Maximum Likelihood Noise Cancellation Using the EM Algorithm

MEIR FEDER, MEMBER, IEEE, ALAN V. OPPENHEIM, FELLOW, IEEE, AND EHUD WEINSTEIN, SENIOR MEMBER, IEEE

Abstract—**Single-microphone speech enhancement systems have typically shown limited performance. Two-microphone systems based on a least-squares error criterion have shown better results in some contexts; however, sometimes the desired (speech) signal is cancelled together with the noise. In this paper we suggest a new approach to the two-microphone speech enhancement problem. Specifically, we formulate a maximum likelihood (ML) problem for estimating the parameters needed for cancelling the noise, and then, we solve this ML problem via the iterative EM (Estimate-Maximize) technique. The resulting procedure shows encouraging results that improve upon the "classical" least-squares approach.**

I. Introduction

THE problem of noise cancellation in single- and multiple-microphone environments has been extensively studied [1]. The performance of the various techniques in the single-microphone case seems to be limited. However, with two or more microphones, the performance of an enhancement system may be improved due to the availability of reference signals.

In this paper, noise cancellation based on a two sensor scenario, as indicated in Fig. 1, is considered. One sensor (the primary microphone) measures a signal that consists of speech embedded in noise. The second sensor (the reference microphone), located away from the speaker, measures a signal that consists mainly of the noise. The signal measured in the reference microphone is used for cancelling the noise in the primary microphone. A reasonably general model for this scenario is shown in Fig. 2.

The most widely used approach to noise cancellation, based on two microphones, was suggested by Widrow *et al.* [2]. In this approach, it is assumed that the system B is zero and that C and D are identity, so that the output of the reference microphone is due only to the noise, and that the noise component in the primary microphone is the output of an unknown linear system with transfer function

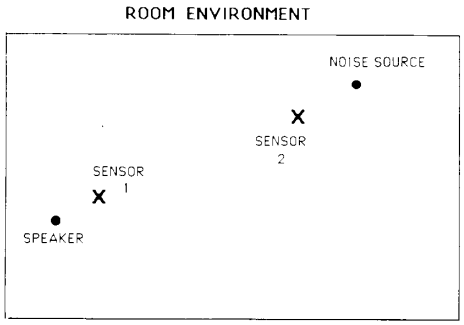


Fig. 1. The acoustic environment.

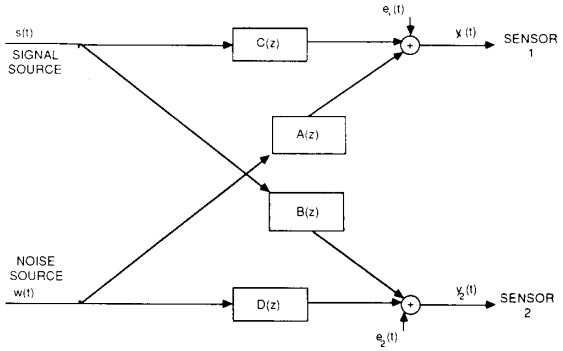


Fig. 2. The noise cancellation problem.

$A(z)$ whose input is the signal measured in the reference microphone. The coefficients of the impulse response of this system are estimated by a least-squares fitting of the reference microphone signal to the primary microphone signal. This method will be referred to in this paper as the least-squares method.

Widrow *et al.* proposed an adaptive solution for this least-squares problem, based on the LMS (Least Mean Square) algorithm. This approach is illustrated in Fig. 3, and has been applied in a speech enhancement context, e.g., [3] and [4]. Adaptive algorithms for solving the linear least-squares problem, based on the RLS (Recursive Least-Squares) algorithm, also exist, e.g., [5] and [6].

A major limitation of the least-squares method, especially when the reference signal is correlated with the desired (speech) signal, is that a portion of the desired signal may be cancelled together with the noise. Since the de-

Manuscript received June 9, 1987; revised May 10, 1988. This work was supported in part by the Advanced Research Projects Agency monitored by ONR under Contract N00014-81-K-0742 and the National Science Foundation under Grant ECS-8407285 at M.I.T., and in part by ONR under Contract N00014-85-K-0272 at WHOI. M. Feder acknowledges the support of Woods Hole Oceanographic Institution.

M. Feder and A. V. Oppenheim are with the Department of Electrical Engineering and Computer Science, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139.

E. Weinstein is with the Department of Electronic Systems, Tel-Aviv University, Tel-Aviv, Israel, and with the Department of Ocean Engineering, Woods Hole Oceanographic Institution, Woods Hole, MA 02543.

IEEE Log Number 8825131.

