimate the IID of a signal in a given frequency band, just as the EE-based mechanism of the Jeffress-Colburn model is typically used to provide an estimate of the ITD of a signal. In addition, the EI mechanism can provide the interaural cancellation in the EC model.

Figure 5.6 is a block diagram that summarizes the processing of Breebaart's model [14] that enables the simultaneous estimation of ITD and IID at a single frequency. In this diagram, units labeled $\Delta\alpha$ insert small attenuations to the signal path, just as the units labelled ΔT insert small time delays, as in the earlier models of Colburn, Blauert, and their colleagues. With this configuration, both the ITD and the IID of a signal component can be inferred by identifying which EI unit exhibits the *minimal* response. Specifically, the IID of the signal determines which *row* of EI units includes the minimal response, and the ITD of the signal determines which *columns* include the minimal response. Note that the response patterns will tend to repeat periodically along the horizontal axis, because the effective input to the network is narrowband after peripheral auditory filtering.

Figure 5.7 is an example of a response of such a network to a pure tone of 500 Hz presented with zero ITD and IID in the absence of internal noise. Note that the network exhibits a *minimum* of response at 0-μs ITD and 0-dB IID, with the minima repeating periodically along the internal delay axis with a period of 2 ms (the period of the 500-Hz tone). Since the pattern shifts horizontally with a change of ITD and vertically with a change of IID, this type of processing maintains an independent representation of ITD and IID for subsequent processing. In contrast, other models (e.g., [9, 46, 70, 110]) combine the effects of ITD and IID at or near the first stage of binaural interaction.

As Breebaart [14] notes, the predictions of this model will be very similar to the predictions of other models based solely on EE processing for many stimuli. Nevertheless, Breebaart also argues that predictions for EI-based models will differ from those of EE-based models in some respects. For example, Breebaart [14] argues that the dependence of binaural detection thresholds on target and masker duration is better described by EI-based processing. Furthermore, he argues that the parallel and independent availability of estimates of ITD and IID (as discussed in the previous paragraph) is likely to be useful for describing particular detection and discrimination results, and that EI-based models are

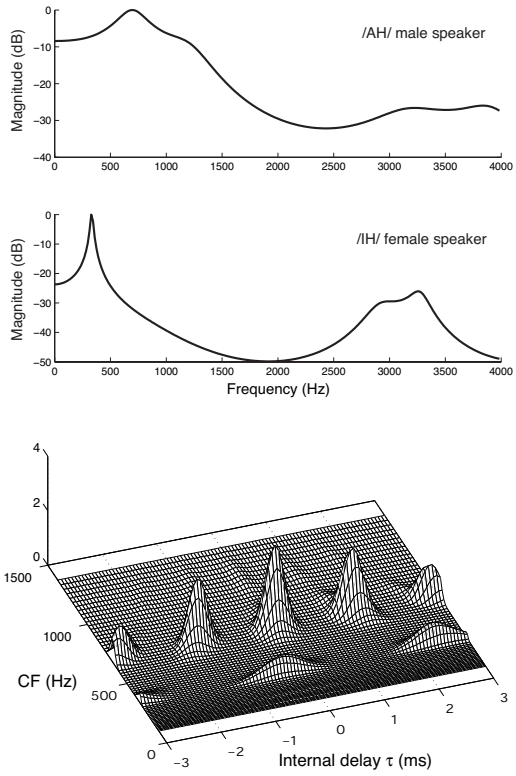

Figure 5.8 Upper and central panels: spectrum of the vowels /AH/ and /IH/ as recorded by a male and female speaker, respectively. Lower panel: response of an implementation of the Jeffress-Colburn model to the simultaneous presentation of the /AH/ presented with a 0-ms ITD and the /IH/ presented with a -0.5-ms ITD.

better able to describe the dependence of binaural detection experiments on overall stimulus level.

5.6 MULTISOURCE SOUND LOCALIZATION

We now consider the ways in which the models discussed above have been applied to describe signals containing multiple sound sources. In particular, we focus on the application of such models within CASA systems, where they often form the basis for localizing and segregating speech that has been contaminated by interfering sounds.

Although the following discussion of multisource localization is based on the type of binaural processing described in the previous section, it should be noted that the problem has been extensively studied in array signal processing (for reviews, see [121, 62]), where the main concern is direction-of-arrival estimation. For example, the classic MUSIC (MUltiple SIgnal Classification) algorithm [100] performs a principal component analysis on the array covariance matrix in order to separate a signal subspace and a noise subspace, which are used in a matrix operation that results in peak responses in the source directions. Although

accurate source localization can be obtained with a large sensor array, the utility of such techniques is limited when the array is limited to only two sensors, as in the case of the auditory system. The MUSIC algorithm, for instance, can localize only one sound source with two microphones. General multisource localization must also consider room reverberation, which introduces multiple reflections of every sound source and complicates the localization problem considerably. The problem of reverberation will be treated extensively in Chapter 7, and hence we do not deal with room reverberation in this section.

As has been described previously, the human binaural system appears to make use of several different types of information to localize and separate signals, including ITDs, IIDs, and the ITDs of the low-frequency envelopes of high-frequency stimulus components. Nevertheless, of these cues, ITD information typically dominates localization (at least at low frequencies) and has been the basis for all of the models described above. As a consequence, the primary stage of binaural interaction in multisource localization algorithms is a representation of interaural timing information based on either an EE-type mechanism as in the Jeffress-Colburn model or an EI-type mechanism as in the Breebaart model. This representation can be subsequently modified to reflect the impact of IIDs and possibly other types of information. Figure 5.8 illustrates how the Jeffress-Colburn mechanism can be used to localize two signals according to ITD. The upper two panels of the figure show the magnitude spectra in decibels of the vowels /AH/ and /IH/ spoken by a male and a female speaker, respectively. The lower panel shows the relative response of the binaural coincidence-counting units when these two vowels are presented simultaneously with ITDs of 0 and -0.5 ms, respectively. The 700-Hz first formant of the vowel /AH/ is clearly visible at the 0-ms internal delay, and the 300-Hz first formant of the vowel /IH/ is seen at the delay of -0.5 ms.

