Effect of Multiple Nonstationary Sources on MVDR Beamformers

Michael E. Lockwood, Douglas L. Jones, Charissa R. Lansing, William D. O'Brien, Jr., Bruce C. Wheeler, Albert S. Feng

Abstract— **MVDR beamformers have been proposed for use in hearing aids. Experiments show poor performance for time-domain methods for more speech sources than sensors, in contrast to the human auditory system. A frequency-domain binaural MVDR algorithm applied in a cocktail-party environment performs significantly better than the Frost or GSC algorithms. Experiments show that the performance advantage of the frequency-domain algorithm can be attributed to rapid, independent adaptation across frequency bands, allowing faster, more accurate tracking of speech sources than time-domain techniques.**

Index Terms—**Adaptive beamforming, MVDR beamforming, nonstationary sources.**

I. Introduction

For signal-processing applications, a major problem is extracting a source in the presence of multiple competing sources. For acoustic processing, the challenge is even greater for speech sounds because they are highly non-stationary in spectrum and intensity, and because the interfering talkers may quickly change position with respect to the listener.

To address this problem, a common approach is to process sampled, time-domain signals from a multi-microphone array with an adaptive beamforming algorithm (see reviews by [1], [2]). Generally, the goal is to extract the target sound from a particular direction with as little distortion as possible while suppressing interference sounds from other directions. Two algorithms that have been used extensively for this application are the iterative-adaptive techniques of Frost [3] and Griffiths and Jim [4]; they generally work well for suppressing statistically stationary interference sources that are uncorrelated with the target source. However, our experience has shown that the performance of these iterative-adaptive algorithms is compromised because they tend to adapt slowly or inaccurately in the presence of multiple, non-stationary interference sources such as speech, especially when there are more sources than sensors.

To improve performance when many sources are present, one solution is to add more microphones [5]. However, this adds to the complexity of both the hardware and the processing algorithm. Accomplishing this in a hearing-aid system is difficult; microphones must be located near the ears in order for the device to be convenient to use and cosmetically acceptable. 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 in the lower to middle frequency range of speech. Thus, the most convenient sensor arrangement is to use one microphone per instrument, located at each ear or in the ear canals, providing greater spatial separation of the microphones (~15 cm).

Several previous studies of adaptive beamformers [5, 6, 7, 8] have avoided including the effects of slow algorithm adaptation by allowing the adaptive filters sufficient time (at least two seconds for these studies) to converge before processing the test signals. However, in a real hearing-aid instrument this pre-adaptation is not possible due to the changing head position of the listener and the (possibly rapid) movement of the interfering sources. The algorithm must adapt quickly to keep up with the changing acoustic scene.

To overcome the two problems of slow adaptation and poor performance when there are more sources than sensors, we implemented a rapidly-adapting, two-channel, frequency-domain MVDR (FMV) algorithm. Its implementation is described in Section I. In Section II, the test environments and signals are described. In Section III, the performance of the FMV algorithm is compared with that of the Frost [3] and GSC [4] algorithms using recordings from real rooms. Examples show differences in the adaptation characteristics of the time- and frequency-domain algorithms. A summary is given in Section IV.

II. Test Procedures

A. Recordings, environment, and test construction

Recordings were made in three different rooms with different reverberation times (RTs: 0.10, 0.37, 0.65 seconds) 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.

Test signals were played from an array of eight loudspeakers that was set up in each room. Each loudspeaker was equidistant (75 cm) from a central point and comprised a single 7.6-cm-diameter driver with frequency response restricted from 200 Hz to ~13 kHz. The two-microphone arrays were located 1.15 m above the floor at the central point of the array, and oriented such that the loudspeakers were at azimuths of 60°, 40°, 20°, 0°, -20°, -40°, -60°, -80° with respect to the broadside array. Therefore, *all sources were located in*

0-7803-8104-1/03/$17.00 ©2003 IEEE

the front half plane.

The test signals were obtained from recordings of real talkers made in a sound-treated room. Three sentences were chosen from four male and four female talkers, for a total of 24 sentences, each approximately 2.5 s in duration. Twelve sentences were used as target signals, and the other twelve for interference. A section of multi-talker babble from the R-SPIN test [9] was also used as an interfering source.