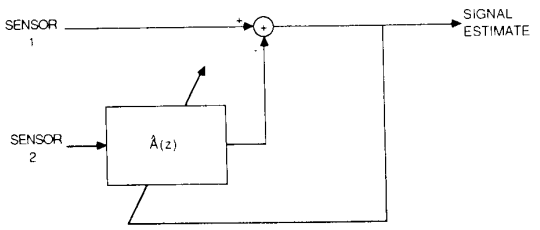


Fig. 3. "Classical" noise cancelling scheme.

sired signal may be cancelled with some delay, the resulting effect is to introduce a reverberant distortion in the output.

In our approach we formulate the problem as a statistical maximum likelihood problem which, as we will show, allows us to take advantage of more *a priori* information than the least-squares formulation. Solving this ML problem directly is difficult, and thus it is solved via a general iterative algorithm for ML that has been introduced by Dempster *et al.* [7] and is referred to as the Estimate-Maximize (EM) algorithm. In the EM algorithm, the observations are considered "incomplete" with respect to more convenient "complete data" measurements. The algorithm iterates between estimating the sufficient statistics of the "complete data" given the observations and a current estimate of the parameters (the E step), and maximizing the likelihood of the complete data using the estimated sufficient statistics (the M step).

It is interesting to note that an algorithm essentially similar to the EM algorithm has been suggested previously in a speech enhancement context. One of the variety of methods that have been suggested for a single microphone case is the iterative enhancement method proposed by Lim and Oppenheim [8]. Although not developed from this point of view, this method can be shown to be an instance of the EM algorithm. In [8] the observations are the desired signal with additive noise and the "complete data" are the signal and noise separately. The unknown parameters are some spectral parameters of the signal (LPC parameters, for speech). The algorithm iterates between Wiener filtering applied to the observations using the current spectral parameters of the signal (the E step), and updating the spectral parameters using the results of the Wiener filter (the M step). In this respect, the procedures presented in this paper may be considered as extensions of the method in [8] to two microphones.

The methods that will be presented in this paper can be used as an alternative to the least-squares method of [2] and its derivatives, e.g., [9] and [10]. Simulation results indicate that the proposed schemes tend to eliminate the reverberant distortion encountered in the least-squares method. We finally note that the proposed scheme can easily be extended to the more general, multiple microphone case.

The paper is organized as follows. In Section II we develop the maximum likelihood formulation of the noise cancellation problem, and in Section III we describe the EM algorithm for obtaining its solution. The application of this iterative algorithm is described in Section IV, for a simplified scenario that basically makes the same assumption as in [2]. In Section V, the algorithm is described for a more general scenario that includes "cross talk," i.e., the coupling of the desired signal into the reference microphone. Simulation results, including results that use a simulated realistic room impulse response, are discussed in Section VI.

II. Maximum Likelihood Formulation of the Two-Sensor Noise Cancellation Problem

The mathematical ML problem encountered in a two-microphone noise cancellation problem is based on the following scenario. A desired (speech) signal source and a noise source both exist in some acoustic environment, say a living room or an office. We have two microphones used in such a way that one microphone is intended to measure mainly the speech source while the other is intended to measure mainly the noise source.

The desired signal and the noise are both coupled into each microphone by the acoustic field in this environment. This situation is illustrated in Fig. 2, and is represented by[1]

$$y_1(t) = C\{s(t)\} + A\{w(t)\} + e_1(t) \quad (1a)$$

$$y_2(t) = B\{s(t)\} + D\{w(t)\} + e_2(t) \quad (1b)$$

where $s(t)$ denotes the desired (speech) signal and $w(t)$ denotes the noise source signal. The systems A, B, C, and D are assumed to be linear systems, representing the acoustic transfer functions between the sources and the microphones. We will assume that these systems are time invariant in our analysis window. The additional noise sources $e_1(t)$ and $e_2(t)$ are included to represent modeling errors, microphone and measurements noise, etc.

Under these assumptions, the observed signals in the analysis window $0 \le t \le N-1$ may be written in the frequency domain as

$$Y_1(\omega) = C(\omega)\,S(\omega) + A(\omega)\,W(\omega) + E_1(\omega) \quad (2a)$$

$$Y_2(\omega) = B(\omega)\,S(\omega) + D(\omega)\,W(\omega) + E_2(\omega) \quad (2b)$$

where $Y_1(\omega)$ and $Y_2(\omega)$ are the Fourier transforms of the the observed signals $y_1(t)$ and $y_2(t)$. Assuming that $y_i(t)$ is a discrete-time signal and that the length of the analysis window is N samples, $Y_i(\omega)$ is given by

$$Y_i(\omega) = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} y_i(t) e^{-j\omega t}. \quad (3)$$

$A(\omega)$, $B(\omega)$, $C(\omega)$, and $D(\omega)$ are the frequency responses of the four linear systems in Fig. 2.

