# Localization of multiple sound sources with two microphones[a)]

Chen Liu,[b)] Bruce C. Wheeler, William D. O'Brien, Jr., Robert C. Bilger,
Charissa R. Lansing, and Albert S. Feng
*Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801*

(Received 10 July 1998; revised 20 July 1999; revised 22 April 2000; accepted 28 June 2000)

This paper presents a two-microphone technique for localization of multiple sound sources. Its fundamental structure is adopted from a binaural signal-processing scheme employed in biological systems for the localization of sources using interaural time differences (ITD). The two input signals are transformed to the frequency domain and analyzed for coincidences along left/right-channel delay-line pairs. The coincidence information is enhanced by a nonlinear operation followed by a temporal integration. The azimuths of the sound sources are estimated by integrating the coincidence locations across the broadband of frequencies in speech signals (the ''direct'' method). Further improvement is achieved by using a novel ''stencil'' filter pattern recognition procedure. This includes coincidences due to phase delays of greater than $2\pi$, which are generally regarded as ambiguous information. It is demonstrated that the stencil method can greatly enhance localization of lateral sources over the direct method. Also discussed and analyzed are two limitations involved in both methods, namely missed and artifactual sound sources. Anechoic chamber tests as well as computer simulation experiments showed that the signal-processing system generally worked well in detecting the spatial azimuths of four or six simultaneously competing sound sources. © *2000 Acoustical Society of America.* [S0001-4966(00)04110-2]

PACS numbers: 43.72.Ew, 43.66.Qp, 43.66.Ts [JLH]

## I. INTRODUCTION

Localization of multiple sound sources is regarded as a challenging task, especially when the signals have overlapping spectra. Yet this capacity is important for the ability to extract signals in acoustically cluttered environments. A variety of signal-processing algorithms has been proposed to attack this problem, most of which revolve around the principle of creating a highly directional receiving system for enhanced detection of signals within a small sector of space, as used in different engineering applications (e.g., sonar systems). The dominant approach is the beamforming technique using an array of spatially distributed microphones (Flanagan *et al.*, 1985). A caveat of the microphone array system is that a relatively large array size and a large number of sensors are required in order to obtain a high degree of spatial selectivity. Furthermore, because one beamformer can only localize one source, multiple systems are necessary to localize different sources simultaneously.

An alternative to the array signal-processing approach is suggested by the neural computational mechanism in biological systems (Cherry, 1953; Bronkhorst and Plomp, 1992). It is well known that human beings as well as other living organisms can communicate effectively by sound in noisy and reverberant environments, in large part due to the advantage of directional hearing conferred by a binaural system (for a review, see Blauert, 1983). A binaural system can effectively compute interaural differences in time and intensity, thereby making it possible to accurately determine the directions of sound sources. Numerous binaural models have been advanced over the last half-century (see reviews by Colburn and Durlach, 1978; Colburn, 1996; and Stern and Trahiotis, 1995, 1997). Of these, the model by Jeffress (1948) has been validated anatomically and physiologically (Konishi *et al.*, 1988; Takahashi and Keller, 1994; Yin and Chan, 1990). Essentially, the Jeffress model involves the creation of a spatial map in the nervous system, i.e., the location of a sound in space is represented by the interaural time difference and the latter is determined by the location of signal coincidence along a dual delay-line neural network.

For the past two decades, the Jeffress model has been incorporated into auditory processors for localization of sound sources (e.g., Colburn, 1973, 1977; Blauert, 1980; Lindemann, 1986; Shamma *et al.*, 1989; Stern and Trahiotis, 1992; Gaik, 1993; Bodden, 1993; Banks, 1993). These processors produce satisfactory results for the localization of two sources but the performance degrades significantly when the signals are speech sounds and the number of sources is greater. In contrast, humans can localize as many as six concurrent sources, if not more (Bronkhorst and Plomp, 1992).

In this paper, we describe a two-microphone signal-processing system based on the Jeffress model capable of detecting and localizing a large number of sound sources in the ambient environment. Two major features of our system are utilization of: (1) a nonlinear procedure for determining the source locations, and (2) the intermicrophone time difference (ITD) information over the *entire* frequency broadband, including phase-ambiguous information at high frequencies. Anechoic chamber tests as well as computer simulation tests showed that four or six speakers could be satisfactorily detected and localized simultaneously. In the next section, we-

[a)] Portions of this paper were presented at the Hearing Aid Research & Development Conference, Bethesda, MD, Sept. 1997.

[b)] Present address: Motorola Labs, 55 Shuman Boulevard, Naperville, IL 60563.

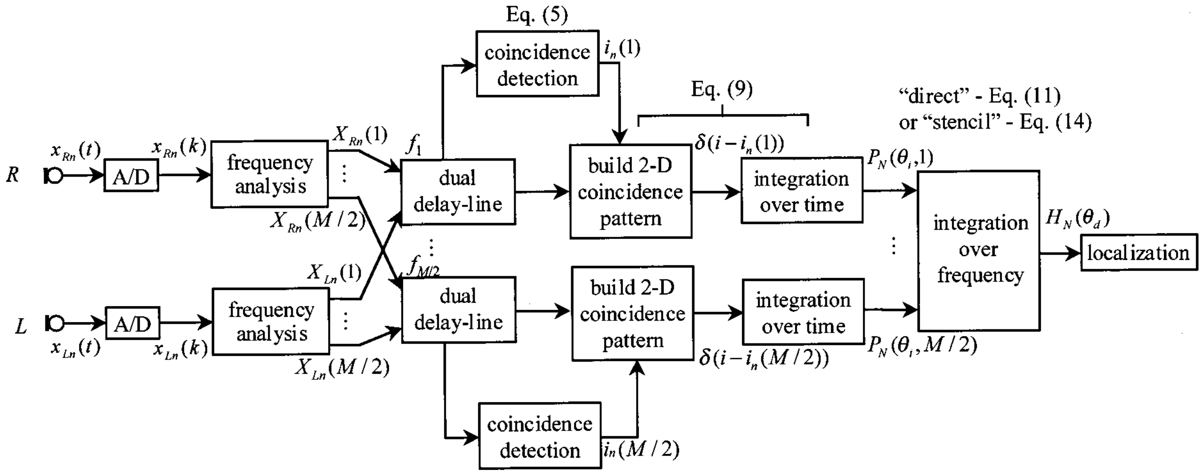

FIG. 1. Block diagram of the broadband localization system.

describe the dual delay-line structure as well as the coincidence detection method. In Sec. III, we describe a ''direct'' broadband computational scheme, operating in the frequency domain, for identifying the ITD information along the dual delay line. In Sec. IV, we describe how the performance of the ''direct'' method can be further improved by using a ''stencil'' filter that exploits both the unambiguous as well as the ambiguous ITD information (i.e., phase delays of greater than $2\pi$). The results of experiments and discussions as well as analyses of limitations of the two methods are given in the final section.

## II. NARROWBAND LOCALIZATION

### A. Dual delay-line structure

The schematic diagram of the signal processing system is shown in Fig. 1. The system assumes two inputs, $x_{Ln}(t)$ and $x_{Rn}(t)$, supplied by two identical omnidirectional microphones. Signal processing is performed in the digital frequency domain, using separate A/D converters. Digitized signals are denoted by $x_{Ln}(k)$ and $x_{Rn}(k)$. The subscripts $L$ and $R$ represent the left and right channels, respectively, and the index $n$ refers to the $n$th time frame of the short-term Fourier transform. Once digitized, the signals are decomposed using the short-term Fourier transform into $X_{Ln}(m)$ and $X_{Rn}(m)$, $m=1,...,M/2$; the corresponding discrete frequencies are $f_m = mf_s/M$, where $f_s$ is the sampling rate.

A ''dual delay-line'' network (Jeffress, 1948) is used for determining the directions of sound sources. For each frequency, the complex signals $X_{Ln}(m)$ and $X_{Rn}(m)$ from the two channels are fed into a pair of delay lines consisting of an odd number of delay units (Fig. 2). The values $\tau_i$ ($i=1,...,I$) of the time delays are assigned *a priori* such that the acoustic space in front of the two microphones is divided uniformly into $I$ sectors in azimuth, and each sector is uniquely mapped to one specific location in the dual delay line. Thus, if we assume there is no acoustic shadowing effect between the two microphones, the value of the delay units may be derived from

$$\tau_i = \frac{\mathrm{ITD}_{\max}}{2}\sin\left(\frac{i-1}{I-1}\pi - \frac{\pi}{2}\right), \quad i=1,...,I, \tag{1}$$

where $\mathrm{ITD}_{\max}$, which equals $D/c$, is the maximum intermicrophone time difference, $D$ is the distance between the two microphones, and $c$ is the speed of sound. The two delay lines delay the signals received by the left and right microphones separately, with progressively longer delays as they propagate through the delay lines. The midpoint of the dual delay line therefore corresponds to an ITD of zero, i.e., to a source located in the midplane. In other words, for a source located directly ahead of the microphones, coincidence of

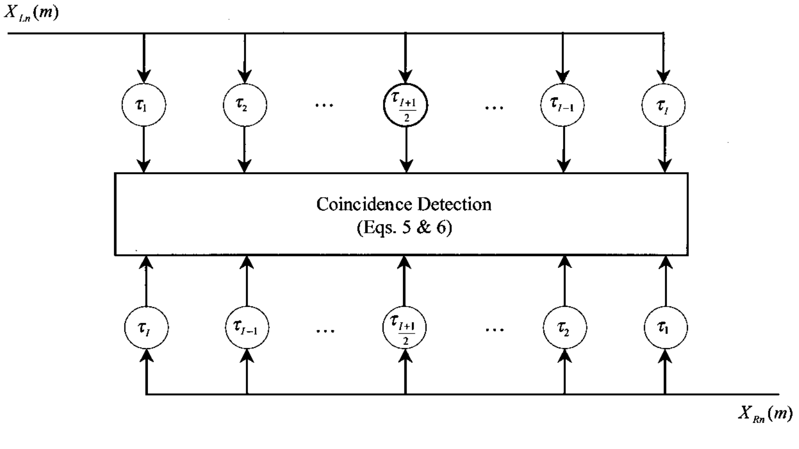

FIG. 2. The dual delay-line structure for narrow-band coincidence detection.

signals from the two microphones takes place in the middle of the dual delay line. A source located in the lateral sound field corresponds to nonzero ITD and thus will be represented on one side of the dual delay line. For example, when an acoustic signal emanates from an azimuth on the right side of the system, coincidence of signals will occur at a location on the left side of the dual delay line (i.e., the signals from the left and right microphones are in phase at this location). It is noted that to maintain symmetry in presenting the theory, Eq. (1) is given in the form of a noncausal structure, namely, the values of delay $\pi_i$ $(i=1,...,I)$ can be negative. Shifting the input signals by $\mathrm{ITD}_{\max}/2$ makes the structure causal. If we replace the index $i$ (namely, the location along the dual delay line) with the corresponding spatial azimuth $\theta_i$ (i.e., the signals in the two channels corresponding to a source at azimuth $\theta_i$ have a zero phase difference at location $i$ in the dual delay line), Eq. (1) can be expressed as

$$\tau_i = -\frac{\mathrm{ITD}_{\max}}{2}\sin\theta_i, \tag{2}$$

where

$$\theta_i = \frac{\pi}{2} - \frac{i-1}{I-1}\pi, \quad i=1,...,I. \tag{3}$$

The intermicrophone time difference $\mathrm{ITD}_i$ of the source located at azimuth $\theta_i$ equals

$$\mathrm{ITD}_i = -(\tau_i - \tau_{I-i+1}) = -[\tau_i - (-\tau_i)] = -2\tau_i, \quad i=1,...,I, \tag{4}$$

where the relationship $\tau_{I-i+1} = -\tau_i$ follows from Eq. (1), and $\mathrm{ITD}_i > 0$ (or $\theta_i > 0$) corresponds to the situation when the source is on the right-hand side of the midline of the microphones. Due to the one-to-one mapping relationship, the parameters $i$, $\tau_i$, $\theta_i$, and $\mathrm{ITD}_i$ are used equivalently in this paper for representing the location in the dual delay line as well as the corresponding azimuth.