Each sentence and the multi-talker babble was played individually from each loudspeaker (azimuth) and recorded at 22.05 kHz with the microphone arrays. Additionally, a white noise signal 10 s in duration was recorded and used to match the microphones in each array with a 43-tap FIR filter, which was long enough to permit sensor equalization but not long enough to equalize room reflections. Recordings of the individual sources were then summed to form test signals. In any test, all sources were different talkers, and only one source was located at a given azimuth.

Eleven spatial arrangements of sources were chosen, four with one interferer, three with two interferers, and two each with three and four interferers. For each arrangement, twelve test signals were produced, using twelve different target sentences and a different combination of interfering sentences. Each of the twelve tests was produced for three different SNRs. For tests with one interferer, each interferer was scaled to have an average power level of –3, 0, or +3 dB, compared to the level of the target signal. For tests with more than one interferer, each interferer had an average power level of –6, –3, and 0 dB, compared to the level of the target signal. (The average power calculation took into account the differing durations of the spoken sentences.)

It is important to point out that the values by which individual interference sources were scaled were calculated using the recordings made with the omnidirectional microphones. The other microphones were either inherently directional (cardioids) or exhibited directional characteristics due to their placement (on KEMAR). Thus, the test signals created from the recordings for all three arrays include the directional effects of the microphones.

After processing with a given algorithm, the performance-metric values were averaged across the three levels, across the twelve tests, and across all configurations with the same number of interferers. Thus, for each algorithm four values of the performance metric are presented corresponding to one, two, three, or four interferers.

B. Performance metric

Processing with the various algorithms was done off-line and the signals from the individual sources before and after processing were known. A metric was chosen that incorporated both interference and signal bias (distortion) error. The output SNR (after processing) is defined as

$$SNR_{OUT} = 10 \cdot \log_{10}\left(\frac{\sum_{v=1}^{V}\left[g_1 t_{u,L}(v) + g_2 t_{u,R}(v)\right]^2}{\sum_{v=1}^{V}\left(y_p(v) - \left(g_1 t_{u,L}(v) + g_2 t_{u,R}(v)\right)\right)^2} \right) \quad (9)$$

where V is the length of the signal in samples, $y_p(v)$ and $t_u(v)$ are the v^{th} samples of the processed output and unprocessed (ideal) target signals, respectively, $t_{u,L}(v)$ 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.

III. Algorithms

A. FMV algorithm

Time-domain input signals are transformed periodically (every L=16 samples) into the frequency domain via a length-N FFT, using a Hamming window. The frequency-domain signals from the two 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 (every L=16 samples) for each frequency bin k using

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

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. Values for parameters N, M and F that produced the best SNR gain for each array are found in Table I. The correlation matrices and FFT buffers were set to zero before processing a signal.

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

$$Y_k = \mathbf{w}_k^H \mathbf{X}_k \quad (2)$$

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

$$\min_{\mathbf{w}_k} E\left\{ |Y_k|^2 \right\} \quad (3a)$$

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

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 [10]. For the minimization goal and constraint given in Eqs. (3a) and (3b), an optimal solution is known ([10], [11]). For each frequency bin k, the optimal weight vector $\mathbf{w}_{opt,k}$ is given by

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

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.

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. It will be demonstrated later that this technique yields faster and more accurate tracking of non-stationary sources than time-domain techniques.

To obtain the time-domain output, the newest optimal weights are applied to buffered FFT data to obtain the output (Eq. (2)). The result is 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. This minimizes the effects of circular convolution that arise due to the FFT-based filtering.

TABLE I
ALGORITHM PARAMETERS FOR BEST PERFORMANCE

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

B. Frost Algorithm

For the Frost algorithm [3], $\mathbf{x}_f$ is a column vector composed of the time-domain input signals from both channels, $\mathbf{P}$ is a pre-computed projection matrix, and $\mathbf{F}$ represents the response constraints. The update equation is

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

where $\mathbf{W}_{old}$ 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. N_F is defined as the length of the adaptive filter. The parameters chosen for best performance with each array are given in Table I.