In the more general case of M microphones and K noise sources, the observed signal may be written (in the frequency domain) as

$$\boldsymbol{Y}(\omega) = \boldsymbol{A}(\omega)\,S(\omega) + \boldsymbol{B}(\omega)\,\boldsymbol{W}(\omega) + \boldsymbol{E}(\omega) \quad (4)$$

where $\boldsymbol{Y}(\omega)$, $\boldsymbol{A}(\omega)$, and $\boldsymbol{E}(\omega)$ are $1 \times M$ vectors, $\boldsymbol{W}(\omega)$ is $1 \times K$ vector, and $\boldsymbol{B}(\omega)$ is $K \times M$ matrix.

[1]The mathematics and the algorithms will be formulated in terms of discrete time signals with the independent variable t representing normalized sample time.

To formulate a statistical maximum likelihood problem, we make the following assumptions. The noise source $w(t)$ is assumed to be a sample from a Gaussian random process. The desired speech signal $s(t)$, in many cases, is modeled as an AR Gaussian random process whose parameters (the LPC parameters) are time varying. For our purposes, in a short analysis window, we assume that those parameters are constant, and thus in the mathematical formulation, the desired signal is also assumed to be a sample from a stationary AR Gaussian process. The error signals $e_1(t)$, $e_2(t)$ are modeled as white Gaussian noise processes. The signals $s(t)$, $w(t)$, $e_1(t)$, and $e_2(t)$ are jointly Gaussian and assumed to be uncorrelated.

The unknown parameters are the coefficients of the various systems, and some spectral parameters of the signals. We denote the power spectra of $s(t)$ and $w(t)$ by $P_s(\omega)$ and $P_w(\omega)$, respectively. $\sigma_{e_1}^2$, $\sigma_{e_2}^2$ will denote the error signals variances.

We formulate the problem in terms of short-time processing so that the signals and the system parameters can be slowly time varying; consequently, a sliding window is applied. As already noted, the window length N must be short enough so that the parameters are constant over its duration. However, we will also assume that it is long enough so that the short-time DFT coefficients of $s(t)$, $w(t)$, $e_1(t)$, and $e_2(t)$ at different frequencies are uncorrelated. Under this assumption, the likelihood of the observations ($y_1(t)$ and $y_2(t)$, $t = 0, \cdots, N-1$) with respect to the parameters above is easily expressed in the frequency domain, and is written as[2]

$$\log f_{y_1 y_2}(y_1(t), y_2(t); \boldsymbol{\theta}) = -\sum_{\omega_l} \left(\log \det \Lambda(\omega_l; \boldsymbol{\theta}) + \boldsymbol{Y}(\omega_l)^{\dagger} \Lambda^{-1}(\omega_l; \boldsymbol{\theta})\, \boldsymbol{Y}(\omega_l)\right) \qquad \omega_l = \frac{2\pi}{N} \cdot l \tag{5}$$

where $\boldsymbol{Y}(\omega)$ is a vector whose components are $Y_1(\omega)$ and $Y_2(\omega)$. The matrix $\Lambda(\omega; \boldsymbol{\theta})$ is the power spectrum matrix, i.e.,

$$\Lambda(\omega; \boldsymbol{\theta}) = E\{\boldsymbol{Y}(\omega)\, \boldsymbol{Y}(\omega)^{\dagger}\} = \begin{bmatrix} C(\omega) P_s(\omega) C^*(\omega) + A(\omega) P_w(\omega) A^*(\omega) + \sigma_{e_1}^2 & C(\omega) P_s(\omega) B^*(\omega) + A(\omega) P_w(\omega) D^*(\omega) \\ B(\omega) P_s(\omega) C^*(\omega) + D(\omega) P_w(\omega) A^*(\omega) & B(\omega) P_s(\omega) B^*(\omega) + D(\omega) P_w(\omega) D^*(\omega) + \sigma_{e_2}^2 \end{bmatrix}. \tag{6}$$

We note that this technique, for representing the likelihood of a stationary Gaussian process with long observation time in the frequency domain, is widely used in many signal processing applications, see, e.g., [11, ch. 4]. It is also justified in [12, ch. 13] and elsewhere.

For the M microphone case, the likelihood function is again (5) where the matrix Λ is now the $M \times M$ power spectrum matrix $E\{\boldsymbol{Y}(\omega)\, \boldsymbol{Y}(\omega)^{\dagger}\}$.

The general maximum likelihood problem, represented by (5) and (6), is not only complicated but it may also be ill posed. The likelihood function depends on the parameters only through the matrix $\Lambda(\omega)$, and all possible solutions that generate the same $\Lambda(\omega)$ have the same likelihood. If indeed all the associated systems and the power spectra are unknown and their structure is allowed to be arbitrary, we expect a nonunique solution, since every value of $\Lambda(\omega)$ may correspond to a set of values for the parameters. Thus, some constraints must be imposed on the parameters. For example, we may assume that some of the parameters are known, or that there is a simple structure to the systems. Of course, the more constraints there are, the more this ML problem becomes well posed mathematically. However, we must limit ourselves to constraints that are consistent with the noise cancellation problem under consideration.

