Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms

Michael E. Lockwood,[a] Douglas L. Jones, Robert C. Bilger, Charissa R. Lansing, William D. O'Brien, Jr., Bruce C. Wheeler, and Albert S. Feng
Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, 405 North Mathews Ave., Urbana, Illinois 61801

(Received 2 November 2002; accepted for publication 22 August 2003)

Extraction of a target sound source amidst multiple interfering sound sources is difficult when there are fewer sensors than sources, as is the case for human listeners in the classic cocktail-party situation. This study compares the signal extraction performance of five algorithms using recordings of speech sources made with three different two-microphone arrays in three rooms of varying reverberation time. Test signals, consisting of two to five speech sources, were constructed for each room and array. The signals were processed with each algorithm, and the signal extraction performance was quantified by calculating the signal-to-noise ratio of the output. A frequency-domain minimum-variance distortionless-response beamformer outperformed the time-domain based Frost beamformer and generalized sidelobe canceler for all tests with two or more interfering sound sources, and performed comparably or better than the time-domain algorithms for tests with one interfering sound source. The frequency-domain minimum-variance algorithm offered performance comparable to that of the Peissig–Kollmeier binaural frequency-domain algorithm, but with much less distortion of the target signal. Comparisons were also made to a simple beamformer. In addition, computer simulations illustrate that, when processing speech signals, the chosen implementation of the frequency-domain minimum-variance technique adapts more quickly and accurately than time-domain techniques. [DOI: 10.1121/1.1624064]

PACS numbers: 43.72.Ew, 43.66.Pn, 43.66.Ts [DOS] Pages: 379–391

I. INTRODUCTION

A major problem for hearing aids, speech recognition, hands-free telephony, teleconferencing, and other acoustic processing applications is extracting, with good fidelity, a target sound in the presence of multiple competing sounds. This is particularly true of speech sounds, which are highly nonstationary in spectrum and intensity, and which may change position with respect to the listener over time. Thus, the cancellation of multiple, nonstationary, interfering speech sources requires fast, accurate tracking of the sources and robustness to reverberation and correlation between sources.

Many interference suppression techniques have been explored to address this problem. The most common approach is the use of an adaptive beamformer to process the sampled, time-domain outputs of a multimicrophone array, such that reception of the target sound from a particular direction is enhanced [see the reviews by Van Veen and Buckley, 1988; Brandstein and Ward, 2001]. These techniques use signals from the array to estimate the gradient of an error function and then iteratively move the filter coefficients closer to an optimal solution in small steps. Two algorithms that have been used extensively to address this problem are the iterative-adaptive techniques of Frost [1972] and Griffiths and Jim [1982]. Other variations of adaptive algorithms include those of Berghe and Wouters [1998] and Welker and Greenberg [1997]. Although these adaptive algorithms generally work well for suppressing statistically stationary interference sources that are uncorrelated with the target source, our experience has shown that they tend to adapt slowly or inaccurately in the presence of multiple, nonstationary interference sources such as speech, especially when there are more sources than sensors. As a result, the performance of the algorithms is compromised.

Greenberg and Zurek [1992] suggested that a solution to the problem of having more speech sources than sensors was to add more microphones. This solution is effective if all microphones are located far enough away from each other that they provide added useful inputs. However, accomplishing this in a hearing-aid system is difficult because locating microphones away from the ears is undesirable. Some current behind-the-ear (BTE) hearing aids contain two microphones per instrument; however, the small separation of the microphones limits the effectiveness of such systems. Likewise, it is impractical to use more than one microphone per instrument in systems that are located in the ear canal. Thus, a preferable system would use two sensors (one per instrument, located at each ear or in the ear canals), providing greater spatial separation of the microphones.

Several previous studies of adaptive beamformers [Greenberg and Zurek, 1992; Kompis and Dillier, 1994; Hoffman *et al.* 1994; Kates and Weiss, 1996] have avoided the problem of slow algorithm adaptation by allowing the adaptive filters sufficient time (at least 2 s for these studies) to converge before processing test signals. To provide more challenging test conditions, Greenberg *et al.* [2003] used a

[a]Electronic mail: melockwo@uiuc.edu

''roving'' interfering sound source that changed location at random times. However, this roving source was not used in all test conditions, and the algorithms in that study were permitted to converge for 1 s before the onset of the target signal. We argue that a beamforming algorithm used in a real hearing-aid instrument cannot be preadapted in an acoustically crowded environment due to the changing head position of the listener and the movement of the interfering talkers, and it may be unable to adapt quickly enough to perform effective interference suppression. Thus, the adaptation rate of a hearing-aid signal-processing algorithm is an important consideration.

To improve the adaptation rate of time-domain algorithms, an alternate approach to an iterative-adaptive technique is direct solution of the optimal beamformer [Capon, 1969]. While this in theory provides rapid convergence and improved interference cancellation, this technique generally requires the inversion of large, time-domain correlation matrices, a process that is inherently unstable and computationally impractical [Golub and Van Loan, 1996].

To suppress multiple, nonstationary interfering sources using only two sensors, frequency-domain beamforming algorithms appear to have distinct advantages compared to time-domain algorithms. For example, the algorithm of Liu *et al.* [1997, 2000, 2001] first determines source locations and strengths, and then performs constrained beamforming in each frequency band to remove interference. This method was shown to adapt quickly and could effectively suppress multiple speech interferers using only two microphones. It demonstrates a distinct improvement in performance over time-domain methods, but it requires intensive computation. The LENS algorithm [Deslodge, 1998] extracts a target signal by placing $n-1$ spatial nulls using n sensors. Greenberg *et al.* [2003] evaluated a four-sensor LENS implementation with up to three interfering sources.

Other frequency-domain algorithms do not perform beamforming, but rather attenuate individual frequency bands that contain interference. The algorithm of Peissig, Kollmeier, and colleagues (Kollmeier, 1997; Kollmeier *et al.*, 1993; Peissig and Kollmeier, 1997; Wittkop *et al.*, 1997), hereafter referred to as the P–K algorithm, uses coherence and phase and amplitude differences between channels to determine a gain for each frequency band. Slyh and Moses [1993] describe another example of this type of algorithm. These techniques have been shown to be effective in attenuating off-axis sources using only two sensors, but they also introduce signal distortion by attenuating part of the target signal.

Frequency-domain minimum-variance distortionless-response (MVDR) beamformers [Cox *et al.*, 1986; 1987] are more computationally efficient than both the technique of time-domain correlation matrix inversion and the algorithm of Liu *et al.* [2000, 2001]. MVDR beamformers pass signals from a target direction with no distortion, assuming the sensors are matched. Kates and Weiss [1996] used adaptive frequency-domain algorithms (preadapted for 2 s) to extract speech in a diffuse noise field using signals from a five-sensor end-fire array. Their study included an MVDR algorithm with a limited adaptation rate.

We hypothesized that a frequency-domain MVDR (FMV) algorithm, specifically implemented for fast adaptation, might be effective for suppressing multiple interfering speech sources using only two sensors. We have implemented such an algorithm and evaluated its performance with computer simulations [Lockwood, 1999; Lockwood *et al.*, 1999]. Initial tests showed that, compared to time-domain adaptive algorithms, the computational cost of the FMV algorithm is similar, but it converges much more quickly. For simulated signals with up to four interfering sources, the FMV algorithm outperformed the algorithms of Frost [1972] and Griffiths and Jim [1982] in terms of SNR gain [Yang *et al.*, 2000].

The focus of the current study is threefold. The first goal is to further evaluate the performance of the FMV algorithm with a two-sensor array in real environments. Although the performance of the FMV algorithm under simulated conditions is promising [Larsen *et al.*, 2001], it has not been evaluated under actual room conditions with real recorded signals. The second goal is to evaluate and compare time- and frequency-domain algorithms in acoustic scenes in which there are more speech sources than sensors. This represents a challenging condition for beamforming algorithms, and a condition that most studies have not explored. Only the study of Kates and Weiss [1996] evaluated algorithms in an environment with more sources than sensors, but the sources were all multitalker babble rather than speech. Finally, because speech sources and acoustic scenes change rapidly in real-world listening environments, the third goal is to develop a better understanding of how adaptation speed affects the performance of these algorithms.

Recordings were made in three different rooms with varying reverberation times (RTs: 0.10, 0.37, 0.65 s) using three different microphone arrays: (1) two microphones coupled to the ear canals of a KEMAR mannequin; (2) two omnidirectional microphones in free field separated by 15 cm; and (3) two cardioid microphones in free field separated by 15 cm. The performances of a two-channel FMV algorithm [Lockwood, 1999; Lockwood *et al.*, 1999], the Frost adaptive beamformer [Frost, 1972], a version of the generalized sidelobe canceler (GSC) [Greenberg, 1998; Griffiths and Jim, 1982], and an implementation of the Peissig–Kollmeier (P–K) [Kollmeier *et al.*, 1993; Wittkop *et al.*, 1997] binaural algorithm were assessed. The signals were also processed with a fixed beamformer. The algorithms were not allowed to preadapt, and their performance was compared in terms of the signal-to-noise ratio (SNR) gain and target signal distortion produced by each. The adaptation characteristics of the FMV, Frost, and GSC algorithms were further evaluated in computer simulation. Finally, the computational costs of the algorithms were examined.

II. ALGORITHM IMPLEMENTATIONS

A. Frequency-domain MVDR (FMV) algorithm

Time-domain input signals, with a sampling rate of 22.05 kHz, are transformed periodically (every $L=16$ samples) into the frequency domain via a length-N FFT, using a Hamming window. For a two-microphone system, the

TABLE I. Algorithm parameters for best performance.

