MICROPHONE ARRAY SPEECH ENHANCEMENT IN OVERDETERMINED SIGNAL SCENARIOS

Raymond E. Slyh[1,2] *Randolph L. Moses*[1]

[1]Department of Electrical Engineering, The Ohio State University, Columbus, Ohio 43210, USA
[2]U. S. Air Force, AL/CFBA, Wright-Patterson AFB, Ohio 45433-6573, USA

ABSTRACT

We consider the problem of enhancing noisy speech by using a microphone array. We present an algorithm, called the Graphic Equalizer (GEQ) array, which trades off signal degradation for additional interference suppression. The algorithm is based upon the concept of directly modifying the short-time spectral magnitude of the sum of the received signals. We include simulation results that illustrate the advantages of using the GEQ array for diffuse-noise scenarios and for scenarios involving more interference signals than array degrees of freedom.

1. INTRODUCTION

In this paper, we examine the problem of enhancing speech by using a microphone-array processing algorithm. A measured speech signal is corrupted by several interference sources, and the goal is to process the microphone array signals to reduce the effect of the interference on the output speech signal. This problem has received attention from several researchers [1, 2, 3] for applications such as hands-free cellular telephony and teleconferencing.

Most methods currently being applied to this problem involve adaptive beamforming techniques such as the Frost array [2, 4, 5] or the closely related Generalized Sidelobe Canceller [1, 3, 6]. In these methods, the microphone array pattern is adaptively adjusted in an attempt to maximize the output SNR subject to array constraints; typically the array constraints are set so as to maintain a desired frequency response in the nominal desired signal direction. These methods have been shown to provide good speech enhancement in the so-called "underdetermined" case, when the number of interference signals is less than or equal to the array degrees of freedom; the number of degrees of freedom in an array is equal to one less than the number of microphones [5]. We note that speech quality is not well correlated with SNR [7], but the SNR gain is quite large in most underdetermined cases, and so speech quality also improves. However, the performance of Frost-type algorithms degrades if the interference signals are diffuse, or if there are more interference signals than array degrees of freedom—a situation that we refer to as "overdetermined". This is due in part because the array cannot successfully cancel all of the interference, and in part because the SNR improvement is not well correlated with speech quality improvement for small SNR gains.

One way to improve the performance of adaptive beamforming algorithms for overdetermined signal scenarios and for diffuse-noise scenarios is to increase the number of microphones in the array; however, this increases the cost of building the array and increases the amount of computation necessary to form the output. Furthermore, it is not always known how many interference signals will be present in a particular environment.

In overdetermined or diffuse-noise signal scenarios in which it is not feasible to use a large number of microphones, one can turn to alternate array signal processing algorithms. In this paper we present one such two-microphone algorithm, which we refer to as the Graphic Equalizer (GEQ) array. The GEQ array is based upon the idea of directly modifying the short-time spectral magnitude of the received speech. This idea has been applied in single-microphone enhancement schemes such as noise spectral subtraction [8]. Allen, Berkley, and Blauert applied a technique closely related to the GEQ array for the purpose of dereverberation [9]. The GEQ array is based upon improving the power function spectral distance measure (PFSDM) instead of SNR; the PFSDM has been shown to be more correlated with human auditory perception than is the SNR [7].

An outline of the paper is as follows. The Frost array is briefly outlined in the next section, while the GEQ array is presented in the third section. Simulation results showing the utility of the GEQ array are presented in the fourth section, and the conclusions are presented in the fifth section.

2. THE FROST ARRAY

Although the Frost array has been developed for an arbitrary number of sensors, we will consider only a two-element Frost array [4, 5]. Each of the microphone elements is followed by a time delay and a finite impulse response (FIR) filter consisting of J adjustable filter weights; we denote the jth impulse response coefficient of the ith sensor by w_{2j+i} for $i \in \{1,2\}$ and $j \in \{0,\ldots,J-1\}$. The time delays are adjusted so that the desired signal is aligned in time in each filter. Finally, the outputs of the FIR filters are summed to form the array output.