We will constrain the systems that represent the room acoustics to be causal, and to have a finite impulse response. Thus, for example, $A(\omega)$ is a frequency response of an FIR filter, i.e.,

$$A(\omega) = \sum_{k=0}^{q} a_k e^{-j\omega k}. \tag{7}$$

As mentioned earlier, we will usually assume that $s(t)$, the desired signal, is a sample from an AR process of order p, and thus its power spectrum $P_s(\omega)$ is of the form

$$P_s(\omega) = \frac{G}{\left|1 - \sum_{i=1}^{p} h_i e^{-j\omega i}\right|^2}. \tag{8}$$

In Sections IV and V, more specific situations will be considered, and additional constraints and assumptions, based on the additional knowledge about the underlying scenario, will be made. In both sections, the resulting ML problem is constrained enough so that it is not ill posed.

We note that even with these assumptions and constraints, the required maximization of the likelihood function (5) with respect to the signal and system parameters is still complicated. Therefore, the EM algorithm will be proposed for its solution. In the cases considered in the next sections, the unavailable desired signal $s(t)$ will be a component of the chosen complete data. Thus, as a by-product of the use of the EM algorithm, an estimate of the desired signal will become explicitly available while implementing the E step.

III. The EM Algorithm for Maximum Likelihood Estimation

The methods proposed in this paper for noise cancellation are based on an iterative solution to the maximum

[2]The symbol † denotes the Hermitian operator, while the symbol * denotes the complex conjugate operator.

likelihood problem of (5). This iterative solution, referred to as the EM algorithm, is briefly summarized in this section.

We denote by $\boldsymbol{Y}$ the data vector with the associated probability density $f_Y(\boldsymbol{y}; \boldsymbol{\theta})$ indexed by the parameter vector $\boldsymbol{\theta}$ where the possible parameter values are contained in a set Θ. Given an observed $\boldsymbol{y}$, the ML estimate $\hat{\boldsymbol{\theta}}_{ML}$ is the value of $\boldsymbol{\theta}$ that maximizes the log likelihood, that is,[3]

$$\hat{\boldsymbol{\theta}}_{ML} = \arg \max_{\boldsymbol{\theta} \in \Theta} \log f_Y(\boldsymbol{y}; \boldsymbol{\theta}). \tag{9}$$

Suppose that the data vector $\boldsymbol{Y}$ can be viewed as being incomplete, and we can specify some "complete" data $\boldsymbol{X}$ related to $\boldsymbol{Y}$ by

$$H(\boldsymbol{X}) = \boldsymbol{Y} \tag{10}$$

where $H(\cdot)$ is a noninvertible (many to one) transformation.

The EM algorithm is directed at finding the solution to (9); however, it does so by making an essential use of the complete data specification. The algorithm is basically an iterative method. It starts with an initial guess $\boldsymbol{\theta}^{(0)}$, and $\boldsymbol{\theta}^{(n+1)}$ is defined inductively by

$$\boldsymbol{\theta}^{(n+1)} = \arg \max_{\boldsymbol{\theta} \in \Theta} E\left\{ \log f_X(\boldsymbol{x}; \boldsymbol{\theta}) / \boldsymbol{y}; \boldsymbol{\theta}^{(n)} \right\} \tag{11}$$

where $f_X(\boldsymbol{x}; \boldsymbol{\theta})$ is the probability density of $\boldsymbol{X}$, and $E\{\cdot / \boldsymbol{y}; \boldsymbol{\theta}^{(n)}\}$ denotes the conditional expectation given $\boldsymbol{y}$, computed using the parameter value $\boldsymbol{\theta}^{(n)}$. The intuitive idea is that we would like to choose $\boldsymbol{\theta}$ that maximizes $\log f_X(\boldsymbol{x}; \boldsymbol{\theta})$, the log likelihood of the complete data. However, since $\log f_X(\boldsymbol{x}; \boldsymbol{\theta})$ is not available to us (because the complete data are not available), we maximize instead its expectation, given the observed data $\boldsymbol{y}$. Since we used the current estimate $\boldsymbol{\theta}^{(n)}$ rather than the actual value of $\boldsymbol{\theta}$ which is unknown, the conditional expectation is not exact. Thus, the algorithm iterates, using each new parameter estimate to improve the conditional expectation on the next iteration cycle (the E step) and then uses this conditional estimate to improve the next parameter estimate (the M step).

The EM algorithm was presented in its general form by Dempster *et al.*[4] in [7]. The algorithm was suggested before, for specific applications, by several authors, e.g., [14]–[16]. The rate of convergence of the algorithm is linear [7], depending on the fraction of the covariance of the complete data that can be predicted using the observed data. If that fraction is small, the rate of convergence tends to be slow, in which case one could use standard numerical methods to accelerate the algorithm.