Microphone	Processing algorithm: FMV	Frost	GSC	P–K
Omnidirectional (Sennheiser MKEII)	$N=1024$ $F=32$	$N_F=401$ $m_F=1.0$	$K_{GSC}=401$ $\alpha=0.15$	$N=1024$ $c_1=5$
	$M=1.03$	$c_F=0.01$		$c_2=1$ $c_3=1$
Cardioids (Sennheiser ME-104)	$N=1024$ $F=32$	$N_F=401$ $m_F=1.0$	$K_{GSC}=401$ $\alpha=0.15$	$N=1024$ $c_1=5$
	$M=1.03$	$c_F=0.01$		$c_2=1$ $c_3=1$
KEMAR (Etymotic ER-1)	$N=1024$ $F=32$	$N_F=401$ $m_F=1.0$	$K_{GSC}=401$ $\alpha=0.07$	$N=1024$ $c_1=1.0$
	$M=1.10$	$c_F=0.01$		$c_2=1.5$ $c_3=1.0$

frequency-domain signals from the sensors are represented by the components of the vector $\mathbf{X}_k=[X_{1k}\ X_{2k}]$, where k indexes the frequency bins. The F most recent FFTs are stored in a buffer, and a correlation matrix $\mathbf{R}_k$ is calculated for each frequency bin k by using:

$$\mathbf{R}_k=\begin{bmatrix} \frac{M}{F}\sum_{i=1}^{F} X_{1k,i}^{*}X_{1k,i} & \frac{1}{F}\sum_{i=1}^{F} X_{1k,i}^{*}X_{2k,i} \\ \frac{1}{F}\sum_{i=1}^{F} X_{2k,i}^{*}X_{1k,i} & \frac{M}{F}\sum_{i=1}^{F} X_{2k,i}^{*}X_{2k,i} \end{bmatrix}, \tag{1}$$

where * represents complex conjugation, and M is a multiplicative ''regularization'' constant slightly greater than 1.00 that helps avoid matrix singularity and improves robustness to sensor mismatch. Cox *et al.* [1987] described the use of additive regularization to control the trade-off between robustness and white-noise gain. Values of N, M, and F are found in Table I. The correlation matrices $\mathbf{R}_k$ are updated every $L=16$ samples, allowing them to quickly track changes of the input signals in all frequency bands. The correlation matrices and FFT buffers were set to zero before processing each signal.

For each frequency band k, the monaural output of the beamformer is

$$Y_k=\mathbf{w}_k^H\mathbf{X}_k, \tag{2}$$

where $\mathbf{w}_k$ is a vector of frequency-domain weights and ^{H} represents the Hermitian transpose of a vector. The optimization goal and constraint are expressed for each frequency band as

$$\min_{\mathbf{w}_k} E\{|Y_k|^2\}, \tag{3a}$$

$$\text{subject to}\quad \mathbf{e}^H\mathbf{w}_k=1, \tag{3b}$$

where min represents the minimization of a function with respect to selected variables (the weights, $\mathbf{w}_k$, in this case), $E\{\}$ represents the expected-value operation, and $\mathbf{e}$ is a vector indicating the desired arrival direction. This general approach is originally attributed to Capon [1969]. For an on-axis target source, both detectors receive the signal at the same time and with the same amplitude, assuming identical detectors. Thus, $\mathbf{e}^H=[1\ 1]$. If the desired receive direction were off-axis, then $\mathbf{e}$ would be complex valued. For the minimization goal and constraint given in Eqs. (3a) and (3b), an optimal solution is known [Capon, 1969; McDonough, 1979; Cox *et al.*, 1987]. For each frequency bin k, the optimal weight vector $\mathbf{w}_{\text{opt},k}$ is given by

$$\mathbf{w}_{\text{opt},k}=\frac{\mathbf{R}_k^{-1}\mathbf{e}}{\mathbf{e}^H\mathbf{R}_k^{-1}\mathbf{e}+\sigma}, \tag{4}$$

where $\mathbf{R}_k$ is defined in Eq. (1), $\mathbf{R}_k^{-1}$ represents the matrix inverse of $\mathbf{R}_k$, and σ is a very small positive quantity that prevents division by zero. Inherent to this solution is the assumption that it is valid only if the inputs are stationary random processes. This is assumed to be true for small time intervals of speech signals in each frequency band.

To respond quickly to changes in $\mathbf{R}_k$, new optimal weights $\mathbf{w}_k$ are calculated for half of the frequency bands every L samples, so all weights are updated every $2L$ samples. This is possible because for a two-sensor system, the matrix inversion for each frequency band is computationally inexpensive. It will be demonstrated in Sec. V that this technique yields faster and more accurate tracking of nonstationary sources than time-domain techniques.

To obtain the time-domain output, the newest optimal weights for each frequency band are applied to buffered FFT data to obtain the output [Eq. (2)]. The resulting N frequency-domain values are then transformed to the time domain using a length-N inverse FFT. This occurs every L samples, and the central L samples of time-domain output are used. As the outer samples of the FFT window are attenuated by the Hamming window, this minimizes the effects of circular convolution which arise due to the FFT-based filtering, while requiring less computation and delay than an overlap-save or overlap-add method [Joho and Moschytz, 2000].

The main consideration in the FMV implementation was frequency resolution, which is determined by the FFT length relative to the sampling rate. For our experiments, a 1024-point FFT was chosen because it provided the best performance. Increasing the FFT length decreases the bandwidth of each frequency bin and should improve FMV performance (for stationary signals), as this provides the more detailed estimates of signal spectra. However, in practice, when the FFT length was too long the performance decreased, likely because the signal was not sufficiently stationary for the interval of the FFT. Also, longer FFTs required more data points, and objectionably increased the system delay.

B. Frost, GSC, P–K, and fixed beamformer implementations

Optimized parameter values (see the next section for details) for three of the algorithms described in this section are shown in Table I. All time-domain adaptive weights were initialized to that of a conventional beamformer, with a value of 0.5 for each channel at an appropriate delay, and zero

otherwise. All time-domain weights were updated each sample. The correlation matrices for the P–K algorithm were also set to zero before processing and were updated every 16 samples (as with the FMV algorithm).

The Frost algorithm was implemented with the update equation

$$\mathbf{W}_{\text{new}} = \mathbf{P} \cdot \left(\mathbf{W}_{\text{old}} - \frac{2}{3} \cdot \frac{\mathbf{W}_{\text{old}}^T \cdot \mathbf{x}_f \cdot \mathbf{x}_f}{m_F \cdot \mathbf{x}_f^T \cdot \mathbf{x}_f + c_F} \right) + \mathbf{F}, \quad (5)$$

where $\mathbf{W}_{\text{old}}$ is the previous set of time-domain filter coefficients, and m_F and c_F are adjustable parameters to control the step size and to prevent divide-by-zero, respectively. Additionally, $\mathbf{x}_f$ is a column vector composed of the time-domain input signals from both channels, $\mathbf{P}$ is a precomputed projection matrix, and $\mathbf{F}$ represents the response constraints, all as per Frost [1972]. The factor of 2/3 facilitates comparison of the step size with a bound derived by Frost. N_F is defined as the length of the adaptive filter.

The generalized sidelobe canceler (GSC) algorithm [Griffiths and Jim, 1982] was implemented as per Greenberg [1998]. This implementation improves performance and reduces target distortion in nonstationary environments when the target signal is strong. The weight update equation was

$$\mathbf{W}_{\text{new}} = \mathbf{W}_{\text{old}} + \frac{\alpha_{\text{sum}}}{K_{\text{GSC}}[\sigma_e^2(n) + \sigma_x^2(n)]} \cdot e(n) \cdot \mathbf{x}_G(n), \quad (6)$$

where α_{sum} is a step size parameter, n is an index of the current sample, $\mathbf{W}_{\text{old}}$ is the previous set of time-domain filter coefficients, e is the processed output, $\mathbf{x}_G$ is a vector of samples of the signal passed by the blocking matrix (mostly interference), K_{GSC} is the filter length, and σ_e^2 and σ_x^2 represent the average powers (updated every sample) of e and $\mathbf{x}_G$, respectively.

The P–K algorithm [Kollmeier *et al.*, 1993; Wittkop *et al.*, 1997] was implemented with three variable parameters to control the attenuation of the signal as a function of the phase and amplitude differences between channels and the coherence between channels. For the kth frequency band, the (real-valued) filter weight G_k was determined using

$$\begin{aligned} G_k &= g_{1k} \cdot g_{2k} \cdot g_{3k}, \\ g_{1k} &= \max\left(0,\ 1 - \frac{c_1 |\angle R_{12}|}{\pi}\right), \\ g_{2k} &= c_2 \cdot \min\left(\frac{R_{11}}{R_{22}}, \frac{R_{22}}{R_{11}}\right), \quad g_{3k} = \left(\frac{\text{Re}[R_{12}^2]}{R_{11} \cdot R_{22}}\right)^{C_3}, \end{aligned} \quad (7)$$

where c_1, c_2, and c_3 are adjustable parameters that control the sensitivity of the algorithm to phase differences between channels, amplitude differences between channels, and coherence between channels, respectively; $\angle$ represents the phase angle in radians, Re[ ] represents the real part of a complex value, and R_{ij} is an element from a frequency-domain correlation matrix [see Eq. (1)], with $M = 1.00$.