### B. Coincidence detection

Perhaps the simplest method for finding coincidence locations is to find the point in the delay line at which there is a minimum in the magnitude of the difference between the two channel signals. It is analogous to the operational principle of a pressure-gradient receiver in biological systems (Feng and Shofner, 1981). We choose to find the minimum magnitude $\Delta X_n^{(i)}(m)$ of the difference between the frequency domain representations $X_{Ln}^{(i)}(m)$ and $X_{Rn}^{(i)}(m)$ at each discrete frequency $m$, yielding $M/2$ potentially different locations. If a spatially coherent source is present, all $M/2$ locations will be consistent across frequency. This operation is described in Eqs. (5)–(8). Thus,

$$i_n(m) = \arg\min_i [\Delta X_n^{(i)}(m)], \quad m=1,...,M/2, \tag{5}$$

where

$$\Delta X_n^{(i)}(m) = |X_{Ln}^{(i)}(m) - X_{Rn}^{(i)}(m)|, \quad i=1,...,I;\ m=1,...,M/2, \tag{6}$$

$$X_{Ln}^{(i)}(m) = X_{Ln}(m)\exp(-j2\pi f_m \tau_i), \quad i=1,...,I;\ m=1,...,M/2, \tag{7}$$

$$X_{Rn}^{(i)}(m) = X_{Rn}(m)\exp(-j2\pi f_m \tau_{I-i+1}), \quad i=1,...,I;\ m=1,...,M/2. \tag{8}$$

When the amplitudes of the two channel signals are identical, the value of $\Delta X_n^{(i)}(m)$ at the coincidence location is equal to zero. In practice, however, there is an intermicrophone intensity difference. Nonetheless, the two channel signals will still be in phase at the same point and the value of $\Delta X_n^{(i)}(m)$, although nonzero, is minimal at this point. Thus, the existence of the intermicrophone intensity difference does not affect the determination of the narrow-band coincidence location. Intensity equalization is unnecessary for the purpose of coincidence detection; its incorporation in fact will offset the space map on the delay line.

A number of methods are available for identifying the in-phase or coincidence point in the dual delay line. For example, Bodden (1993) used a cross-correlation method by computing a running integration on the 24 critical bands in the time domain. However, since our signals have already been transformed to the frequency domain it is computationally efficient to simply perform the subtractions in Eq. (6) and to look for minima using Eq. (5). On a frequency-by-frequency basis, the point of minimum magnitude of difference between two channels is the same as the point of minimum phase difference. As mentioned earlier, this calculation is robust against intermicrophone intensity differences. As described next, this technique gives us the ability to estimate azimuth by integrating the coincidence locations.

## III. LOCALIZATION USING THE "PRIMARY" CONTOUR (THE "DIRECT" METHOD)

### A. Spectral integration

Our coincidence detection method described in the preceding section can be viewed as a narrow-band operation because it is performed for each frequency. When there is one source and the signal $x_n(k)$ is broadband, the direction of the source estimated based on the frequency component $f_m$ in the $n$th time frame corresponds to the coincidence location $i_n(m)$, which satisfies Eq. (5). However, where there are multiple sources emitting spectrally overlapping sounds, interactions between sources in the $n$th time frame may lead to coincidences corresponding to phantom sources (e.g., at the midpoint between two identical sources). To overcome this problem, we integrate coincidence locations over both time (described later) and frequency. Our strategy takes advantage of the following facts: (a) normal conversation includes a large number of pauses (Flanagan, 1972, p. 386) which provide opportunities for competing sounds to be detected, and (b) when several independent sources emit sounds simultaneously, usually one can find frequency bins which are dominated by one source. Banks (1993) implicitly used a similar assumption for localization of two sound sources.

The broadband coincidence detection developed here assumes that the intermicrophone time delay is independent of

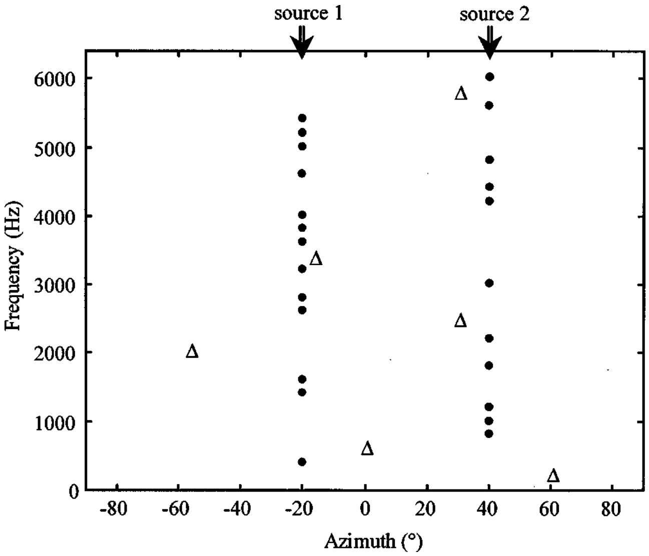

FIG. 3. A schematic depiction of the instantaneous 2D coincidence pattern illustrating the typical pattern of narrow-band coincidence detection and the effect of interaction between two sources using the direct method. Filled circles are data points showing correct localization. Data points associated with phantom detection are shown in open triangles. Note that the high-frequency ambiguity is neglected here.

frequency. When the two microphones are mounted on a structure where there is little diffraction, this assumption is satisfied. The importance of the consistency of the interaural timing information across frequency in the lateralization of broadband binaural stimuli was previously indicated and exploited in the model of Stern and Trahiotis (1992). However, if the diffraction is not negligible (such as due to head shadowing), the delay is not completely independent of frequency (Durlach and Colburn, 1978; Blauert, 1983). Therefore, care in this regard should be taken in implementation.

The logical basis for the ''direct'' broadband integration is illustrated in a schematic figure in Fig. 3, wherein the locations of coincidence are plotted against the signal frequency ($y$ axis) and azimuth ($x$ axis) for a two-source stimulus arrangement. The graph suggests that there are two sources plus some misidentified locations at frequencies where both sources have significant energy. Obviously, by integrating the coincidence pattern across frequency we are able to gain more robust and reliable estimates of the locations of real sources. Previously, Banks (1993) adopted an image processing technique for detecting short vertical straight segments, which, however, restricts the binaural information to a narrow spectral band of the signal. In contrast, our localization algorithm utilizes the ITD information throughout the entire spectral band. Localization based on broadband coincidence detection has also been demonstrated neurophysiologically in barn owls (Takahashi and Keller, 1994).

Our method accumulates the locations of coincidence, not the value of the function used to determine the minima. This is equivalent to applying a Kronecker delta function, $\delta(i-i_n(m))$, to $\Delta X_n^{(i)}(m)$ in Eq. (6) prior to the integration across frequency. This enhances the expression of coincidence location in the 2D pattern for each frequency $f_m$. An alternative way of representing the localization results with impulses can be formulated by incorporating an idea from Colburn's model (Colburn, 1977; Colburn *et al.*, 1990). In Colburn's localization model, the input signals are transformed into trains of impulses simulating the firing of neural spikes prior to coincidence detection.

## B. Temporal integration

Integration over time is achieved by the use of a forgetting average of the instantaneous 2D coincidence patterns (such as the example in Fig. 3) acquired over the preceding set of time frames. When the signals are not correlated at each frequency, mutual interference between signals can be gradually attenuated by temporal integration. The interaction between sources as well as the effect of temporal integration will be illustrated with an example in the next section.

## C. Algorithm

Our computational strategy, including both temporal and spectral integrations, is as follows:

(a) The instantaneous 2D coincidence patterns at each frequency $f_m$ are integrated over time

$$P_N(\theta_i,m)=\sum_{n=1}^{N}\beta^{N-n}\delta(i-i_n(m)),$$

$$i=1,...,I;\quad m=1,...,M/2, \tag{9}$$

where $\beta$, restricted to be between 0 and 1, is a weighting coefficient which exponentially de-emphasizes (or forgets) the effect of previous coincidence results, $\delta(\cdot)$ is the Kronecker delta function, $i$ represents the location in the dual delay line corresponding to spatial azimuth $\theta_i$ [Eq. (3)], and $N$ refers to the current time frame. We chose $\beta$ in the range of 0.85–0.99 for our tests. A smaller value of $\beta$ takes insufficient advantage of spatial information in past coincidence patterns, while a larger value, by overemphasizing past patterns, may make the system less sensitive to source dynamics such as onset, offset, or movement. Thus, one can choose an optimum value for $\beta$, trading off localization enhancement for higher sensitivity to the source movement. We let $\beta=0.98$ in our following experiments, in which all the sources were stationary.

(b) The temporally integrated coincidence pattern $P_N(\theta_i,m)$ is then thresholded so as to eliminate the majority of spurious coincidence detections

$$P_N(\theta_i,m)=\begin{cases}P_N(\theta_i,m), & P_N(\theta_i,m)\geqslant\Gamma\\ 0, & \text{otherwise}\end{cases}. \tag{10}$$

Here $\Gamma$ is set greater than or equal to zero. A greater value of $\Gamma$ can remove phantom coincidences that last longer. In our tests, we set $\Gamma=1$, which removed the phantom coincidences lasting for one time frame only.

(c) In the ''direct'' method the temporally integrated and thresholded coincidence patterns are integrated over frequency

$$H_N(\theta_d)=\sum_m P_N(\theta_d,m),\quad d=1,...,I. \tag{11}$$

Formally, the integration is conducted along the primary contour (a vertical line for each azimuthal angle as suggested in Fig. 3). In the next section the contour is extended to account for phase ambiguity at high frequencies.

Ideally, the peaks in $H_N(\theta_d)$ occur at the azimuthal angles at which the sound sources are located. Other peaks may be due to noise or due to interaction of multiple sources. The former generally gives rise to peaks that are significantly higher than the latter after broadband integration. Hence, a simple clustering of the peaks of $H_N(\theta_d)$ by height into two groups can extract the peaks corresponding to the real sources. First, we derive the envelope of $H_N(\theta_d)$ by applying a low-pass filter on the function. Then, all the local minima of the envelope are located. The local maximum of $H_N(\theta_d)$ between each adjacent pair of minima is determined as a valid peak. (In our implementation the low-pass filter cutoff was defined as 1/32 with 1 corresponding to half the sample rate.) In comparison with the narrow-band localization, the broadband localization not only enhances the accuracy of localization, it also enables localization of multiple sources.

The low and high frequencies play complementary roles in the broadband localization. First, the estimation of coincidence locations is more accurate at high frequencies, which produces more consistent peaks in the 3D coincidence pattern, and eventually narrower and sharper peaks after spectral integration. The estimation error at low frequencies will result in a wide spread of response peaks, thereby producing broader and lower peaks after spectral integration. The reason for this estimation error will be discussed in detail in Sec. V B 1. Second, high-frequency components in speech (mostly from consonants) occupy a smaller proportion of time when compared to the low-frequency sounds (mostly from vowels) and hence are much less likely to coincide temporally in normal conversation. However, because lower-frequency sounds (vowels) occupy a larger proportion of time in conversation (Flanagan, 1972), a localization scheme based solely on high-frequency components may not work all the time. Hence, the broadband scheme confers the advantages of both the low- and high-frequency components, with the low-frequency information providing the estimate of source localization most of the time and the high-frequency information enhancing the accuracy of localization.

The energy information is removed prior to spectral integration [Eq. (9)]. Therefore, the height of a peak in $H_N(\theta_d)$ only indirectly reflects the energy of a sound; instead, it is influenced by factors such as the energy of the signal relative to the energy of the other signals in each frequency bin, and the number of frequency bins as well as the duration over which the signal is dominant. Since each frequency bin contributes equally in the broadband integration (in contrast to time-domain correlation algorithms), high-energy signals are less likely to mask low-energy signals, especially if they occupy distinct frequency bands during at least part of the integration time. However, the drawback is that very narrow-band and short-duration signals could be missed.

Our method has the potential to localize sound sources dynamically as they move in space, although we have not tested this capability. Movement trajectories could be estimated from the sets of locations computed at each time window, e.g., every 5 ms in our tests.

## IV. LOCALIZATION WITH THE ''STENCIL'' FILTER

### A. High-frequency ambiguity

In practice, when the entire band is exploited for localization, the 3D coincidence pattern is not as simple as the vertical straight traces shown in Fig. 3 because for high frequencies there is *no* one-to-one map of auditory space using time or phase cues. For mammals including humans, it has been postulated (Mills, 1972; Gourevitch, 1987) that localization of sound azimuth depends on two cues: interaural time differences at low frequencies, and interaural intensity differences at high frequencies. The primary argument is the fact that the interaural time difference is ambiguous for high-frequency sounds because the wavelengths are shorter than the separation between the two ears.

Figure 4(A) illustrates a theoretical broadband coincidence pattern $P_N(\theta_i, m)$ calculated by using the coincidence detection [Eqs. (5)–(8)], for an intermicrophone distance of 144 mm over the frequency range of 0–6400 Hz. The solid traces in Fig. 4(A) represent the coincidence points of a source at $-60°$ azimuth over frequency, the dotted traces the coincidence pattern for a source at 45° azimuth. The phase ambiguities at frequencies $>1200$ Hz are shown as the curved traces which shall be referred to as the ''secondary'' traces, to distinguish them from the vertical (or ''primary'') trace.