5.6.1 Estimating source azimuth from interaural cross-correlation

The first computational system for joint localization and source separation was proposed by Lyon in 1983 [76]. His system begins with a cochleagram of a sound mixture, which is essentially a representation that incorporates cochlear filtering and mechanical-to-neural transduction of the auditory nerve (see Sec. 1.3). The system subsequently computes the cross-correlation between the left and the right cochleagram responses, resulting in a representation similar to that shown in the lower panel of Fig. 5.8. Lyon termed this the *cross-correlogram*. He suggests summing the cross-correlation responses over all frequency channels, leading to a summary cross-correlogram in which prominent peaks indicate the ITDs of distinct sound sources. The idea of performing peak detection in the summary cross-correlogram for multisource localization has since been adopted in many subsequent systems, although Lyon's original study did not evaluate the idea in a systematic way.

A more systematic study of multisource localization was conducted by Bodden [10], again in the context of location-based source separation. The binaural processor used in Bodden's system follows the extension to Jeffress's model by Blauert and his colleagues as detailed in the previous section; in particular it incorporates contralateral inhibition and adapts to HRTFs. As a result, his model for localization is based not only on ITD but also on IID. His system analyzes the acoustic input using a filterbank with 24 channels, intended to simulate the critical bands in human hearing. A mixture of broadband signals such as speech may have large spectral overlap. In other words, different sources of this kind are usually not well separated in frequency as shown in Fig. 5.8. As a result, binaural responses to different sources interact in a complex way in the cross-correlogram, so that peaks in the

cross-correlogram no longer reliably indicate ITDs of individual sounds. Hence, spectral overlap between multiple sound sources creates a major difficulty in localization.

To deal with the problem of spectral overlap, the Bodden model incorporates a number of computational stages. First, the cross-correlation function within each frequency band is converted from an internal-delay axis to an azimuth axis. This conversion is performed in a supervised training stage using white noise presented at various arrival angles between -90° to 90° in the frontal horizontal plane. The mapping between peak positions on the cross-correlation axis to the corresponding azimuths is first established within each frequency band, and linear interpolation is used to complete the conversion. Bodden has observed some frequency dependency in the conversion, which is consistent with the observed frequency dependence of physical measurements of ITD by Shaw and others, as summarized in Eq. 5.1. To perform localization, the model sums converted cross-correlation patterns across different frequency channels. Rather than simply adding them together as is done in Lyon's model, the Bodden system introduces another supervised training stage in order to determine the relative importance of different bands; this is done in a similar way as in the conversion to the azimuth axis. This second training stage provides a weighting coefficient for each critical band, and the weighted sum across frequency is performed on the binaural responses in different frequency channels. To further enhance the reliability of multisource localization, his system also performs a running average across time within a short window (100 ms). These steps together result in a summary binaural pattern indexed by azimuth and running time.

A peak in the resulting summary pattern at a particular time is interpreted as a candidate for the azimuth of an acoustic event at that time. Bodden's model decides whether an azimuth candidate corresponds to a new event by tracking the time course of the amplitude of the summary binaural pattern – a new event should have an accompanying amplitude increase. Results on two-source localization were reported, and when the two sources are well separated in azimuth, the model gives good results. It should be noted that the sound sources used in Bodden's experiments were combined digitally from recordings of the individual sources in isolation (Bodden, personal communication); it is more difficult to obtain comparable improvements using sound sources recorded in a natural environment, especially if reverberation is a factor. This concern also applies to a number of subsequent studies (e.g., [94]).

5.6.2 Methods for resolving azimuth ambiguity

When the azimuths of sound sources are not far apart, the Bodden model has difficulty in resolving them separately because, as discussed earlier, different sources may interact within individual frequency channels to produce misleading peaks in the binaural pattern. For example, two sounds may interact to produce a broad cross-correlogram peak that indicates a ghost azimuth in the middle of the two true azimuths. Several methods have been proposed to sharpen the cross-correlation modes along the internal-delay axis and resolve ambiguities in source azimuth, including the previously-mentioned second level of coincidences across frequency proposed by Stern and Trahiotis [114, 116] and the contralateral-inhibition mechanism proposed by Lindemann [70, 71].

One approach that has been effective in computational systems for source separation is the computation of a 'skeleton' cross-correlogram [88, 94], which is motivated by the observation that peaks in a cross-correlogram function are too broadly tuned to resolve small differences in azimuth (see Fig. 5.8). The basic idea is to replace the peaks in a cross-correlogram response by a Gaussian function with a narrower width. Specifically, each local

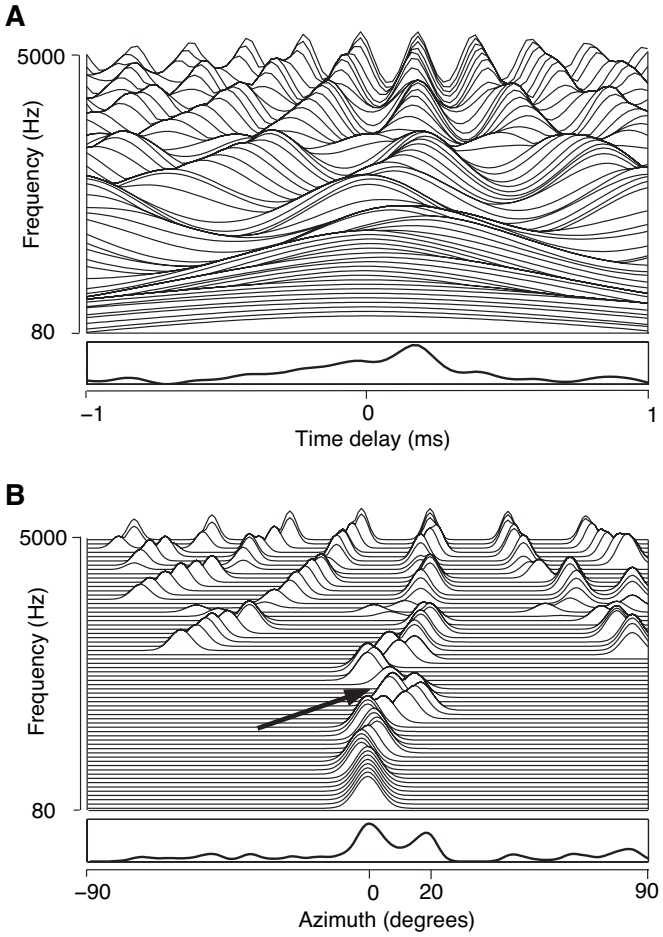

Figure 5.9 Azimuth localization for a mixture of two utterances: a male utterance presented at 0° and a female utterance at 20°. A. Cross-correlogram for a time frame 400 ms after stimulus onset. The lower panel shows the summary cross-correlogram. B. Skeleton cross-correlogram for the same time frame. The lower panel shows a summary plot along the azimuth axis. The arrow points to channels that contain roughly equal energy from the target and the interference. Reproduced with permission from [94].

peak in the cross-correlogram is reduced to an impulse of the same height. The resulting impulse train is then convolved with a Gaussian, whose width is inversely proportional to the center frequency of the corresponding filter channel. The resulting summary skeleton cross-correlogram is more sharply peaked, and therefore represents multiple azimuths more distinctly.