A simple fixed beamformer (referred to as a conventional beamformer) was implemented by averaging the signals from the two microphones after a matching filter was applied.

C. Optimization

All algorithms and metrics were implemented using MATLAB 6.5 (The MathWorks, Inc., Natick, MA). Floating-point calculations were performed with 64-bit precision. The algorithm parameters (Table I) were adjusted for best performance in terms of the SNR metric [Sec. III F, Eq. (9)]. The test signals were processed with many different sets of parameters. For the frequency-domain algorithms, the effect of changing the FFT length was generally independent of the effects of changing other parameters, so this was set first. Additionally, the additive constant c_F in the Frost algorithm was found to have little effect on performance so long as it was above a minimum value. This narrowed the parameter space to two variables for the FMV, Frost, and GSC algorithms, and to three variables for the P–K algorithm.

Many sets of values were chosen for the remaining free parameters, and the signals were processed and results obtained for all. It was found that performance varied with the number of interfering sources. Because this study emphasizes the effects of multiple interfering sources (more sources than sensors), the parameter set for each algorithm that produced the best overall performance for tests with multiple interferers was chosen as optimal. This was an *ad hoc* decision made by plotting the performance on a graph and choosing the line that was highest for the multiple-interferer tests. For the time-domain algorithms, special care was taken not to choose parameter sets that caused instability, as it was possible to have good multiple-interferer performance while the algorithm became unstable for the one-interferer test signals.

III. EXPERIMENTAL METHODS

A. Test materials

A series of high-context sentences by eight talkers (four females and four males) from the revised R-SPIN test [Bilger *et al.*, 1984] were recorded on digital audio tape (DAT) at a sampling rate of 48 kHz, quantized to 16 bits. Recordings were made in a sound-treated studio (model: Studio 7×5×5 ft., Acoustic Systems, Austin, TX). The recorded sentences were downsampled to 44.1 kHz. Three different sentences were chosen from each talker. This provided a total of 24 different sentences, 12 by male talkers and 12 by female talkers. A section of multitalker babble from the R-SPIN test was also used as an interfering signal in several test configurations. Its sampling rate was also 44.1 kHz.

B. Recording techniques and setup

Each of the 24 sentences, the multitalker babble signal, and 10 s of white noise were played back from eight loudspeakers housed in a semicircular enclosure (SPATS, Sensimetrics Corp., Somerville, MA). Each loudspeaker was equidistant (75 cm) from a central point and comprised a single 7.6-cm-diameter driver with frequency response restricted to 200-Hz to ~13 kHz. The two-microphone arrays were located at the central point of the array, 1.15 m above the floor,

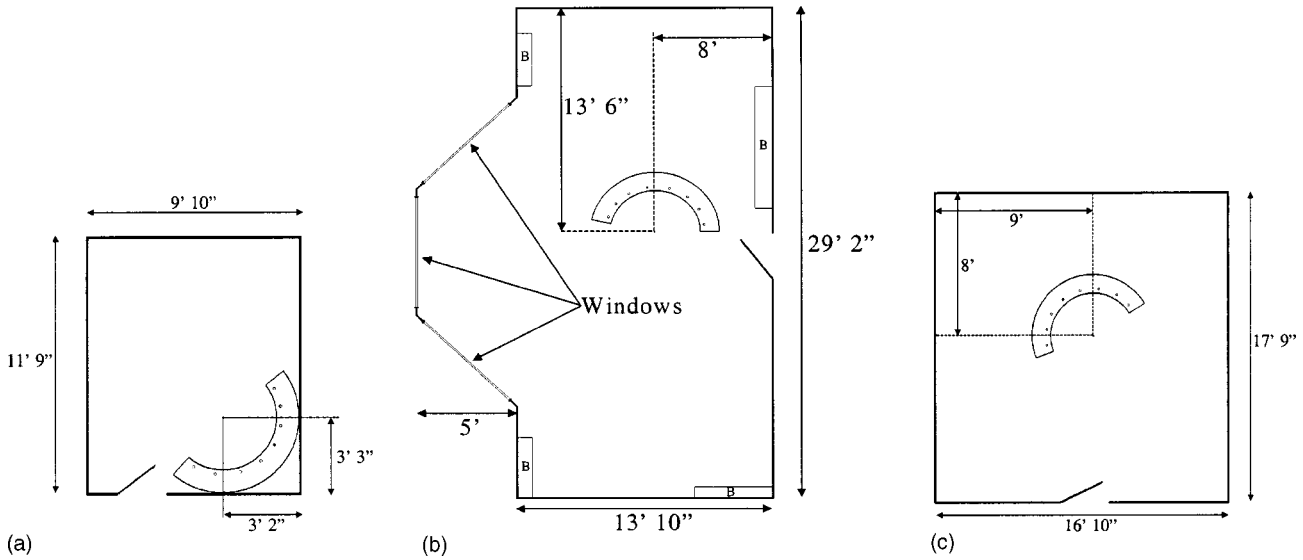

FIG. 1. (A) Diagram of room 1 (treated with acoustic foam) with loudspeaker array. (B) Diagram of room 2 (conference room with windows) with loudspeaker array. (C) Diagram of room 3 (conference room, bare walls) with loudspeaker array. Filled circle represents target loudspeaker, open circle represents interferer; distances to central point of the array are shown.

and oriented such that the loudspeakers were at azimuths of 60°, 40°, 20°, 0°, −20°, −40°, −60°, −80° with respect to the broadside array.

Recordings were made with three sets of microphones: (1) Sennheiser MKEII omnidirectional microphones spaced 15 cm apart in free field; (2) Sennheiser ME104 cardioid microphones spaced 15 cm apart in free field; and (3) Etymotic ER-1 microphones mounted in the ears of a KEMAR mannequin (Knowles Electronics, Itasca, IL). The omnidirectional and cardioid microphones were connected directly to a microphone preamplifier (Millennia Media HV-3B, Placerville, CA) that was connected to the inputs of the Aark 24 system. The KEMAR microphones were connected to their own dedicated preamps, and then to the Millennia Media preamplifier. Recording and playback were done with a sampling rate of 44.1 kHz. The recorded data were downsampled to 22.05 kHz prior to being saved to hard disk.

C. Room descriptions

Recordings were made in three rooms with different reverberation characteristics. In all rooms, the ceiling had acoustic tile suspended at a height of 9 ft. Above it was a concrete ceiling. The floors of all rooms were covered with short carpet. The dimensions of the rooms and the positioning of the loudspeaker and microphone arrays within them are shown in Figs. 1(a), (b), and (c). Room 1 was an acoustically controlled space with 6.4-cm-thick foam (SONEX Valueline, Illbruck Corp., Minneapolis, MN) attached to all wall surfaces; Room 2 was a rectangular conference room with bay windows and bookshelves on two walls; Room 3 was a rectangular conference room with painted gypsumboard walls.

The reverberation times of the rooms were measured with a sound-level meter (Bruel & Kjaer, model 2260). A more suitable source of an impulse was not available, so a hand clap was used as a stimulus after it was found that it provided consistent results over several measurements. Five measurements were taken in each room at the central point of the loudspeaker array and the results were averaged. The meter calculated the T_{60} times for 1/3-octave bands between 200-Hz and 10 kHz. The values for each band were averaged over the five measurements, and then averaged across frequency bands to obtain the reverberation time. The estimated average T_{60} values for rooms 1, 2, and 3 were 0.10, 0.37, and 0.65 s, respectively.

D. Compensation, energy equalization, test creation

To match the microphones, the recorded white noise from the loudspeaker at 0° was filtered with a 43-tap FIR filter adapted by an LMS algorithm to match the responses of the microphones for the target direction. For a sampling rate of 22.05 kHz, this filter has a 2-ms delay, corresponding to a sound propagation distance of approximately 0.7 m. Given the dimension of the rooms and location of the microphones, direct reflections should require more than 2 ms to reach the microphones; the filters should thus compensate mostly for the differences in frequency response of the microphones, rather than responding to reflections from the room.

All acoustic recordings (except for the recordings of white noise) were of the same duration; however, the duration of the spoken sentences varied. Therefore, the average power of each recorded sentence was calculated (taking into account the varying durations) and all sentences were scaled to have the same average power. This is referred to as the normalized energy of a single interfering source. The amplitude of each interfering source could then be adjusted to provide a specified relative amount of interference with respect to the target sentence.

To construct a test, the recordings of each source from its appropriate location (loudspeaker) were scaled and added, keeping the left and right channels separated, thereby producing a binaural test signal. This assumed that the acoustic addition of the sources and the various components in the recording and playback system all act in a linear manner.

TABLE II. Summary of test configurations.

Configuration number	Azimuth angle of source				
	+60°	+20°	0°	−40°	−80°
1	Interferer[a]		Target		
2		Interferer	Target		
3	MTB[a]		Target		
4		MTB	Target		
5	Interferer		Target		Interferer
6	Interferer		Target	Interferer	
7		Interferer	Target	Interferer	
8	Interferer	Interferer	Target		Interferer
9	Interferer	Interferer	Target	Interferer	
10	Interferer	Interferer	Target	Interferer	Interferer
11	Interferer or MTB	Interferer or MTB	Target	Interferer or MTB	Interferer or MTB

[a]''Interferer'' and ''MTB'' indicate interference from the angle listed by a single talker and multitalker babble source, respectively.