The existence of the ambiguous, secondary traces in $P_N(\theta_i, m)$ will generate artifactual peaks in $H_N(\theta_d)$. When there are several sources in the ambient environment, superposition of secondary traces from several sources produces a noisier $H_N(\theta_d)$. The artifact peaks, when far away from the peaks of any real sources, may result in a false detection of nonexistent sources, and, when close to the peaks of real sources, they may affect both the detection and estimation of the azimuths of real sources. We will give a detailed discussion and analysis in the experiment section.

### B. Stencil filter corresponding to the uniform-azimuth dual delay line

The advantage of short, high-frequency consonants in sound localization (described in Sec. III C) is normally offset by high-frequency phase ambiguity. To reduce the ambiguity effect, a weighting scheme can be used to reduce the importance of the secondary traces relative to the primary traces (Stern *et al.*, 1988; Stern and Trahiotis, 1992). We have chosen to take a different approach to accounting for this issue. We have created a ''stencil'' filtering method, which takes full advantage of the information in the secondary traces. In essence, both the primary trace and the secondary traces are utilized in the directional estimation. In principle, this approach should help significantly in cases where one source is heavily masked spectrally by other sources and its primary traces are distorted by overlap from the secondary traces associated with other sources. A justification of this method is the fact that each sound azimuth is uniquely associated with a specific phase-coincidence trace pattern [see Fig. 4(A) for

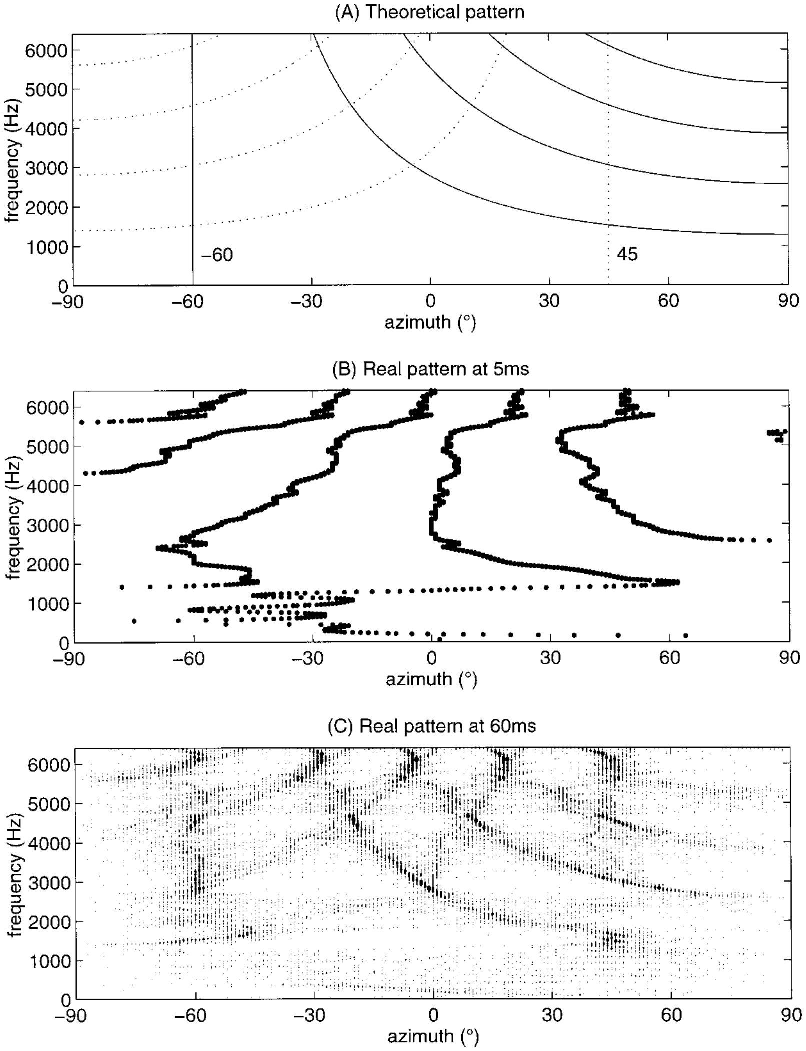

FIG. 4. (A) Primary (vertical traces) and secondary contours (curved traces) from theoretical broadband coincidence patterns associated with a source at −60° in azimuth and a source at 45° in azimuth. The solid lines are the pattern corresponding to −60° azimuth and the dotted lines to 45°. In (B) and (C) are shown the actual broadband coincidence patterns generated by two concurrent sources at −60° and 45° azimuths, at 5 ms (B) and 60 ms (C) after the onset. The temporal weighting factor $\beta=0.98$. The window length is 20 ms with consecutive windows overlapped by 15 ms.

coincidence patterns associated with −60° (solid lines) and 45° (dotted lines)]. For example, it is possible to construct a filter applied on $P_N(\theta_i, m)$, such that when deriving $H_N(\theta_d = -60°)$ we integrate the peaks not only along the primary trace (at $\theta_i = -60°$), but also along the secondary traces for −60°. Similarly, we can derive the $H_N(\theta_d)$ for *all* the azimuthal directions $\theta_d$ $(d=1,...,I)$. For each $\theta_d$, the filter looks like a stencil window with a shape defined by the ideal pattern $P_N(\theta_i, m)$ $(i=1,...,I; m=1,...,M/2)$.

Below, the concepts of stencil filtering and temporal integration (deferred from Sec. III B) are illustrated using a real example [Figs. 4(B) and (C)]. In this example, a female talker was located at −60° azimuth and a male talker at 45° azimuth. They concurrently uttered a different word at identical average intensity. Figure 4(B) shows the broadband coincidence pattern for the first time frame (5 ms after onset). It can be seen that, due to the interaction between the two sources, the coincidence pattern was highly irregular and did not provide a good indication of the locations of the sources. After 60 ms, however, both sources were clearly revealed in the pattern [Fig. 4(C)].

Corresponding to the coincidence detection [Eqs. (5)–(8)], a mathematical expression of the stencil filter is

$$\sin\theta_i - \sin\theta_d = \frac{\gamma_{d,m}}{\mathrm{ITD}_{\max} f_m}, \tag{12}$$

where, as previously defined, $\theta_i$ represents the place along the dual delay line by using the corresponding spatial azimuth, and $\theta_d$ is the azimuthal direction for which the broadband integration is conducted. Please note that since Eq. (12) was derived based on the 3D coincidence trace pattern $P_N(\theta_i, m)$, it can be used both for characterizing the integration contours of the stencil filter, and for describing the 3D coincidence pattern of the sound sources. For differentiating the two uses, we refer to the former as the *integration contours* and the latter as *coincidence traces*. The parameter $\gamma_{d,m}$ is an integer, equal to the number of intersections of the integration contours with the horizontal line $f=f_m$. The range of $\gamma_{d,M/2}$ ($M/2$ is the highest digital frequency, and $f_{M/2}=f_s/2$) also equals the number of the integration contours associated with the direction $\theta_d$ in the 3D pattern $P_N(\theta_i, m)$. The primary contour corresponds to $\gamma_{d,m}=0$. For a specific azimuthal direction $\theta_d$, the range of valid $\gamma_{m,d}$, or the number of the integration contours, is

$$-\mathrm{ITD}_{\max} f_m (1+\sin\theta_d) \leq \gamma_{d,m} \leq \mathrm{ITD}_{\max} f_m (1-\sin\theta_d). \tag{13}$$

See Appendix B for derivation of Eqs. (12) and (13).

By employing the stencil filter [Eq. (12)], the broadband integration procedure in Eq. (11) becomes

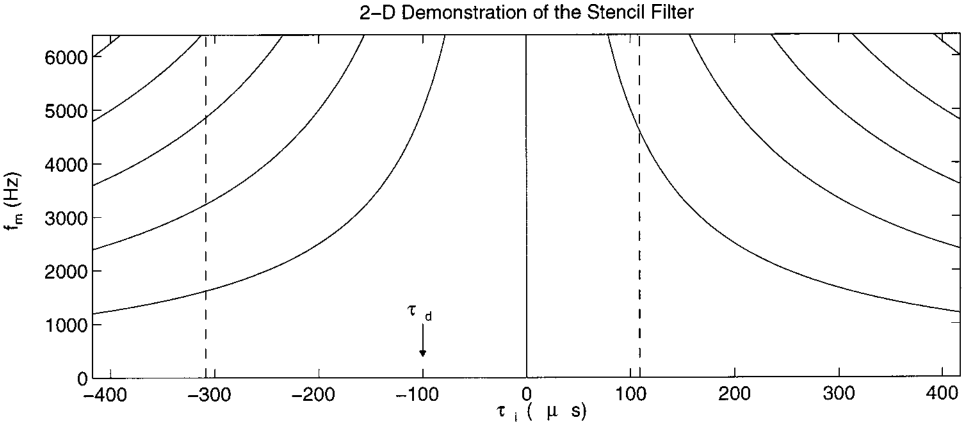

FIG. 5. Two-dimensional representation of the stencil filter corresponding to dual delay line with a uniform $\tau_i$. The $x$ axis represents location along the dual delay line by the corresponding delay unit $\tau_i$ in the left channel. Within the window defined by the two vertical dashed-line markers is shown the stencil pattern that is associated with the direction $\theta_d = 28.63°$ azimuth [the direction corresponds to location $\tau_d = -100\ \mu\text{s}$ in dual delay line by using Eq. (2)].

$$H_N(\theta_d) = \frac{1}{A(\theta_d)} \sum_m P_N\left[\sin^{-1}\left(\sin\theta_d + \frac{\gamma_{d,m}}{\text{ITD}_{\max} f_m}\right), m\right],$$

$$d = 1,\ldots,I, \tag{14}$$

where $A(\theta_d)$ denotes the number of points involved in the summation. The normalization is used because the number of discrete points contained in the integration contours may vary with the targeted azimuth $\theta_d$. It can be seen that the direct method [Eq. (11)] is a special case of the stencil method [Eq. (14)], i.e., when $\gamma_{d,m} = 0$. The block diagram of the whole system is given in Fig. 1.

### C. Stencil filter corresponding to uniform time delay in a dual delay line

A disadvantage of the stencil pattern in Eq. (12) is that it varies with azimuth $\theta_d$. Thus, in implementation, a large memory is required for storing the stencil filter patterns. This shortcoming can be overcome, but with a tradeoff. As shown in Eq. (2), both variables $\theta_i$ and $\tau_i$ are equivalent and represent the same position along the dual delay line. The difference between them is that $\theta_i$ indicates location in the dual delay line by virtue of its spatial azimuth $\theta_i$, whereas $\tau_i$ denotes location by virtue of the value $\tau_i$ of the corresponding time-delay unit. The stencil pattern is simpler if it is expressed with $\tau_i$. Namely,

$$\tau_d - \tau_i = \frac{\gamma_{d,m}}{2 f_m}, \tag{15}$$

where $\tau_d$ relates to $\theta_d$ through Eq. (2). For a specific $\tau_d$, the range of valid $\gamma_{m,d}$ is

$$-(\text{ITD}_{\max} - 2\tau_d) f_m \leqslant \gamma_{d,m} \leqslant (\text{ITD}_{\max} + 2\tau_d) f_m,$$

$$\gamma_{d,m} \text{ is an integer.} \tag{16}$$

Obviously, changing the value of $\tau_d$ only shifts the coincidence pattern (or stencil pattern) along the $\tau_i$ axis but without changing its shape. A 2D display of the stencil filter function defined by Eqs. (15) and (16) ($0 \leqslant f_m \leqslant 6.4$ kHz and $\text{ITD}_{\max} = 417.4\ \mu\text{s}$) is given in Fig. 5, in which the actual stencil pattern corresponding to a $\tau_d$ is within the region between the two dashed-line markers centered at $\tau_d$ (e.g., $\tau_d = -100\ \mu\text{s}$ in Fig. 5). A caveat of Eq. (15) is that, because the scaling of the delay units $\tau_i$ is linearly uniform along the dual delay line, the partition of the frontal azimuth by the dual delay line is not uniform, with the regions close to the midplane having higher azimuthal resolution. Therefore, to obtain an equivalent resolution in azimuth, using a uniform $\tau_i$ would require a much larger number $I$ of delay units than using a uniform $\theta_i$ (see Fig. 6).

