A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers[a)]

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

(Received 28 June 2000; revised 14 June 2001; accepted 19 September 2001)

This paper describes algorithms for signal extraction for use as a front-end of telecommunication devices, speech recognition systems, as well as hearing aids that operate in noisy environments. The development was based on some independent, hypothesized theories of the computational mechanics of biological systems in which directional hearing is enabled mainly by binaural processing of interaural directional cues. Our system uses two microphones as input devices and a signal processing method based on the two input channels. The signal processing procedure comprises two major stages: (i) source localization, and (ii) cancellation of noise sources based on knowledge of the locations of all sound sources. The source localization, detailed in our previous paper [Liu *et al.*, J. Acoust. Soc. Am. **108**, 1888 (2000)], was based on a well-recognized biological architecture comprising a dual delay-line and a coincidence detection mechanism. This paper focuses on description of the noise cancellation stage. We designed a simple subtraction method which, when strategically employed over the dual delay-line structure in the broadband manner, can effectively cancel multiple interfering sound sources and consequently enhance the desired signal. We obtained an 8–10 dB enhancement for the desired speech in the situations of four talkers in the anechoic acoustic test (or 7–10 dB enhancement in the situations of six talkers in the computer simulation) when all the sounds were equally intense and temporally aligned. [DOI: 10.1121/1.1419090]

PACS numbers: 43.72.Ar, 43.72.Dv, 43.60.Bf [DOS]

I. INTRODUCTION

Selective hearing is a useful mechanism for extracting desired signals in complex acoustic environments such as a cocktail party. This so-called ''cocktail party'' effect has been studied psychophysically (Cherry, 1953; Blauert, 1983; Bregman, 1990; Bronkhorst and Plomp, 1992). The ability to hear in complex acoustic environments is largely attributed to the capacity to discern the spatial origins of sound sources. The neural circuitry and the underlying mechanisms for sound localization are fairly well established (Konishi *et al.*, 1988; Takahashi and Keller, 1994; Yin and Chan, 1990). Sound localization involves binaural processing of minute differences in time, intensity, and spectrum between the two ears. However, although we know the capacity of the auditory system to selectively attend to sounds originating from one source and suppress the other sounds in the ambiance, the underlying mechanisms for doing so are largely unknown. Therefore, designing an artificially intelligent system today to achieve selective hearing is still largely based on our relatively rich knowledge of the physical world (e.g., signal processing techniques) plus our limited knowledge of the biological world.

One of the prominent noise suppression concepts is the EC or equalization-and-cancellation scheme of Durlach (Durlach, 1960, 1972). It requires two inputs followed by a two-stage signal processing: (i) equalization that makes the noise components identical in both channels; and (ii) cancellation or subtraction of the noise components in one channel from those in the other channel. Actually most two-microphone-based noise cancellation techniques to date (e.g., Widrow *et al.*, 1975; Strube, 1981; Chabries *et al.*, 1982; Chazan *et al.*, 1988; Weiss, 1987; Peterson *et al.*, 1987) are essentially variants of the EC scheme and differ primarily in the procedures by which the filter parameters are adapted. Thus far, these have rendered satisfactory noise reduction only for situations in which there are one desired source and one noise source.

Our noise cancellation technique described herein also falls in this category. However, it is devised so as to cancel multiple noise sources more efficiently by capitalizing on the knowledge of the spatial directions of the sound sources in the environment. For the purpose of sound localization, we have designed a system (Liu *et al.*, 2000) based on a broadband ''dual delay-line'' structure and the coincidence detection principle of Jeffress (1948). Our noise cancellation technique also adopts the dual delay-line as the infrastructure.

So far the Jeffress model has been studied and various modifications have been developed to account for different psychological observations (see reviews in the book chapters by Colburn and Durlach, 1978; Colburn, 1996; and Stern and Trahiotis, 1995, 1997). It was only recently that the Jeffress model began to be considered for use in the extraction of

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

[b)]Present address: Motorola Labs, 1141 Opus Place, Downers Grove, IL 60515.
[c)]Electronic mail: a-feng@uiuc.edu

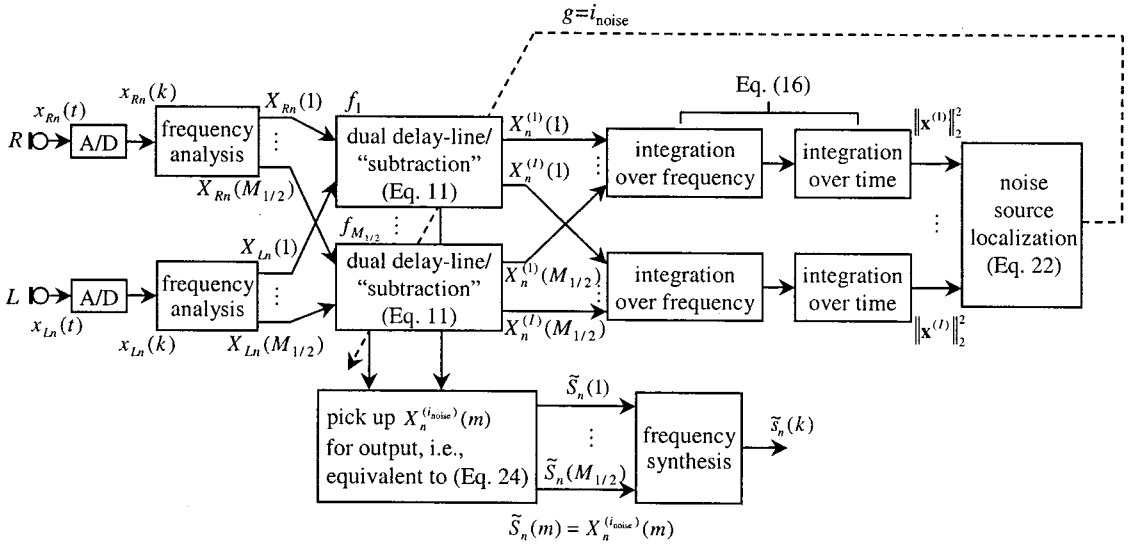

FIG. 1. The block diagram of System I for extraction of the desired source, whose location is known *a priori*, in the presence of one noise source whose location is estimated by the system.

speech in noise (e.g., Bodden, 1993; Banks, 1993). Since the model maps the acoustic space into a network, one appealing feature is the potential for detecting the number and the azimuths of sound sources present in the auditory space. This provides the mechanism by which the system can focus on, and extract the signal from, one desired source direction, while at the same time suppressing the sounds arising from the other directions. In his acoustic processor, Bodden (1993, 1996) basically took the Jeffress' coincidence sound localization models, as implemented by Lindemann (1986) and Gaik (1993), and added a time-variant Wiener filter for noise cancellation after sounds had been localized. However, since it is impossible to obtain an accurate estimate of the power density spectra of both the desired and noise signals, the result will always have residual noise and some cancellation and distortion of the desired signal.

The work described herein was motivated by the need to find a general solution for signal extraction in real world situations where there are multiple (>2) concurrent sound sources. Our signal extraction technique evolved from a subtraction procedure. Note that, interestingly, subtraction is also employed in the directional hearing mechanism with a pressure-gradient receiver (Feng and Shofner, 1981). Theoretically, a conventional noise cancellation system using a two-microphone array performs well when there are two sources but its performance degrades rapidly as the number of sources increases. To attack this problem, we developed a broadband noise cancellation strategy, making the two-microphone array subtraction approach more effective by taking advantage of the dual delay-line structure.

In this paper, we first introduce a subtraction method, which is the core of our noise cancellation technique. The subtraction procedure is then extended via the broadband dual delay-line structure for cancellation of multiple sources. In Sec. II A, we describe the subtraction procedure in the context of extracting a desired source at a known location in the presence of one interfering source at an arbitrary location. The subtraction operation is mathematically analyzed in Sec. II B. Section II C gives a beamforming interpretation of the subtraction method. In Sec. II D, the method is generalized to situations in which neither the location of the desired source nor that of the interference is known. Section III describes a strategy for extending the method to a system suitable for cancellation of multiple interfering sources. Section IV presents the experimental results and analysis. Discussion of several practical issues is given in Sec. V.

II. INTRODUCTION TO THE NEW CANCELLATION SCHEME

A. Cancellation algorithm based on the dual delay-line structure

In this section we will describe a new noise cancellation algorithm. It is fundamentally a subtraction operation applied on the two input signals. The signals are received by two microphones, which are paired with a fixed inter-microphone distance. The subtraction is conducted based on the infrastructure of the dual delay-line network in the frequency domain. A block diagram of the basic signal processing system (System I) is shown in Fig. 1. The two inputs, $x_{Ln}(t)$ and $x_{Rn}(t)$, are digitized, their digital versions being $x_{Ln}(k)$ and $x_{Rn}(k)$, respectively. Their spectra, $X_{Ln}(m)$ and $X_{Rn}(m)$, $m=1,\ldots,M$, are obtained through discrete Fourier transform (DFT). The subscripts L and R denote left and right channels, and n the frame index of the short-term Fourier analysis.