For the desired signal, the time-alignment procedure causes the array to appear as an FIR filter. By constraining the weights as

$$w_{2j+1} + w_{2j+2} = f_j,$$

for f_j fixed and $j \in \{0,\ldots,J-1\}$, we constrain the frequency response of the array in the look direction (i.e. the direction of arrival of the desired signal). We can write these constraint equations in matrix form as

$$\mathbf{C}^T\mathbf{w} = \mathbf{f}, \tag{1}$$

where $\mathbf{w}$ is the vector of array weights, $\mathbf{f}$ is the vector containing the f_j's, $\mathbf{C}$ is the constraint matrix, and ^{T} denotes the transpose operator.

The development of the Frost array proceeds by minimizing the expected value of the output power of the array over the weights subject to the given constraints in the look direction [4]. The array output, $y(k)$, can be written as

$$y(k) = \mathbf{w}^T\mathbf{x}(k) = \mathbf{x}^T(k)\mathbf{w},$$

where $\mathbf{x}(k)$ is the vector of the signals at the inputs to the weights. The Frost array beamforming problem is then given as

0-7803-0946-4/93 $3.00 © 1993 IEEE

Problem 1 *Minimize* $\mathbf{w}^T\mathbf{R_{xx}}\mathbf{w}$ *subject to* $\mathbf{C}^T\mathbf{w} = \mathbf{f}$,

where $\mathbf{R_{xx}}$ is the correlation matrix of $\mathbf{x}$. The optimal weights are given by

$$\mathbf{w}_{opt} = \mathbf{R_{xx}^{-1}}\mathbf{C}\left[\mathbf{C}^T\mathbf{R_{xx}^{-1}}\mathbf{C}\right]^{-1}\mathbf{f}. \qquad (2)$$

Equivalently, the optimal weights can be expressed as

$$\mathbf{w}_{opt} = \mathbf{R_{nn}^{-1}}\mathbf{C}\left[\mathbf{C}^T\mathbf{R_{nn}^{-1}}\mathbf{C}\right]^{-1}\mathbf{f}, \qquad (3)$$

where $\mathbf{R_{nn}}$ is the correlation matrix of the interference and noise.

It is assumed in Frost's derivation [4] that the correlation matrix $\mathbf{R_{xx}}$ is unknown *a priori*. For this reason, an adaptive algorithm is proposed. If we define $\mathbf{g}$ and $\mathbf{P}$ as

$$\begin{aligned} \mathbf{g} &\triangleq \mathbf{C}\left(\mathbf{C}^T\mathbf{C}\right)^{-1}\mathbf{f} \\ \mathbf{P} &\triangleq \mathbf{I}-\mathbf{C}\left(\mathbf{C}^T\mathbf{C}\right)^{-1}\mathbf{C}^T, \end{aligned}$$

then an adaptive algorithm that converges to the weights given in Equation 2 can be written as

$$\begin{aligned} \mathbf{w}(0) &= \mathbf{g} \\ \mathbf{w}(k+1) &= \mathbf{P}\left[\mathbf{w}(k) - \mu y(k)\mathbf{x}(k)\right] + \mathbf{g}, \end{aligned}$$

where μ is a constant that controls the adaptation rate.

It is easy to show that the Frost array problem stated in Problem 1 is equivalent to maximizing the output SNR subject to the hard constraint on the weights given by Equation 1.

It is well known, however, that the SNR is not a very good objective speech quality measure [7]. In fact, in studies examining the ability of the SNR to predict diagnostic acceptablity measure (DAM) scores, the SNR had a correlation coefficient no better than 0.31 [7]. There are several measures, such as the power function spectral distance measure (PFSDM) outlined in [7], that correlate better with human auditory perception than does the SNR. The PFSDM has been shown to have a correlation coefficient of 0.71 with DAM scores.