E. Test terminology and guidelines

A test configuration was defined as one particular spatial arrangement of target and interfering sources. Eleven test configurations were used; Table II describes the positioning of the target and interferer(s). Each configuration contained one on-axis target source (always a single talker) presented from the loudspeaker at 0°, and one to four off-axis interfering sources presented from the loudspeakers at possible azimuth angles of +60°, +20°, −40°, or −80°, each a single talker or multitalker babble source. Multitalker babble was used in test configurations 3 and 4 as the only interferer. For test configuration 11, a multitalker babble was placed at one of the four interference locations and single-talker interferers were placed at the other three.

All sources were located in the front half-plane, and thus the effects of the directionality of the cardioid microphones was minimal. Additionally, a front–back ambiguity did not exist as there was no source located greater than 90° from the target signal.

For each spatial test configuration, 12 tests were constructed. A test was defined as one permutation of a target sentence and interfering sentence(s). Twelve different sentences, three sentences each from two male and two female talkers, were used exclusively as target sentences. (Thus, for each test configuration, six target talkers were males and six were females.) The remaining 12 sentences were used exclusively as interferers. Additionally, in any test, each source was a different talker, and all interferers had the same energy.

Each test was constructed and processed at three different SNRs. For configurations 1–4 (one interferer), three normalized energies were used: −3, 0, and +3 dB (corresponding to each interference source having 3 dB less, the same, or 3 dB more average power, respectively, than the target source). For test configurations 5–11, three lower normalized energies (−6, −3, 0 dB) were used. Therefore, the entire battery of tests consisted of 11 test configurations, 12 tests per configuration, and 3 SNR levels per test, for a total of 396 test signals. Each test was approximately 2.5 s long, for a total of 16.5 minutes of test signals. Results for each test configuration are presented in Sec. IV, and were obtained by averaging the performance metrics over the 12 tests and three SNR levels for each configuration.

F. Performance metrics

Processing with the various algorithms was done off-line and the signals from the individual sources before and after processing were known, greatly simplifying the calculation of performance metrics. Because the target signal could be distorted by significant amounts during processing, and because distortion is detrimental to speech perception, a signal-to-noise ratio (SNR) metric that incorporated both interference and signal bias (distortion) error was chosen. The output SNR (after processing) is defined as

$$\mathrm{SNR}_{\mathrm{OUT}} = 10 \cdot \log_{10}\left(\frac{\Sigma_{v=1}^{V} t_u(v)^2}{\Sigma_{v=1}^{V} (y_p(v) - t_u(v))^2} \right), \tag{8}$$

where $y_p(v)$ and $t_u(v)$ are the vth samples of the processed output and unprocessed (ideal) target signals, respectively, and V is the length of the signal in samples. In this work, the unprocessed signal is binaural and the output is monaural. Thus, the expression is modified to be

$$\mathrm{SNR}_{\mathrm{OUT}} = 10 \cdot \log_{10}\left(\frac{\Sigma_{i=1}^{V} [g_1 t_{u,L}(v) + g_2 t_{u,R}(v)]^2}{\Sigma_{i=1}^{V} (y_p(v) - (g_1 t_{u,L}(v) + g_2 t_{u,R}(v)))^2} \right), \tag{9}$$

where $t_{u,L}(v)$ and $t_{u,R}(v)$ are the unprocessed target signals from the left and right microphones, respectively, and g_1 and g_2 represent the gains applied by the algorithm (effectively the steering vector) to the target signal in each channel. In this study, g_1, $g_2 = 0.5$ because filters are used to match the responses of the microphones in the target direction (see Sec. III D). If the microphones are not matched for the target direction, the gains must be replaced with filters.

The input SNR for one channel is defined as

$$\mathrm{SNR}_{\mathrm{IN}} = 10 \cdot \log_{10}\left(\frac{\Sigma_{v=1}^{V} t_{u,L}(v)^2}{\Sigma_{v=1}^{V} i_{u,L}(v)^2} \right), \tag{10}$$

where $i_{u,L}(v)$ is the interference signal received by the left microphone. Because all signals were recorded individually, the target and interference signals received by the micro-

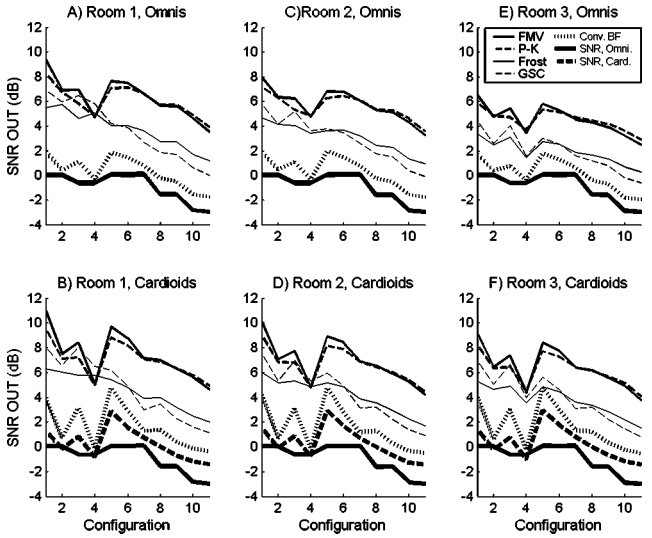

FIG. 2. (A), (C), (E) Processed and unprocessed SNR values for the left omnidirectional microphone data in rooms 1, 2, and 3, respectively. (B), (D), (F) Processed and unprocessed SNR values for the left cardioid microphone data in rooms 1, 2, and 3, respectively.

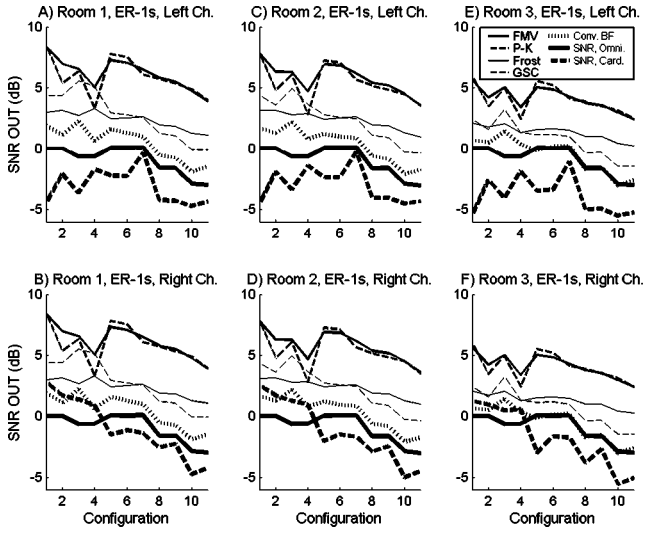

FIG. 3. (A), (C), (E) Processed and unprocessed SNR values for left KEMAR microphone data in rooms 1, 2, and 3, respectively. (B), (D), (F) Processed and unprocessed SNR values for right KEMAR microphone data in rooms 1, 2, and 3, respectively.

phones are known and the $\mathrm{SNR}_{\mathrm{IN}}$ calculation yields the SNR after microphone reception.

A second metric used was a measure of the target distortion. Access to the processed target signal allowed this distortion to be quantified. The FMV, Frost, and GSC algorithms are constrained to pass the target with zero distortion, and are referred to as distortionless response beamformers. However, microphone mismatch and reverberation mean that only a portion of the target signal satisfies the constraint, so target distortion may occur. The P–K algorithm attenuates frequency bands that contain interference, and thus almost always cancels some of the target signal. We define a measure of target distortion as

$$T_D = \frac{\Sigma_{v=1}^{V}(t_p(v) - [g_1 t_{u,L}(v) + g_2 t_{u,R}(v)])^2}{\Sigma_{v=1}^{V}[g_1 t_{u,L}(v) + g_2 t_{u,R}(v)]^2}, \qquad (11)$$

where $t_p(v)$ is the processed (monaural) target signal. When T_D is greater than zero, the target signal has been distorted. Note that this metric does not distinguish between attenuation and phase distortion.

IV. RESULTS AND DISCUSSION

A. Algorithm comparisons

Figure 2 (results for omnidirectional and cardioid microphones, left channel) and Fig. 3 (results for KEMAR microphones, both channels) show the average $\mathrm{SNR}_{\mathrm{OUT}}$ produced by the algorithms after processing. The $\mathrm{SNR}_{\mathrm{IN}}$, as received by the omnidirectional, cardioid, and KEMAR (ER-1) microphones, is labeled as (SNR, omni), (SNR, card), and (SNR, ER-1), respectively. To facilitate the comparisons in Sec. IV C, (SNR, omni) is also shown with the results for the cardioid and KEMAR microphones. In all three rooms and for all microphone types, the FMV and P–K algorithms consistently outperform the Frost and GSC algorithms in terms of $\mathrm{SNR}_{\mathrm{OUT}}$ for all test configurations with more than one interfering source (configurations 5–11). This is the main advantage conferred by the frequency-domain algorithms.

For the one-interferer tests (configurations 1–4), the FMV and P–K algorithms perform better than the time-domain algorithms in terms of $\mathrm{SNR}_{\mathrm{OUT}}$ for configurations 1 and 2. For configurations 3 and 4, the time-domain algorithms perform similarly or slightly better in room 1, and in room 2 with cardioid microphones. (Though the algorithm parameters were not optimized for these one-interferer tests, the improvement in performance that can be obtained by optimizing the parameters for these tests is not very large.)