## V. EXPERIMENTS AND ANALYSES

### A. Method

The signal-processing system was evaluated both in computer simulation and in acoustic tests using physical devices in an anechoic chamber. In computer simulation, we evaluated the performance of the system under two experimental conditions: one-talker and two-talker tests. In anechoic chamber tests, we evaluated its performance under the condition of four-talker tests. The speech materials consisted of spondaic words spoken by native speakers of American English. All the speech recordings were equalized in average intensity. Unless otherwise stated, the words in each experimental condition were temporally aligned, i.e., all the talkers started speaking at the same time; this represented a challenging listening condition. The intermicrophone distance was 144 mm. The speech material was presented in free field, and various azimuthal configurations were used for each experimental condition. No diffraction or shadowing effect existed between the two microphones, and the intermicrophone intensity difference was set to zero across all frequencies for the tests in the computer simulation.

The signals were low-pass filtered at 6 kHz and sampled at a 12.8-kHz rate with 16-bit quantization. In the short-term spectral analysis, a 20-ms segment of signal was weighted by a Hamming window, padded with zeros to 2048 points, and Fourier transformed with frequency resolution of about 6 Hz. Consecutive frames overlapped by 15 ms. The values of the time delay units $\tau_i$ ($i = 1,\ldots,I$) in the dual delay line corresponded to a uniform azimuthal resolution of 0.5° (namely, $I = 361$). Both the direct and stencil methods were tested; the stencil method used a uniform-azimuth dual delay line [Eqs. (12) and (13)]. The value of the weighting factor $\beta$ in the temporal integration [Eq. (9)] was 0.98. The threshold $\Gamma$ [for cleaning the sporadic impulses in the 3D coincidence pattern in Eq. (10)] was 1.

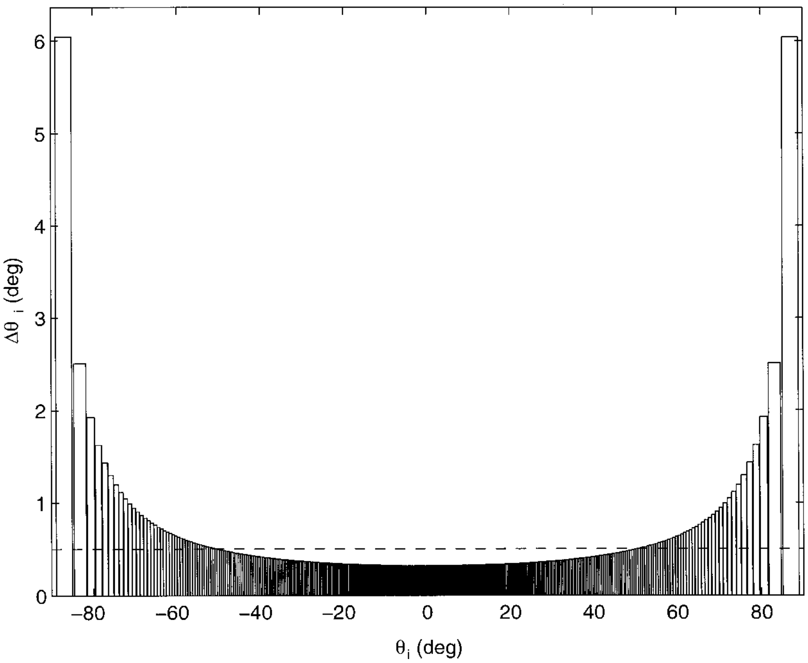

FIG. 6. The azimuthal resolution of dual delay line with a uniform $\tau_i$. The $x$ axis is azimuth. The $y$ axis shows the interval of mapping azimuths between adjacent points in the dual delay line. As a reference, the horizontal dashed line indicates the resolution of dual delay line with a uniform $\theta_i$ having the same number of delay units $(I=361)$.

## B. Results and analyses

### 1. Acuity of localization in one-talker tests

The purpose of the one-talker computer-simulation tests was to evaluate the acuity of the localization peak of a single source in the absence of any interfering sound. Figure 7 shows the value of $H_N(\theta_d)$ at 30 ms (or $N=6$) after the beginning of a spondaic word ''pancake'' by a female speaker from several different azimuths. When the source was close to the 0° azimuth, $H_N(\theta_d)$ had a distinct and sharp peak corresponding to the source azimuth with both the direct and stencil methods. For a source positioned more laterally in the sound field, the peak was progressively lower and broader. The reason is that an estimation error, although very small, always exists in frequency analysis due to the finite length of the time window. In the current analysis, we employed a time window of 20 ms by characterizing speech as a quasistationary signal. The error is larger for low frequencies because fewer periods of sine waves are present in the time window for low frequencies than for high frequencies.

However, even for the same frequency, due to the nonlinear temporal resolution associated with the dual delay line with uniform azimuth, this estimation error has a varying effect on $H_N(\theta_d)$ along the dual delay line. This estimation error includes errors in the estimates of both amplitude and phase while the width of the peak in $H_N(\theta_d)$ is attributed to the error in the phase estimate. Because the intermicrophone time difference is a sine function of the azimuth $\theta_i$ [Eq. (2)], the difference $\Delta\tau_i$ (namely, $\tau_i-\tau_{i-1}$) between adjacent delay units is greater for $i$ in the middle than toward the two ends of the dual delay line. For the dual delay line used in our tests, the maximum value of $\Delta\tau_i$ is 1821.2 ns (in the middle of the dual delay line) and the minimum value is 7.9465 ns (at the both ends of the dual delay line). Therefore, the same phase error in the frequency analysis may result in impulses deviating more in the 3D coincidence pattern $P_N(\theta_i,m)$ when the source is located laterally than when the source is close to the midline. Consequently, the direct broadband integration [Eq. (11)] produces a lower and broader peak in $H_N(\theta_d)$ for a more lateral $\theta_d$.

It is noted that when $i$ is close to the midpoint of the dual delay line, the value $\Delta\tau_i$ of delay resolution is larger than the phase error. Therefore, when a source is located laterally, its primary trace lies away from the midpoint, and is inherently low in spatial consistency, while one or more of its secondary traces are more consistent in the middle of the dual delay line. Consequently, the resultant peak of integration in $H_N(\theta_d)$ is more pronounced and sharper when using the stencil method (that performs integration along both the primary and secondary contours) than using the direct method (that performs integration along the primary contour only). This is evident as shown by comparing Figs. 7(E) and (F).

In principle, the low acuity in the lateral sound field will affect resolution of two closely adjacent sources located laterally, as will be shown in the next subsection. However, it is noted that this kind of resolution should be distinguished from the resolution of the dual delay line itself. In other words, increasing the number $I$ of the delay units in a dual delay line provides a finer azimuthal partition and hence higher spatial precision, but the ability to resolve two adjacent sources is compromised by the low acuity, especially in the lateral field, due to the phase estimation error.

### 2. Interaction of sources in two-talker tests

We conducted a series of two-talker tests in the computer simulation to analyze the interactions between sources. We compared the performances of the direct and stencil methods, for talkers situated in different positions, and for sources having different relative intensities, and onset times. We also used the two-talker tests as an example to analyze the origin as well as the location of the artifacts that became problematic in localization of multiple sources with both

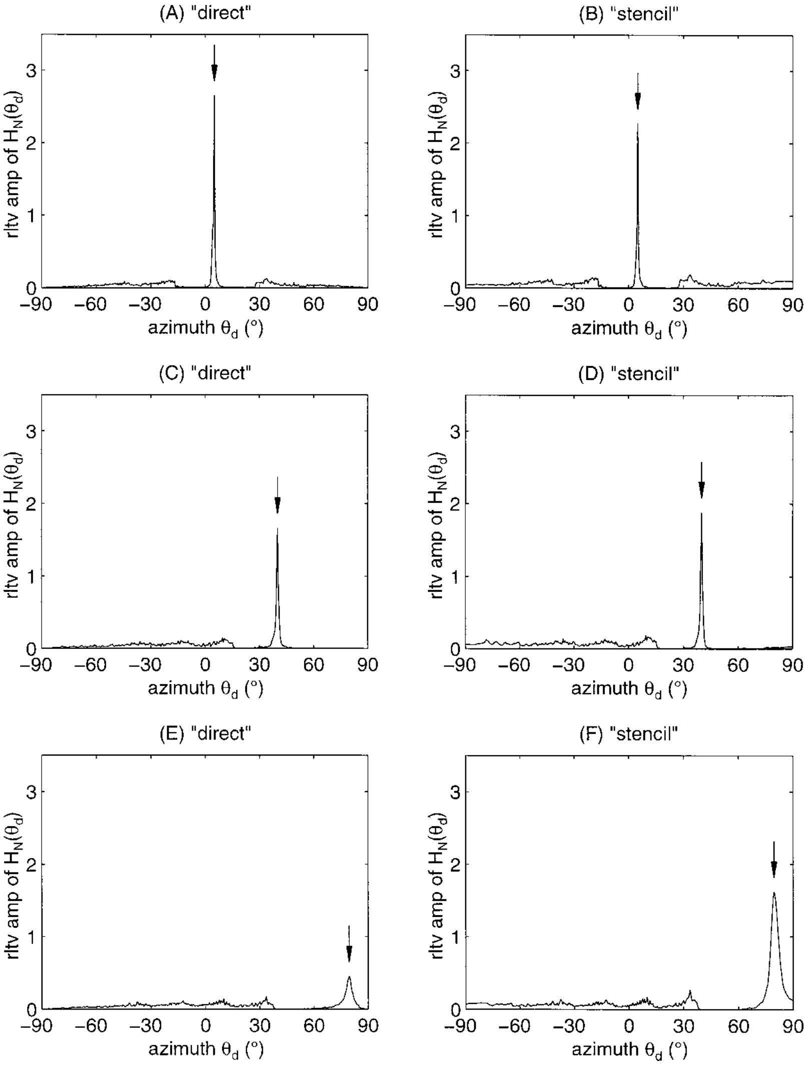

FIG. 7. Localization results $H_N(\theta_d)$ of one source (spondaic word ‘‘pancake’’ spoken by a female speaker) at different azimuths. The $x$ axis is azimuth ($\theta_d$) of the talker, and the $y$ axis represents the relative amplitude of $H_N(\theta_d)$ taken at $N=6$ (or 30 ms) after onset of speech. Left panels (A), (C), (E) are results from the direct method, while right panels (B), (D), (F) are from the stencil method. The talker was located at 5° (A), (B), 40° (C), (D), or 80° (E), (F). The arrow in each panel indicates the real location. The localization peak is high, narrow, and sharp with both methods when the source is close to midline (A), (B), (C), (D). The acuity reduced in the lateral sound field with the direct method (E); the stencil filter restores considerable lateral acuity (F).

methods. For these tests, we used the speech materials from a male talker (M2 saying ‘‘playground’’) and a female talker (F1 saying ‘‘pancake’’).

*a. Different spatial configurations.* The nine configurations used in our tests are shown in Table I. The result from a typical configuration (Configuration #1: M2 at 0° and F1 at 40°) is shown in Fig. 8. Both methods gave accurate estimates of the locations of the two talkers with no significant difference between the two methods. Note that the localization computation was conducted once for each frame, i.e., every 5 ms, but the display of the $H_N(\theta_d)$ function in Fig. 8 (and all the 3D graphics in the experiment section) is shown once every 50 ms for clarity.

Configurations #2, #3, and #4 (see Table I) were designed to evaluate the system resolution for two closely adjacent sources (5° apart) at different azimuthal directions. The results given in Fig. 9 show that there was little difference in performance between the direct and stencil methods. Both methods could easily and clearly distinguish the two closely adjacent sources when they were in front of the microphones [Figs. 9(A)–(D)]; distinguishing the two sources became increasingly more difficult as the two sources were positioned more laterally [Figs. 9(E) and (F)]. As analyzed in the preceding subsection, this is due to the lower and broader peaks in $H_N(\theta_d)$ for sources located in the lateral sound field.