For clarity, we shall focus on the system description for an arbitrary frequency ω_m. For each frequency, the complex signals from the two channels are fed into a pair of delay-lines (Fig. 2), both of which are composed of I delay units with delay values τ_i $(i=1,\ldots,I)$ given by

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

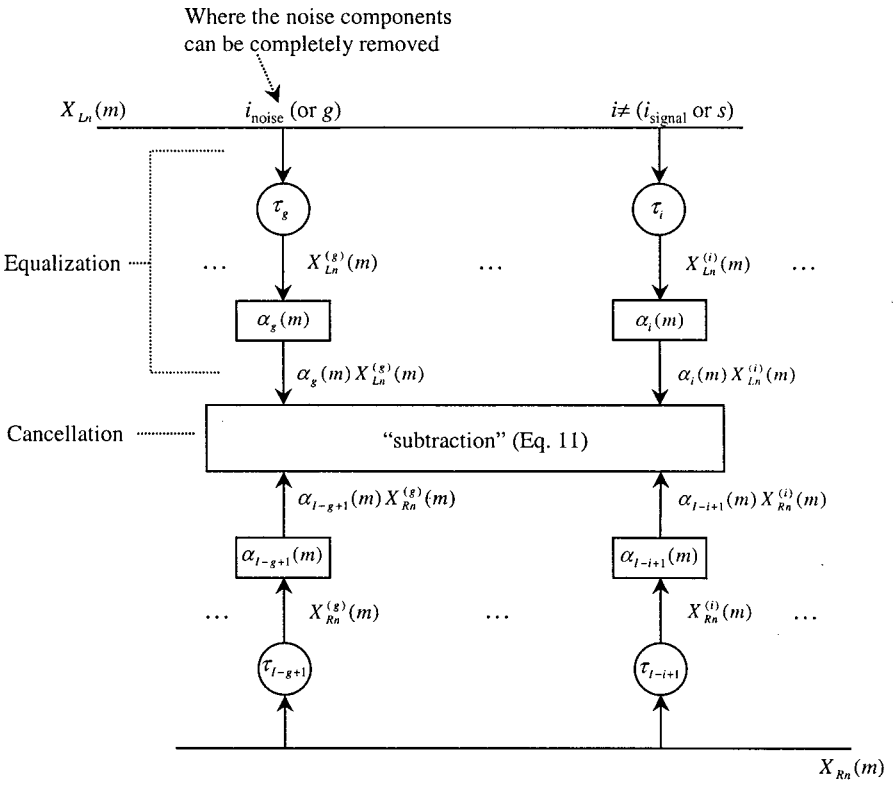

FIG. 2. The dual delay-line used as the basic structure in our system.

where $\mathrm{ITD}_{\max}=D/c$ is the maximum inter-microphone time difference, D is the inter-microphone distance, c is the speed of sound, and I is an odd number greater than 1. By using Eq. (1), it can be shown that the time delays are antisymmetric with respect to the midpoint $(I+1)/2$, i.e.,

$$\tau_{I-i+1}=-\tau_i, \quad i=1,\ldots,(I+2)/2-1. \tag{2}$$

It is noted that if there is no diffracting object between the two microphones, the horizontal plane can be uniformly divided into I sectors with the azimuth of each sector being

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

Therefore, the azimuths can be mapped one-to-one onto the corresponding positions in the delay line

$$\tau_i=-\frac{\mathrm{ITD}_{\max}}{2}\sin\,\theta_i, \quad i=1,\ldots,I. \tag{4}$$

Note that the resolution of the dual delay-lines representing the spatial azimuth is determined by the values of the time-delay units τ_i that have a time unit such as millisecond. As will be described in detail next, the delays are applied to the left and right input signals at each frequency in the frequency domain. Thus the dual delay-line works like rotating two separate phasors in the opposite directions in the complex plane [Eqs. (9) and (10)] until they are in phase, i.e., the so-called coincidence operation. The step size of the rotation, $\exp(-j\omega_m\tau_i)$, can be arbitrarily small in the frequency domain. Therefore, the azimuthal resolution of the dual delay-lines is not controlled by the sampling rate. Some other relevant discussions will be given in Sec. II C.

Figure 2 shows that the dual delay-line structure is similar to that adopted previously for sound localization (Liu *et al.*, 2000) except that a compensation element $\alpha_i(m)$ has been added following each delay unit. These elements, which compensate for differences in the intensity of noise at the two microphones, are functions of both azimuth and frequency. Appendix A derives the compensation values for the ideal case of point sources with distance-dependent amplitude decline, in a lossless medium. In practice, however, all the values of $\alpha_i(m)$ and τ_i $(i=1,\ldots,I)$ are to be adjusted empirically the same time when the system is being calibrated to compensate for asymmetries between the two microphones. The compensation factors remain fixed so long as the asymmetries are not changed. This fixed interaural intensity difference (IID) corresponding to each interaural time difference (ITD) mimics that observed in humans (Gaik, 1993). In the anechoic chamber tests reported below, the values of ITD units τ_i $(i=1,\ldots,I)$ were set uniformly while the values of IID $\alpha_i(m)$ $(i=1,\ldots,I)$ were determined empirically.

In this subsection let us suppose the direction of the desired source is known *a priori* and we use $i_{\mathrm{signal}}=s$ to denote the in-phase (coincident) position along the dual delay-line for the desired signal components. We use $i_{\mathrm{noise}}=g$ to denote the in-phase position for the noise signal components. Note that the position index along the dual delay-line is coincident with the index of the delay units in the left channel. After equalization, the in-phase desired signal components are identical in the left and right channels at $i_{\mathrm{signal}}=s$, which is assumed to be $S_n(m)=A_s\exp[j(\omega_m t+\phi_s)]$; likewise, the in-phase noise signal component is identical in the left and right channels at $i_{\mathrm{noise}}=g$, which is assumed to be $G_n(m)=A_g\exp[j(\omega_m t+\phi_g)]$, where ϕ_s and ϕ_g are the initial phases for signal and noise, respectively. Based on

these assumptions, the left and right channel input (microphone) signals are, respectively,

$$X_{Ln}(m)=\frac{1}{\alpha_s(m)}S_n(m)\exp(j\omega_m\tau_s)+\frac{1}{\alpha_g(m)}G_n(m)\exp(j\omega_m\tau_g) \quad (5)$$

and

$$X_{Rn}(m)=\frac{1}{\alpha_{I-s+1}(m)}S_n(m)\exp(j\omega_m\tau_{I-s+1})+\frac{1}{\alpha_{I-g+1}(m)}G_n(m)\exp(j\omega_m\tau_{I-g+1}). \quad (6)$$

Then, we can find the mathematical representation for the equalized signals $\alpha_i(m)X_{Ln}^{(i)}(m)$ for the left channel, and $\alpha_{I-i+1}(m)X_{Rn}^{(i)}(m)$ for the right channel at any arbitrary point i (except $i=s$), along the dual delay-line. They are

$$\alpha_i(m)X_{Ln}^{(i)}(m)=\frac{\alpha_i(m)}{\alpha_s(m)}S_n(m)\exp[j\omega_m(\tau_s-\tau_i)]+\frac{\alpha_i(m)}{\alpha_g(m)}G_n(m)\exp[j\omega_m(\tau_g-\tau_i)] \quad (7)$$

and

$$\alpha_{I-i+1}(m)X_{Rn}^{(i)}(m)=\frac{\alpha_{I-i+1}(m)}{\alpha_{I-s+1}(m)}S_n(m)\times\exp[j\omega_m(\tau_{I-s+1}-\tau_{I-i+1})]+\frac{\alpha_{I-i+1}(m)}{\alpha_{I-g+1}(m)}G_n(m)\times\exp[j\omega_m(\tau_{I-g+1}-\tau_{I-i+1})], \quad (8)$$

where

$$X_{Ln}^{(i)}(m)=X_{Ln}(m)\exp(-j\omega_m\tau_i) \quad (9)$$

and

$$X_{Rn}^{(i)}(m)=X_{Rn}(m)\exp(-j\omega_m\tau_{I-i+1}). \quad (10)$$

The subtraction step in the algorithm performs the following operation on each signal pair, $\alpha_i(m)X_{Ln}^{(i)}(m)$ and $\alpha_{I-i+1}(m)X_{Rn}^{(i)}(m)$, for $i=1,\ldots,I$, at any location along the delay line except the location where $i=s$:

$$X_n^{(i)}(m)=\frac{\alpha_i(m)X_{Ln}^{(i)}(m)-\alpha_{I-i+1}(m)X_{Rn}^{(i)}(m)}{[\alpha_i(m)/\alpha_s(m)]\exp[j\omega_m(\tau_s-\tau_i)]-[\alpha_{I-i+1}(m)/\alpha_{I-s+1}(m)]\exp[j\omega_m(\tau_{I-s+1}-\tau_{I-i+1})]}, \quad \text{for } i\neq s. \quad (11)$$

A caveat in using Eq. (11) is that if the value of the denominator is too small, a small positive constant ϵ is added to limit the magnitude of $X_n^{(i)}(m)$.