Figure 5.9 illustrates the utility of the skeleton cross-correlogram for a mixture of a male speech utterance presented at 0° and a female utterance presented at 20°. Fig. 5.9A shows the cross-correlogram for a time window of 20 ms in response to the stimulus, where a 64-channel gammatone filterbank with filter center frequency ranging from 80 Hz to 5 kHz is used to perform peripheral analysis. The cross-correlogram is shown for the range [-1 ms, 1 ms]. For frequency bands where energy from one source is dominant, the binaural response

indicates the ITD of the true azimuth. On the other hand, for channels where the energy from each source is about the same, peaks in the binaural response deviate from the true ITDs. In this case, the summary cross-correlogram shown in the lower panel of Fig. 5.9A does not provide an adequate basis for resolving the two azimuths. Figure 5.9B shows the corresponding skeleton cross-correlogram with sharper peaks (note that a conversion from the delay axis to the azimuth axis has been performed in this figure). Integrating over frequency yields a summary response that clearly indicates the two underlying azimuths, as shown in the lower panel of Fig. 5.9B. A skeleton cross-correlogram could be produced by applying a lateral inhibition mechanism to a conventional cross-correlogram; lateral inhibition should be contrasted with contralateral inhibition used in Lindemann's model [70].

Motivated by the psychoacoustical evidence that human performance in localizing multiple sources is enhanced when the sources have different onset times, Braasch [12] introduced a model, called interaural cross-correlation difference, for localizing a target source in the presence of background noise. When the target and the background noise are uncorrelated, the cross-correlogram of the mixture is the sum of the cross-correlogram of the target and that of the noise. If the cross-correlogram of the background is known, it can then be used to subtract from the mixture cross-correlogram, producing the cross-correlogram of the target alone. As a result, target-localization performance is better than that which is obtained using the traditional cross-correlogram. The basic idea behind Braasch's difference model is similar to that of spectral subtraction commonly used in speech enhancement (see Sec. 1.5). Consequently, we can expect that Braasch's model will work reasonably well for signals that are in the presence of stationary background noise, but not so well when the interference is nonstationary in nature, as in the case of a competing talker.

It is well known that the energy distribution of a speech utterance varies a great deal with respect to time and frequency. Hence, for a mixture of speech sources – or any other sources with such characteristic such as musical sounds – a single source tends to dominate the responses within individual time-frequency units, each corresponding to a specific frequency channel and a specific time frame. This property has been well established in earlier parts of this book (see Chapter 1 and Chapter 4). Figure 5.8 illustrates such a situation for a mixture of two concurrent vowels. Because the two vowels show energy peaks in different spectral regions, they are well separated in the binaural pattern shown in the lower panel of the figure. Capitalizing on this observation, Faller and Merimaa [41] recently proposed a method to automatically select the time-frequency (T-F) units dominated by a single source. The ITD and IID cues provided by such units resemble those elicited by a single source. Faller and Merimaa assume that only reliable binaural cues derived from such units are fed to the later stages of the auditory pathway.

Faller and Merimaa [41] propose an interaural coherence (IC) measure to select reliable T-F units. The IC is estimated from a normalized cross-correlation function, which may be defined for a specific frequency channel at time t as

$$\hat{R}(t,\tau) = \frac{\int_{-\infty}^{t} x_L(\alpha) x_R(\alpha-\tau) w(t-\alpha) d\alpha}{\sqrt{\int_{-\infty}^{t} x_L^2(\alpha) w(t-\alpha) d\alpha} \sqrt{\int_{-\infty}^{t} x_R^2(\alpha-t) w(t-\alpha) d\alpha}} \tag{5.5}$$

The notations in the above equation are the same as those in Eq. 5.2. $\hat{R}(t,\tau)$ is evaluated in the range [-1 ms, 1 ms]. The IC at time t is then given by

$$IC(t) = \max_{\tau} \hat{R}(t,\tau) \tag{5.6}$$

This measure of interaural coherence, which in many ways may be regarded as an amplification and implementation of an earlier approach by Allen *et al.* [1], is effective for the following reasons. When a single source dominates a frequency band at a particular time, the left and the right signal will be similar except for an ITD, which leads to a high IC value. On the other hand, if multiple sources have significant energy within the same time-frequency unit, the left and the right signals will be incoherent, leading to a low IC value. In the Faller and Merimaa model, binaural cues from T-F units with high IC values are retained, and the cues from other T-F units are discarded. Their evaluation results show that the selection of binaural cues based on interaural coherence yields sharper peaks in a joint ITD-IID feature space, which implies more robust localization of multiple sources. Faller and Merimaa have also evaluated their method on localizing reverberant sounds, and more discussion on this will be given in Chapter 7.

A major cause of azimuth ambiguity is the occurrence of multiple peaks in the high-frequency range of the cross-correlogram in response to a broadband source (e.g. speech), as shown in Fig. 5.9A. This is because at high frequencies the wavelength is shorter than the physical distance between the two ears. When multiple sources are active, their interaction may lead to a summary cross-correlogram in which the peaks no longer reflect the true ITDs, as discussed by Stern *et al.* [114, 117]. Liu *et al.* [73] divide the peaks into those on the primary trace that coincides with the true ITDs and those on the secondary traces that do not indicate true ITDs. Furthermore, they observe that the primary and secondary traces form a characteristic pattern on the cross-correlogram and devise a template matching method that integrates peaks across frequency on the basis of the characteristic correlation pattern. They call their spectral integration method the 'stencil' filter, and have reported good results for localizing up to four speech sources. A more detailed description of the stencil filter will be given in Chapter 6, in the context of location-based sound separation.