The FMV and P–K algorithms perform similarly in

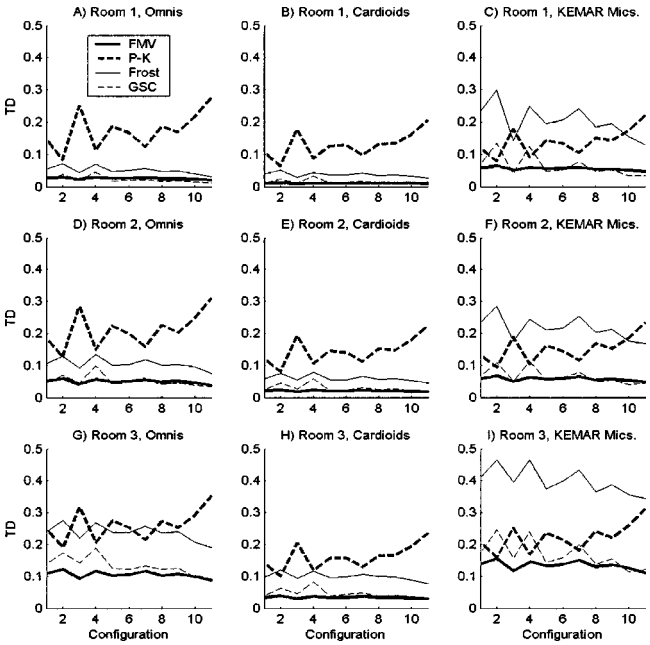

FIG. 4. (A), (D), (G) Target distortion for omnidirectional microphones in rooms 1, 2, and 3, respectively. (B), (E), (H) Target distortion from cardioid microphones in rooms 1, 2, and 3, respectively. (C), (F), (I) Target distortion for KEMAR microphones in rooms 1, 2, and 3, respectively.

terms of SNR_{OUT}, but the P–K produces more target distortion, as can be seen in Fig. 4. This is to be expected, as the FMV algorithm uses constrained spatial filtering to remove interference without distorting the target signal (with matched sensors and no reverberation), while the P–K algorithm simply attenuates frequency bands that appear to be dominated by interference. Therefore, target distortion is accepted in exchange for an improved SNR_{OUT}.

The differences between the GSC and the Frost algorithm are generally small. Griffiths and Jim [1982] showed that their GSC structure converges to the same solution as the Frost beamformer, although their paths to convergence may be different. In the tests conducted here, the GSC algorithm appears to have a slight performance advantage over the Frost algorithm for the single-interferer tests, while the Frost algorithm appears to perform better than the GSC as the number of interferers rises. This difference is more apparent with the KEMAR microphones (Fig. 3), for which the GSC has a more pronounced advantage in the one-interferer tests (up to 3 dB) and for which the Frost algorithm performs better in the multiple-interferer tests (by up to 2 dB). The cause of these differences is not known.

The FMV and GSC algorithms generally produce less distortion (Fig. 4) of the target signal than the Frost and P–K algorithms. The FMV and GSC distortion figures are often comparable, but the FMV has an advantage in the more reverberant environments. The high distortion figures of the P–K algorithm are expected, but those for the Frost algorithm are not. They suggest that the Frost algorithm is more sensitive to the amount of reverberation and the type of microphone being used than the GSC or FMV algorithms are.

The conventional beamformer, which averages the two inputs, performs more poorly than all other algorithms, as shown in Figs. 2 and 3.

B. Effects of reverberation

The performance, in terms of SNR_{OUT} (Figs. 2 and 3), of all algorithms generally decreases by 1 dB or less for the tests in room 2 ($RT=0.37$ s) as compared to room 1 ($RT=0.10$ s). This decrease is fairly consistent across the different types of microphones. Performance differences between room 2 and room 3 ($RT=0.65$ s) are slightly more pronounced, including a drop of 1–2 dB for the FMV and P–K algorithms across all configurations. Frost and GSC algorithm performance is reduced by a similar amount.

In general, the performance of all algorithms is decreased by similar amounts as the reverberation time of the room increases. Importantly, this implies that the performance advantage of the frequency-domain algorithms is maintained as the reverberation time increases. For the one-interferer configurations, the advantage of the frequency-domain algorithms actually increases with reverberation time; this implies that the advantages of these algorithms are not limited to the configurations in which there are more sources than sensors.

It is important to point out that distortion for distortionless-response beamformers (FMV, Frost, GSC) is due to sensor mismatch (including mis-steering) and reverberation. The matching filters were chosen carefully, but they were kept sufficiently short so that they did not compensate for reflections and reverberation. Therefore, the distortion figures generally reveal how sensitive the algorithms are to reverberation.

Increasing the amount of reverberation consistently increases target distortion. Figure 4 reveals that the target distortion for the FMV, Frost, and GSC algorithms approximately doubles from room 1 to room 2, and doubles again from room 2 to room 3. The FMV and GSC generally have the lowest distortion for all rooms. The distortion of the P–K algorithm rises by approximately 0.05 for both rooms 2 and 3. The P–K algorithm has by far the highest distortion in rooms 1 and 2 with omnidirectional and cardioid microphones; Frost has the second highest. The distortion of the Frost algorithm is comparable to that of the P–K algorithm for room 3. While the FMV, Frost, and GSC algorithms appear to be more sensitive to increases in reverberation time than the P–K algorithm, for the range of RTs used in this study, the distortion of the FMV and GSC algorithms is still notably less than that of the P–K algorithm.

C. Microphone effects

The directionality of the cardioid microphones accounts for up to a 3-dB improvement in the SNR_{IN} (SNR, card) over that of the omnidirectional microphones (SNR, omni), as can be seen in Fig. 2. This appears to account for the approximately 1–2-dB improvement in the SNR_{OUT} that is observed for all algorithms for the multiple-interferer tests in all three rooms. Overall, the FMV algorithm with cardioid microphones performs best, albeit by a small margin.

The KEMAR microphones (Fig. 3) cause a reduction in the SNR_{IN} for the left channel and an increase in the SNR_{IN} (SNR, ER-1) for the right channel for single-interferer tests. This is because the interfering source for these tests is from the left of the array, at $+60°$ or $+20°$, and thus the interference is stronger in the left ear of the KEMAR than the right. For multiple interferer tests, the SNR_{IN} is lower than with the other microphones. Thus, processing the KEMAR signals yields 1–2-dB lower SNR_{OUT} than with omnidirectional or cardioid microphones. Another notable effect of the KEMAR microphones is the dramatic increase in distortion [Figs. 4(c),(f),(i)] for the Frost algorithm, likely caused by sensitivity to reverberation or microphone mismatch. The distortion for the GSC algorithm remains low.

For the experiments in this study, all sound sources were placed in the front half-plane. This reduced the benefits obtained by using the cardioid or KEMAR microphones. If sources had been placed in the back half-plane, using these microphones would have resulted in higher values of SNR_{IN}. However, for this study, the effect of the directionality of the microphones that is most interesting is the ability to reduce the reverberation in the recorded signals. It was hoped that the directional microphones would make all algorithms more robust to reverberation effects. Compared to the processed signals from the omnidirectional microphones, processing the signals from the cardioid microphones produces a somewhat higher output SNR (for all three rooms) and generally lower target distortion, while processing signals from the

TABLE III. Frequencies of sinusoidal interferers and positions as a function of time.

Time interval	Sinusoid frequencies (Hz)									
	500	611	682	769	921	1016	1095	1187	1331	1448
	Azimuth angles									
0.00–0.75 s	−65°	−55°	−45°	−35°	−25°	20°	30°	40°	50°	60°
0.75–1.50 s	20°	30°	40°	50°	60°	−65°	−55°	−45°	−35°	−25°
1.50–2.25 s	60°	50°	40°	30°	20°	−25°	−35°	−45°	−55°	−65°

KEMAR microphones produces slightly lower SNRs and higher distortion. Overall, the performances of the algorithms were improved by the use of the cardioid directional microphones, and only slightly reduced by the KEMAR microphones.

D. Best parameter sets

The optimal parameter sets for each algorithm (Table I) were found to be the same for both omnidirectional and cardioid microphones, but different for the KEMAR microphones. As compared to the best parameter sets for the free-field microphones, the FMV algorithm performs best with a larger regularization value M, the GSC algorithm requires a smaller step size α to remain stable in multiple interferer tests, and the P–K algorithm performs best with values of c_1 and c_2 that weight the amplitude difference most heavily instead of the phase difference.

V. TIME-VERSUS FREQUENCY-DOMAIN MVDR BEAMFORMERS

A. Overview

The results in Sec. IV show the performance advantage of the FMV over conventional time-domain algorithms. The Frost and GSC algorithms differ from the FMV in that they are iterative-adaptive time-domain techniques, while the FMV calculates optimal solutions directly for many frequency bands. However, all are classified as MVDR beamformers, because they have identical optimization goals and constraints (minimize output power, pass target source undistorted). These MVDR beamformers (time- and frequency-domain) will all asymptotically converge to identical solutions for inputs that are stationary random processes. Therefore, differences in SNR performance are due to different adaptation characteristics in the presence of nonstationary signals such as speech.

Using specific simulated signals, we now examine the reasons for the performance advantage of the FMV over the other algorithms. The test signals are not true binaural recordings as in the previous experiments, rather they are simulated. The signals for each source in each channel differ only by time delay. Therefore, no sensor mismatch is present, the FMV, Frost, and GSC algorithms produce no target distortion, and performance differences result only from the varying adaptation characteristics of the algorithms. The absence of sensor mismatch also allows accurate beam patterns to be calculated to observe these characteristics.