C. GSC Algorithm

The implementation of the generalized sidelobe canceller (GSC) algorithm [4] used the modifications proposed in [12], which improves performance and reduces target distortion in non-stationary environments when the target signal is strong. The update equation was

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

where α_{sum} is a step-size parameter, n is an index of the current sample, $\mathbf{W}_{old}$ 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 parameters chosen for best performance with each array are given in Table I.

IV. RESULTS

A. Algorithm Comparisons

Fig. 1 (results for omnidirectional and cardioid microphones, left channel) and Fig. 2 (results for KEMAR microphones, both channels) show the average SNR_{OUT} after processing. The SNR_{IN}, as received by the omnidirectional microphones (SNR, Omni), provides the unprocessed SNR. In all three rooms and for all microphone types, the FMV algorithm consistently outperforms the Frost and GSC algorithms in terms of SNR_{OUT} for all numbers of interferers.

The performances of the Frost and the GSC algorithms are similar. The GSC algorithm [4] was shown to converge to the same solution as the Frost algorithm [3], although the paths to convergence might be different. The GSC appears to have a slight performance advantage for the single-interferer tests, while the Frost appears to perform better as the number of interferers rises. These differences are slightly more pronounced with the KEMAR microphones, and their cause is not known.

The FMV and GSC algorithms generally produce less distortion (Fig. 4) of the target signal than the Frost algorithm. The high distortion figures for the Frost algorithm are not expected, and suggest that the Frost algorithm is most sensitive to the amount of reverberation and the type of microphone being used.

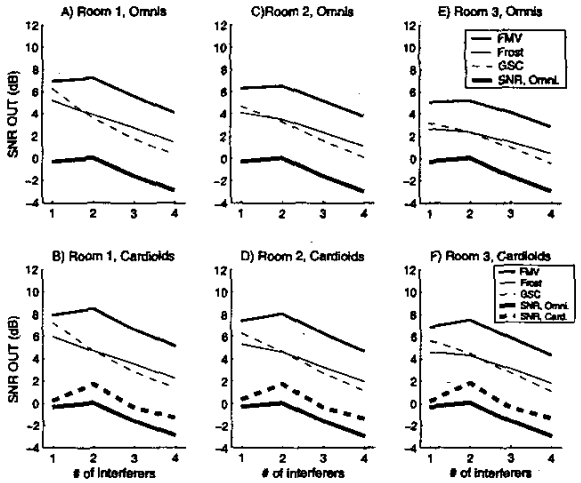

Figure 1. Performance of FMV, Frost and GSC algorithms for 1 to 4 interfering sources for omni and cardioid microphone arrays in three rooms.

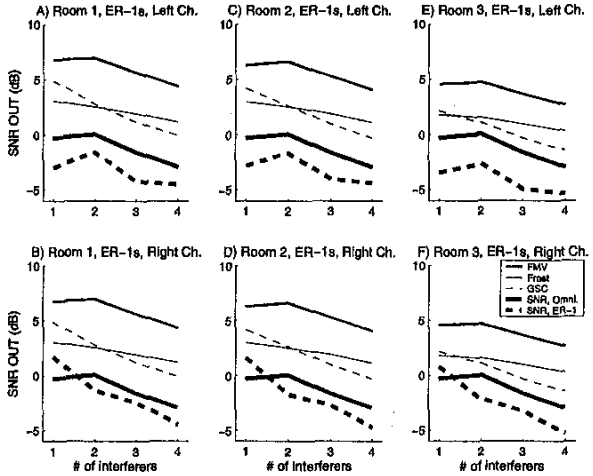

Figure 2. Performance of FMV, Frost and GSC algorithms for 1 to 4 interfering sources for both ears of the KEMAR microphone array in three rooms.

B. Reverberation effects

The performance, in terms of SNR_{OUT} of all algorithms, decreases by about 1 dB for the tests in room 2 (RT = 0.37 s) as compared to room 1 (RT = 0.10 s) for all microphone arrays. Performance differences between room 2 and room 3 (RT = 0.65 s) are slightly more pronounced for the KEMAR microphones.