B. Physical meaning of the delay-line subtraction operation

To analyze the operation, Eq. (11) can be expressed in the following form via substitution of Eqs. (7) and (8):

$$X_n^{(i)}(m)=S_n(m)+G_n(m)v_{s,g}^{(i)}(m), \quad i\neq s, \quad (12)$$

where

$$v_{s,g}^{(i)}(m)=\frac{[\alpha_i(m)/\alpha_g(m)]\exp[j\omega_m(\tau_g-\tau_i)]-[\alpha_{I-i+1}(m)/\alpha_{I-g+1}(m)]\exp[j\omega_m(\tau_{I-g+1}-\tau_{I-i+1})]}{[\alpha_i(m)/\alpha_s(m)]\exp[j\omega_m(\tau_s-\tau_i)]-[\alpha_{I-i+1}(m)/\alpha_{I-s+1}(m)]\exp[j\omega_m(\tau_{I-s+1}-\tau_{I-i+1})]}, \quad i\neq s. \quad (13)$$

Equations (11) and (13) can be simplified when the antisymmetric relationship in Eq. (2) is used. Thus,

$$X_n^{(i)}(m)=\frac{\alpha_i(m)X_{Ln}^{(i)}(m)-\alpha_{I-i+1}(m)X_{Rn}^{(i)}(m)}{[\alpha_i(m)/\alpha_s(m)]\exp[j\omega_m(\tau_s-\tau_i)]-[\alpha_{I-i+1}(m)/\alpha_{I-s+1}(m)]\exp[-j\omega_m(\tau_s-\tau_i)]}, \quad \text{for } i\neq s, \quad (14)$$

and

$$v_{s,g}^{(i)}(m)=\frac{[\alpha_i(m)/\alpha_g(m)]\exp[j\omega_m(\tau_g-\tau_i)]-[\alpha_{I-i+1}(m)/\alpha_{I-g+1}(m)]\exp[-j\omega_m(\tau_g-\tau_i)]}{[\alpha_i(m)/\alpha_s(m)]\exp[j\omega_m(\tau_s-\tau_i)]-[\alpha_{I-i+1}(m)/\alpha_{I-s+1}(m)]\exp[-j\omega_m(\tau_s-\tau_i)]}, \quad i\neq s. \quad (15)$$

When ignoring the compensation factors $\alpha_i(m)$, an interesting observation of the subtraction [Eq. (11) or (14)] is that it computes the difference between each pair of taps at the ith location divided (shifted) by a factor that is determined only by the difference in time delay between that location and the location corresponding to the desired signal. Next we will show that Eq. (11) performed at the location i in the dual delay-line corresponding to the noise source will cancel the noise signal and provide an estimate of the desired signal. Moreover, the location can be found using an energy quantity.

A signal vector containing all the frequency components for the preceding N time frames is $\mathbf{x}^{(i)}=(X_1^{(i)}(1), X_1^{(i)}(2), \ldots, X_1^{(i)}(M), X_2^{(i)}(1), \ldots, X_2^{(i)}(M), \ldots, X_N^{(i)}(1), \ldots, X_N^{(i)}(M))^T$, $i=1, \ldots, I$, where T denotes vector transposition. The energy $E[\mathbf{x}^{(i)}]$ of vector $\mathbf{x}^{(i)}$ is

$$E[\mathbf{x}^{(i)}]=\|\mathbf{x}^{(i)}\|_2^2=\sum_{n=1}^{N}\sum_{m=1}^{M}|X_n^{(i)}(m)|^2 = \sum_{n=1}^{N}\sum_{m=1}^{M}|S_n(m)+G_n(m)v_{s,g}^{(i)}(m)|^2, \quad i=1, \ldots, I, \tag{16}$$

where the energy of the signal $X_n^{(i)}(m)$ is

$$E[X_n^{(i)}(m)]=|X_n^{(i)}(m)|^2=|S_n(m)+G_n(m)v_{s,g}^{(i)}(m)|^2. \tag{17}$$

To separate the complex signal into the desired signal and noise, we define the following vectors in the similar manner

$$\mathbf{s}=(S_1(1), S_1(2), \ldots, S_1(M), S_2(1), \ldots, S_2(M), \ldots, S_N(1), \ldots, S_N(M))^T,$$

and

$$\mathbf{g}^{(i)}=(G_1(1)v_{s,g}^{(i)}(1), G_1(2)v_{s,g}^{(i)}(2), \ldots, G_1(M)v_{s,g}^{(i)}(M), G_2(1)v_{s,g}^{(i)}(1), \ldots, G_2(M)v_{s,g}^{(i)}(M), \ldots, G_N(1)v_{s,g}^{(i)}(1), \ldots, G_N(M)v_{s,g}^{(i)}(M))^T,$$

where $i=1, \ldots, I$. The energy of $\mathbf{s}$ and $\mathbf{g}^{(i)}$ are, respectively,

$$E[\mathbf{s}]=\|\mathbf{s}\|_2^2=\sum_{n=1}^{N}\sum_{m=1}^{M}|S_n(m)|^2 \tag{18}$$

and

$$E[\mathbf{g}^{(i)}]=\|\mathbf{g}^{(i)}\|_2^2=\sum_{n=1}^{N}\sum_{m=1}^{M}|G_n(m)v_{s,g}^{(i)}(m)|^2, \quad i=1, \ldots, I. \tag{19}$$

In general, the desired signal and the noise signal are independent. Thus, vectors $\mathbf{s}$ and $\mathbf{g}^{(i)}$ are orthogonal. According to the Pythogoras Theorem, we would have the following relationship:

$$E[\mathbf{x}^{(i)}]=\|\mathbf{x}^{(i)}\|_2^2 = \|\mathbf{s}+\mathbf{g}^{(i)}\|_2^2 = \|\mathbf{s}\|_2^2+\|\mathbf{g}^{(i)}\|_2^2=E[\mathbf{s}]+E[\mathbf{g}^{(i)}], \quad i=1, \ldots, I. \tag{20}$$

Because $\|\mathbf{g}^{(i)}\|_2^2 \geqslant 0$,

$$E[\mathbf{x}^{(i)}]=\|\mathbf{x}^{(i)}\|_2^2 \geqslant \|\mathbf{s}\|_2^2=E[\mathbf{s}], \quad i=1, \ldots, I. \tag{21}$$

The equality in Eq. (21) is satisfied, or equivalently $\min E[\mathbf{x}^{(i)}]$ occurs, only when $E[\mathbf{g}^{(i)}]=\|\mathbf{g}^{(i)}\|_2^2=0$, which happens in either of the following two conditions:

(a) When $G_n(m)=0$, i.e., the noise source is silent. In this case, there is no need for doing localization of the noise source and noise cancellation.

(b) When $v_{s,g}^{(i)}(m)=0$, Eq. (15) indicates that this case corresponds to $i=g=i_{\text{noise}}$. Therefore, $E[\mathbf{x}^{(i)}]$ has its minimum at $i=g=i_{\text{noise}}$ and the minimum value, according to Eq. (21), is $E[\mathbf{s}]$. Thus,

$$E[\mathbf{s}]=E[\mathbf{x}^{(i_{\text{noise}})}]=\min_i E[\mathbf{x}^{(i)}]. \tag{22}$$

When $i=i_{\text{noise}}$, Eq. (12) provides

$$\tilde{S}_n(m)=X_n^{(i_{\text{noise}})}(m) = S_n(m)+G_n(m)v_{s,g}^{(i_{\text{noise}})}(m)=S_n(m). \tag{23}$$

In other words, in the presence of one desired source and one noise source, the subtraction operation [Eq. (11)] applied at the location $i=g(=i_{\text{noise}})$ in the dual delay-line structure can produce an accurate estimate of the desired signal. Namely,

$$X_n^{(g)}(m)=\frac{\alpha_g(m)X_{Ln}^{(g)}(m)-\alpha_{I-g+1}(m)X_{Rn}^{(g)}(m)}{[\alpha_g(m)/\alpha_s(m)]\exp[j\omega_m(\tau_s-\tau_g)]-[\alpha_{I-g+1}(m)/\alpha_{I-s+1}(m)]\exp[j\omega_m(\tau_{I-s+1}-\tau_{I-g+1})]}. \tag{24}$$

The above analysis with energy also suggests a simple method to estimate the location $g=i_{\text{noise}}$ of the noise source in the two-source situation where the direction of the desired source is known *a priori*. Specifically, localization of the noise source can be conducted by finding the location i_{noise} along the dual delay-line that produces the minimum value of $E[\mathbf{x}^{(i)}]$ [Eqs. (11), (16), and (22)]. Once the location i_{noise} is determined, the azimuth of the noise source is easily determined by using Eq. (3). The estimated noise location i_{noise} can be fed back to the dual delay-line for noise cancellation

and extraction of the desired signal using Eq. (24).

In Fig. 1, the blocks labeled ''Integration over Time'' and ''Integration over Frequency'' together calculate the energy $E[\mathbf{x}^{(i)}]$ defined in Eq. (16). The block labeled ''Noise Source Localization'' locates the minimum point of $E[\mathbf{x}^{(i)}]$, and then supplies this as the estimate of i_{noise} to the dual delay-line. Since all the components $X_n^{(i)}(m)$, $i=1,\ldots,I$, have been computed at the localization step, now we only need to pick up the appropriate component $X_n^{(i_{\text{noise}})}(m)$, i.e., $\tilde{S}_n(m)$; Eq. (24) does not need to be actually executed in this case. Note that all the frequency computations so far are conducted on the first half ($m=1,\ldots,M_{1/2}$) of the whole bandwidth. The block labeled ''Frequency Synthesis'' derives the second half ($m=M_{1/2}+1,\ldots,M$) by means of the symmetry property of the inverse discrete Fourier transform (IDFT) and then conducts the IDFT to generate the time-domain signal $\tilde{s}_n(k)$.