The algorithms were all optimized to produce the best SNR gain for the simulated signals. The first example illustrates the behavior of MVDR beamformers with statistically stationary interference. The second example shows that the FMV is able to adapt more quickly and accurately to rapid changes in interfering sound sources, and thus it more effectively attenuates multiple-interfering sound sources.

B. Simulation example 1: Multiple spatially separated sinusoidal interferers

MVDR beamforming algorithms may, in theory, attenuate multiple, statistically stationary narrow-band interference signals at different frequencies. To demonstrate this, a two-channel test signal was synthesized containing a single on-axis speech source and ten off-axis sinusoidal signals, each with a different frequency and azimuth. The sinusoid frequencies ranged from 500 to 1450 Hz, with approximately 100-Hz spacing. Neglecting frequency smearing due to win-

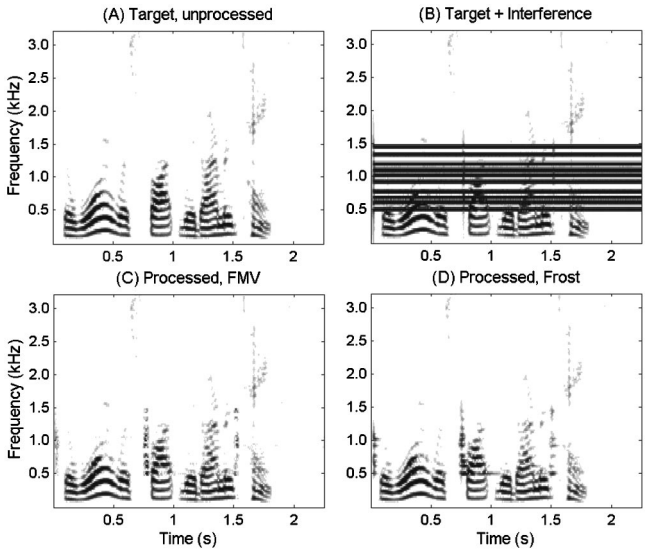

FIG. 5. (A) Spectrogram of the target speech signal in example 1 (''The war was fought with armored tanks.'') before processing. (B) Spectrogram of the combined target and sinusoidal interference signals in example 1 before processing. The target signal is at 0° azimuth, and ten sinusoidal interferers originate from angles between ±65° azimuth. (C), (D) Spectrograms of example 1 after processing with the FMV algorithm (C) and the Frost algorithm (D).

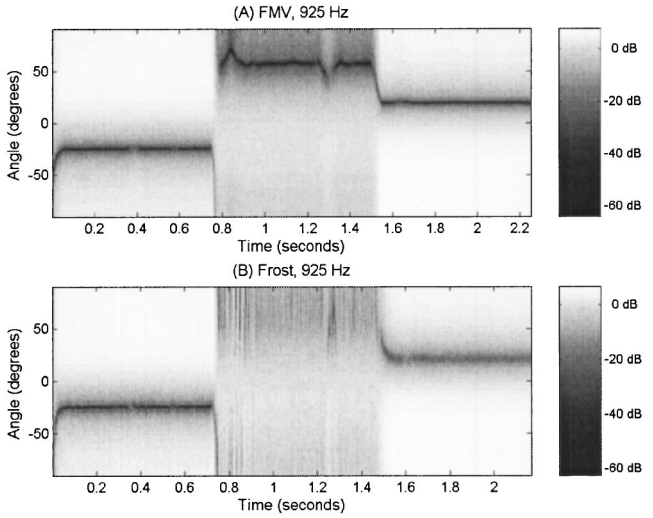

FIG. 6. (A) Magnitude (in decibels) of the beam pattern for the FMV algorithm for the 925-Hz FFT bin. Note the null at $-25°$ ($t=0.00$ to 0.75 s), $\approx 60°$ ($t=0.75$ to 1.50 s), and 20° ($t=1.50$ to 2.25 s). (B) Magnitude (in decibels) of the beam pattern for the Frost algorithm for the 925-Hz FFT bin. The null location varies considerably, but on average is near $-25°$ ($t=0.00$ to 0.75 s), 60° ($t=0.75$ to 1.50 s), and 20° ($t=1.50$ to 2.25 s).

dowing, there is no spatial or frequency overlap of the interferers. So that the adaptation rate of both algorithms can be observed, the sinusoids instantaneously change locations at 0.75 and 1.5 s. The sinusoid frequencies and azimuths as a function of time are listed in Table III.

Spectrograms of the target and combined signals before processing are shown in Figs. 5(a) and (b), respectively. The SNR before processing is -7.35 dB. The spectrograms of the processed output from the FMV and Frost algorithms are shown in Figs. 5(c) and (d), respectively, and the SNR gains were 18.56 and 16.14 dB, respectively.

While processing this signal, all of the frequency-domain weights calculated by the FMV algorithm for the FFT bin centered on 925 Hz were recorded. All of the time-domain filter coefficients from the Frost algorithm were transformed to frequency-domain coefficients and recorded. Using these coefficients, the beam pattern for the 925-Hz FFT bin (the gain, in decibels, for signals as a function of azimuth angle and time) was calculated for both algorithms, and is shown in Fig. 6(a). A two-element beamformer (with a constraint) can place at most one spatial null per frequency band, and the null direction changes with time as it adapts to the input signal. The null is easily visible, occurring at $-25°$, near 60°, and at 20° for time intervals [0.0, 0.75 s], [0.75, 1.5 s], and [1.5, 2.25 s], respectively. Table III shows that a sinusoidal interferer with a frequency of 921 Hz was at these locations at the above time intervals.

Figure 6(a) also shows that the FMV algorithm null deviates from the direction of the interferer between 0.75 and 1.5 s. Examining Fig. 5(b) reveals that the spectra of the target (speech) and interference (sinusoidal) sources overlap in the 925-Hz FFT bin between 0.75 and 1.5 s. The optimal filter weights are calculated assuming that there is no correlation between target and interfering sources; in practice, this is rarely true over short time intervals. Thus, the correlation matrix estimates contain some error. (To reduce this error, correlation matrices may be formed using data from longer time intervals with the trade-off of slower adaptation.)

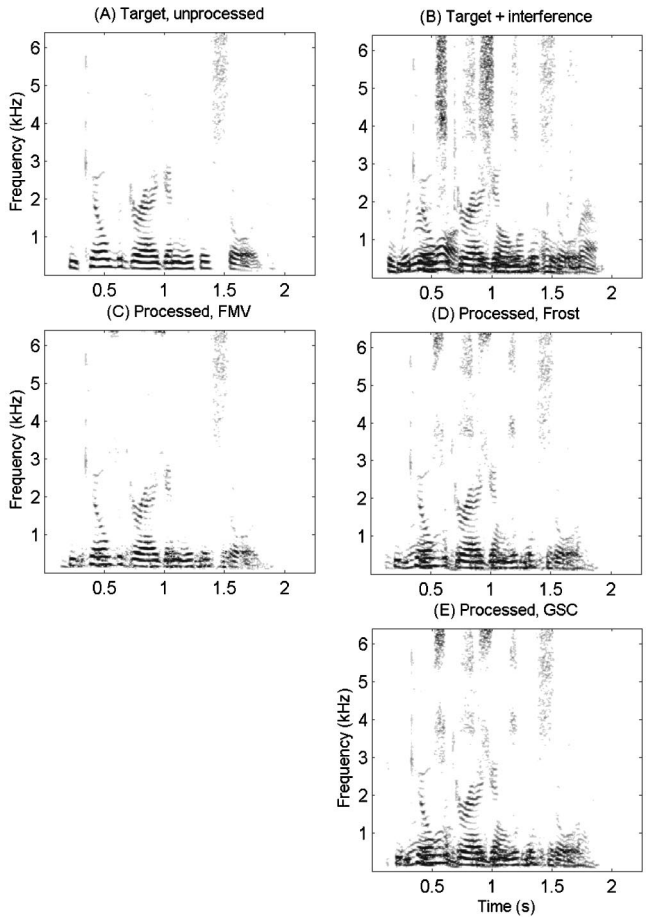

FIG. 7. (A) Spectrogram of target speech signal in example 2 (''He killed the dragon with his sword.'') before processing. (B) Spectrogram of combined target and two interfering speech signals in example 2, before processing. The target signal is at 0° azimuth and the interference sources are at $\pm 45°$. (C), (D), (E) Spectrogram of example 2 after processing with the FMV (C), Frost (D), and GSC (E) algorithms.

Figure 6(b) shows the beam pattern for the Frost algorithm for the 925-Hz FFT bin. The direction of the null placed by the Frost algorithm is erratic when the signals overlap (between 0.75 and 1.5 s). This occurs because a noisy time-domain estimate of the gradient is used to adapt the filter weights. This misadjustment error can be reduced by lowering the adaptation rate, but only at the expense of slower convergence, thus leading to poorer performance in the presence of multiple interferers.

In this example the FMV and Frost algorithms show nonideal effects when target and interference signals overlap. However, both algorithms effectively attenuate multiple, nonoverlapping statistically stationary interferers using signals from two sensors.