The performance of all algorithms generally decreased by similar amounts as the reverberation time of the room increased. This implies that the performance advantage of the FMV is maintained as the reverberation time increases. For the one-interferer arrangements, the FMV advantage increases with reverberation time; this implies that the advantages of the FMV are not limited to the situations with more sources than sensors.

C. Microphone effects

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

The KEMAR microphones (Fig. 2) cause a reduction in the SNR_{IN}, except for the right channel for single-interferer tests. (This is because the interfering source for these tests is to the left of the array, and thus the interference is stronger in the left ear of the KEMAR.) For multiple-interferer tests, the SNR_{IN} is lower than with the other microphones. Thus, processing these signals yields 1 – 2 dB lower SNR_{OUT} than with the other microphones.

V. Time- Versus Frequency-Domain MVDR Beamformers

A. Overview

The results from the previous section 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, even if there are more sources than sensors. Therefore, differences in SNR performance are due to different adaptation characteristics in the presence of non-stationary signals such as speech, which negatively impact the ability of the algorithms to cancel interference.

Using *simulated* binaural signals, we now examine the reasons for the performance advantage of the FMV over the other algorithms. Performance differences result only from the varying adaptation characteristics of the algorithms, and not from sensor mismatch, etc. This allows accurate beampatterns to be calculated to observe these characteristics.

B. Simulation example: multiple speech interferers

The test signal used in this example has a target signal (speech) at 0° azimuth, and two interferers (also speech) at +45° and –45° azimuth (simulated purely with time delays between channels). Because all the sources are speech, they are continuously changing, and this example offers the opportunity to examine how the beampatterns of the algorithms change. Examining time and frequency regions in which the sources overlap reveals the adaptation characteristics of the algorithms.

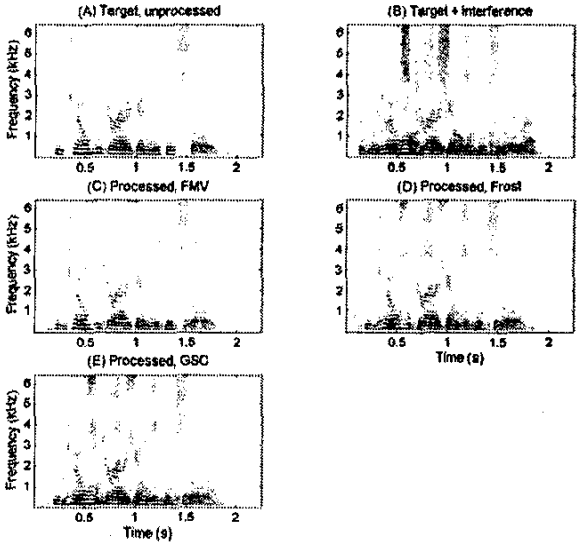

Fig. 3. Spectrograms of (A) target, (B) target+interference, and processed output of FMV (C), Frost (D) and GSC (E) algorithms.

Figs. 3a and 3b show the spectrograms of the target and target plus interference signals, respectively. Figs. 3c, 3d, and 3e show the processed output from the FMV, Frost, and GSC

algorithms, respectively. The SNR gains produced by processing are 9.14 dB (FMV), 5.41 dB (Frost), and 4.30 dB (GSC). Thus, the FMV most effectively reduced the interference. This is supported by the spectrograms; it is evident that the FMV output more closely resembles the target signal in both the high- and low-frequency regions.

We now examine time-varying beampatterns for the FMV, Frost, and GSC algorithms. For a frequency of 800 Hz, these beampatterns (Figs. 4a-c) reveal key differences between the algorithms. The nulls placed by the FMV algorithm (Fig 4a) converge consistently and quickly to ±45°, the location of the interfering sources. In contrast, the Frost algorithm's nulls rarely converge to exact the directions of the interference sources (Fig 4b), and the GSC algorithm's convergence is even less precise (Fig 4c).