5.6.3 Localization of moving sources

Sound localization in real-world environments must consider the movement of sound sources. For example, a speaker often turns his or her head and walks while talking. Source movement introduces yet another dimension of complexity; in particular, it limits the window length for temporal integration, which has proven to be very useful for accurate localization. Few studies have addressed this important problem. Bodden [10] made an attempt to test his model on localizing two moving sources that cross each other in their azimuthal paths. However, the two sources in his evaluation are active alternatively with little temporal overlap, presumably because of the inability of the model to deal with simultaneously active sources.

Roman and Wang [93] recently presented a binaural method for tracking multiple moving sources. They extend a hidden Markov model (HMM) for multipitch tracking [127] to the domain of multisource localization and tracking. Specifically, they extract ITD and IID cues and integrate these cues across different frequency channels to produce a likelihood function in the target space. The HMM is subsequently employed to form continuous azimuth tracks and detect the number of active sources across time. Their model combines instantaneous binaural information and an explicit dynamic model that simulates the motion trajectory of a sound source. They have reported good results for tracking three active sources whose azimuth paths may cross each other. A further evaluation shows a favorable comparison with a Kalman-filter based approach [92].

5.7 GENERAL DISCUSSION

The ability to localize sound sources is a crucial aspect of auditory perception; it creates a sense of space for the listener, and provides information about the spatial location of objects in the environment which can be critical for survival. The psychoacoustical studies reviewed here reveal that human binaural hearing can accurately localize sounds, particularly in the horizontal median plane, and that this capacity remains largely intact in the presence of other interfering sounds. Reproducing this ability in computational systems is a significant challenge, and an important one for the development of CASA systems.

Binaural sound localization is largely based on comparisons between the signals arriving at the two ears. Several cues contribute in this respect, including ITD, IID, as well as the ITD of the low-frequency envelopes of high-frequency components of the signals. Cues related to frequency-dependent filtering by the pinnae also play a role, particularly for localization in the vertical plane and for resolving front-back ambiguities. Although ITD appears to play a dominant role in the determination of sound azimuth, it is clear that all of these cues interact in complex ways. At present, these interactions are not well understood; similarly, there is much that we do not know about how binaural cues are employed to process multiple sound sources.

The Jeffress coincidence mechanism [59], postulated almost 60 years ago, has dominated computational modeling of sound localization. However, the Jeffress scheme and its variants (see Sec. 5.5.2) were principally developed to explain the localization of simple stimuli such as a single sound source in an anechoic environment. While these models have been successful in modeling the binaural perception of isolated sounds, there have been relatively few attempts to quantitatively evaluate them using multiple sound sources. Nonetheless, the key representation in Jeffress' scheme – referred to by some as the cross-correlogram – has been used as the basis for a number of multisource localization and sound separation systems. Within such systems, integration of cross-correlation responses across frequency and time is necessary for good localization performance. Computational efforts in multisource localization have mostly concentrated on how to resolve azimuth ambiguity introduced by the presence of multiple sound sources.

Spatial location is an important cue for auditory scene analysis (see Sec. 1.1 and Bregman [17]), and many computational studies have addressed the problem of location-based grouping – the topic of the next chapter. In such studies, sound localization is often considered to be a prerequisite step. This, plus the fact that spectro-temporal integration is usually needed for accurate localization, raises an interesting conceptual issue of whether sound separation depends on localization or sound localization depends on separation. Experimental evidence discussed in Sec. 5.4.1 suggests that a target sound in background noise that is clearly audible is also easily localizable. This indicates a close link between sound localization and separation.

On the one hand, the robust MLD effect due to spatial separation (see Sec. 5.4.2) and other binaural advantages in sound separation would strongly imply the contribution of location-based grouping to auditory scene analysis. On the other hand, it is well known that binaural cues such as ITD and IID are susceptible to intrusion by background noise and room reverberation (see Chapter 7). Such intrusions make reliable estimates of binaural cues very difficult, if not impossible, in local time-frequency regions.

Indeed, the notion that the human auditory system groups sources according to ITD is far from universally accepted. For example, the results of several experiments using earphones indicate that listeners are unable to achieve separate identification of simultaneously-presented vowel-like bandpass-noise sounds solely on the basis of their ITDs (e.g., [30, 31,

57]). Culling and Summerfield (and others) believe that these results indicate that the human auditory system does not make use of ITD information in achieving an initial grouping of signal components according to sound source. They suggest that signal components are extracted independently for each frequency using an interaural cancellation mechanism similar to that used in the EC model. In contrast, Stern *et al.* have found that identification according to ITD is easily accomplished if similar noise bands are modulated in amplitude or frequency in a fashion that is similar to the natural modulations of speech signals [91, 112]. One plausible interpretation of these results is that the human auditory system first groups signal components according to common modulation in amplitude and/or frequency (as well as other information such as the harmonicity of periodic signal components), and that ITD information is then used to segregate the grouped components, perhaps abetted by IID information to provide additional help in interpreting ambiguous inputs. Certainly, sound localization and sound separation are likely to interact in ways that we do not yet fully understand.

While the question of how human listeners utilize ITDs to segregate sounds remains an interesting one, it is not necessary for a CASA system to mimic every aspect of auditory processing. As a practical matter, signal segregation is generally more easily achieved computationally on the basis of ITD information than it is based on common amplitude or frequency modulation, at least for anechoic signals. Hence, the use of ITD information as the basis for signal separation remains appealing for computational systems.

Future research in multisource localization will need to address the issue of room reverberation, which few computational studies have tackled directly (see Chapter 7). More consideration also needs to be given to moving sound sources, which frequently occur in natural acoustic scenes. Background interference, reverberation, and motion conspire to create a tremendous problem for computational sound localization systems that operate in real-world environments, and they will present a challenge for many years to come.