Configuration #5 represents a situation in which the two sources are far apart from each other (160° in this case) such that one source is at each lateral side. As mentioned in the preceding subsection, when sound sources were located laterally, the ambiguous secondary traces in $P_N(\theta_i, m)$ showed high spatial consistency. As a result, when using the direct integration, the secondary traces resulted in a number of artifact peaks in $H_N(\theta_d)$ which might be more prominent than the peaks resulting from integration along the primary contours corresponding to the real sources [Fig. 10(A)]. In con-

TABLE I. Configurations in the two-talker tests for analysis of the performance of the dual delay line.

| Configuration index | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Source | M2 | 0° | 0° | 40° | 75° | −80° | −40° | 0° | 40° | −40° |
| Azimuth | F1 | 40° | 5° | 45° | 80° | 80° | 80° | 80° | 80° | 45° |

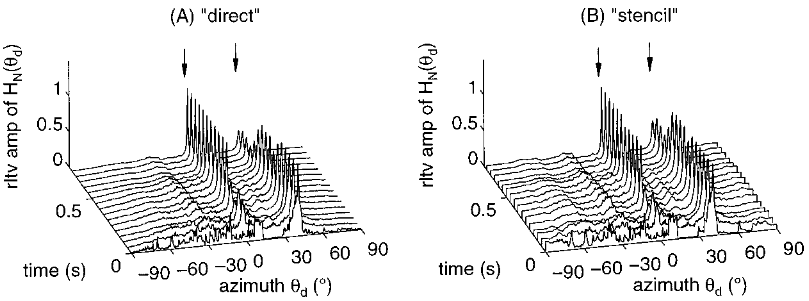

FIG. 8. Localization of two talkers using direct (A) and stencil (B) methods. Two talkers were located at 0° and 40° (indicated by arrows). The talkers spoke different words that had the same average intensity and were temporally aligned. The localization computation was conducted once every 5 ms, but the display is shown every 50 ms for clarity.

trast, the stencil method, which utilized both the primary and secondary traces for localization, enhanced the peaks of real sources. At the same time, the artifact peaks were dramatically suppressed, allowing reliable localization of the real sources [Fig. 10(B)].

Configurations #6, #7, and #8 illustrate another merit of using the stencil method. Similar to the one-talker tests, the localization peak was much lower and obscured as the source approached ±90° with the direct method. However, using the stencil filter amplified the height of the response peak of a lateral source [e.g., see the source at 80° in Figs. 10(C) and (D) for configuration #7].

*b. Different relative rms intensities.* Tests were conducted to examine the effect of relative intensity on localization. In particular, we sought to determine the ability of the algorithm to locate a weaker sound in the presence of a more intense interfering sound. The tests were conducted with three configurations (configurations #1, 2, and 9 in Table I) using three different relative intensities (the rms intensity of M2 was +5 dB, +10 dB, and +15 dB relative to F1). To determine whether the effect of intensity was independent of speaker and location, three more tests were conducted for configuration #9 where the rms intensity of M2 was −5 dB, −10 dB, and −15 dB relative to F1, respectively. Both the direct and stencil methods were tested for each situation. Results show that, with both methods, the two sources could be localized for an intensity difference of up to 10 dB (Fig. 11). The peak associated with the weaker source (at −40°) grew higher with time; this was attributed to the temporal integration [Eq. (9)]. For intensity difference greater than 10

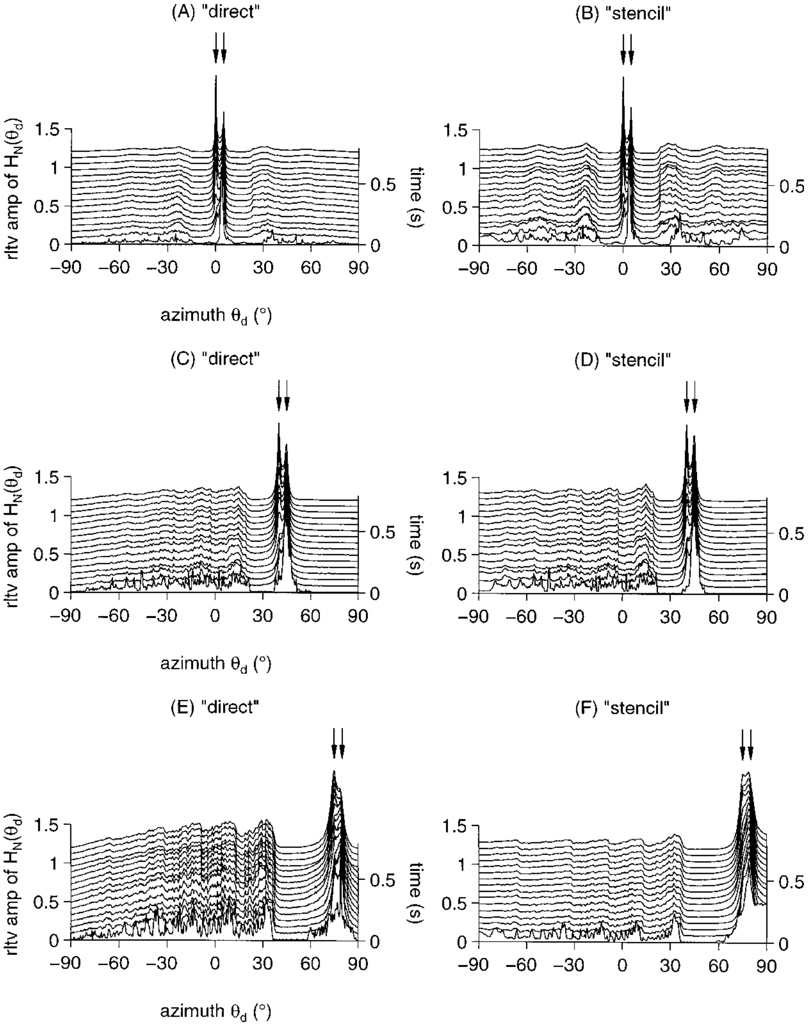

FIG. 9. Location of two closely adjacent talkers (separated by 5°) at various azimuths. The locations of the talkers were: 0° and 5° in (A), (B), 40° and 45° in (C), (D), 75° and 80° in (E), (F). The real locations of the talkers are indicated by arrows. The left column is from direct method; right column from stencil method. Resolution of the sources becomes difficult when they are located laterally (E), (F). The localization computation was conducted once every 5 ms, but the display is shown every 50 ms for clarity.

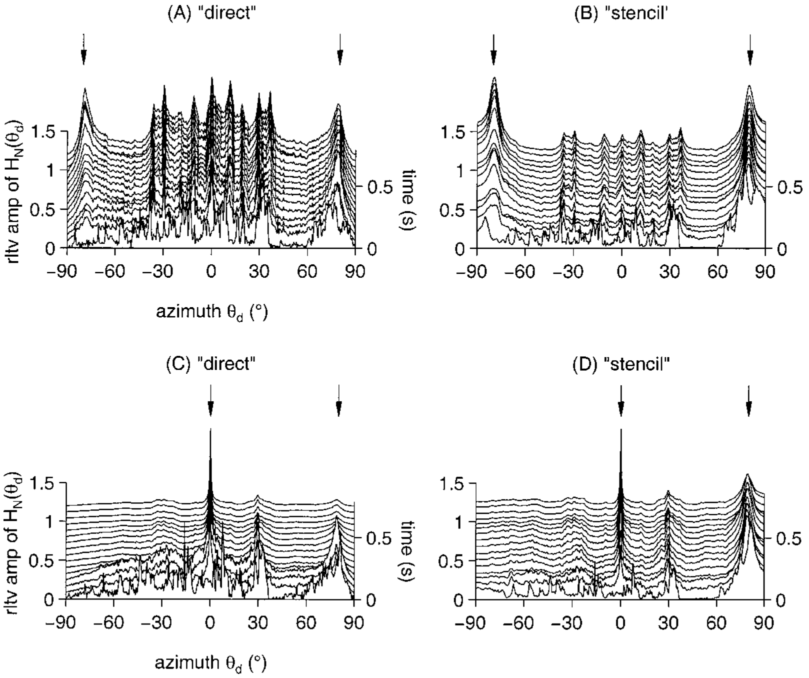

FIG. 10. Localization of two lateral sources. In the first case (A), (B), the two sources were located at −80° and 80° (shown by arrows). The artifactual peaks were very high with the direct method (A), whereas the stencil method (B) enhanced the height of the real peaks and suppressed the artifactual peaks. In the second case (C), (D), the talkers were located at 0° and 80°. The talker at 80° was poorly detected and localized by the direct method (C). The stencil method improved its localization (D). The localization computation was conducted once every 5 ms, but the display is shown every 50 ms for clarity.

dB, the weaker source could not be detected.

*c. Different onset times.* These tests were conducted to examine the dynamic performance of the algorithm when one sound source is present and a second source goes on or off. The words ''playground'' spoken by Talker M2 and ''pancake'' spoken by Talker F1 (approximately 750 and 670 ms in duration, respectively) were presented using the same three spatial configurations as in the preceding subsection. For each configuration, the speech signal from F1 was delayed relative to M2 by 97.7, 195.3, and 293.0 ms, respectively. Figure 12 illustrates the system performance for the case with 195.3-ms delay (M2 at −40° and F1 at 45°). Both the direct and stencil methods could respond to the onset and offset of each talker. The responses lagged behind the onset or offset by about 100 ms. The response lag was a side effect of temporal integration in Eq. (9), which although enhanced and stabilized localization of sound sources, produced an inertia in the response to the onset and offset of the sound.

*d. Location of the artifacts.* Some of the above two-talker tests revealed an artifact issue in $H_N(\theta_d)$, such as those shown in Fig. 10(A), introduced from the broadband integration operation. This phenomenon exists in both the direct and stencil methods. We illustrate the issues with an example using the direct method, although the analysis applies to the stencil method and to the situations involving more talkers. Since $H_N(\theta_d)$ is the result of broadband integration of the 3D coincidence pattern $P_N(\theta_i,m)$ along the principal contour [Eq. (11)], the artifact peaks are generated by integration of the coincidence peaks lying on the secondary traces contributed by some other source(s). This is a result of high-frequency ambiguity. These can also be seen in Fig. 7 for the one-talker case. However, this kind of peak is generally extremely low, compared to the peaks associated with real sources, in $H_N(\theta_d)$ after our broadband integration. Nonetheless, there are two general contributing factors to artifactual peaks (neither of which is sufficient for the occurrence of artifactual peaks).

(a) In the 3D coincidence pattern $P_N(\theta_i,m)$, the peaks of coincidence are more pronounced at the intersections of the secondary coincidence traces. In Appendix C we give an example showing the intersections between the coincidence patterns for two talkers from 80° and −80°, and how they produce the artifactual peaks in Fig. 10(A), after integration along the primary contours. This is the main contributor to the artifactual peaks in all our tests. However, often this problem does not exist; for instance, our solution of Eq. (C3) indicates that the secondary coincidence traces of two talkers have no intersections when they are less than 20° apart from each other, and this separation can increase up to 50° when the two talkers are in a lateral sound field on the same side. The number of intersections increases when the two talkers are farther away from each other, and the maximal number

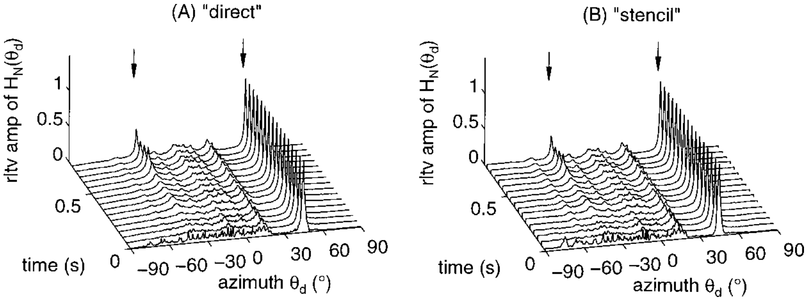

FIG. 11. Effects of relative intensity on localization: a weak talker at −40° and a stronger talker (10 dB higher in average intensity) at 45°. Both the direct (A) and stencil (B) methods can barely detect the weaker source. However, localization becomes easier as the weak source lasts longer. The localization computation was conducted once every 5 ms, but the display is shown every 50 ms for clarity.

of intersections under our experimental setup is ten.

(b) Another cause of the artifactual peak is high coincidence peaks at the high-frequency end of the secondary traces. (As discussed in Sec. IV B 1, due to increased phase estimation error, each secondary trace is high and narrow at the high frequency end and gradually becomes lower and broader toward the low-frequency end.) Sometimes these peaks are almost as conspicuous as the coincidence peaks in $H_N(\theta_d)$ associated with real sources [e.g., Fig. 7(E) and Fig. 9(E)], thereby making it difficult to distinguish which are real and which are artifacts. Yet another cause is when a sound has an unusually large amount of energy concentrated in selected frequencies. This sound will elicit prominent coincidence peaks at those frequencies in both the primary and secondary traces—the latter can produce artifactual peaks in $H_N(\theta_d)$. However, this situation is not applicable to a wide-band signal like speech.

A single coincidence peak can lie only on one principal contour but on multiple secondary contours. Hence, each peak contributes to the calculation of $H_N(\theta_d)$ at multiple azimuths in the stencil integration method and could contribute to artifactual peaks in $H_N(\theta_d)$ at these azimuths. However, these artifactual peaks are generally of low magnitude and readily removed by thresholding methods.