For signal scenarios in which there are enough degrees of freedom to theoretically null the interference signals independently, the Frost array concept generally works well. The SNR gain for these scenarios is large, and the quality of the speech is improved. For overdetermined and diffuse-noise signal scenarios, there are not enough degrees of freedom in the array to null each interference signal and/or noise signal independently. For these scenarios, the Frost array chooses weights which trade off attenuation of one signal for attenuation of another signal based upon how this trade-off affects the output SNR. We have already noted that the SNR is not a good measure of speech quality; thus, the Frost array may not yield perceptually optimal weights.

The performance degradation for overdetermined and diffuse-noise signal scenarios is compounded by the fact that the Frost array places a hard constraint on the frequency response of the array in the look direction (see Problem 1). This means that the Frost array does not allow any degradation in the frequency response of the array for the desired signal, even though such degradation could produce better interference suppression.

In the next section, we present an alternate array processing algorithm called the GEQ array. The GEQ array does not place a hard constraint on the array weights, nor does it attempt to maximize the SNR.

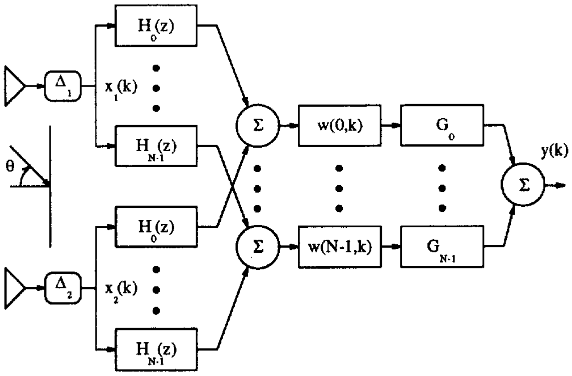

Figure 1. Block diagram of the GEQ array.

3. THE GEQ ARRAY

In this section, we present the Graphic Equalizer (GEQ) array, which is based upon the concept of the graphic equalizer common to many stereo systems. A GEQ is a bank of bandpass filters in which each filter has a separate adjustable gain. The gain of each channel can be adjusted independently on a short-time basis so as to trade off signal degradation and noise suppression; this eliminates the hard constraint of the Frost array (see Equation 1) and improves the power function spectral distance measure (PFSDM).

A block diagram of the GEQ algorithm is shown in Figure 1. The received signals are first aligned in time using the steering delays Δ_1 and Δ_2. Each time-aligned signal is then filtered through a bank of bandpass filters by using a short-time discrete cosine transform (STDCT) [10]. The filters are denoted by the $H_i(z)$'s in Figure 1. The corresponding channel signals of the two sensors are added and then weighted by a gain which is a function of both channel number and time. Finally, the weighted sums are inverse-transformed by the G_i coefficients to form the output signal.

We compute the STDCT using IIR filters, since we use a rectangular sliding window on the data [10]. We use a window centered about the current data point. The advantage of calculating the STDCT in this fashion is that the computations are of $O(N)$ in complexity, where N is the length of the sliding window. In addition, the STDCT yields real channel signals as opposed to the complex signals that would arise from the use of a short-time discrete Fourier transform; this lowers the amount of computation required to calculate the channel gains. As a result, the computational complexity of the GEQ array is only twice that of the Frost array, provided that the length of the sliding window, N, used in the GEQ array is equal to the number of filter taps, J, used in the Frost array.

Let $x_i(k)$ be the received signal at sensor i for time index k, and let $\tilde{x}_i(n,k)$ be the output at time index k of channel n of the STDCT. The STDCT outputs are given by

$$\begin{aligned} \tilde{x}_i(0,k) &= \tilde{x}_i(0,k-1) + x_i\left(k+N-\left\lfloor\frac{N-1}{2}\right\rfloor-1\right) \\ &\quad -x_i\left(k-\left\lfloor\frac{N-1}{2}\right\rfloor-1\right) \\ \tilde{x}_i(n,k) &= 2\cos\left(\frac{\pi n}{N}\right)\tilde{x}_i(n,k-1) - \tilde{x}_i(n,k-2) \\ &\quad +x_i\left(k-\left\lfloor\frac{N-1}{2}\right\rfloor-2\right) \\ &\quad -x_i\left(k-\left\lfloor\frac{N-1}{2}\right\rfloor-1\right) \end{aligned}$$

$$+(-1)^n x_i\left(k+N-\left\lfloor\frac{N-1}{2}\right\rfloor-1\right)$$
$$-(-1)^n x_i\left(k+N-\left\lfloor\frac{N-1}{2}\right\rfloor-2\right),$$

for $n \in \{1,2,\ldots,N-1\}$, where $\lfloor\cdot\rfloor$ denotes the "floor" operator.