C. Beamforming interpretation of the delay-line subtraction operation

This system can be understood conceptually by an equivalent beamforming procedure. Equation (11) can be expressed in the following form:

$$X_n^{(i)}(m)=w_{Ln}^{(i)}(m)X_{Ln}^{(i)}(m)+w_{Rn}^{(i)}(m)X_{Rn}^{(i)}(m), \tag{25}$$

where $w_{Ln}^{(i)}(m)$ and $w_{Rn}^{(i)}(m)$ are beamforming weights. That is, for each location along the dual delay-line at each frequency, a specific nulling pattern is generated with the null pointed toward the direction corresponding to the delay-line location while the gain in the presumed target direction is kept unity. Figure 3(a) shows a broadband intelligibility-weighted beampattern (for definition, see Appendix B) for selected nulling directions at −80°, −60°, −40°, −20°, 20°, 40°, 60°, and 80° azimuth (labeled A through H, respectively) with the desired source at 0° azimuth. It can be seen that Eq. (11) actually positions a null in the direction of the noise source while keeping the broadband gain always unity in the direction of the desired source. Since each nulling pattern uses only 2 degrees of freedom, i.e., $w_{Ln}^{(i)}(m)$ and $w_{Rn}^{(i)}(m)$, to satisfy the two constraints (directions of null and unity-gain), the null patterns are fixed and there is no room to play optimization on the pattern shape. As will be presented in Sec. III, this study, by taking advantages of the dual delay-line network, the estimated source locations, as well as the broadband characteristics of dialog speech, sought to find an appropriate strategy [which is a nonlinear one as shown in Eq. (27)] to utilize the simple null patterns for target extraction among multiple interferers.

To extend the discussion on the azimuthal resolution of the dual delay-lines in Sec. II A, let us look at a distinguished feature of Eq. (11). In the numerator of Eq. (11), the signals in the two channels can be phase-shifted by any arbitrary (small) values in the frequency domain. However, the denominator eliminates the effect and thus $X_n^{(i)}(m)$ contains an intact component of the desired signal $S_n(m)$. Moreover, at the location $i=g=i_{\text{noise}}$ where the noise component is completely cancelled, only the desired signal is left in the result. If interpreted as a beamformer, Eq. (11) operated on the dual delay-line in the frequency domain enables a null steering precisely in any arbitrary direction regardless of the sampling rate.

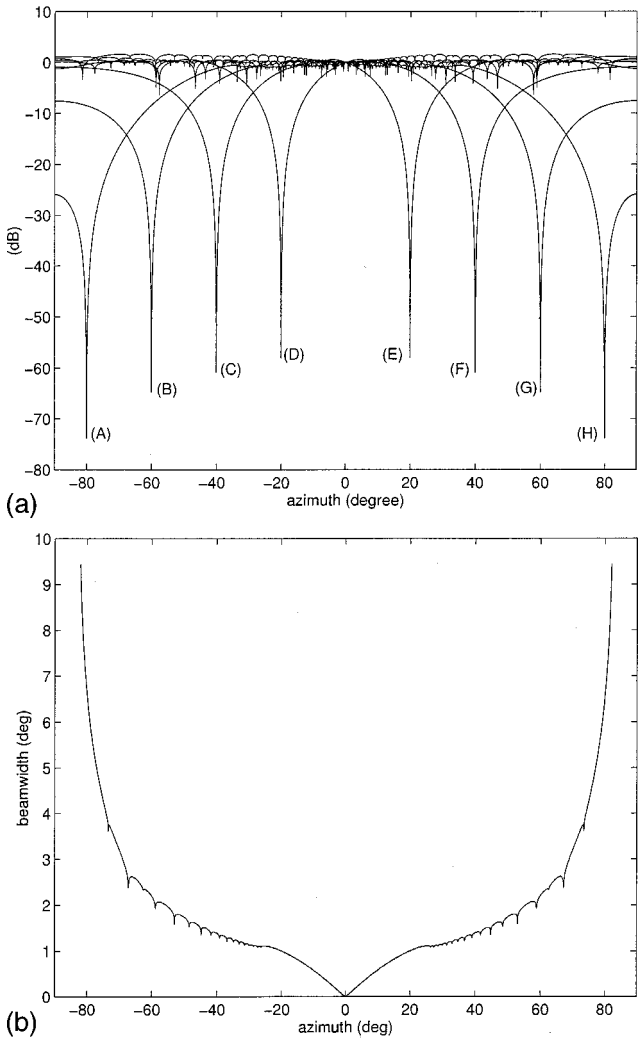

FIG. 3. (a) The intelligibility-weighted beampattern created by Eq. (11) for the cases where the desired source was always at 0° azimuth while the noise source was at −80° (A), −60° (B), −40° (C), −20° (D), 20° (E), 40° (F), 60° (G), and 80° (H) azimuth, respectively. The inter-microphone distance in this example was 144 mm. (b) The null-width of the intelligibility-weighted beampattern at −30 dB as a function of azimuth.

D. Extended application

The method suggested in the preceding subsection for localization and cancellation of the noise source is valid only when the direction of the desired source is known *a priori*. It cannot be directly applied in the situations where the direction of the desired source is also unknown. Therefore, we designed another system (System II in Fig. 4). The operation of this system is divided into two steps: it localizes both the desired source and noise source, and then selectively extracts the signal from the desired direction. The localization step employs an efficient localization method comprising dual delay-line coincidence detection followed by a nonlinear operation and then temporal and spectral integrations. The method was described in detail in a previous paper (Liu *et al.*, 2000) in which it was shown to accurately localize

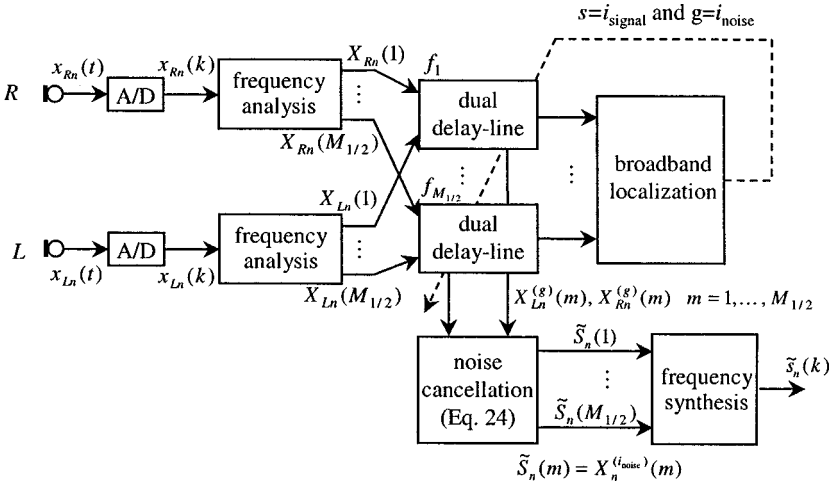

FIG. 4. The block diagram of System II for extraction of one desired source in the presence of a noise source when both source locations have to be estimated by the system. See Liu *et al.* (2000) for details about the block "Broadband Localization."

four talkers in an anechoic room and six talkers in computer simulation. Thus the localization block in Fig. 4 determines the in-phase positions, $i_{\text{signal}}=s$ and $i_{\text{noise}}=g$, of both the desired and noise sources along the dual delay-line, which were then used by the subtraction in Eq. (24) for extracting the desired signal $\tilde{S}_n(m)$. That is, except for the separate source localization step, System II employs the same noise cancellation method as described in the preceding subsection [Eq. (24)].

In comparison with System I, System II (without the assumption of direction of the desired source) is functionally more flexible. For example, in a situation with two talkers, there is no need to align the dual microphones physically toward one talker, and either talker can be taken as the desired one. The user can choose between the two sources at any time by using an electronic switch instead of changing the pointing direction of the microphones. Actually the microphones do not necessarily point to either of the sources.

We presented System I in the preceding subsection mainly for illustrating the mechanism of the dual delay-line subtraction [Eq. (11)] and shows its capacity for both noise-localization and desired-extraction. However, the operation in System I is computationally expensive because Eq. (11) must be applied to each tap in the dual delay-line for localization. Moreover, its use is limited to a two-talker (with the direction of the desired talker known *a priori*) situation. In comparison, the coincidence detection scheme for localization employed by System II is simpler in computation. What is more important is that, as we will show in the next section, System II can be further extended to situations with multiple interfering talkers.

Although our localization method worked quite well in a multiple-source environment, we normally observed relatively larger and more frequent localization errors for the lateral sources (Liu *et al.*, 2000). The robustness of the noise cancellation to localization errors can be roughly estimated by looking at the null-width of the nulls in Fig. 3(a). For example, the null-width evaluated at -30 dB is shown in Fig. 3(b). It shows that the width is wider when the direction of the null is farther away from the midline; that is, the noise cancellation method can tolerate bigger localization errors for lateral sources. Therefore, the greater localization errors for lateral sources do not degrade the system performance in terms of noise cancellation.

III. STRATEGY FOR BROADBAND MULTIPLE-SOURCE CANCELLATION

The greatest challenge associated with extension of the noise cancellation method from two-source situations to multiple-source (>2) situations is that a two-input system in theory can only effectively cancel the sound from *one* interfering source. This is due to the fact that only one null can be generated in the beampattern when using a two-microphone array. In the narrow-band situation, an apparent solution is to adaptively steer the null toward the most intense noise source at each moment. In the broadband situation, since the input signal is decomposed into its frequency components, one can formulate a separate one-null beampattern for each frequency. When there is one noise source as described in the preceding section, the nulls at all frequencies are steered in the same direction of the single noise source. However, when there are more than two sources, each frequency bin can be treated separately so that its beampattern null is adaptively steered at each moment toward the noise source that is emitting the most intense energy at that frequency, while maintaining unity gain toward the desired source. It is actually a dynamic application of the subtraction operation in Eq. (24). This noise cancellation strategy is based on the following rationale:

(a) Natural speech has many pauses and silent intervals, both of which usually occupy 60%–65% of the total time (Flanagan, 1972, p. 386). Therefore, when multiple talkers speak simultaneously, there are always a number of short temporal gaps present. The number of overlapping talkers at each moment is usually smaller than the total number of talkers.