Figs. 4d-4f show instantaneous beampatterns 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 ±45°), especially for the -45° source between 2300 and 5000 Hz.

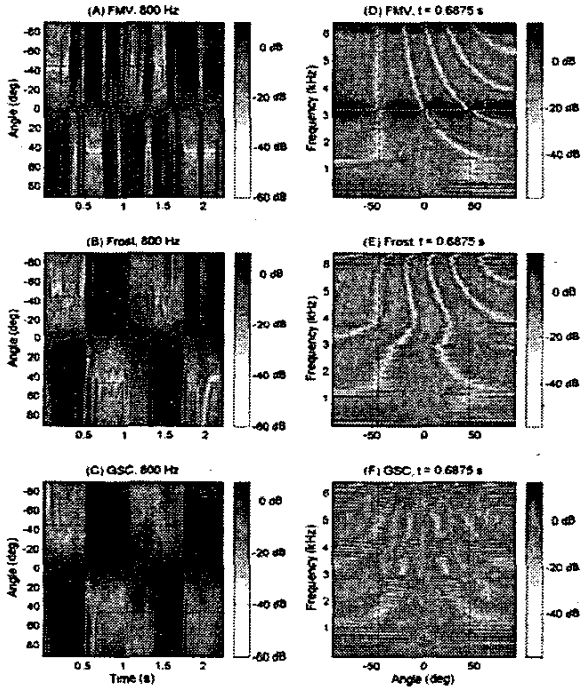

Fig. 4. Beampatterns for the FMV (A), Frost (B), and GSC (C) algorithms for 800 Hz. Beampatterns for FMV (D), Frost (E), and GSC (F) for t = 0.6875 s.

This example illustrates the behavior of the algorithms only for very small portions of the signal. However, these characteristics may be present at any time and frequency, and severely handicap the performance of the time-domain algorithms. The FMV algorithm exhibits faster, more accurate adaptation than the Frost and GSC algorithms. This results in better interference cancellation when the sources are nonstationary, and when the signals include more nonstationary interfering sources than sensors.

VI. Conclusion

To successfully cancel multiple speech interferers using signals from only two sensors, fast adaptation is of primary concern. A frequency-domain MVDR beamformer (FMV) was implemented specifically to exhibit fast adaptation in each frequency band, and this algorithm outperforms the time-domain Frost and GSC algorithms, in terms of SNR gain, for a wide variety of tests with multiple interfering speech sources. The ability of the FMV algorithm to place nulls more precisely and more quickly than the Frost and GSC algorithms is demonstrated via an example using simulated signals.

More information regarding this study can be found in the manuscript [13].

Acknowledgment

This work was supported in part by grants from the National Institute on Deafness and Other Communication Disorders (R21DC-04840) and from DARPA. Additional support was provided by Phonak USA, Inc.

References

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

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

[3] Frost, O. L. (**1972**), "An algorithm for linearly constrained adaptive array processing," Proc. IEEE, **60** (8), 926-935.

[4] Griffiths, L. J., and Jim, C. W. (**1982**), "An alternative approach to linearly constrained adaptive beamforming," IEEE Trans. Antennas and Propagation, AP-**30**(1), 27-34.

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

[6] 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.

[7] 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.

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

[9] 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.

[10] Capon, J. (**1969**), "High-resolution frequency-wavenumber spectrum analysis," Proc. IEEE, **57**(8), 1408-1419.

[11] Cox, H., Zeskind, R. M. and Kooij, T. (**1986**), "Practical supergain," IEEE Trans. on Acoust., Speech and Signal Processing, ASSP-**34**, 3, 393-398.

[12] Greenberg, J. E. (**1998**), "Modified LMS algorithms for speech processing with an adaptive noise canceller," IEEE Trans. on Speech and Audio Proc., (**6**), 338-351.

[13] Lockwood, M. E., Jones, D. L., Bilger, R. C., Lansing, C. R., O'Brien Jr., W. D., Wheeler, B. C., Feng, A. S., (**2003**), "Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms", J. Acoust. Soc. Am., in press.