The output, $y(k)$, is formed from the inverse STDCT of the weighted channel sums; we use the filter bank summation method for computing the inverse STDCT. Let $w(n,k)$ be the weight for channel n at time index k, then

$$\begin{aligned} y(k) &= \frac{1}{N}\tilde{y}(0,k) \\ &+ \sum_{n=1}^{N-1} \frac{2}{N}\cos\left(\frac{\pi n}{2N}\right)\cos\left(\alpha(n)\right)\tilde{y}(n,k), \end{aligned}$$

where

$$\begin{aligned} \alpha(n) &= \frac{\left(2\left\lfloor\frac{N-1}{2}\right\rfloor+1\right)\pi n}{2N}, \\ \tilde{y}(n,k) &= w(n,k)\left[\tilde{x}_1(n,k)+\tilde{x}_2(n,k)\right]. \end{aligned}$$

The main task is to choose a gain function, $w(n,k)$, where n is the channel index and k is the time index. Since the PFSDM is more highly correlated with human auditory perception than is the SNR, we want to choose a gain that will improve the PFSDM of the output of the array.

The PFSDM is a time-frequency spectral magnitude measure based upon critical band filtering [7]. Let $s_d(k)$ be a desired speech signal, and let $s_p(k)$ be a processed speech signal. The PFSDM is given as

$$d_{PF}(s_d,s_p) = \frac{1}{LM}\sum_{l=0}^{L-1}\sum_{m=0}^{M-1}\left|S_d^{0.2}(m,l)-S_p^{0.2}(m,l)\right|, \quad (4)$$

where L is the length of $s_d(k)$ in time frames of 10 ms duration, M is the number of critical bands in the filter bank, $S_d(m,l)$ is the RMS value of the output of the mth filter over the lth time frame given the desired signal as the filter input, and $S_p(m,l)$ is the RMS value of the output of the mth filter over the lth time frame given the processed signal as the filter input.

Even though we use the STDCT in the GEQ array for computational efficiency as opposed to using critical band filtering, the definition of the PFSDM in Equation 4 yields insight into some of the qualities that the gain function should possess. We see that the gain function should pass relatively unattenuated those time-frequency bins which consist of mostly the desired signal, since the RMS magnitude of these bins should be close to the desired RMS magnitude. The gain function should also greatly attenuate those time-frequency bins which consist of very little signal and a large amount of interference. Although it's not immediately clear what properties the gain should possess between these two extremes, we can choose a family of gain functions, and determine the best parameter setting experimentally.

We propose the following family of gain functions:

$$w(n,k) = \left(\frac{|\phi_{12}(n,k)|}{\sqrt{\phi_{11}(n,k)\phi_{22}(n,k)}}\right)^{b(n)}$$

where $b(n)$ is an adjustable channel-dependent exponent, $\phi_{12}(n,k)$ is the short-time cross correlation (at lag zero) between channel n of sensor 1 and channel n of sensor 2, and $\phi_{ii}(n,k)$ is the short-time autocorrelation (at lag zero) for channel n of sensor i. Thus, $\phi_{ii}(n,k)$ is the short-time energy of the total signal (desired and interference) received at sensor i in frequency band n, and $\phi_{12}(n,k)$ is an estimate of the short-time energy of the desired signal in frequency band n, since the signals were aligned in time. We use a rectangular sliding window in the calculation of $\phi_{ij}(n,k)$; thus,

$$\phi_{ij}(n,k) = \frac{1}{2L+1}\sum_{l=-L}^{L}\tilde{x}_i(n,k+l)\tilde{x}_j(n,k+l),$$

where L controls the window length. In order to acheive enough interference suppression at low frequencies, we use $b(n) = B/f_n$, where B is an adjustable parameter, and f_n is the center frequency of channel n in Hertz. Through simulations, we have found that the setting of $B = 3.5\times10^5$, $L = 10$, and $N = 512$ works well for signals sampled at 16 kHz.