(b) Even when multiple talkers speak at the same moment, different talkers likely dominate at different frequency bins at each moment due to the differences in articulation such as voicing and pitch. There are about ten phonemes per second in conversational speech, more than 60% of which are low-energy, high-frequency consonants, and less than 40% of which are high-

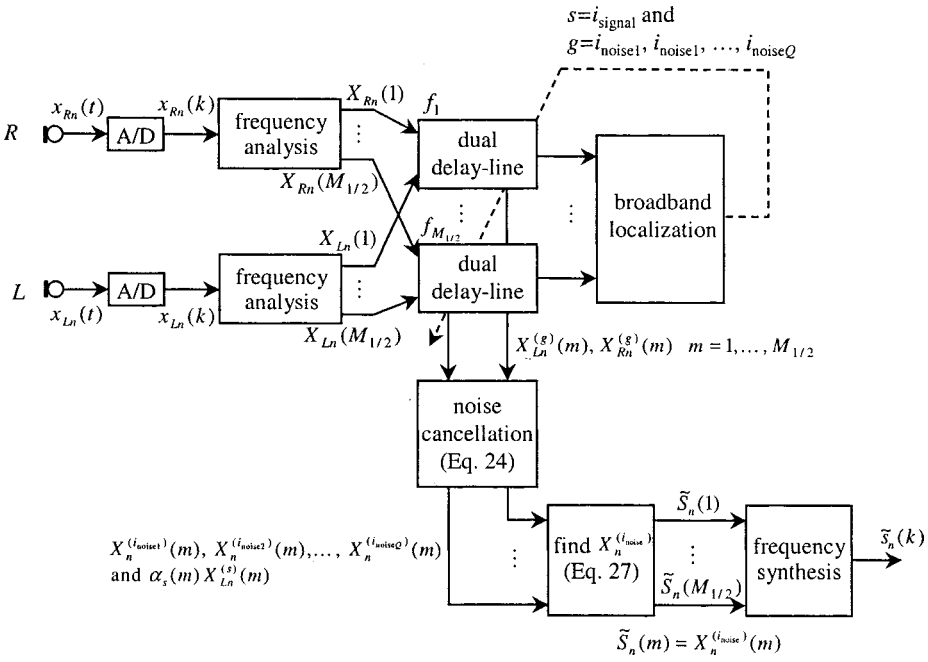

FIG. 5. The block diagram of System III for extraction of one desired source in the presence of more than one noise source when all the source locations have to be estimated by the system. See Liu *et al.* (2000) for details about the block "Broadband Localization."

energy, low-frequency vowels (Flanagan, 1972, p. 4). In the presence of multiple talkers, the talkers who are articulating the high-energy vowels are dominant in both the localization (Liu *et al.*, 2000) and cancellation [Eq. (27)] and hence more easily removed. On the other hand, due to the asymmetry of the filtering response of the human ear, the masking effect of low frequencies on high frequencies is much stronger than the reverse (Jeffress, 1970, p. 95). In other words, cancellation of a talker at his/her strongest frequency components, which are likely the major components of a vowel, may effectively cancel the masking effect of the talker.

To obtain the information about location of each source for the noise cancellation algorithm, the localization algorithm in Liu *et al.* (2000) is employed. Suppose there are Q noise sources with corresponding locations in the dual delay-line being $i_{\text{noise1}}, i_{\text{noise2}}, \ldots, i_{\text{noiseQ}}$. By applying Eq. (24), we obtain $X_n^{(i_{\text{noise1}})}(m), X_n^{(i_{\text{noise2}})}(m), \ldots, X_n^{(i_{\text{noiseQ}})}(m)$ for each frequency ω_m. If the localization is accurate, they all should include the component of the desired signal at frequency ω_m as well as components from interfering sources other than the one to be canceled. In order to determine the particular noise source to be canceled, the energies of $X_n^{(i_{\text{noise1}})}(m), X_n^{(i_{\text{noise2}})}(m), \ldots, X_n^{(i_{\text{noiseQ}})}(m)$ are calculated and compared. The minimum $X_n^{(i_{\text{noise}})}(m)$ is taken as output $\tilde{S}_n(m)$:

$$\tilde{S}_n(m)=X_n^{(i_{\text{noise}})}(m), \tag{26}$$

where $X_n^{(i_{\text{noise}})}(m)$ satisfies the following condition:

$$|X_n^{(i_{\text{noise}})}(m)|^2=\min\{|X_n^{(i_{\text{noise1}})}(m)|^2, |X_n^{(i_{\text{noise2}})}(m)|^2, \ldots, |X_n^{(i_{\text{noiseQ}})}(m)|^2, |\alpha_s(m)X_{Ln}^{(s)}(m)|^2\}. \tag{27}$$

By referring to the energy analysis in Sec. II B, it is easy to see that this strategy is logically consistent with Eq. (22). It is noted that, in Eq. (27), we also include the original signal $\alpha_s(m)X_{Ln}^{(s)}(m)$ for the following reason. The beampattern designed above sometimes may amplify other less intense noise sources. When the amount of noise amplification is larger than the amount of cancellation of the most intense noise source, it may be better to keep the input signal at that frequency at that moment unchanged. An extended system (System III in Fig. 5) was developed using System II (Fig. 4) as the foundation. In comparison with System II, it identifies multiple (>2) source directions and tentatively cancels each noise source; specifically it cancels the instantaneously most intense source on a frequency-by-frequency basis [Eq. (27)].

The cancellation step relies on the localization step to provide azimuth information for each source, which is usually a difficult task especially in the presence of multiple sources. However, as shown in our previous paper (Liu *et al.*, 2000), our localization system can satisfactorily localize four sources in an anechoic room and six sources in simulation, if not more. In addition, the cancellation step does not have rigid requirement that all the sources must be localized accurately. As a matter of fact, our strategy is to cancel the strongest noise component at each frequency bin—this is usually emitted from one of those momentarily relatively intense noise sources, which are easy to localize compared with other relatively less intense sources.

IV. EXPERIMENT

For the case of two talkers, once the locations of the talkers are determined, the sound from one talker can be removed by using System I or System II with essentially no residual noise while the estimated desired signal is distortionless. This was clearly supported in theory and also pre-

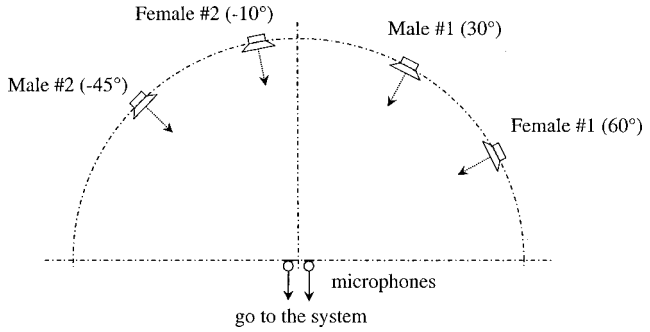

FIG. 6. Top view of the spatial configuration of one of our experimental setups. This experimental setup corresponds to test configuration #2 in Tables I and II.

viously demonstrated by Banks (1993). In this paper we present the results of four-talker experiments using the noise cancellation of System III.

The experiments employed a record-and-play procedure with four high-fidelity loudspeakers (ADS 200LC) and were conducted in an anechoic room and in a moderately quiet conference room with a reverberation time constant of approximately 400 ms. The speech materials consisted of spondaic words spoken by native speakers of American English; these were recorded in a sound studio at the Beckman Institute. All the speech recordings were equalized in average intensity and played through the loudspeakers. The words in each experimental condition were temporarily aligned, i.e., it was equivalent to all the talkers starting at the same time. The inter-microphone distance was 144 mm. All the loudspeakers were at a fixed equal distance of 1.0 m (unless otherwise stated) from the midpoint between the two microphones, and all the loudspeakers and microphones were at the same elevation ($\sim$1 m from the floor). Correspondingly, the compensation factors in Eq. (11) were determined for that distance.

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 τ_i ($i=1,\ldots,I$) were determined such that the dual delay-line has a uniform azimuthal resolution of 0.5° (i.e., $I=361$).

Two groups of talkers were used in our tests. Each group consisted of four different talkers speaking different spondaic words. Five tests were conducted for each group; each test adopted a different azimuthal arrangement of the sources, with the separation between adjacent sources ranging 10°–75°. Figure 6 illustrates one of the configurations. Each test consisted of four subtests in which each talker was taken in turn as the desired source with all the other talkers as the noise sources. The localization of the talkers was conducted using both the ''direct'' and ''stencil'' methods in Liu *et al.* (2000).

The system performance was evaluated using an objective intelligibility-weighted measure, whose concept was first proposed by Peterson (1989) and described in detail in Liu and Sideman (1996). Specifically, we used intelligibility-weighted signal cancellation, intelligibility-weighted noise cancellation, and net intelligibility-weighted gain (see Appendix B for definition).