The artifact issue is less important when there are only a few (e.g., less than four) talkers. In that case, the chance of generating artifactual peaks is small, and these peaks are generally so weak that most can be separated from the real peaks using thresholding methods. However, as the number of talkers increases, it becomes more difficult to reject artifacts.

### 3. Multitalker tests in an anechoic room

To determine the system performance in environments close to real-world situations, we carried out on-site acoustic tests where loudspeakers delivered prerecorded sound from multiple talkers. Spondaic words recorded from four talkers (two males and two females) were presented from an equal distance to the input microphones at equal average intensity in an anechoic room (the Frances Searle Hall at Northwestern University) under several configurations (see Table II). The four speech segments were presented simultaneously. The instantaneous response of our system initially showed many scattered response peaks due to random interaction between sources. Some of these indicated the locations of the sources and others were artifacts. Within 100–200 ms, the response peaks associated with the real sources remained or became more prominent but the scattered artifactual peaks were attenuated. The direct and stencil methods could generally determine the locations of three out of the four sources accurately during the first 50 ms. However, it usually took 200–300 ms for the system to determine the presence and locations of all four sources.

For the purpose of illustration, the result of one test (configuration #3 from Table II) is shown in detail in Fig. 13. Each slice of the 3D graphic at instant $n=N$ represents the integration result $H_N(\theta_d)$. The localization computation was conducted once every 5 ms, but the display is shown every 50 ms for clarity. Figure 13(A) shows the result of the direct method. The sources F1 at −50° and M1 at 10° were easily detected. The presence of the source F2 at 45° was detected with an estimation error of 8° (indicated by the arrow with clear face color at 37°). The error was due to the influence of the nearby intersection of the coincidence patterns of the sources at 10° and −80°. An artifact source showed up at −18° (indicated by arrow ⇓), which was caused by both the intersection of the coincidence patterns of the sources at 10° and −50° and the intersection between sources at 10° and 45°. The source M2 at −80° was completely missed by the direct method. In contrast, Fig. 13(B) shows that all four sources were successfully detected by the stencil method. The magnitude of the peak corresponding to F2 at 45° was enhanced due to incorporation of the secondary contours, although the estimation error remained large. The most significant improvement is that the lateral source M2 at −80° becomes more conspicuous.

The results for the other tests are outlined in text.

(a) Configuration #1: Both the direct and stencil methods were able to detect three sources (0°, 20°, 75°). Both methods began to target the three sources correctly at the 30th ms. This correct detection stayed until the end. The peak at 75° was relatively low for the first 500 ms because the source energy was low, but was more pronounced for the last 400 ms. The direct method failed to identify the source at 75° for 60 ms at 290th ms. The source at 20° was localized with 3° error for the first 500 ms, but the error grew to 11° for the last 400 ms due to the increasingly stronger interaction between the sources at 0° and 75° that occurred near 20°. Both methods missed the −75° source entirely. Both methods had two consistent artifact peaks, one at −21° (due to strong ambiguous coincidence peaks of the 0° source at high frequencies) and one at −54° (due to ambiguous coincidence peaks of the 20° source at high frequencies).

(b) Configuration #2: This result was the best among the five configurations tested. Both methods successfully detected all four sources for their entire duration. With both methods, the localization errors were smaller than 2° except that the 30° source had a localization error of 8° for the first 360 ms, perhaps due to the interaction between the 0° and 60° sources. The direct method had an artifactual peak at 2° for the last 400 ms due to ambiguous coincidence peaks of the −45° source at high frequencies, while the stencil method did not have any artifacts.

(c) Configuration #4: Both methods localized the 5° source entirely and the 15° source after the first 230 ms. Both methods also detected the −30° source but with an 8° error for the first 600 ms, due presumably to the interaction between the sources at 15° and −60°. The direct method detected the −60° source entirely but with a 5° error. The stencil method detected the same source only for the first 850 ms but with a 2° error; then the −60° peak was replaced by an artifact peak around −72° for the last 600 ms, due to the interaction between sources at 15°, −30°, and −60°. Both methods had artifactual peaks at 31° (from interaction between sources −60° and 5°) and 42° (from interaction between sources 15°, −30°, and −60° and between sources −30°, and 5°), for the first 1.1 s (direct method) and first 850 ms (stencil method). The direct method also had an artifac-

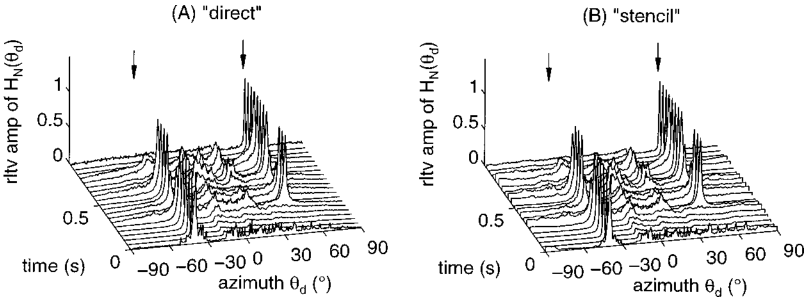

FIG. 12. Effects of different onset times on localization. The onset time of speech from source at 45° lagged behind that from source at −40° by 195 ms. Both the direct (A) and stencil (B) methods can respond to the onset and offset of the signals. The localization computation was conducted once every 5 ms, but the display is shown every 50 ms for clarity.

tual peak around −6° for the last 500 ms (from interaction between sources 15°, −30°, and −60°).

(d) Configuration #5: Both methods localized the sources at −70°, −25°, and 10° correctly, but missed the 80° source entirely. Both methods had artifactual peaks around 61° (due to the interaction between the sources at −70° and 25°) and −2° (due to the interaction between the sources at 25° and −25).

In summary, of the five configurations in four-talker tests, almost all the sources are successfully localized by the two methods with the following exceptions: they missed the −75° source entirely in configuration #1 and the 80° source entirely in configuration #5, and the direct method missed the −80° source entirely in configuration #3. There were on average 1–2 artifact peaks in each test.

One important issue related to the performance evaluation is how we define hits and misses. We evaluated the system performance using three criteria for hits (±2°, ±5°, or ±10°). Several results are notable: (a) The azimuth estimations were generally quite accurate (for more than 60% of our trials, the error was less than 2°). (b) The accuracy of azimuth estimations was highest for sources located near the midline, and the estimate was less accurate for sources located more laterally. (c) As the hit threshold was relaxed, more sources were detected and localized, but the errors increased. (d) There was no significant difference between the direct and stencil methods in terms of estimation accuracy. However, most of the lateral sources missed by the direct method could be reliably localized by the stencil method.

A further evaluation of misses is shown in Fig. 14. The evaluation was made over the period of 200–600 ms (i.e., the middle segment of test signals) when the localization was most stable. Three different hit criteria (±2°, ±5°, and ±10°) were used, and the hit/miss was determined once for each frame, i.e., every 5 ms. We note that the frequency of misses was very small for sources close to the midline. For sources located more laterally, however, the source detection degraded as the frequency of misses increased. As before, the criterion for hits also played a role in the result. As the threshold was relaxed, the number of misses decreased. When comparing the left and right columns, it can be seen that the stencil method is more effective in localizing the lateral sources. This result was consistent with our analysis with the one-talker and two-talker computer-simulation tests.

TABLE II. Summary of experimental setups for four-talker anechoic chamber tests.

| Speaker<br>Word | | M1<br>*''armchair''* | M2<br>*''playground''* | F1<br>*''pancake''* | F2<br>*''woodwork''* |
|---|---|---|---|---|---|
| Config. idx | 1 | −75° | 0° | 20° | 75° |
| | 2 | 30° | −45° | 60° | −10° |
| | 3 | 10° | −80° | −50° | 45° |
| | 4 | −30° | 15° | 5° | −60° |
| | 5 | −25° | 25° | −70° | 80° |

To evaluate whether the system performance was talker- and/or word dependent, we conducted four talker tests under the same experimental conditions in the anechoic chamber using a different group of spondaic words from a different group of talkers. These tests revealed that the system performance showed no systematic difference across talkers, or speech materials, in terms of source detection and localization accuracy. As before, the stencil method was superior in detecting and localizing the sources at the lateral side.

## C. Discussion

### 1. Artifacts

In terms of sound localization, there were two kinds of artifacts: (i) classification artifacts, i.e., those response peaks that corresponded to a real source but were classified as artifacts due to the strict criterion for hits (e.g., the arrow with clear face color at 37° in Fig. 13), and (ii) phantom responses, i.e., the responses that did not correspond to real sources (e.g., arrow ⇓ at −18° in Fig. 13). In Sec. V B 2 d, we have given analyses about the origin and location of the artifacts in the situation of two talkers. The principle is easily extended to the situations with more talkers. For example, the artifact at −18° had a dual origin: the intersection between the coincidence patterns of the sources at 10° and −50° and the intersection between the sources at 10° and 45°.

Elimination of the artifacts can be accomplished by using a more sophisticated method for recognizing the 3D coincidence pattern $P_N(\theta_i,m)$. However, this generally comes with a price of more intensive computation. A modified stencil method is being pursued in our continuing research.

### 2. Maximal number of talkers

Due to the constraints of our experimental facility, evaluation of the system performance for more than four talkers was conducted only in the computer simulation. In the computer simulation, results from six-talker situations were not very different from results from four-talker situations. Specifically, of our 7 six-talker configurations, all the sources could be successfully localized (except that in two configurations the direct method missed a lateral source) and

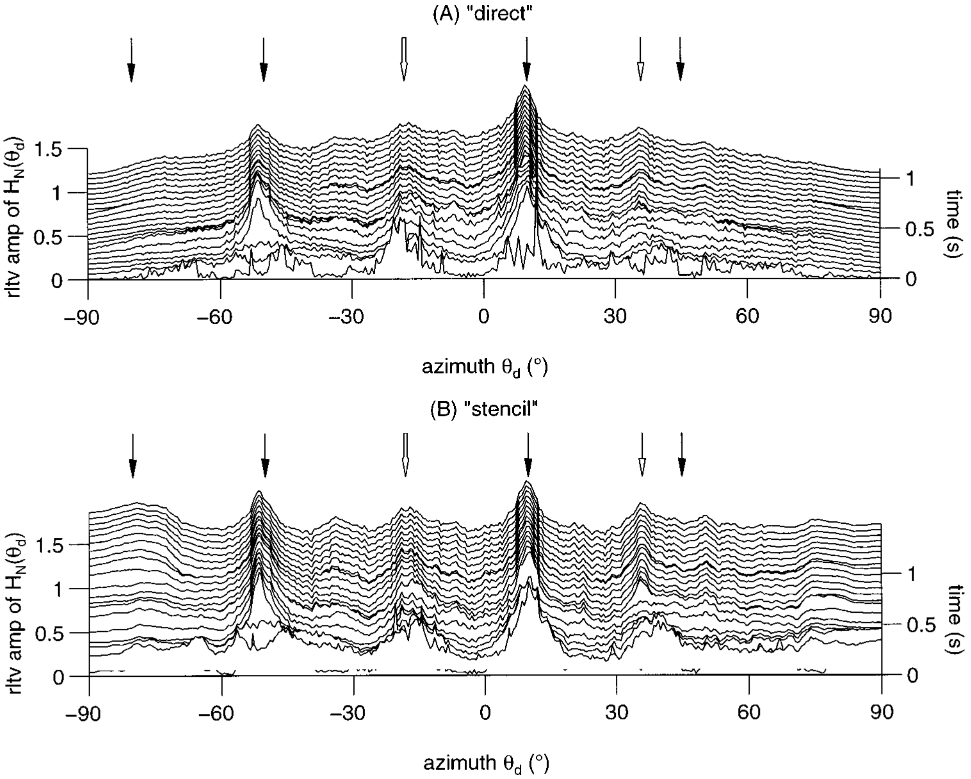

FIG. 13. Localization of four talkers (configuration #3 in Table II) in an anechoic chamber using the direct (A) and stencil (B) methods. The actual locations are shown by single tail, black face-color arrows. The single tail, clear face-color arrows (at 37°) indicate the estimated location of source F2 at 45°. The arrows ⇓ indicate the artifactual peaks that do not correspond to any of the real sources. The direct method missed the lateral source M2 at −80° (A), while the stencil method successfully localized all four sources (B). The localization computation was conducted once every 5 ms, but the display is shown every 50 ms for clarity.