4. EXAMPLES

For the first example, consider a scenario which consists of a desired signal and white noise received at each of two sensors. The desired signal has an arrival angle, θ, of 0° (see Figure 1 for the definition of θ). For the desired speech signal, we use the TIMIT database sentence "Don't ask me to carry an oily rag like that." spoken by a male. The noise is such that the SNR of the received signal at the first sensor is 1.7 dB. We use a sensor spacing of 2 cm.

It is easy to show using Equation 3 that the Frost array reduces to a delay-and-sum beamformer (DSBF) for this scenario. A DSBF is an array which time shifts each received sensor signal and then adds them so that the desired signal components add coherently. For this scenario, the Frost array should theoretically yield only a 3 dB improvement in the SNR of the output over the SNR of one of the sensor signals. We are interested in how the Frost array performs in terms of a speech quality measure such as the PFSDM. In addition, we want to compare the Frost array results with those of the GEQ array.

The results of the experiment are outlined in Table 1. As expected, the Frost array improves the SNR by 3 dB. The GEQ array only improves the SNR by 0.8 dB. The results for the PFSDM, however, are quite the opposite. The Frost array only shows an improvement of 12.80% in the PFSDM, while the GEQ array shows an improvement of 38.18% in the PFSDM. Informal listening tests indicate that the GEQ array suppresses the noise more than does the Frost array; however, the GEQ array yields an output speech signal with a more synthetic quality as compared to the output of the Frost array.

For the second example, we consider the following set of three-source cases. The desired signal is the same as in the previous example. The first interference signal is the TIMIT database sentence "She had your dark suit in greasy wash water all year." spoken by a female. The second interference signal is the TIMIT database sentence "Growing well-kept gardens is very time-consuming." spoken by a male. The arrival angle of the second interference signal is fixed at -40°, while the arrival angle of the first interference signal,

Table 1. Results of White-Noise Experiment

Processing	SNR (in dB)	PFSDM	Improvement in PFSDM
Unprocessed	1.700	0.922	—
Frost Array	4.730	0.804	12.80%
GEQ Array	2.553	0.570	38.18%

θ_1, is stepped from -90° to 90° in 10° increments. The SNR of the received signal at the first sensor is -6.19 dB, while the PFSDM is 0.707.

Now, using the case with $\theta_1 = 10^\circ$, we tune the parameters of the Frost array in order to achieve the best performance in terms of the PFSDM; the resulting best parameter settings are $\mu = 2.0 \times 10^{-8}$ and $J = 64$. The f_i's are set to zeros, except for f_{31} which is set to 1; this results in an all-pass frequency response for the desired signal. We tune the parameters of the GEQ array in the same manner, obtaining $B = 3.5 \times 10^5$, $L = 10$, and $N = 512$ as the best parameter setting. We hold the parameter settings of both arrays constant for the other signal scenarios involving different values of θ_1.

The improvement in SNR for these scenarios is shown in Figure 2, while the improvement in the PFSDM for these scenarios is shown in Figure 3. We note that, for the $\theta_1 = 0^\circ$ case, the first interference source appears to be part of the desired signal for both algorithms; thus, any gain in SNR or PFSDM by either algorithm is due solely to suppression of the second interference signal. We also note that, for the $\theta_1 = -40^\circ$ case, both interference signals arrive from the same direction; thus, both algorithms act as if there is only one interference signal. We expect the Frost array to do well in this case, since this scenario is an underdetermined signal scenario. For values of θ_1 close to -40°, the Frost array places a broad null in the direction of the two interference sources; however, the broader the null, the less deep it is. This is why the Frost array performance peaks at $\theta_1 = -40^\circ$. The Frost array yields more improvement in the PFSDM than the GEQ array over the range $\theta_1 = -65^\circ$ to -30° due to the broad nulling behavior of the Frost array in this region; the GEQ array yields better PFSDM performance for all other values of θ_1. It is apparent from Figure 2 that the Frost array does not always yield a better SNR than the GEQ array yields, even though the Frost array attempts to maximize SNR (see Problem 3); the reason for this is that the Frost array is limited in its performance by the hard constraint on the weights.