As mentioned above, an array of tests was conducted with a number of variables such as different talkers, different spondaic words, different azimuthal arrangements, different localization methods, and different combinations of the variables. However, it is not necessary to present all our results since most of the variables, as they turned out to be, have no statistically significant effect on the final noise cancellation performance. Specifically, the experimental results showed no statistical difference due to talkers and words spoken. It also showed no significant effect from using the ''direct'' method versus the ''stencil'' method for source localization (Liu *et al.*, 2000). Therefore, we only present the results from Group #1 with the location information derived using the ''stencil'' method. As mentioned above, it contained five tests corresponding to five different spatial configurations. For each test, we present result from only one (arbitrarily chosen) of the four subtests since the location of the desired source has no obvious effect on noise cancellation. Table I shows typical results chosen from the anechoic chamber test

TABLE I. Experiment results attained from the anechoic room using System III. The results shown were derived from five tests (configurations) from Group #1 including two male speakers (M1 and M2) and two female speakers (F1 and F2). The spondaic word spoken by each talker is given in italic. The values in parentheses are cancellation of the desired sources in dB. Configuration test #2 is shown in Fig. 6.

	Intelligibility-weighted signal cancellation (dB)				Intelligibility-weighted noise cancellation (dB)	Net intelligibility-weighted gain (dB)
Test	M1 *"armchair"*	M2 *"playground"*	F1 *"pancake"*	F2 *"woodwork"*		
#1	−75°	0°	20°	75°		
	8.04	(0.15)	4.98	3.07	9.25	9.09
#2	30°	−45°	60°	−10°		
	8.34	4.71	4.12	(0.67)	8.38	7.71
#3	10°	−80°	−50°	45°		
	(0.55)	6.90	5.57	3.83	8.56	8.00
#4	−30°	15°	5°	−60°		
	10.53	2.07	(1.14)	6.35	8.27	7.13
#5	−25°	25°	−70°	80°		
	8.09	(0.34)	5.82	4.46	8.78	8.44

TABLE II. Same as Table I except that the recordings were made in a moderately quiet conference room with a 400 ms reverberation time [RT was derived using Schroeder's method; see J. Acoust. Soc. Am. **37**, 409–412 (1965)].

Test	Intelligibility-weighted signal cancellation (dB)				Intelligibility-weighted noise cancellation (dB)	Net intelligibility-weighted gain (dB)
	M1 *"armchair"*	M2 *"playground"*	F1 *"pancake"*	F2 *"woodwork"*		
#1	−75°	0°	20°	75°		
	4.82	(0.55)	4.07	2.06	6.73	6.18
#2	30°	−45°	60°	−10°		
	6.27	4.18	3.09	(0.58)	7.26	6.69
#3	10°	−80°	−50°	45°		
	(1.12)	3.85	2.91	2.71	5.75	4.63
#4	−30°	15°	5°	−60°		
	6.29	(0.85)	0.91	3.61	6.16	5.25
#5	−25°	25°	−70°	80°		
	5.70	(0.69)	4.28	2.92	6.97	6.29

while Table II the results from the conference room test. In the tables, the numbers in parentheses represent the degree of cancellation in dB of the desired source (which should ideally be 0 dB) and the other numbers represent the degree of noise cancellation for each noise source. Because we had separate recordings of speech signals from each talker, we applied the same processing both on the complex signal and synchronously on each signal corresponding to each talker as well. As such, we were able to tell the effect of processing on each signal involved. The next to the last column in the tables show the degree of cancellation for all the noise sources lumped together, while the last column gives the net intelligibility-weighted improvement (which considers both noise cancellation and loss in the desired signal). Our results from the anechoic room show that, in the intelligibility-weighted measure, the cancellation strategy was able to cancel each noise source by 3–11 dB, while the degradation in the desired source was very small (mostly smaller than 0.5 dB). The total noise cancellation was between 8 and 10 dB. For the conference room, the cancellation was roughly 2 dB less, indicating that the room reverberation degraded the system performance somewhat. In spite of the drop in system performance the system still produced a sizable gain in speech intelligibility.

In order to get an insight into the effect of the signal processing on each talker, we choose one subtest example (anechoic room; desired source: F1 at 60°; noise sources: M1 at 30°, M2 at −45°, and F2 at −10°). We display the signal waveform of each talker as well as the complex signal of all the four talkers, before [Fig. 7(A)] and after [Fig. 7(B)] the signal processing. Comparison of the two panels shows a great attenuation of the interfering talkers (M1, M2, and F2) while the desired signal (F1) is essentially not attenuated and the distortion of the desired talker is unperceivable. A moment-by-moment comparison shows that the momentarily strongest noise source was always reduced, indicating that the system adapted rapidly. The last trace in Fig. 7(B) is the system output, which turned out to be cleaner and closer to the desired speech [F1 in Fig. 7(A)] than the noisy unprocessed signal [the last trace in Fig. 7(A)].

In an informal listening experiment with normal hearing listeners, we found the unprocessed signal to be impossible to understand, even when the spatial cues were retained. After the processing, however, the extracted speech from a desired source was easily understandable.

Limited by our experimental facility, we only conducted on-site acoustic tests for four-talker situations. However, our computer simulation results for six-talker situations were quite similar. To avoid redundancy, we omit presentation of the details. Basically, we obtained a 7–10 dB enhancement in the intelligibility-weighted signal-to-noise ratio when there were six equally loud, temporally aligned speech sounds originating from six different sources.

V. DISCUSSION

There are three key differences between the algorithm proposed in this paper and conventional adaptive beamformers such as the Frost and Griffiths-Jim beamformers (Van Veen and Buckley, 1988), namely, (i) direct frequency-domain null steering, (ii) explicit source localization, and (iii) implicit utilization of dialogue characteristics. The frequency-domain null-steering algorithm described herein does rapid, independent steering of the beampattern at each frequency. Independent steering allows rapid steering of the single null at each frequency to the dominant interferer at that time and frequency. It provides a maximum potential to cancel intense components emitted from multiple interferers with only two inputs available. What distinguishes this method from other methods is that this independent steering can be implemented with no time delay when it is provided with instant localization information. When processing signals with strong, rapidly varying time-frequency structure such as speech, the net effect is to allow significant cancellation of several *simultaneous* interferers by exploiting differences in their instantaneous time-frequency structures. In contrast, slowly adapting time-domain algorithms such as the Frost (Frost, 1972) and Griffiths-Jim (Griffiths and Jim, 1982) beamformers are unable to track the nonstationary structure rapidly enough to achieve significant cancellation of more than a single interferer. This claim is clearly dem-

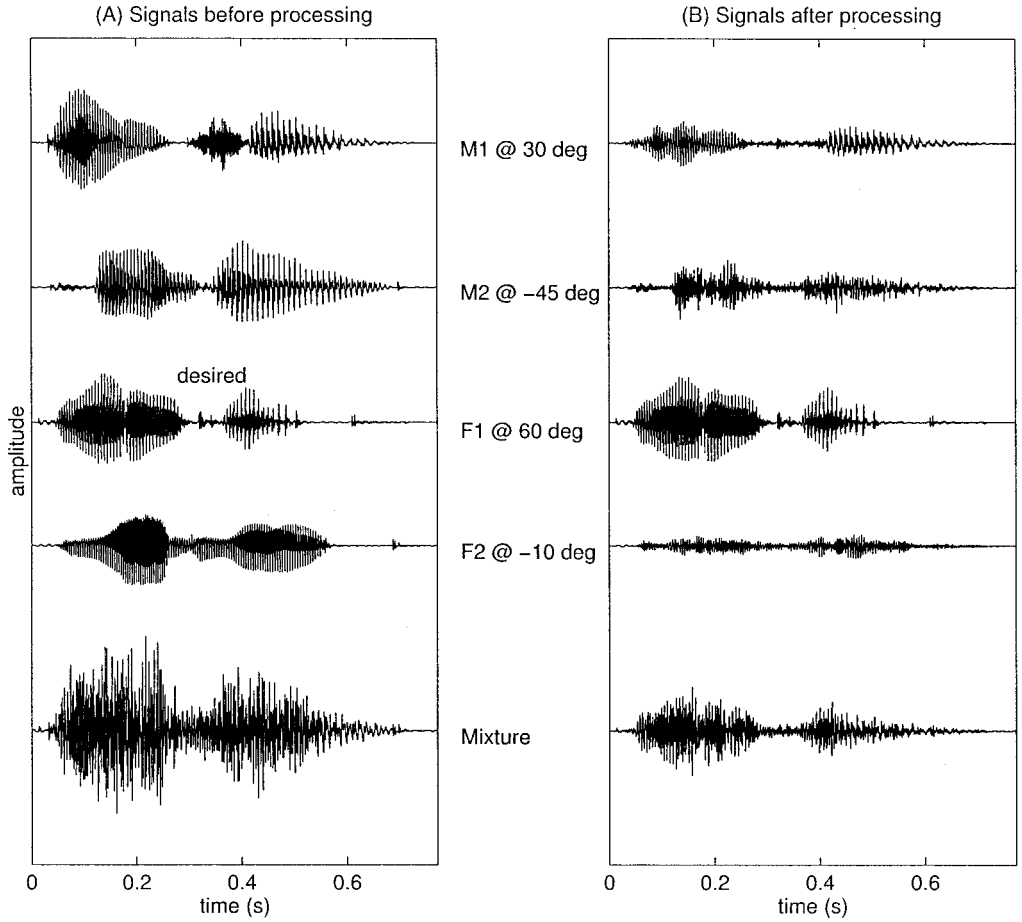

FIG. 7. The signal waveform of each talker as well as the complex signal of all the four talkers before (A) and after (B) the signal processing. See Fig. 6 for the test configuration.

onstrated in the results of comparison experiment presented in Yang *et al.* (2000) using complete sentences under a variety of signal-to-noise conditions.