ACKNOWLEDGMENTS

Preparation of this chapter was supported by the National Science Foundation (Grant IIS-0420866) for Richard Stern. Guy Brown was supported by EPSRC grant GR/R47400/01, and DeLiang Wang was supported by an AFOSR grant (FA9550-04-01-0117) and an AFRL grant (FA8750-04-1-0093). The first author is especially grateful for many discussions over many years with Steve Colburn and Tino Trahiotis, which have been very instrumental in shaping his attitudes about many theoretical and experimental aspects of binaural hearing. Substantial portions of the discussion about binaural models in this chapter are based on previous chapters written with Trahiotis [115, 116].

REFERENCES

1. J. B. Allen, D. A. Berkley, and J. Blauert. Multimicrophone signal-processing technique to remove room reverberation from speech signals. *Journal of the Acoustical Society of America*, 62(4):912–915, 1977.
2. D. Batteau. The role of the pinna in human localization. *Proc R Soc London B Biol. Sci.*, 168:158–180, 1967.
3. L. R. Bernstein and C. Trahiotis. Lateralization of low-frequency complex waveforms: the use of envelope-based temporal disparities. *Journal of the Acoustical Society of America*, 77:1868–1880, 1985.

4. L. R. Bernstein and C. Trahiotis. Detection of interaural delay in high-frequency sinusoidally amplitude-modulated tones, two-tone complexes, and bands of noise. *Journal of the Acoustical Society of America*, 95:3561–3567, 1994.

5. L. R. Bernstein and C. Trahiotis. Enhancing sensitivity to interaural delays at high frequencies by using "transposed stimuli". *Journal of the Acoustical Society of America*, 112:1026–1036, 2002.

6. J. Blauert. Modeling of interaural time and intensity difference discrimination. In G. van den Brink and F. Bilsen, editors, *Psychophysical, Physiological, and Behavioural Studies in Hearing*, pages 412–424. Delft University Press, Delft, 1980.

7. J. Blauert. Introduction to binaural technology. In R. H. Gilkey and T. R. Anderson, editors, *Binaural and Spatial Hearing in Real and Virtual Environments*, chapter 28, pages 593–610. Lawrence Erlbaum Associates, Mahwah, NJ, 1997.

8. J. Blauert. *Spatial Hearing*. MIT Press, Cambridge, MA, 1997. revised edition.

9. J. Blauert and W. Cobben. Some considerations of binaural cross-correlation analysis. *Acustica*, 39:96–103, 1978.

10. M. Bodden. Modelling human sound-source localization and the cocktail party effect. *Acta Acustica*, 1:43–55, 1993.

11. J. C. Boudreau and C. Tsuchitani. Binaural interaction in the cat superior olive S segment. *J. Neurophysiol.*, 31:442–454, 1968.

12. J. Braasch. Localization in the presence of a distractor and reverberation in the frontal horizontal plane: II. model algorithms. *Acustica/Acta Acustica*, 88:956–969, 2002.

13. J. Braasch. Modelling of binaural hearing. In J. Blauert, editor, *Communication Acoustics*, chapter 4, pages 75–108. Springer-Verlag, Berlin, 2005.

14. J. Breebaart, S. van de Par, and A. Kohlrausch. Binaural processing model based on contralateral inhibition. I. Model structure. *Journal of the Acoustical Society of America*, 110:1074–1088, 2001.

15. J. Breebaart, S. van de Par, and A. Kohlrausch. Binaural processing model based on contralateral inhibition. II. Dependence on spectral parameters. *Journal of the Acoustical Society of America*, 110:1089–1103, 2001.

16. J. Breebaart, S. van de Par, and A. Kohlrausch. Binaural processing model based on contralateral inhibition. III. Dependence on temporal parameters. *Journal of the Acoustical Society of America*, 110:1125–1117, 2001.

17. A. S. Bregman. *Auditory scene analysis*. MIT Press, Cambridge, MA, 1990.

18. M. D. Burkhard and R. M. Sachs. Anthroponetric manikin for acoustic research. *Journal of the Acoustical Society of America*, 58(1):214–222, 1974.

19. D. Caird and R. Klinke. Processing of binaural stimuli by cat superior olivary complex neurons. *Exp. Brain Res.*, 52:385–399, 1983.

20. R. Carhart. Monaural and binaural discrimination against competing sentences. *Int. Audiol.*, 1:5–10, 1965.

21. R. Carhart, T. W. Tillman, and K. R. Johnson. Release from masking for speech through interaural time delay. *Journal of the Acoustical Society of America*, 42:124–138, 1967.

22. C. E. Carr and M. Konishi. Axonal delay lines for time measurement in the owl's brainstem. *Proc. natl. Acad. Sci. USA*, 85:8311–8315, 1988.

23. C. Cherry. Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America*, 26:554–559, 1953.

24. E. C. Cherry. Two ears – but one world. In W. A. Rosenblith, editor, *Sensory Communication*, pages 99–117. MIT Press, Cambridge MA, 1961.

25. H. S. Colburn. Theory of binaural interaction based on auditory-nerve data. I. general strategy and preliminary results on interaural discrimination. *Journal of the Acoustical Society of America*, 54:1458–1470, 1973.

26. H. S. Colburn. Theory of binaural interaction based on auditory-nerve data. II. detection of tones in noise. *Journal of the Acoustical Society of America*, 61:525–533, 1977.

27. H. S. Colburn. Computational models of binaural processing. In H. Hawkins and T. McMullen, editors, *Auditory Computation*, Springer Handbook of Auditory Research, chapter 9, pages 332–400. Springer-Verlag (New York), 1996.

28. H. S. Colburn and N. I. Durlach. Models of binaural interaction. In E. C. Carterette and M. P. Friedmann, editors, *Hearing*, volume IV of *Handbook of Perception*, chapter 11, pages 467–518. Academic Press, New York, 1978.

29. H. S. Colburn and A. Kulkarni. Models of sound localization. In R. Fay and T. Popper, editors, *Sound Source Localization*, Springer Handbook of Auditory Research, chapter 8, pages 272–316. Springer-Verlag, 2005.

30. J. F. Culling and Q. Summerfield. Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay. *Journal of the Acoustical Society of America*, 98(2):785–797, 1995.

31. C. J. Darwin and R. W. Hukin. Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity. *Journal of the Acoustical Society of America*, 102(4):2316–2324, 1997.

32. E. E. David, N. Guttman, and W. A. van Bergeijk. On the mechanism of binaural fusion. *Journal of the Acoustical Society of America*, 30:801–802, 1958.