We note that the EM algorithm is not uniquely defined. The transformation $H(\cdot)$ relating $\boldsymbol{X}$ to $\boldsymbol{Y}$ can be *any* noninvertible transformation. Obviously, there are many possible "complete" data specifications that will generate the observed data. Thus, the EM algorithm can be implemented in many possible ways. The way $H(\cdot)$ is specified (i.e., the choice of the "complete" data) may critically affect the complexity and the rate of convergence of the algorithm.

[3] "arg max" denotes the argument of the maximization.

[4] In [7] it is shown that each iteration increases the likelihood. However, there is an error in the convergence proof (theorem 2 of [7]), pointed out Wu [13]. The proper conditions that guarantee the convergence of the algorithm to a stationary point of the likelihood are given in [13].

IV. A Simplified Scenario

In this section, a simplified version of the problem is assumed, corresponding to restricting $B(z)$ to be zero and both $C(z)$ and $D(z)$ to be unity in Fig. 2, so that Fig. 2 reduces to Fig. 4. This scenario is assumed (at least implicitly) by Widrow *et al.* in [2]. In this scenario, one microphone measures the desired (speech) signal with additive noise, while the second microphone measures a reference noise signal, which is correlated with the noise component of the signal measured in the first microphone, but has no signal component present. In Section V we consider a more general configuration.

A. The ML Problem

As indicated in Fig. 4, the observed signals are $y_1(t)$ and $y_2(t)$, $A(z)$ is an FIR filter, $e(t)$ is Gaussian white noise, and $s(t)$ is the desired signal.

Specifically, then

$$y_1(t) = s(t) + n(t) \tag{12}$$

where the noise component in the primary microphone is

$$n(t) = \sum_{k=0}^{q} a_k y_2(t-k) + e(t) \tag{13}$$

and it incorporates the coupling of $w(t)$ into the first sensor and the additional error $e(t)$. Equivalently, (12) and (13) may be written as

$$y_1(t) = s(t) + \sum_{k=0}^{q} a_k w(t-k) + e(t) \tag{14}$$

$$y_2(t) = w(t). \tag{15}$$

As described above, we assume that the desired signal $s(t)$ can be represented as a sample function from a stationary Gaussian process whose spectrum is known up to some parameters. The unknown parameters $\boldsymbol{\theta}$, in this case, are the system coefficients $\{a_k\}$, the spectral parameters of $s(t)$ (which will be denoted $\boldsymbol{\varphi}$), and σ^2, the variance of $e(t)$.

The likelihood of the observation is again expressed in the frequency domain. This case is simpler than the general case discussed in the previous section. The likelihood may be obtained without referring to the general formula of (5).

Specifically, since under the assumptions made above, the Fourier coefficients of the signals in different frequencies are uncorrelated, the likelihood of the observation may be written in the frequency domain as

$$\begin{aligned} L(\boldsymbol{\theta}) &= \log f_{Y_1, Y_2}(y_1(t), y_2(t); \boldsymbol{\theta}) \\ &= \sum_{\omega_l} \log f_{Y_1, Y_2}(Y_1(\omega_l), Y_2(\omega_l); \boldsymbol{\theta}). \end{aligned} \tag{16}$$

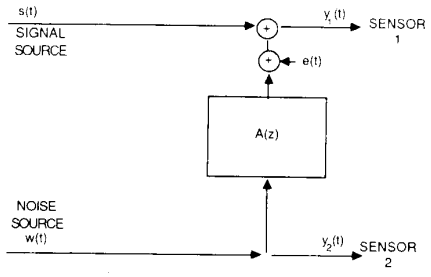


Fig. 4. The observations; simplified scenario.

Now, at each frequency ω_l

$$\log f_{Y_1,Y_2}(Y_1(\omega_l), Y_2(\omega_l); \theta) = \log f_{Y_1/Y_2}(Y_1(\omega_l)/Y_2(\omega_l); \theta) + \log f_{Y_2}(Y_2(\omega_l)) \tag{17}$$

where $\log f_{Y_2}(Y_2(\omega_l))$ is independent of θ. The conditional density of $Y_1(\omega_l)$ given $Y_2(\omega_l)$ is given by

$$\log f_{Y_1/Y_2}(Y_1(\omega_l)/Y_2(\omega_l); \theta) = -\left[\log \pi(P_s(\omega_l) + \sigma^2) + \frac{|Y_1(\omega_l) - A(\omega_l) \cdot Y_2(\omega_l)|^2}{P_s(\omega_l) + \sigma^2}\right]. \tag{18}$$

Thus, maximizing the likelihood (16) in this case is equivalent to minimizing

$$\sum_{\omega_l}\left[\log (P_s(\omega_l) + \sigma^2) + \frac{|Y_1(\omega_l) - A(\omega_l) \cdot Y_2(\omega_l)|^2}{P_s(\omega_l) + \sigma^2}\right] \tag{19}$$

with respect to σ^2 and the coefficients of $P_s(\omega)$ and $A(\omega)$.