The performance of this algorithm is comparable to the conventional beamformers for the case of a single interferer, but markedly better for cases involving more than one interferer. The comparisons conducted in Zheng *et al.* (2001) were made in computer simulation where up to four interferers were used at four different SNR settings (-6, -3, 0, $+3$ dB). In the presence of two or more interferers, the present method provided 6–7 dB of SNR gains, while the Frost beamformer and the Griffiths-Jim beamformer had SNR gains in the 2–5 dB range.

The second difference between the conventional beamformers and the proposed method is that the latter explicitly identifies the spatial directions of the target and interferers via a nonlinear, cross-frequency localization procedure (Liu *et al.*, 2000) and exploits this information to optimally steer the null pattern in each frequency bin. The localization is conducted on a frame-by-frame basis and the results are used immediately by the cancellation on the same frame. Therefore, as mentioned above, the adaptation time is virtually zero. This feature is especially important when processing signals with rapidly varying time-frequency structure such as speech. Explicit source localization also offers several other potential advantages, including the ability to steer toward a spatially moving target, better and more robust estimation of signal and interference locations from which to optimize the beam patterns, and the ability (not explored here) to perform additional useful tasks such as auditory scene characterization. The results in Zheng *et al.* (2001) suggest that these characteristics may indeed be advantageous in many situations (with different number of interferers, different spatial configurations, and different SNRs), particularly when the interferers are in close azimuthal proximity to the target.

The third, and most unique, difference is that our method takes full advantage of the characteristics and masking effect of human dialogue as detailed in Sec. III. That strategy makes it possible to utilize a limited resource (two inputs only) to obtain maximum speech intelligibility enhancement benefits such as effective cancellation of multiple interfering sources.

The improvement in signal quality reported in Tables I and II is encouraging but preliminary. The algorithm's performance in anechoic conditions (8–10 dB cancellation) is sufficient to justify further research, while the performance in the conference room (2 dB less cancellation) raises the question as to whether, when used in a real-time environment, the quality of the cancellation will degrade so as to no longer be useful. Practical computational limitations restricted the work reported here, although improvements have allowed off-line analysis over a wider range of materials (Zheng *et al.*, 2001). A related frequency domain beamformer (Lockwood *et al.*, 1999) has been implemented in

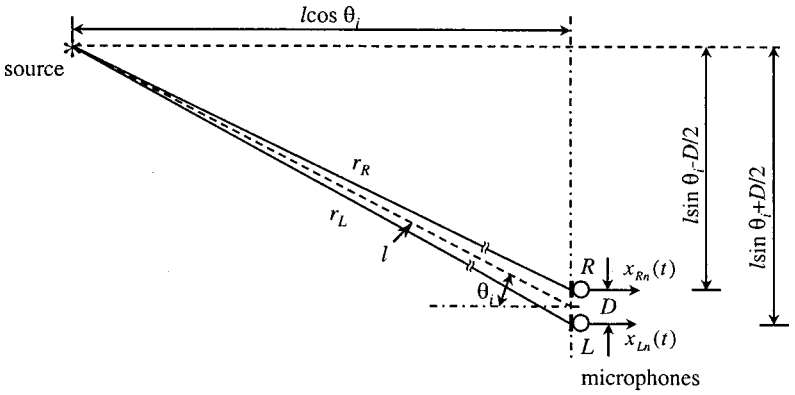

FIG. 8. Top view of the geometry of the source-microphone distance.

real-time (Elledge, 2000) with highly satisfying subjective sound improvement and quality (Larsen *et al.*, 2001). A real-time version of the present algorithm is in the process of implementation; it should permit subjective evaluations to determine whether the technique is viable for hearing aid and other applications.

One practical issue is that when the source to microphone distance is very short (e.g., 2 m or less), it is important to compensate for left–right differences in channel intensity; indeed preliminary tests indicated degradation of about 1 dB in the total net gain without compensation. However, for larger source-microphone distances (e.g., >2 m), the difference with and without compensation was insignificant.

VI. SUMMARY AND CONCLUSION

In this paper, we have presented the technique and experimental results that illustrate the performance of signal processing systems designed for effective extraction of a desired signal in the presence of multiple competing talkers. The signal processing technique is based on dual delay-line structure, a well-known biological network for binaural hearing. The entire system consists of two steps: localization of all sources and extraction of the desired source. Our anechoic chamber tests showed an 8–10 dB of speech enhancement in the presence of four equally loud, temporally aligned talkers; our computer simulation showed a 7–10 dB of speech enhancement in the presence of six equally loud, temporally aligned talkers. The system can localize all the sources present and allow the user to selectively extract any one of them, hence it is more flexible than assuming that the desired source is always straight ahead. It can be applied in many applications such as radar, sonar, communications, and robots.

It is noted that in the present study we focused on separating out a particular talker from all the other competing talkers, i.e., selective hearing. It is technically straightforward to convert the present system to a simulator that can capture the source to which the gaze of the listener is directed at any time instant. It is also possible to simply use multiple noise-cancellation components following the localization so as to extract each of the sources within the environment, i.e., to achieve separation of multiple signals.

The dual delay-line structure implies that the computation is highly parallel. That, and the repeated use of the Fourier transform, made it practical to implement the algorithm by means of VLSI for a fast, miniature device.

Our future work includes evaluation using formal tests in normal listening rooms with human subjects with real-time versions of the algorithm. We will also extend our algorithms to compensate for reverberant environments.

ACKNOWLEDGMENTS

This research was supported by grants from the Beckman Institute, UIUC and from the National Institute for Deafness and other Communication Disorders of the NIH (R21 DC 04870). The work was conducted when the first author (CL) was a Beckman Fellow. A number of people from the Beckman Institute generously donated their time to participate in the recording of speech materials used for our tests. Their generosity is acknowledged. We thank Dean Garstecki for permission to use the anechoic room at the Northwestern University. We also wish to thank the two anonymous reviewers for their constructive comments to an earlier version of the paper.

APPENDIX A: CALCULATION OF THE AMPLITUDE FACTORS

In this appendix we illustrate the calculation of the amplification factors $\alpha_i(m)$, which are used to compensate for differences in the amplitudes of signals arriving at the two microphones. In this example we model the sound source as a simple point source, ignore the absorption of energy by the media, and assume the amplitude variation is solely dependent on the differences in the distance from source to the microphones. In this case, the amplitude compensation is independent of frequency.

The amplitude of the received sound pressure $|\mathbf{p}|$ varies with the source–receiver distance r:

$$|\mathbf{p}| \propto \frac{1}{r} \tag{A1}$$

or

$$\frac{|\mathbf{p}_L|}{|\mathbf{p}_R|} = \frac{r_R}{r_L}, \tag{A2}$$

where $|\mathbf{p}_L|$ and $|\mathbf{p}_R|$ are the amplitude of sound pressures at the two microphones (Kinsler *et al.*, 1982, p. 168). According to the geometry in Fig. 8, the distances from the source

to the left and right microphones r_L and r_R are, respectively,

$$r_L = \sqrt{(l \sin \theta_i + D/2)^2 + (l \cos \theta_i)^2} = \sqrt{l^2 + lD \sin \theta_i + D^2/4} \quad \text{(A3)}$$

and

$$r_R = \sqrt{(l \sin \theta_i - D/2)^2 + (l \cos \theta_i)^2} = \sqrt{l^2 - lD \sin \theta_i + D^2/4}. \quad \text{(A4)}$$

For a pair of taps in the dual delay-line in Fig. 2, in order to equalize the signals at the tap outputs, the compensation factors $\alpha_i(m)$ and $\alpha_{I-i+1}(m)$ must satisfy the following condition:

$$|\mathbf{p}_L| \alpha_i(m) = |\mathbf{p}_R| \alpha_{I-i+1}(m). \quad \text{(A5)}$$

Substituting Eq. (A2) into Eq. (A5), the above condition becomes

$$\frac{r_L}{r_R} = \frac{\alpha_i(m)}{\alpha_{I-i+1}(m)}. \quad \text{(A6)}$$

We define the value of $\alpha_i(m)$ to be equal to

$$\alpha_i(m) = K\sqrt{l^2 + lD \sin \theta_i + D^2/4}, \quad \text{(A7)}$$

where K has unit of inverse length and is chosen for a convenient amplitude level. Applying the definition in Eq. (A7), the value of $\alpha_{I-i+1}(m)$ will be

$$\alpha_{I-i+1}(m) = K\sqrt{l^2 + lD \sin \theta_{I-i+1} + D^2/4} = K\sqrt{l^2 - lD \sin \theta_i + D^2/4}, \quad \text{(A8)}$$

where the relationship $\sin \theta_{I-i+1} = -\sin \theta_i$ can be obtained by substituting $I-i+1$ into i in Eq. (3). By substituting Eqs. (A7) and (A8) into Eq. (A6), one can verify that the values assigned to $\alpha_i(m)$ in Eq. (A7) satisfy the condition in Eq. (A6).