33. D. D. Dirks and R. H. Wilson. The effect of spatially-separated sound sources. *J. Speech Hearing Res.*, 12:5–38, 1969.

34. R. H. Domnitz and H. S. Colburn. Lateral position and interaural discrimination. *Journal of the Acoustical Society of America*, 61:1586–1598, 1977.

35. R. Drullman and A. W. Bronkhorst. Multichannel speech intelligibility and speaker recognition using monaural, binaural and 3D auditory presentation. *Journal of the Acoustical Society of America*, 107:2224–2235, 2000.

36. H. Duifhuis. Consequences of peripheral frequency selectivity for nonsimultaneous masking. *Journal of the Acoustical Society of America*, 54:1471–1488, 1973.

37. N. I. Durlach. Equalization and cancellation theory of binaural masking level differences. *Journal of the Acoustical Society of America*, 35(8):1206–1218, 1963.

38. N. I. Durlach. Binaural signal detection: Equalization and cancellation theory. In J. V. Tobias, editor, *Foundations of Modern Auditory Theory*, volume 2, pages 369–462. Academic Press, New York, 1972.

39. N. I. Durlach and H. S. Colburn. Binaural phenomena. In E. C. Carterette and M. P. Friedman, editors, *Hearing*, volume IV of *Handbook of Perception*, chapter 10, pages 365–466. Academic Press, New York, 1978.

40. N. I. Durlach, K. J. Gabriel, H. S. Colburn, and C. Trahiotis. Interaural correlation discrimination: Ii. relation to binaural unmasking. *Journal of the Acoustical Society of America*, 79:1548–1557, 1986.

41. C. Faller and J. Merimaa. Sound localization in complex listening situations: Selection of binaural cues based on interaural coherence. *Journal of the Acoustical Society of America*, 116(5):3075–3089, 2004.

42. H. Fletcher and R. H. Galt. The perception of speech and its relation to telephony. *Journal of the Acoustical Society of America*, 22:89–151, 1950.

43. S. H. Foster, E. M. Wenzel, and R. M. Taylor. Real time synthesis of complex acoustic environments. In *Proceedings of the ASSP (IEEE) Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 1991.

44. N. R. French and J. C. Steinberg. Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America*, 19:90–119, 1947.

45. K. J. Gabriel and H. S. Colburn. Interaural correlation discrimination: I. bandwidth and level dependence. *Journal of the Acoustical Society of America*, 69:1394–1401, 1981.

46. W. Gaik. Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling. *Journal of the Acoustical Society of America*, 94:98–110, 1993.

47. B. Gardner and K. Martin. HRTF measurements of a KEMAR dummy-head microphone. Technical Report 280, MIT Media Lab Perceptual Computing Group, 1994. Available online at http://sound.media.mit.edu/KEMAR.html.

48. M. D. Good and R. H. Gilkey. Sound localization in noise: the effect of signal-to-noise ratio. *Journal of the Acoustical Society of America*, 99:1108–1117, 1996.

49. G. G. Harris. Binaural interactions of impulsive stimuli and pure tones. *Journal of the Acoustical Society of America*, 32:685–692, 1960.

50. M. L. Hawley, R. Y. Litovsky, and H. S. Colburn. Speech intelligibility and localization in a multi-source environment. *Journal of the Acoustical Society of America*, 105:3436–3448, 1999.

51. J. Hebrank and D. Wright. Spectral cues used in the localization of sound sources in the median plane. *Journal of the Acoustical Society of America*, 56:1829–1834, 1974.

52. G. B. Henning. Detectability of interaural delay in high-frequency complex waveforms. *Journal of the Acoustical Society of America*, 55:84–90, 1974.

53. R. M. Hershkowitz and N. I. Durlach. Interaural time and amplitude jnds for a 500-hz tone. *Journal of the Acoustical Society of America*, 46:1464–1467, 1969.

54. I. J. Hirsch. The influence of interaural phase on interaural summation and inhibition. *Journal of the Acoustical Society of America*, 20:150–159, 1948.

55. I. J. Hirsch. Relation between localization and intelligibility. *Journal of the Acoustical Society of America*, 22:196–200, 1950.

56. P. M. Hofman and A. J. van Opstal. Spectro-temporal factors in two-dimensional human sound localization. *Journal of the Acoustical Society of America*, 103:2634–2648, 1998.

57. R. W. Hukin and C. J. Darwin. Effects of contralateral presentation and of interaural time differences in segregating a harmonic from a vowel. *J Acoust. Soc. Amer.*, 98:1380–1387, 1995.

58. T. Jacobsen. Localization in noise. Technical Report 10, Technical University of Denmark Acoustics Laboratory, 1976.

59. L. A. Jeffress. A place theory of sound localization. *J. Comp. Physiol. Psych.*, 41:35–39, 1948.

60. W. E. Kock. Binaural localization and masking. *Journal of the Acoustical Society of America*, 22:801–804, 1950.

61. W. Koenig. Subjective effects in binaural hearing. *Journal of the Acoustical Society of America*, 22:61–62(L), 1950.

62. H. Krim and M. Viberg. Two decades of array signal processing research: The parametric approach. *IEEE Signal Processing Magazine*, 13:67–94, 1996.

63. K. D. Kryter. Methods for the calculation and use of the articulation index. *Journal of the Acoustical Society of America*, 34:1689–1697, 1962.

64. G. F. Kuhn. Model for the interaural time differences in the azimuthal plane. *Journal of the Acoustical Society of America*, 62:157–167, 1977.

65. G. F. Kuhn. Physical acoustics and measurements pertaining to directional hearing. In W. A. Yost and G. Gourevitch, editors, *Directional Hearing*. Springer-Verlag, 1987.

66. S. Kuwada, R. Batra, and D. C. Fitzpatrick. Neural processing of binaural temporal cues. In R. H. Gilkey and T. R. Anderson, editors, *Binaural and Spatial Hearing in Real and Virtual Environments*, chapter 20, pages 399–425. Lawrence Erlbaum Associates (Mahwah, New Jersey), 1997.

67. E. H. A. Langendijk, D. J. Kistler, and F. L. Wightman. Sound localization in the presence of one or two distractors. *Journal of the Acoustical Society of America*, 109:2123–2134, 2001.

68. H. Levitt and L. R. Rabiner. Predicting binaural gain in intelligibility and release from masking for speech. *Journal of the Acoustical Society of America*, 42:820–829, 1967.