there were at most 1–2 artifact peaks in a configuration (Liu *et al.*, 1997). In general, the performance in the more complex environment was slightly inferior to the results from the less complex environment. What, then, is the maximum number of talkers that can be localized by the system? This is difficult to determine because, in addition to the number of talkers, the system performance also depends on the relative time of occurrence, the spectrum, and intensity of concurrent phonemes, the number and duration of pauses in the speech, as well as the relative locations of the talkers in the auditory space. The limit of the system is mainly attributed to two kinds of problems. The first is related to the coincidence

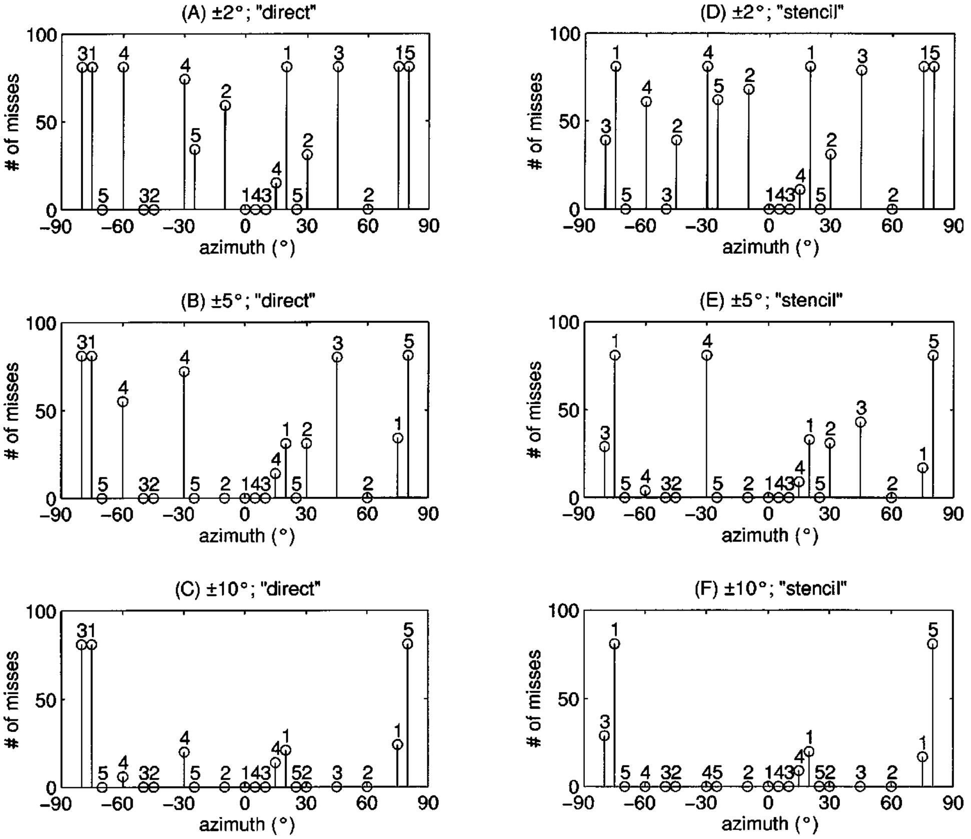

FIG. 14. Number of misses incurred in the four-talker anechoic chamber tests. The hit/miss was determined once for each frame (every 5 ms). The left panels (A), (B), (C) are from the direct method; the right panels (D), (E), (F) from the stencil method. The hit threshold was ±2°, ±5°, and ±10°, respectively, for each row. The number above each stem denotes the index of the test configuration. As the hit threshold is higher, the number of misses decreases. For both methods, the sources closer to the midline were more easily detected while lateral sources were more often missed. However, the stencil method can localize more sources than the direct method, especially when the hit threshold is large [compare (B) and (E), (C) and (F)].

detection step. When the duration of one source masked by others is longer, the spectrum of one source masked by others is wider, or the intensity of one source is lower relative to others, it is more difficult for the source to be detected in the narrow-band coincidence detection. The second is related to the recognition of the 3D coincidence pattern. When a source is located very laterally, when the number of sources is large, or when the sound duration is short, it is difficult for the source to be detected by temporal and broadband integration. All these factors are being studied for a deeper understanding of the method proposed.

### 3. Head-shadow effect

The present system was applied without any object between the two microphones. However, any diffraction from an object between the two microphones, such as the head-shadow effect, will complicate our theory. The diffraction will introduce both sound-pressure change and phase shift, as functions of size of the object, frequency, and the direction of incidence. It can be interpreted as filtering with the transfer function characteristic of the diffraction. Theoretically this effect can be removed by filtering with the reciprocal of the diffraction transfer function. This filter can be incorporated into the delay units of the dual delay line. Although it is possible to calculate the filter coefficients based on the shape of the object, we suggest that the values be measured empirically. However, the added filtering will likely degrade the computational performance of the system. How well the filtering will compensate for the deviation from ideal introduced by head-shadow diffraction is not known.

## VI. SUMMARY AND CONCLUSION

In this research, we developed a signal-processing system for determining the spatial locations of multiple sound sources with two microphones. The system is based on a dual delay-line structure, which detects the intermicrophone time difference (which depends on the source azimuth) using a coincidence detection method. The direction information is mapped, in a nonlinear manner, from the dual delay line onto a 3D coincidence pattern. Both temporal integration and spectral integration are employed to take full advantage of the broad bandwidth such that the source azimuth can be estimated reliably in adverse noisy conditions. We also developed a novel stencil filter to enhance the localization of sound sources in lateral fields.

Results show that our broadband localization technique works well in complex auditory scenes involving four talkers in an anechoic space or six talkers in computer simulation. The limitations of the method require further investigation. Work in progress involves extending the results to real rooms over a range of reverberation times.

A number of problems remain to be addressed. First, because we do not have the elevation asymmetry, which in the case of humans results from the filtering effects of head, body, and pinna, the present version of the system cannot localize vertically. However, simple geometry argues that a second pair of microphones oriented vertically could identify elevation. Second, as shown in the experimental results, some artifacts in the localization need to be removed by a more sophisticated pattern recognition method. Third, the dual delay-line architecture, as well as parallel nature of computation is suited for very large-scale integrated (VLSI) implementation and miniaturization for eventual use. For instance, Mead *et al.* (1991) has reported a VLSI implementation of neural network architecture for a silicon cochlea. These problems are the focus of ongoing investigation. In our research, the overall goal is to not only localize the sources of speech sounds but also to selectively extract the speech from one (or more) talker(s). What is shown in this paper is the ability of the system to detect and locate the sources. In a separate paper, we shall describe how the system can extract the speech from one of the talkers with high fidelity.

## ACKNOWLEDGMENTS

This research was supported by grants from the Beckman Institute at the University of Illinois at Urbana-Champaign, and the Mary Jane Neer Research Fund. The work was conducted while the first author (C.L.) was a Beckman Fellow. We thank Jiri Jonas for his enthusiasm for this project, Terry Takahashi for suggestions in the early phase of the research, Dean Garstecki for permission to use the anechoic chamber at Northwestern University, Susanne Aref for advice on data analysis, and Benjamin Grosser, John Prato, Youngmin Kim, and Huiming Zhang for assistance in some of the experiments. A number of staff members in the Beckman Institute voluntarily participated in the recording of speech materials used in the tests; their participation is appreciated. The facilities of the Visualization Laboratory at the Beckman Institute were employed in part of the experiments.

## APPENDIX A

In order to illustrate the coincidence detection method, we consider a simplified dual delay line at frequency $\omega_m = 2\pi f_m$ (Fig. A1). Illustrated is a case where there are two sources: source 1 and source 2. It will be seen that the conclusion applies to cases involving more sources. Suppose that the coincidence locations of source 1 and source 2 are at $i_{\text{source 1}} = s$ and $i_{\text{source 2}} = g$ along the dual delay line, respectively, and that the in-phase signal source 1 at $i_{\text{source 1}} = s$ is $A_s \exp[j(\omega_m t + \phi_s)]$ and the in-phase signal source 2 at $i_{\text{source 2}} = g$ is $A_g \exp[j(\omega_m t + \phi_g)]$. Thus, at an arbitrary location $i$ along the dual delay line, the signals from the left and right channels are, respectively,

$$X_L^{(i)}(m) = A_s \exp j[\omega_m(t + \tau_s - \tau_i) + \phi_s] + A_g \exp j[\omega_m(t + \tau_g - \tau_i) + \phi_g], \tag{A1}$$

and

$$X_R^{(i)}(m) = A_s \exp j[\omega_m(t + \tau_{I-s+1} - \tau_{I-i+1}) + \phi_s] + A_g \exp j[\omega_m(t + \tau_{I-g+1} - \tau_{I-i+1}) + \phi_g], \tag{A2}$$

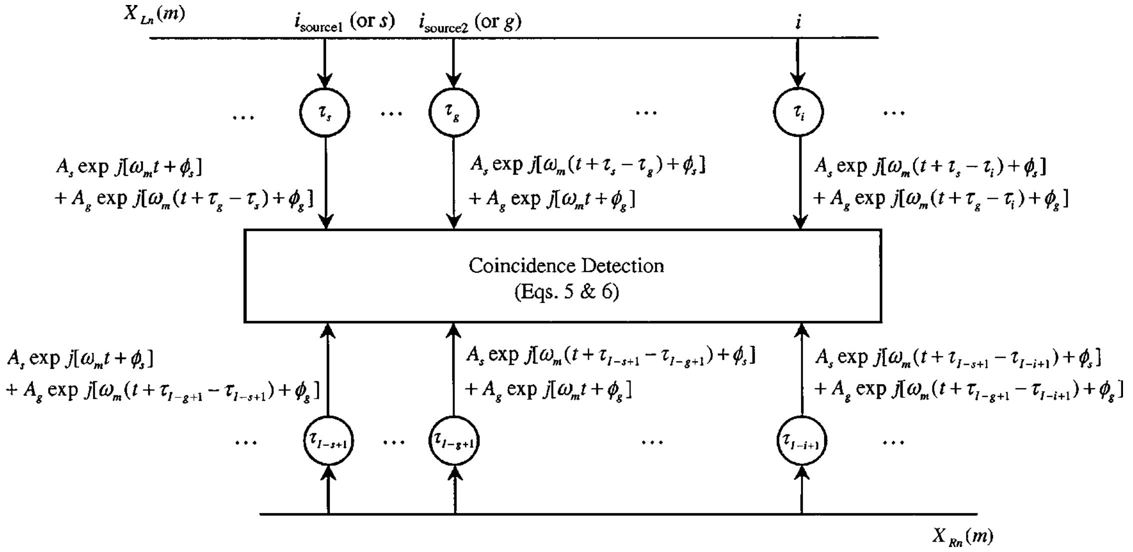
$X_{Ln}(m)$
$i_{\rm source1}$ (or $s$)
$i_{\rm source2}$ (or $g$)
$i$
$\tau_s$
$\tau_g$
$\tau_i$
$A_s\exp j[\omega_m t+\phi_s] + A_g\exp j[\omega_m(t+\tau_g-\tau_s)+\phi_g]$
$A_s\exp j[\omega_m(t+\tau_s-\tau_g)+\phi_s] + A_g\exp j[\omega_m t+\phi_g]$
$A_s\exp j[\omega_m(t+\tau_s-\tau_i)+\phi_s] + A_g\exp j[\omega_m(t+\tau_g-\tau_i)+\phi_g]$
Coincidence Detection
(Eqs. 5 & 6)
$A_s\exp j[\omega_m t+\phi_s] + A_g\exp j[\omega_m(t+\tau_{I-g+1}-\tau_{I-s+1})+\phi_g]$
$A_s\exp j[\omega_m(t+\tau_{I-s+1}-\tau_{I-g+1})+\phi_s] + A_g\exp j[\omega_m t+\phi_g]$
$A_s\exp j[\omega_m(t+\tau_{I-s+1}-\tau_{I-i+1})+\phi_s] + A_g\exp j[\omega_m(t+\tau_{I-g+1}-\tau_{I-i+1})+\phi_g]$
$\tau_{I-s+1}$
$\tau_{I-g+1}$
$\tau_{I-i+1}$
$X_{Rn}(m)$

FIG. A1. The dual delay-line structure.

where $\tau_s$, $\tau_{I-s+1}$, $\tau_g$, $\tau_{I-g+1}$, $\tau_i$, and $\tau_{I-i+1}$ are values of delay units at locations *s*, *g*, and *i*, in the left and right channels, respectively.