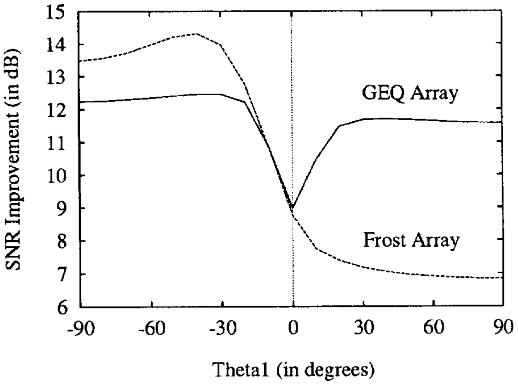

Figure 2. SNR improvement of the Frost and GEQ arrays for the three source case versus θ_1.

5. CONCLUSIONS

We have presented a speech enhancement algorithm that we call the Graphic Equalizer (GEQ) array. The GEQ array works by directly modifying the short-time spectral magnitude of the sum of the received sensor signals. The resulting algorithm acts like a delay-and-sum beamformer followed by a graphic equalizer; the graphic equalizer works to trade off signal degradation for additional noise suppression. The GEQ array was shown to outperform the Frost array for diffuse-noise and overdetermined signal scenarios in which the Frost array was not capable of attenuating all of the interference with one broad null. While the GEQ array did not always yield a better SNR than did the Frost array for these scenarios, it often outperformed the Frost array in terms of the power function spectral distance measure, a measure that is more highly correlated with human auditory perception than is the SNR.

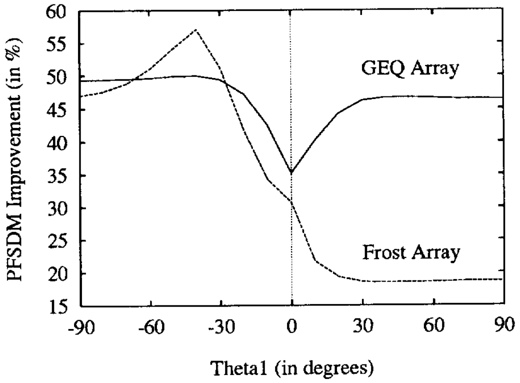

Figure 3. PFSDM improvement of the Frost and GEQ arrays for the three source case versus θ_1.

REFERENCES

[1] S. Oh, V. Viswanathan, and P. Papamichalis, "Hands-free voice communication in an automobile with a microphone array," in *Proc. IEEE ICASSP*, pp. 281–284, 1992.

[2] K. Farrell, R. J. Mammone, and J. L. Flanagan, "Beamforming microphone arrays for speech enhancement," in *Proc. IEEE ICASSP*, pp. 285–288, 1992.

[3] Y. Grenier, "A microphone array for car environments," in *Proc. IEEE ICASSP*, pp. 305–308, 1992.

[4] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, pp. 926–935, Aug. 1972.

[5] R. T. Compton, Jr., *Adaptive Antennas: Concepts and Performance.* Englewood Cliffs, NJ: Prentice-Hall, 1988.

[6] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained beamforming," *IEEE Trans. Ant. and Prop.*, vol. 30, pp. 27–34, Jan. 1982.

[7] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective Measures of Speech Quality.* Englewood Cliffs, NJ: Prentice-Hall, 1988.

[8] J. S. Lim, ed., *Speech Enhancement.* Englewood Cliffs, NJ: Prentice-Hall, 1983.

[9] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acous. Soc. Amer.*, vol. 62, pp. 912–915, Oct. 1977.

[10] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comp.*, vol. 23, pp. 90–93, Jan. 1974.