69. J. C. R. Licklider. Three auditory theories. In S. Koch, editor, *Psychology: A Study of a Science*, pages 41–144. McGraw-Hill, New York, 1959.

70. W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition. I. simulation of lateralization for stationary signals. *Journal of the Acoustical Society of America*, 80:1608–1622, 1986.

71. W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition. II. the law of the first wavefront. *Journal of the Acoustical Society of America*, 80:1623–1630, 1986.

72. R. Y. Litovsky, S. H. Colburn, W. A. Yost, and S. J. Guzman. The precedence effect. *Journal of the Acoustical Society of America*, 106:1633–1654, 1999.

73. C. Liu, B. C. Wheeler, W. D. O'Brien, R. C. Bilger, C. R. Lansing, and A. S. Feng. Localization of multiple sound sources with two microphones. *Journal of the Acoustical Society of America*, 108(4):1888–1905, 2000.

74. C. Lorenzi, S. Gatehouse, and C. Lever. Sound localization in noise in normal-hearing listeners. *Journal of the Acoustical Society of America*, 105:1810–1820, 1999.

75. R. F. Lyon. A computational model of filtering, detection and compression in the cochlea. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1282–1285, Paris, May 1982.

76. R. F. Lyon. A computational model of binaural localization and separation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 1148–1151, 1983.

77. N. W. MacKeith and R. R. A. Coles. Binaural advantages in hearing of speech. *J. Laryng. and Otol.*, 85:213–232, 1985.

78. D. McAlpine, D. Jiang, and A. R. Palmer. Interaural delay sensitivity and the classification of low best-frequency binaural responses in the inferior colliculus of the guinea pig. *Hearing Research*, 97:136–152, 1996.

79. D. McFadden and E. G. Pasanen. Lateralization of high frequencies based on interaural time differences. *Journal of the Acoustical Society of America*, 59:634–639, 1976.

80. R. Meddis and M. J. Hewitt. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i. pitch identification. *Journal of the Acoustical Society of America*, 89(6):2866–2882, 1991.

81. S. Mehrgardt and V. Mellert. Transformation charactersitics of the external human ear. *Journal of the Acoustical Society of America*, 61:1567–1576, 1977.

82. J. C. Middlebrooks, J. C. Makous, and D. M. Green. Directional sensitivity of sound-pressure levels in the human ear canal. *Journal of the Acoustical Society of America*, 86:89–108, 1989.

83. P. Minnaar, S. K. Olesen, F. Christensen, and H. Møller. Localization with binaural recordings from artificial and human heads. *Journal of the Audio Engineering Society*, 49(5):323–336, May 2001.

84. H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi. Binaural technique: do we need individual recordings? *Journal of the Audio Engineering Society*, 44:451–469, June 1996.

85. B. C. J. Moore. *An introduction to the psychology of hearing*. Academic Press, London, fifth edition, 2003.

86. J. W. Strutt (Third Baron of Rayleigh). On out perception of sound direction. *Philosoph. Mag.*, 13:214–232, 1907.

87. A. R. Palmer. Neural signal processing. In B. C. J. Moore, editor, *Hearing*, Handbook of Perception and Cognition, chapter 3, pages 75–121. Academic (New York), 1995.

88. K. J. Palomäki, G. J. Brown, and D. L. Wang. A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. *Speech Communication*, 43(4):361–378, 2004.

89. R. D. Patterson, M. H. Allerhand, and C. Giguere. Time-domain modeling of peripheral auditory processing: a modular architecture and a software platform. *Journal of the Acoustical Society of America*, 98:1980–1984, 1995.

90. I. Pollack and W. J. Trittipoe. Binaural listening and interaural cross correlation. *Journal of the Acoustical Society of America*, 31:1250–1252, 1959.

91. A. M. Ripepi. Lateralization and identification of simultaneously-presented whispered vowels and speech sounds. Master's thesis, Carnegie Mellon University, Pittsburgh, PA, May 1999.

92. N. Roman. *Auditory-based algorithms for sound segregation in multisource and reverberant environments*. PhD thesis, Ohio State University Department of Computer Science and Engineering, 2005. Available at http://www.cse.ohio-state.edu/pnl/theses.html.

93. N. Roman and D. L. Wang. Binaural tracking of multiple moving sources. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume V, pages 149–152, 2003.

94. N. Roman, D. L. Wang, and G. J. Brown. Speech segregation based on sound localization. *Journal of the Acoustical Society of America*, 114(4):2236–2252, 2003.

95. J. E. Rose, N. B. Gross, C. D. Geisler, and J. E. Hind. Some neural mechanisms in the inferior colliculus of the cat which may be relevant to localization of a sound source. *J. Neurophysiol.*, 29:288–314, 1966.

96. B. McA. Sayers. Acoustic-image lateralization judgments with binaural tones. *Journal of the Acoustical Society of America*, 36:923–926, 1964.

97. B. McA. Sayers and E. C. Cherry. Mechanism of binaural fusion in the hearing of speech. *Journal of the Acoustical Society of America*, 36:923–926, 1957.

98. B. McA. Sayers and P. A. Lynn. Interaural amplitude effects in binaural hearing. *Journal of the Acoustical Society of America*, 44:973–978, 1968.

99. B. McA. Sayers and F. E. Toole. Acoustic-image judgments with binaural transients. *Journal of the Acoustical Society of America*, 36:1199–1205, 1964.

100. R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34:276–280, 1986.

101. M. R. Schroeder. New viewpoints in binaural interactions. In E. F. Evans and J. P. Wilson, editors, *Psychophysics and Physiology of Hearing*, pages 455–467. Academic Press (London), 1977.

102. C. L. Searle, L. D. Braida, D. R. Cuddy, and M. F. Davis. Binaural pinna disparity: another auditory localization cue. *Journal of the Acoustical Society of America*, 57:448–455, 1975.

103. S. Seneff. A joint synchrony/mean-rate model of auditory speech processing. *J. Phonetics*, 15:55–76, 1988.

104. S. A. Shamma, N. Shen, and P. Gopalaswamy. Binaural processing without neural delays. *Journal of the Acoustical Society of America*, 86:987–1006, 1989.