We will assume that $A(\omega)$ is the frequency response of an FIR filter of a given order q, i.e., it is of the form of (7). Also, we will assume that $s(t)$ is an AR process of order p with coefficients $\{h_i\}_{i=1}^{p}$ and gain G, so that its power spectrum $P_s(\omega)$ is given by (8).

In some applications, like LPC vocoding and speech recognition of noisy data, we will be interested mainly in the spectral parameters of the speech signal. In this case, solving this ML problem explicitly provides these desired parameters. In other applications, we will be interested in the speech signal itself. So, using the estimated signal parameters, $\{a_k\}$, we will cancel the noise in the primary microphone and obtain an enhanced speech signal. As mention above, this speech signal estimate will be available as a by product of the EM algorithm suggested below, while implementing the E step.

B. Solution via the EM Algorithm

Direct minimization of (19) is complicated; therefore, we consider the use of the EM algorithm. In this approach, the complete data are chosen to be the set of signals $\{s(t), n(t), y_2(t)\}$. This choice of complete data is motivated by the simple maximum likelihood solution available if indeed $s(t)$, $n(t)$, and $y_2(t)$ are observed separately.

Loosely speaking, if these complete data are available, the maximum likelihood estimate of $\{a_k\}$ and σ^2 is achieved by least squares fitting of $y_2(t)$ to $n(t)$. The spectral parameters of $s(t)$ are also easily estimated by solving, for example, the normal equation using the sample correlation of $s(t)$, for LPC parameters.

More specifically, the likelihood of the complete data, $L_c(\theta)$, satisfies

$$L_c(\theta) = \log f_{s,n,y_2}(s(t), n(t), y_2(t); \theta) = \log f_{s,n/y_2}(s(t), n(t)/y_2(t); \theta) + \log f_{y_2}(y_2(t)) \tag{20}$$

where $\log f_{y_2}(y_2(t))$ is independent of θ. Also, given $y_2(t)$, the signals $s(t)$ and $n(t)$ are statistically independent, and thus,

$$\log f_{s,n/y_2}(s(t), n(t)/y_2(t); \theta) = \log f_{s/y_2}(s(t)/y_2(t); \theta) + \log f_{n/y_2}(n(t)/y_2(t); \theta). \tag{21}$$

Now, $\log f_{n/y_2}(n(t)/y_2(t); \theta)$ depends only on $\{a_k\}$ and σ^2, and it is defined by the p.d.f. of $e(t)$, i.e.,

$$\log f_{n/y_2}(n(t)/y_2(t); \theta) = -\sum_{t=0}^{N-1}\left[\log \sigma^2 + \frac{1}{\sigma^2}\left(n(t) - \sum_{k=0}^{q} a_k y_2(t-k)\right)^2\right]. \tag{22}$$

In general, the signal $y_2(t)$ may be related to $s(t)$. However, this relation is arbitrary and unknown. Therefore, we will assume that the probability distribution of $s(t)$ given $y_2(t)$ is the *a priori* distribution of $s(t)$. This probability distribution is the distribution of a stationary random process with power spectrum $P_s(\omega)$ and it depends only on the spectral parameters, φ, of $s(t)$, thus,

$$\log f_{s/y_2}(s(t)/y_2(t); \theta) = \log f_s(s(t); \varphi) = \sum_{\omega_l}\left[\log P_s(\omega_l; \varphi) + \frac{|S(\omega_l)|^2}{P_s(\omega_l; \varphi)}\right] \tag{23}$$

where $S(\omega)$ is the Fourier transform of $s(t)$, i.e.,

$$S(\omega) = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} s(t) e^{-j\omega t}.$$

Thus, estimating θ by maximizing the likelihood of the complete data is equivalent to estimating σ^2 and $\{a_k\}$ by

minimizing

$$\frac{1}{\sigma^2}\sum_{t=0}^{N-1}\left(n(t)-\sum_{k=0}^{q}a_k y_2(t-k)\right)^2+N\cdot\log\sigma^2 \tag{24}$$

and estimating the spectral parameters φ by minimizing

$$\sum_{\omega_l}\left[\log P_s(\omega_l;\varphi)+\frac{|S(\omega_l)|^2}{P_s(\omega_l;\varphi)}\right]. \tag{25}$$

Note that when $s(t)$ is assumed to be an AR process, we show in the Appendix that minimizing (25) is equivalent to solving the Yule–Walker equation, using the sample autocorrelation of $s(t)$.