APPENDIX B: DEFINITION OF THE INTELLIGIBILITY-WEIGHTED MEASURE

For any signal s, the intelligibility-weighted measure $\Gamma(s)$ is calculated by (Link and Buckley, 1993)

$$\Gamma(s) = \int_{BW} W_{\mathrm{AI}}(f) 20 \log_{10} \mathrm{rms}_{1/3}(|S(f)|) df, \quad \text{(B1)}$$

where

$$W_{\mathrm{AI}}(f) = \frac{1/[1 + (f/1925)^2]}{\int_{BW} 1/[1 + (f/1925)^2] df}, \quad \text{(B2)}$$

$$\mathrm{rms}_{1/3}(|S(f)|) = \left[\frac{\int_{2^{-1/6}f}^{2^{1/6}f} |S(f')|^2 df'}{(2^{1/6} - 2^{-1/6}) f} \right]^{1/2} \quad \text{(B3)}$$

and BW denotes the frequency range. The system improvement in the intelligibility-weighted measure for the target signal T and interference I are, respectively,

$$\Delta\Gamma(T) = \Gamma(T_o) - \Gamma(T_i) \quad \text{(B4)}$$

and

$$\Delta\Gamma(I) = \Gamma(I_i) - \Gamma(I_o), \quad \text{(B5)}$$

where the subscripts i and o denote the input and output, respectively. The overall (or net) intelligibility-weighted gain, G_I, is the sum of the two measures, thus

$$G_I = \Delta\Gamma(T) + \Delta\Gamma(I). \quad \text{(B6)}$$

In our experiment, since we had separate recordings of the target and noise signals, i.e., T_i and I_i (the latter might include more than one interfering talker), we were able to apply the same processing framewise on either of them and obtain the results T_o and I_o, respectively. The framewise processing [Eq. (11)], however, was determined based on the target and noise signals mixed together as would be encountered in the real situation. It is noted that the spectrum $S(f)$ was computed based on the full-length signal, which in our case was a whole spondaic word. We used the full long-term spectrum, as opposed to a frame-by-frame spectrum, for two reasons: (1) It was consistent with the way the intelligibility-weighted measure was applied in other papers published in the area such that our results could be compared directly with earlier results; (2) Since our system has virtually no adaptation time (i.e., it almost always is successful in localizing the strongest interference within a few milliseconds), there is no advantage to computing with the short-term spectra.

Since the intelligibility-weighted measure was constructed as an estimate of the subjective improvement based on the objective calculation, it deliberately emphasized the low frequency domain according to the "critical-band" theory. However, because the low frequency domain is always the hardest to clean with the approaches using multi-microphone arrays of limited size, the intelligibility-weighted measure usually has a smaller value (~1 dB difference in our experiment) than the non-weighting counterpart, i.e., SNR. Nonetheless, this effect does not change the overall picture of the performance; especially the comparison of our system with others, as given in the paper, remains valid.

Similarly, if we denote the array beampattern as $E(f, \theta)$, where f is frequency and θ is the incident direction, the intelligibility-weighted beampattern, $\overline{E}(\theta)$, is defined by (Liu and Sideman, 1996)

$$\overline{E}(\theta) = \int_{BW} W_{\mathrm{AI}}(f) E(f, \theta) df. \quad \text{(B7)}$$

Banks, D. (**1993**). "Localization and separation of simultaneous voices with two microphones," IEE Proc. I **140**, 229–234.

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

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

Bodden, M. (**1996**). "Auditory demonstration of a cocktail-party-processor," Acust. Acta Acust. **82**, 356–357.

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (The MIT Press, Cambridge, MA).

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.

Chabries, D. M., Christiansen, R. W., Brey, R., and Robinette, M. (**1982**). "Application of the LMS adaptive filter to improve speech communication in the presence of noise," Proc. IEEE ICASSP, 148–151.

Chazan, D., Medan, Y., and Shvadron, U. (**1988**). ''Noise cancellation for hearing aids,'' IEEE Trans. Acoust., Speech, Signal Process. **36**, 1697–1705.
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., 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. (**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.
Durlach, N. I. (**1960**). ''Note on the equalization and cancellation theory of binaural masking level differences,'' J. Acoust. Soc. Am. **32**, 1075–1076.
Durlach, N. I. (**1972**). ''Binaural signal detection: Equalization and cancellation theory,'' in *Foundations of Modern Auditory Theory*, edited by V. T. Jerry (Academic, New York), pp. 369–462.
Elledge, M. E., Lockwood, M. E., Bilger, R. C., Feng, A. S., Jones, D. L., Lansing, C. R., O'Brien, W. D., and Wheeler, B. C. (**2000**). ''Real-time implementation of a frequency-domain beamformer on the TI C62X EVM,'' the 10th Ann. DSP Technol. Educ. Res. Conf., Houston, TX, August 2–4, 2000.
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. (**1972**). *Speech Analysis, Synthesis and Perception* (Springer-Verlag, Berlin).
Frost, O. L. (**1972**). ''An algorithm for linearly constrained adaptive array processing,'' Proc. IEEE **60**, 926–935.
Gaik, W. (**1993**). ''Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling,'' J. Acoust. Soc. Am. **94**, 98–110.
Griffiths, L. J., and Jim, C. W. (**1982**). ''An alternative approach to linearly constrained adaptive beamforming,'' IEEE Trans. Antennas Propag. **AP-30**, 27–34.
Jeffress, L. A. (**1948**). ''A place theory of sound localization,'' J. Comp. Physiol. Psychol. **41**, 35–39.
Jeffress, L. A. (**1970**). ''Masking,'' in *Foundations of Modern Auditory Theory*, edited by J. V. Tobias (Academic, New York), pp. 85–114.
Kinsler, L. E., Frey, A. R., Coppens, A. B., and Sanders, J. V. (**1982**). *Fundamentals of Acoustics* (Wiley, New York).
Konishi, M., Takahashi, 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.
Larsen, J., Lockwood, M. E., Lansing, C. R., Bilger, R. C., O'Biren, W. D., Jones, D. L., Wheeler, B. C., and Feng, A. S. (**2001**). ''Performance of a frequency-based minimum variance beamforming algorithm for normal and hearing impaired listeners,'' J. Acoust. Soc. Am. **109**, 2494.
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.
Link, M. J., and Buckley, K. M. (**1993**). ''Prewhitening for intelligibility gain in hearing aid arrays,'' J. Acoust. Soc. Am. **93**, 2139–2145.
Liu, C., and Sideman, S. (**1996**). ''Simulation of fixed microphone arrays for directional hearing aids,'' J. Acoust. Soc. Am. **100**, 848–856.
Liu, C., Wheeler, B. C., O'Brien, Jr., W. D., Bilger, R. C., Lansing, C. R., and Feng, A. S. (**2000**). ''Localization of multiple sound sources with two microphones,'' J. Acoust. Soc. Am. **108**, 1888–1905.
Lockwood, M. E., Jones, D. L., Elledge, M. E., Bilger, R. C., Feng, A. S., Goueygou, M., Lansing, C. R., Liu, C., O'Brien, Jr., W. D., an Wheeler, B. C. (**1999**). ''A minimum variance frequency-domain algorithm for binaural hearing aid processing,'' J. Acoust. Soc. Am. **106**, 2278.
Peterson, P. M. (**1989**). ''Adaptive array processing for multiple microphone hearing aids,'' Ph.D. Dissertation, Dept. Elect. Eng. and Comp. Sci., MIT; Res. Lab. Elect. Tech. Rept. 541, MIT, Cambridge, MA.
Peterson, P. M., Durlach, N. I., Rabinowitz, W. M., and Zurek, P. M. (**1987**). ''Multimicrophone adaptive beamforming for interference reduction in hearing aids,'' J. Rehabil. Res. Dev. **24**, 103–110.
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 (Lawrence Erlbaum Associates, Mahwah, NJ), pp. 499–532.
Strube, H. W. (**1981**). ''Separation of several speakers recorded by two microphones (cocktail-party processing),'' Signal Process. **3**, 355–364.
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.
Van Veen, B. D., and Buckley, K. M. (**1988**). ''Beamforming: a versatile approach to spatial filtering,'' IEEE ASSP Mag., April 1988, 4–24.
Weiss, M. (**1987**). ''Use of an adaptive noise canceler as an input preprocessor for a hearing aid,'' J. Rehabil. Res. Dev. **24**, 93–102.
Widrow, B., Glover, J. R., McCool, J. M., Kaunitz, J., Williams, C. S., Hearn, R. H., Zeidler, J. R., Dong, E., and Goodlin, R. C. (**1975**). ''Adaptive noise cancelling: Principles and applications,'' Proc. IEEE **63**, 1692–1716.
Yang, K. L., Lockwood, K. L., Elledge, M. E., and Jones, D. L. (**2000**). ''A comparison of beamforming algorithms for binaural acoustic processing,'' Proceedings of the 9th IEEE Digital Signal Processing Workshop, Hunt, TX, October 15–18, 2000.
Yin, T. C. T., and Chan, J. C. K. (**1990**). ''Interaural time sensitivity in medial superior olive of cat,'' J. Neurophysiol. **64**, 465–488.
Zheng, Y., Lockwood, M. E., Wheeler, B. C., Jones, D. L., Feng, A. S., O'Brien, W. D., Bilger, R. C., and Lansing, C. R. (**2001**). ''Comparison of binaural beamformers for speech extraction in complex auditory scenes,'' J. Acoust. Soc. Am. **109**, 2494.