105. E. A. G. Shaw. Transformation of sound pressure level from the free fild to the ear drum in the horizontal plane. *Journal of the Acoustical Society of America*, 39(465–470), 1974.

106. E. A. G. Shaw. Acoustical features of the human external ear. In R. H. Gilkey and T. R. Anderson, editors, *Binaural and Spatial Hearing in Real and Virtual Environments*, chapter 2, pages 25–47. Lawrence Erlbaum Associates (Mahwah, New Jersey), 1997.

107. W. M. Siebert. Frequency discrimination in the auditory system: Place or periodicity mechanisms. *Proc. IEEE*, 58:723–730, 1970.

108. M. Slaney. *Auditory Toolbox (V.2)*, 1998. http://www.slaney.org/malcolm/pubs.html.

109. T. R. Stanford, S. Kuwada, and R. Batra. A comparison of the interaural time sensitivity of neurons in the inferior colliculus and thalmus of the unaesthetized rabbit. *J. Neurosci.*, 12:3200–3216, 1992.

110. R. M. Stern and H. S. Colburn. Theory of binaural interaction based on auditory-nerve data. IV. a model for subjective lateral position. *Journal of the Acoustical Society of America*, 64:127–140, 1978.

111. R. M. Stern and H. S. Colburn. Lateral-position based models of interaural discrimination. *Journal of the Acoustical Society of America*, 77:753–755, 1985.

112. R. M. Stern, A. M. Ripepi, and C. Trahiotis. Fluctuations in amplitude and frequency fluctuations in amplitude and frequency enable interaural delays to foster the identification of speech-like stimuli. In P. Divenyi, editor, *Dynamics of Speech Production and Perception*. IOS Press, 2005.

113. R. M. Stern and G. D. Shear. Lateralization and detection of low-frequency binaural stimuli: effects of distribution of internal delay. *Journal of the Acoustical Society of America*, 100:2278–2288, 1996.

114. R. M. Stern and C. Trahiotis. The role of consistency of interuaral timing over frequency in binaural lateralization. In Y. Cazals, K. Horner, and L. Demany, editors, *Auditory physiology and perception*, pages 547–554. Pergamon Press, Oxford, 1992.

115. R. M. Stern and C. Trahiotis. Models of binaural interaction. In B. C. J. Moore, editor, *Hearing*, Handbook of Perception and Cognition, chapter 10, pages 347–386. Academic (New York), 1995.

116. R. M. Stern and C. Trahiotis. Models of binaural perception. In R. Gilkey and T. R. Anderson, editors, *Binaural and Spatial Hearing in Real and Virtual Environments*, chapter 24, pages 499–531. Lawrence Erlbaum Associates, 1996.

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

118. R. M. Stern, T. Zeppenfeld, and G. D. Shear. Lateralization of rectangularly-modulated noise: explanations for counterintuitive reversals. *Journal of the Acoustical Society of America*, 90:1901–1907, 1991.

119. C. Trahiotis, L. R. Bernstein, R. M. Stern, and T. N. Buell. Interaural correlation as the basis of a working model of binaural processing: An introduction. In R. Fay and T. Popper, editors, *Sound Source Localization*, Springer Handbook of Auditory Research, chapter 7, pages 238–271. Springer-Verlag, Heidelberg, 2005.

120. C. Trahiotis and R. M. Stern. Lateralization of bands of noise: effects of bandwidth and differences of interaural time and intensity. *Journal of the Acoustical Society of America*, 86:1285–1293, 1989.

121. B. D. van Veen and K. M. Buckley. Beamforming: a versatile approach to spatial filtering. *IEEE ASSP Magazine*, pages 4–24, April 1988.

122. G. von Békésy. *Experiments in Hearing*. McGraw Hill (New York); reprinted by the Acoustical Society of America, 1989, 1960.

123. E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. Localization using nonindividualized head-related transfer functions. *Journal of the Acoustical Society of America*, 94:111–123, 1993.

124. F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening. I: Stimulus synthesis. *J. Acoustic. Soc. Amer.*, 85:858–867, 1989.

125. F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening. II: Psychophysical validation. *Journal of the Acoustical Society of America*, 87:868–878, 1989.

126. F. L. Wightman and D. J. Kistler. Factors affecting the relative salience of sound localization cues. In R. H. Gilkey and T. R. Anderson, editors, *Binaural and Spatial Hearing in Real and Virtual Environments*, pages 1–23. Lawrence Erlbaum Associates, 1997.

127. M. Wu, D. L. Wang, and G. J. Brown. A multipitch tracking algorithm for noisy speech. *IEEE Transactions on Speech and Audio Processing*, 11(3):229–241, 2003.

128. T. C. T. Yin, P. X. Joris, P. H. Smith, and J. C. K. Chan. Neuronal processing for coding interaural time disparities. In R. H. Gilkey and T. R. Anderson, editors, *Binaural and Spatial Hearing in Real and Virtual Environments*, chapter 21, pages 427–445. Lawrence Erlbaum Associates (Mahwah, New Jersey), 1997.

129. T. C. T. Yin and S. Kuwada. Neuronal mechanisms of binaural interaction. In G. M. Edelman, W. E. Gall, and W. M. Cowan, editors, *Dynamic Aspects of Neocortical Function*, page 263. Wiley (New York), 1984.

130. W. A. Yost. Lateral position of sinusoids presented with intensitive and temporal differences. *Journal of the Acoustical Society of America*, 70:397–409, 1981.

131. W. A. Yost, R. H. Dye, and S. Sheft. A simulated 'cocktail party' with up to three sources. *Perception and Psychophysics*, 58:1026–1036, 1996.

132. S. R. Young and E. W. Rubel. Frequency-specific projections of individual neurons in chick brainstem auditory nuclei. *J. Neurosci.*, 3:1373–1378, 1983.

133. X. Zhang, M. G. Heinz, I. C. Bruce, and L. H. Carney. A phenomenologial model for the response of auditory-nerve fibers: I. nonlinear tuning with compression and suppresion. *Journal of the Acoustical Society of America*, 109:648–670, 2001.

134. P. M. Zurek. Binaural advantages and directional effects in speech intelligibility. In G. A. Studebaker and I. Hochberg, editors, *Acoustical Factors Affecting Hearing Aid Performance*. Allyn and Bacon, Boston, 1993.