We observe from (24) and (25) that the required statistics of the complete data are $n(t)$, $|S(\omega)|^2$, and also $n^2(t)$ if we need to estimate σ^2. Thus, the E step of the algorithm requires the following expectations:

$$\hat{n}(t)=E\left\{n(t)/y_1(t),\,y_2(t);\,\theta^{(n)}\right\} \tag{26}$$

and

$$M_S(\omega)=E\left\{|S(\omega)|^2/Y_1(\omega),\,Y_2(\omega);\,\theta^{(n)}\right\} \tag{27}$$

where $\theta^{(n)}$ denotes the parameters $\{a_k\}$, σ^2, and φ in the nth iteration. If we also need to estimate σ^2, we have to take the expectation

$$\hat{e^2}(t)=E\left\{\left(n(t)-\sum_{k=0}^{p}a_k y_2(t-k)\right)^2\Big/ y_1(t),\,y_2(t);\,\theta^{(n)}\right\}. \tag{28}$$

The E and the M steps of the EM algorithm for minimizing (19) may now be stated explicitly. Recall that we denote by $\theta^{(n)}$ [or by $\{a_k^{(n)}\}$, $(\sigma^2)^{(n)}$, and $P_s^{(n)}(\omega)$] the current estimate of the parameters.

- *The E Step, the nth Iteration:*
  - Generate a signal $x(t)$

$$x(t)=y_1(t)-\sum_{k=0}^{q}a_k^{(n)}y_2(t-k). \tag{29}$$

Note that if the true coefficients $\{a_k\}$ were known, then $x(t)=s(t)+e(t)$.

- Apply a Wiener filter of $x(t)$ to obtain the conditional expectation or the minimum mean square error estimate of $s(t)$ [or $S(\omega_l)$] and $|S(\omega_l)|^2$. Specifically, for all ω_l, generate an estimate of $S(\omega_l)$, $E(\omega_l)$ and the quadratic terms $|S(\omega_l)|^2$ and $|E(\omega_l)|^2$ as

$$\hat{S}(\omega_l)=\frac{P_s^{(n)}(\omega_l)}{P_s^{(n)}(\omega_l)+(\sigma^2)^{(n)}}\cdot X(\omega_l) \tag{30a}$$

$$\hat{E}(\omega_l)=X(\omega_l)-\hat{S}(\omega_l) \tag{30b}$$

$$M_S(\omega_l)=|\hat{S}(\omega_l)|^2+\frac{P_s^{(n)}(\omega_l)\cdot(\sigma^2)^{(n)}}{P_s^{(n)}(\omega_l)+(\sigma^2)^{(n)}} \tag{30c}$$

$$M_E(\omega_l)=|\hat{E}(\omega_l)|^2+\frac{P_s^{(n)}(\omega_l)\cdot(\sigma^2)^{(n)}}{P_s^{(n)}(\omega_l)+(\sigma^2)^{(n)}} \tag{30d}$$

where $X(\omega)$ is the Fourier transform of $x(t)$ and $\hat{E}(\omega)$ is the Fourier transform of the signal $\hat{e}(t)$.

- The conditional expectation (estimate) of $n(t)$ is

$$\hat{n}(t)=\sum_{k=0}^{q}a_k^{(n)}y_2(t-k)+\hat{e}(t). \tag{31}$$

- *The M Step, the nth Iteration:* Substitute the conditional expectations of (30) and (31) into (24) and (25). Specifically,
  - update $\{a_k\}$ by solving the least-squares problem of (24) with (31) substituted for $n(t)$, i.e.,

$$\{a_k^{(n+1)}\}=\arg\min_{\{a_k\}}\sum_{n=0}^{N-1}\left(\sum_{k=0}^{q}(a_k^{(n)}-a_k)\cdot y_2(t-k)+\hat{e}(t)\right)^2 \tag{32}$$

  - update σ^2 as

$$(\sigma^2)^{(n+1)}=\frac{1}{N}\sum_{t=0}^{N-1}e^2(t) \tag{33}$$

where $e^2(t)$ defined in (28) is the inverse Fourier transform of $M_E(\omega)$, calculated in the E step.

- update the spectral parameters by solving (2b) with $M_S(\omega_l)$ substituted for $|S(\omega_l)|^2$. For LPC parameters, solve the Yule–Walker equation using the correlation values obtained by inverse Fourier transforming $M_S(\omega)$.

The EM algorithm above iterates, until some convergence criterion is met. This algorithm is summarized in Fig. 5.

V. A More General Scenario

The modeling of the two-microphone noise cancellation situation in the previous section ignores the possible coupling of the desired signal $s(t)$ into the reference microphone, as is present in Fig. 2 and equation (1). In the classical least-squares approach, this coupling results in a reverberant quality to the output because the desired signal is partially cancelled together with the noise. Since the ML problem of the previous section also ignores this coupling, the EM noise cancelling algorithm, developed in Section IV, has a similar problem.