The coincidence detection calculates the value of $|X_L^{(i)}(m)-X_R^{(i)}(m)|$ [Eq. (6)]. After substituting Eqs. (A1) and (A2), we obtain the following:

$$|X_L^{(i)}(m)-X_R^{(i)}(m)|=2\Big\{A_s^2\sin^2\frac{\omega_m}{2}[(\tau_s-\tau_i)-(\tau_{I-s+1}-\tau_{I-i+1})]+A_g^2\sin^2\frac{\omega_m}{2}[(\tau_g-\tau_i)-(\tau_{I-g+1}-\tau_{I-i+1})]$$
$$+2A_sA_g\sin\frac{\omega_m}{2}[(\tau_s-\tau_i)-(\tau_{I-s+1}-\tau_{I-i+1})]\sin\frac{\omega_m}{2}[(\tau_g-\tau_i)-(\tau_{I-g+1}-\tau_{I-i+1})]\cos\kappa_{s,g}(m)\Big\}^{1/2}, \tag{A3}$$

where $\kappa_{s,g}(m)=(\omega_m/2)(\tau_s-\tau_g+\tau_{I-s+1}-\tau_{I-g+1})+(\phi_s-\phi_g)$ in the cross term.

Consider the relation between the values $\tau_i$ and $\tau_{I-i+1}$ of any paired delay units. By employing Eq. (1), we have

$$\tau_{I-i+1}=\frac{\mathrm{ITD}_{\max}}{2}\sin\left[\frac{(I-i+1)-1}{I-1}\pi-\frac{\pi}{2}\right]$$
$$=-\frac{\mathrm{ITD}_{\max}}{2}\sin\left(\frac{i-1}{I-1}\pi-\frac{\pi}{2}\right)=-\tau_i. \tag{A4}$$

Similarly, we have $\tau_{I-s+1}=-\tau_s$ and $\tau_{I-g+1}=-\tau_g$. Using these relations, Eq. (A3) can be further simplified such that

$$|X_L^{(i)}(m)-X_R^{(i)}(m)|$$
$$=2[A_s^2\sin^2\omega_m(\tau_s-\tau_i)+A_g^2\sin^2\omega_m(\tau_g-\tau_i)+2A_sA_g$$
$$\times\sin\omega_m(\tau_s-\tau_i)\sin\omega_m(\tau_g-\tau_i)\cos\kappa_{s,g}(m)]^{1/2}. \tag{A5}$$

It is easy to see that Eq. (A5) has the form

$$\mathrm{term}(s-i)+\mathrm{term}(g-i)+\mathrm{cross\ \ term}. \tag{A6}$$

It is the first and second terms that indicate the directions of the two sources.

## APPENDIX B

In order to obtain the expression of the stencil pattern, we focus on either of the first two terms in Eq. (A5). We use the subscript *d* to denote the source being focused. Referring to Eq. (5), the minimum point of $|\sin^2\omega_m(\tau_d-\tau_i)|$ should correspond to the direction of the source *d*, thus,

$$\arg\min_i|\sin^2\omega_m(\tau_d-\tau_i)|=\{i:\omega_m(\tau_d-\tau_i)=\gamma_{d,m}\pi,$$
$$\gamma_{d,m}\ \text{is an integer}\}. \tag{B1}$$

After substituting Eq. (2) for $\tau_d$ and $\tau_i$ into Eq. (B1) and considering that $\omega_m=2\pi f_m$, we can get the expression Eq. (12) of the stencil pattern. Considering $-1\leqslant\sin\theta_i\leqslant1$, the range of $\gamma_{d,m}$ in Eq. (13) can be derived from Eq. (12).

## APPENDIX C

Let us calculate the intersections between the coincidence trace patterns of two sources. We assume the azimuths of the two sources are $\theta_d'$ and $\theta_d''$, respectively. Since the coincidence pattern of each source is determined by Eq. (13), the intersections of the two patterns can be obtained by solving the following system:

TABLE CI. The calculated intersection points between the coincidence traces of the two sound sources located at 80° and −80°, respectively. The values in boldface are the intersections by the secondary traces.

| $\gamma''_{d,m}$ $\gamma'_{d,m}$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| −5 | −80° | ⋯ | ⋯ | ⋯ | ⋯ | ⋯ |
| −4 | −80° | **−36.2°** | ⋯ | ⋯ | ⋯ | ⋯ |
| −3 | −80° | **−29.5°** | **−11.4°** | ⋯ | ⋯ | ⋯ |
| −2 | −80° | **−19.2°** | **0°** | **11.4°** | ⋯ | ⋯ |
| −1 | −80° | **0°** | **19.2°** | **29.5°** | **36.2°** | ⋯ |
| 0 | ⋯ | 80° | 80° | 80° | 80° | 80° |

$$\begin{aligned} \sin\theta'_i - \sin\theta'_d &= \frac{\gamma'_{d,m}}{\mathrm{ITD}_{\max} f_m} \\ \sin\theta''_i - \sin\theta''_d &= \frac{\gamma''_{d,m}}{\mathrm{ITD}_{\max} f_m}. \end{aligned} \tag{C1}$$

The solution is

$$\begin{aligned} \sin\theta'_i = \sin\theta''_i &= \frac{\gamma''_{d,m}\sin\theta'_d - \gamma'_{d,m}\sin\theta''_d}{\gamma''_{d,m} - \gamma'_{d,m}} \\ f_m &= \frac{\gamma''_{d,m} - \gamma'_{d,m}}{\mathrm{ITD}_{\max}(\sin\theta'_d - \sin\theta''_d)}. \end{aligned} \tag{C2}$$

The number of the intersection points is not only determined by the validity of Eq. (C2) but also the ranges of $\gamma'_{d,m}$ and $\gamma''_{d,m}$ [Eq. (13)]. Thus, the contraints are

$$\begin{aligned} &-1 \leqslant \frac{\gamma''_{d,m}\sin\theta'_d - \gamma'_{d,m}\sin\theta''_d}{\gamma''_{d,m} - \gamma'_{d,m}} \leqslant 1 \\ &0 \leqslant \frac{\gamma''_{d,m} - \gamma'_{d,m}}{\mathrm{ITD}_{\max}(\sin\theta'_d - \sin\theta''_d)} \leqslant f_s/2 \\ &-\mathrm{ITD}_{\max} f_m (1+\sin\theta'_d) \leqslant \gamma'_{d,m} \leqslant \mathrm{ITD}_{\max} f_m (1-\sin\theta'_d) \\ &-\mathrm{ITD}_{\max} f_m (1+\sin\theta''_d) \leqslant \gamma''_{d,m} \leqslant \mathrm{ITD}_{\max} f_m (1-\sin\theta''_d). \end{aligned} \tag{C3}$$

The number of intersections can be obtained by numerically solving the above system.

We calculate for the intersections of the two sources in Fig. 10(A); thus $\theta'_d = -80°$ and $\theta''_d = 80°$. The result is displayed in Table CI. There are 20 intersection points altogether between the coincidence traces of the two speakers, among which 10 intersections might contribute the artifacts at 9 different places through the direct integration.

Banks, D. (**1993**). ''Localisation and separation of simultaneous voices with two microphones,'' IEE Proc.-I Commun. Speech Vision **140**, 229–234.

Blauert, J. (**1980**). ''Modelling of interaural time and intensity difference discrimination,'' in *Psychophysical, Physiological and Behavioural Studies in Hearing: Proceedings of the 5th International Symposium on Hearing*, edited by G. V. D. Brink and F. A. Bilsen (Delft University Press, Delft, Noordwijkerhout, The Netherlands), pp. 421–424.

Blauert, J. (**1983**). *Spatial Hearing: The Psychophysics of Human Sound Localization*, translated by John S. Allen (The MIT Press, Cambridge, MA).

Bodden, M. (**1993**). ''Modeling human sound source localization and the cocktail-party-effect,'' Acta Acust. (China) **1**, 43–55.

Bronkhorst, A. W., and Plomp, R. (**1992**). ''Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing,'' J. Acoust. Soc. Am. **92**, 3132–3139.

Cherry, E. C. (**1953**). ''Some experiments on the recognition of speech, with one and with two ears,'' J. Acoust. Soc. Am. **25**, 975–979.

Colburn, H. S. (**1973**). ''Theory of binaural interaction based on auditory-nerve data. I. General strategy and preliminary results on interaural discrimination,'' J. Acoust. Soc. Am. **54**, 1458–1470.

Colburn, H. S. (**1977**). ''Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise,'' J. Acoust. Soc. Am. **61**, 525–533.

Colburn, H. S. (**1996**). ''Computational models of binaural processing,'' in *Auditory Computation*, edited by H. L. Hawkins, T. A. McMullen, A. N. Popper, and R. R. Fay (Springer, New York), pp. 332–400.

Colburn, H. S., and Durlach, N. I. (**1978**). ''Models of binaural interaction,'' in *Handbook of Perception, IV: Hearing*, edited by E. C. Carterette and P. F. Morton (Academic, New York).

Colburn, H. S., Han, Y., and Culotta, C. P. (**1990**). ''Coincidence model of MSO responses,'' Hear. Res. **49**, 335–346.

Durlach, N. I., and Colburn, H. S. (**1978**). ''Binaural phenomena,'' in *Handbook of Perception, IV: Hearing*, edited by E. C. Carterette and P. F. Morton (Academic, New York).

Feng, A. S., and Shofner, W. P. (**1981**). ''Peripheral basis of sound localization in anurans: Acoustic properties of the frog's ear,'' Hear. Res. **5**, 201–216.

Flanagan, J. L., Johnston, J. D., Zahn, R., and Elko, G. W. (**1985**). ''Computer steered microphone arrays for sound transduction in large rooms,'' J. Acoust. Soc. Am. **78**, 1508–1518.

Flanagan, J. L. (**1972**). *Speech Analysis, Synthesis and Perception* (Springer, Berlin).

Gaik, W. (**1993**). ''Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling,'' J. Acoust. Soc. Am. **94**, 98–110.

Gourevitch, G. (**1987**). ''Binaural hearing in land mammals,'' in *Directional Hearing*, edited by W. A. Yost and G. Gourevitch (Springer, New York), pp. 226–246.

Jeffress, L. A. (**1948**). ''A place theory of sound localization,'' J. Comp. Physiol. Psychol. **41**, 35–39.

Konishi, M., Takahashi, T., T., Wagner, H., Sullivan, W. E., and Carr, C. E. (**1988**). ''Neurophysiological and anatomical substrates of sound localization in the owl,'' in *Auditory Function: Neurobiological Bases of Hearing*, edited by M. E. Gerald, W. E. Gall, and W. M. Cowan (Wiley, New York), pp. 721–745.

Lindemann, W. (**1986**). ''Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals,'' J. Acoust. Soc. Am. **80**, 1608–1622.

Liu, C., Feng, A. S., Wheeler, B., O'Brien, W. D., Bilger, R. C., and Lansing, C. (**1997**). ''A binaurally based auditory processor effectively extracts speech in the presence of multiple competing sounds,'' Hearing Aid Research and Development Conference, NIH, Bethesda, MD, Sept., 1997.

Mead, C. A., Arreguit, X., and Lazzaro, J. (**1991**). ''Analog VLSI model of binaural hearing,'' IEEE Trans. Neural Netw. **2**, 230–236.

Mills, A. W. (**1972**). ''Auditory localization,'' in *Foundations of Modern Auditory Theory*, edited by V. T. Jerry (Academic, New York), pp. 301–348.

Shamma, S. A., Shen, N., and Gopalaswamy, P. (**1989**). ''Stereausis: Binaural processing without neural delays,'' J. Acoust. Soc. Am. **86**, 989–1006.

Stern, R. M., and Trahiotis, C. (**1992**). ''The role of consistency of interaural time over frequency in binaural lateralization,'' in *Auditory Physiology and Perception, Proceedings of the 9th International Symposium on Hearing*, edited by Y. Cazals, L. Demany, and K. Horner (Pergamon, Oxford, Carcens, France), pp. 547–554.

Stern, R. M., and Trahiotis, C. (**1995**). ‘‘Models of binaural interaction,’’ in *Hearing*, edited by B. C. J. Moore (Academic, San Diego), pp. 347–386.

Stern, R. M., and Trahiotis, C. (**1997**). ‘‘Models of binaural perception,’’ in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. H. Gilkey and T. R. Anderson (Erlbaum, Mahwah, NJ), pp. 499–532.

Stern, R. M., Zeiberg, A. S., and Trahiotis, C. (**1988**). ‘‘Lateralization of complex binaural stimuli: A weighted-image model,’’ J. Acoust. Soc. Am. **84**, 156–165.

Takahashi, T. T., and Keller, C. H. (**1994**). ‘‘Representation of multiple sound sources in the owl’s auditory space map,’’ J. Neurosci. **14**, 4780–4793.

Yin, T. C. T., and Chan, J. C. K. (**1990**). ‘‘Interaural time sensitivity in medial superior olive of cat,’’ J. Neurophysiol. **64**, 465–488.