C. Simulation example 2: Multiple speech interferers

This example uses statistically nonstationary interference sources (speech) to compare the speed and accuracy of convergence of the FMV, Frost, and GSC algorithms. An on-axis speech source and two off-axis interfering speech sources (at $\pm 45°$) are simulated. Figures 7(a) and (b) show

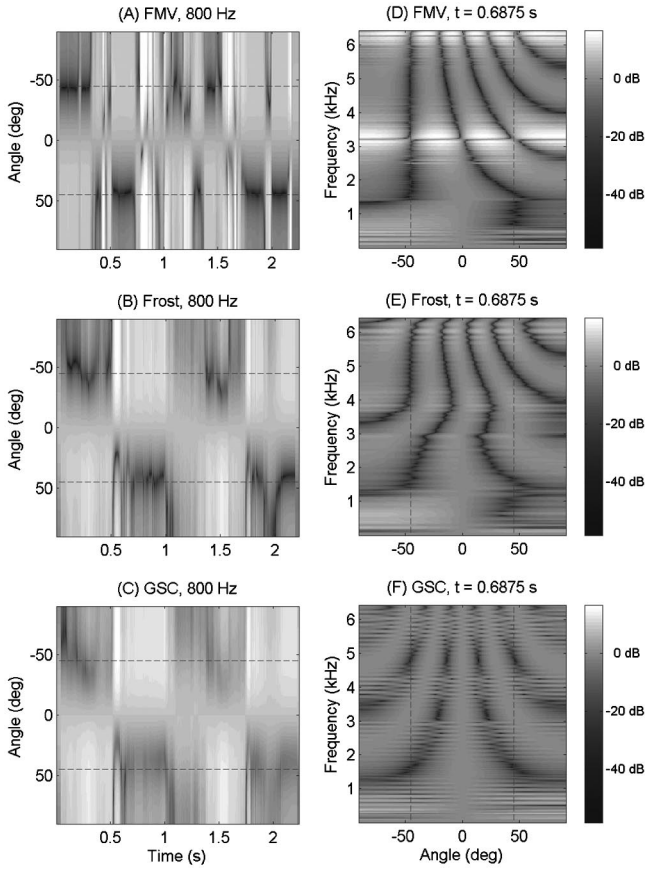

FIG. 8. (A), (B), (C) Magnitude (in decibels) of the beam patterns for the FMV (A), Frost (B), and GSC (C) algorithms for the 800-Hz FFT bin. Nulls near $\pm 45°$ correspond to cancellation of interfering speech sources at $\pm 45°$. (D), (E), (F) Magnitude (in decibels) of the beam patterns at $t = 0.6875$ s for the FMV (D), Frost (E), and GSC (F) algorithms.

the spectrograms of the target and target plus interference signals, respectively. The SNR gains produced by processing with the FMV, Frost, and GSC algorithms are 9.14, 5.41, and 4.30 dB, respectively, indicating that the FMV most effectively reduced the interference. This is supported by Figs. 7(c)–(e), which show the spectrograms of the outputs of the FMV, Frost, and GSC algorithms, respectively. It is evident that the FMV output [Fig. 7(c)] more closely resembles the target signal [Fig. 7(a)].

The better performance of the FMV in terms of the SNR gain may be explained by examining the time-varying beam patterns for the FMV, Frost, and GSC algorithms, shown in Figs. 8(a), (b), and (c), respectively. The nulls placed by the FMV converge consistently and quickly to $\pm 45°$. In contrast, the Frost algorithm's nulls rarely converge to exact the directions of the interference sources, and the GSC algorithm's convergence is even less precise. Figures 8(d), (e), and (f) show instantaneous beam patterns for the FMV, Frost, and GSC algorithms. At this point in time, the FMV algorithm placed nulls closer to the directions of the interferers (at $\pm 45°$), especially for the $-45°$ source between 2300 and 5000 Hz.

This example illustrates the typical behavior of the algorithms for signals containing multiple speech sources. When compared to the Frost and GSC algorithms, the FMV algorithm exhibits faster, more accurate adaptation. This results in better interference cancellation.

D. Computational complexity

The computational complexity of the FMV algorithm is derived after some simplifications of the computation. Time-domain microphone data are real-valued, so weights for only $N/2+1$ frequency bins are calculated (where N is the number of FFT bins), and the remaining $N/2-1$ bins are obtained by utilizing the conjugate-symmetry property of the FFT. Because cross-correlation terms in $\mathbf{R}$ are conjugate symmetric, only three of the four terms need to be calculated. For an n-microphone system, weights for only $n-1$ channels

TABLE IV. Computational expense for various parts of the FMV algorithm.

Part of algorithm	Complex adds	Complex multiplies	Complex divisions	Real multiplies	Time
Windowing[a]	0	0	0	$n \cdot N$	$\frac{L}{f_s}$
FFTs	$(n+1)\cdot\frac{N}{2}\log_2 N$	$(n+1)\cdot\frac{N}{4}\log_2 N$	0	0	$\frac{L}{f_s}$
Correlation	$\frac{(n^2+n)\left(\frac{N}{2}+1\right)}{2}$	$\frac{(n^2+n)\left(\frac{N}{2}+1\right)}{2}$	0	0	$\frac{L}{f_s}$
Weight application	$(n-1)\left(\frac{N}{2}+1\right)$	$n\left(\frac{N}{2}+1\right)$	0	0	$\frac{L}{f_s}$
Matrix inversion	$\frac{n^3}{3}$	$\frac{n^3}{3}$	0	0	$\frac{BL}{f_s}$
Weight calculation (not including matrix inversion)	$(n^2+n-2)\left(\frac{N}{2}+1\right)$	$(n^2+n)\left(\frac{N}{2}+1\right)$	$(n-1)\left(\frac{N}{2}+1\right)$	0	$\frac{BL}{f_s}$

[a]For example, the windowing operation requires $n \cdot N$ real multiplies every L/f_s seconds, where L is the number of samples between FFTs, and f_s is the sampling rate in Hz.

need be calculated using Eq. (4), because Eq. (3b) may be utilized to obtain weights for the nth channel with less computation.

The computational complexity of the FMV is a function of f_s, the sampling frequency, n, the number of sensors, B, the block length (the number of FFTs, for a single channel, that are calculated between each computation of new filter weights), N, the number of points in the FFT, and L, the number of output samples obtained from each IFFT. (It is assumed that there is no overlap of groups of output samples.) Assumptions made for this analysis are: (1) a radix-2-real-valued FFT is calculated every L samples; (2) calculating a matrix inverse requires $n^3/3$ complex multiplications and $n^3/3$ complex additions; and (3) the second weight is obtained using Eq. (3b).

Table IV summarizes the costs of each part of the algorithm in terms of complex additions, complex multiplications, complex divisions, real multiplications, and the time in which each must be completed. These cost estimates do not include the cost of overhead within a computer program, such as retrieving data from memory; they reflect only the mathematical operations required to execute the algorithm.

Assuming that two real additions for each complex addition, four real multiplications and two real additions for each complex multiplication, and two complex multiplications and two real divisions for each complex division are required, the total number of operations per second (OPS) required for the algorithm is

$$\mathrm{OPS}=\frac{f_s}{L}\left[N\left(n+\frac{5}{2}(n+1)\log_2 N\right)+(N+2)\left(\frac{4}{3B}n^3+\left(2+\frac{4}{B}\right)n^2+\left(20+\frac{4}{B}\right)n-15-\frac{2}{B}\right)\right]. \quad (12)$$

Additionally, the cost of the Frost and GSC algorithms are, respectively,

$$\mathrm{OPS}_F=f_s\cdot(2K_F+4K_F\cdot n), \quad (13)$$

$$\mathrm{OPS}_{\mathrm{GSC}}=f_s\cdot(n^2+4K_{\mathrm{GSC}}\cdot(n-1)). \quad (14)$$

For the best parameter sets, the FMV and P–K algorithms require 178 million OPS, the Frost algorithm requires 88 million OPS, and the GSC requires 36 million OPS. As implemented, the FMV requires about five times more computation than the GSC and about twice as much as the Frost. By increasing L from 16 to 32 in the FMV and P–K algorithms, the cost is roughly cut in half. In practice, we have found that by increasing the interval between calculation of FFTs, the computational cost of the FMV algorithm can be made approximately equal to that of the GSC or Frost algorithms with only a small decrease in performance.

The P–K algorithm implementation was very similar to that of the FMV, because it also performs FFTs periodically and uses correlation matrices and frequency-domain filtering. Its cost was considered to be comparable to that of the FMV algorithm. Thus, with similar computational cost, the FMV and P–K algorithms provide performance superior to that of the Frost and GSC algorithms in terms of SNR gain. By comparison, the cost of the direct inversion of a time-domain correlation matrix (as per Capon [1969]) of dimension 401 (the length of the Frost and GSC filters), done every 32 samples (as with the calculation of the FMV weights), is approximately 15 billion OPS.

VI. CONCLUSION

Test signals containing multiple speech sources were created from recordings made with three different types of two-microphone arrays in three rooms with varying reverberation times. The performance of a frequency-domain minimum-variance distortionless-response (FMV) beamformer, the Frost adaptive beamformer, the generalized sidelobe canceler (GSC), and an implementation of the Peissig–Kollmeier (P–K) binaural algorithm were evaluated. The FMV and P–K algorithms outperform the time-domain Frost and GSC algorithms in terms of output SNR. A pair of cardioid microphones yields the best performance and lowest target distortion. A pair of omnidirectional microphones performs almost as well, followed by microphones mounted in each of the ear canals of a KEMAR.

The performances of the FMV and P–K algorithms are very similar in terms of the SNR of the output signal, but the P–K algorithm generates significantly more distortion of the target signal. In the tests conducted in this study, filters were used to match approximately the response of the microphones for sounds received from the target direction, but some inherent error is still present in the test signals. This error means that even the distortionless response beamformers could distort the target signal, as is evidenced in the plots of the target distortion. (This issue has been addressed by Elledge [2000].) The FMV generally produces the smallest amount of distortion, while the Frost and P–K algorithms generally produce the largest distortion figures. The impact of the distortion on speech perception and quality has not been determined with human listeners.

Informal observations indicate that the P–K algorithm, as implemented here, produces highly intelligible, though distorted output. This suggests that if distortionless response is a requirement (such as in a high-fidelity hearing aid), then the FMV algorithm is preferred, but if maximal intelligibility is the principal goal, then the P–K algorithm (as implemented here) is also promising. A hybridization of the FMV and P–K algorithms may prove beneficial if it can preserve some of the benefits of each approach.

The two-channel FMV has been implemented in a real-time system [Elledge *et al.*, 2000]. Preliminary tests show that the real-time system helps normal-hearing listeners understand an on-axis talker amidst many interfering sources in controlled environments as compared to conditions with dichotic unprocessed signals and diotic signals (channels summed) [Larsen *et al.*, 2001; Feng, 2002]. Informal tests indicate that the FMV improves intelligibility in rooms that vary widely in size and reverberation characteristics. The output of the real-time system sounds natural and is of high fidelity.

This study uses recordings that were made at a distance of 75 cm from the loudspeaker, where the direct-to-reverberant ratio is relatively high. Even though reverberation was not dominant, the performance of all the algorithms

is compromised to varying degrees, with the FMV and P–K algorithms showing the smallest degradation. Our goal for the present study was to determine the relative effectiveness of these algorithms in rooms having different reverberant characteristics, but not to quantify the effect of the reverberation *per se*. Future work is required to evaluate the performance of the algorithms with microphones farther from the sources where the direct-to-reverberant ratio is lower. Additionally, the 7.6-cm-diameter loudspeakers in the array are quite directional at high frequencies due to the directivity of the driver, thereby reducing the amount of high-frequency reverberation. In the future, loudspeakers with less directional responses should be used in tests that involve sources at greater distances.

ACKNOWLEDGMENTS

This research was supported by a grant from the National Institute on Deafness and Other Communication Disorders (R21DC-04840) of the National Institutes of Health and by grants from the University of Illinois at Urbana-Champaign (the Beckman Institute, the Mary Jane Neer Research Fund, and the Charles M. Goodenberger Fund). We thank Chen Liu and Kong Yang for their earlier work on algorithm simulations, Mark Elledge and Jeffrey Larsen for their participation in the group, and the many staff who generously donated their time to participate in the recording of the speech materials used in our tests.

Berghe, J. V., and Wouters, J. (**1998**). ''An adaptive noise canceller for hearing aids using two nearby microphones,'' J. Acoust. Soc. Am. **103**, 3621–3626.

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., and Rzeczkowski, C. (**1984**). ''Standardization of a test of speech perception in noise,'' J. Speech Hear. Res. **27**, 32–48.

Brandstein, M., and Ward, D. (**2001**). *Microphone Arrays: Signal Processing Techniques and Applications* (Springer, Berlin).

Capon, J. (**1969**). ''High-resolution frequency-wavenumber spectrum analysis,'' Proc. IEEE **57**(8), 1408–1419.

Cox, H., Zeskind, R. M., and Kooij, T. (**1986**). ''Practical supergain,'' IEEE Trans. Acoust., Speech, Signal Process. **ASSP-34**(3), 393–398.

Cox, H., Zeskind, R. M., and Owen, M. M. (**1987**). ''Robust adaptive beamforming,'' IEEE Trans. Acoust., Speech, Signal Process. **ASSP-35**(10), 1365–1376.

Deslodge, J. G. (**1998**). ''The location-estimating null-steering (LENS) algorithm for adaptive microphone-array processing,'' Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Elledge, M. E. (**2000**). ''Real-time implementation of a frequency-domain array processor,'' M.S. thesis, University of Illinois, Urbana, IL.

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,'' presented in 10th Annual DSP Technology Education and Research Conference Houston, TX, 2–4 August.

Feng, A. S. (**2002**). ''Biologically inspired binaural hearing aid algorithms: Design principles and effectiveness,'' J. Acoust. Soc. Am. **111**, 2354.

Frost, O. L. (**1972**). ''An algorithm for linearly constrained adaptive array processing,'' Proc. IEEE **60**(8), 926–935.

Golub, G., and Van Loan, C. (**1996**), *Matrix Computations*, 3rd ed. (The Johns Hopkins University Press Ltd., London).

Greenberg, J. E., and Zurek, P. M. (**1992**). ''Evaluation of an adaptive beamforming method for hearing aids,'' J. Acoust. Soc. Am. **91**, 1662–1676.

Greenberg, J. E. (**1998**). ''Modified LMS algorithms for speech processing with an adaptive noise canceller,'' IEEE Trans. Speech Audio Process. **6**, 338–351.

Greenberg, J. E., Deslodge, J. G., and Zurek, P. M. (**2003**). ''Evaluation of array-processing algorithms for a headband hearing aid,'' J. Acoust. Soc. Am. **113**, 1646–1657.

Griffiths, L. J., and Jim, C. W. (**1982**). ''An alternative approach to linearly constrained adaptive beamforming,'' IEEE Trans. Antennas Propag. **AP-30**(1), 27–34.

Hoffman, M. W., Trine, T. D., Buckley, K. M., and Van Tasell, D. J. (**1994**). ''Robust adaptive microphone array processing for hearing aids: Realistic speech enhancement,'' J. Acoust. Soc. Am. **96**, 759–770.

Joho, M., and Moschytz, G. S. (**2000**). ''Connecting partitioned frequency-domain filters in parallel or in cascade,'' IEEE Trans. Circuits Syst. **47**, 685–698.

Kates, J. M., and Weiss, M. R. (**1996**). ''A comparison of hearing-aid array-processing techniques,'' J. Acoust. Soc. Am. **99**, 3138–3148.

Kollmeier, B. (**1997**). *Psychoacoustics, Speech and Hearing Aids* (World Scientific, Singapore).

Kollmeier, B., Peissig, J., and Hohmann, V. (**1993**). ''Real-time multiband dynamic compression and noise reduction for binaural hearing aids,'' J. Rehabil. Res. Dev. **30**(1), 82–94.

Kompis, M., and Dillier, N. (**1994**). ''Noise reduction for hearing aids: Combining directional microphones with an adaptive beamformer,'' J. Acoust. Soc. Am. **96**, 1910–1913.

Larsen, J. B., Lockwood, M. E., Lansing, C. R., Bilger, R. C., Wheeler, B. C., O'Brien, Jr., W. D., Jones, D. L., 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.

Liu, C., Feng, A. S., Wheeler, B. C., O'Brien, W. D., Bilger, R. C., and Lansing, C. R. (**1997**). ''Speech enhancement via directional hearing based on the place theory,'' 2nd Biennial Hearing Aid Research & Development Conference, Washington, DC, Sept. 1997, 58A (II-P-29).

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.

Liu, C., Wheeler, B. C., O'Brien, Jr., W. D., Lansing, C. R., Bilger, R. C., Jones, D. L., and Feng, A. S. (**2001**). ''A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers,'' J. Acoust. Soc. Am. **110**, 3218–3231.

Lockwood, M. E. (**1999**). ''Development and testing of a frequency-domain minimum-variance algorithm for use in a binaural hearing aid,'' M.S. thesis, University of Illinois, Urbana, IL.

Lockwood, M. E., Jones, D. L., Bilger, R. C., Elledge, M. E., Feng, A. S., Gouegou, M., Lansing, C. R., Liu, C., O'Brien, Jr., W. D., and Wheeler, B. C. (**1999**). ''A minimum-variance frequency domain algorithm for binaural hearing aid processing,'' J. Acoust. Soc. Am. **106**, 2278.

McDonough, R. N. (**1979**). ''Application of the maximum-likelihood method and the maximum-entropy method to array processing,'' in *Nonlinear Methods of Spectral Analysis: Topics in Applied Physics*, edited by S. Haykin (Springer, New York), pp. 181–243.

Peissig, J., and Kollmeier, B. (**1997**). ''Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners,'' J. Acoust. Soc. Am. **101**, 1660–1670.

Slyh, R. E., and Moses, R. L. (**1993**). ''Microphone array speech enhancement in overdetermined signal scenarios,'' ICASSP-93, **2**, 347–350.

Van Veen, B. D., and Buckley, K. M. (**1988**). ''Beamforming: A versatile approach to spatial filtering,'' IEEE ASSP Magazine, April, 4–24.

Welker, D. P., and Greenberg, J. E. (**1997**). ''Microphone-array hearing aids with binaural output II. A two-microphone adaptive system,'' IEEE Trans. Speech Audio Process. **5**(6), 543–551.

Wittkop, T., Albani, S., Hohmann, V., Peissig, J., Woods, W. S., and Kollmeier, B. (**1997**). ''Speech processing for hearing aids: Noise reduction motivated by models of binaural interaction,'' Acustica **83**, 684–699.

Yang, K. L., Lockwood, M. E., Elledge, M. E. and Jones, D. L. (**2000**). ''A comparison of beamforming algorithms for binaural acoustic processing,'' Proc. 9th IEEE Digital Signal Processing Workshop, Hunt, TX.