In the ML approach considered in this section, this coupling is taken into account. Specifically, we now include the presence of the system B in Fig. 2, but still assume that $C\equiv 1$ and $D\equiv 1$ corresponding to the assumption that sensor 1 is close to the signal source and sensor 2 is close to the noise source. The resulting model is shown in Fig. 6. We also assume that $A(z)$ and $B(z)$ are both FIR systems. These assumptions are important because without them, the problem is ill posed. For example, if A, B, C, and D are arbitrary, intuitively one sees that there is a symmetry to the problem that precludes the algorithm

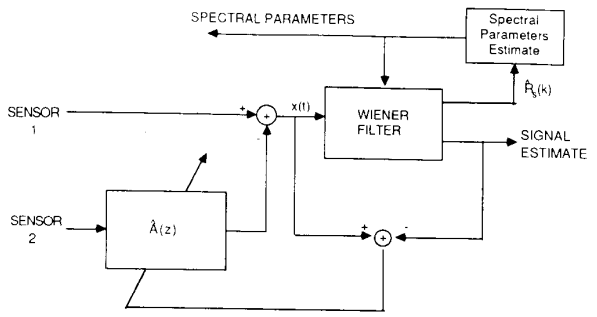


Fig. 5. The EM algorithm; simplified scenario.

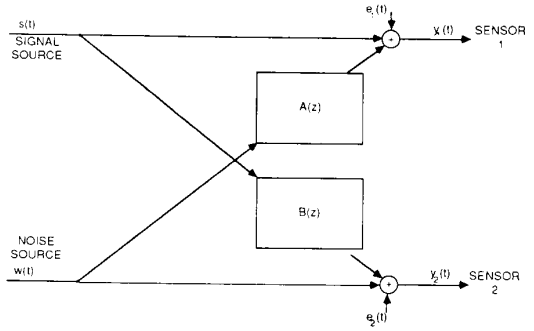


Fig. 6. The observations; more general scenario.

distinguishing between the signal and the noise components in each sensor. With the stated assumptions this symmetry is removed.

We will start by explicitly presenting the ML problem for this scenario. We will then present an EM algorithm for maximizing this likelihood, where the complete data will be composed of the desired speech signal $s(t)$ and the noise source signal $w(t)$, in addition to the observed signals $y_1(t)$ and $y_2(t)$.

A. The ML Problem

The situation assumed in this section is indicated in Fig. 6. The mathematical model that corresponds to the situation is given by

$$y_1(t) = s(t) + A\{w(t)\} + e_1(t) \tag{34a}$$

$$y_2(t) = B\{s(t)\} + w(t) + e_2(t) \tag{34b}$$

where, as before, $s(t)$ is the desired signal, $w(t)$ is the noise source signal, and $e_1(t)$, $e_2(t)$ are the measurement and modeling error signals in the two microphones. As in the general problem, $s(t)$ and $w(t)$ are assumed to be sample signals from Gaussian random processes. The error signals $e_1(t)$, $e_2(t)$ are white Gaussian noise processes. The unknown parameters $\boldsymbol{\theta}$ are the impulse response coefficients $\{a_k\}$ and $\{b_k\}$ of the systems A and B, the spectral parameters of the signals $s(t)$ and $w(t)$ denoted φ_s and φ_w, respectively, and the variances σ_{e_1} and σ_{e_2} of the noises $e_1(t)$ and $e_2(t)$.

With these assumptions, the likelihood of the observations is given again by (5). However, with $C(\omega) \equiv D(\omega) \equiv 1$, the power spectrum matrix $\Lambda(\omega)$ is simplified to

$$\Lambda(\omega) = E\{\boldsymbol{Y}(\omega)\,\boldsymbol{Y}(\omega)^{\dagger}\} = \begin{bmatrix} P_s(\omega) + A(\omega)\,P_w(\omega)\,A^*(\omega) + \sigma_{e_1}^2 & P_s(\omega)\,B^*(\omega) + A(\omega)\,P_w(\omega) \\ B(\omega)\,P_s(\omega) + P_w(\omega)\,A^*(\omega) & B(\omega)\,P_s(\omega)\,B^*(\omega) + P_w(\omega) + \sigma_{e_2}^2 \end{bmatrix}. \tag{35}$$

We will assume again that $A(\omega)$ and $B(\omega)$ are frequency responses of FIR filters, i.e., their structure is given by (7). The orders of those FIR filters are assumed to be known, and are denoted q_a, q_b, respectively. The desired signal is assumed to be a sample from an AR process of a given order p, and thus $P_s(\omega)$ will have the structure of (8). We further assume that $w(t)$ is a white noise signal, i.e., $P_w(\omega)$ is constant. Even with these assumptions, the underlying ML problem is complicated, and again we will use the EM algorithm for its solution.

For applications such as LPC vocoding, where only the spectral parameters of the speech signals are required, solving this ML problem will explicitly provide these desired parameters. For applications where the speech signal is required, the MMSE estimate of the speech signal using the ML estimate of the parameters will be suggested. This MMSE estimate will be available for each current parameter value, as a byproduct, while implementing the E step of the suggested EM algorithm.

B. Solution via the EM Algorithm

The complete data suggested for defining the EM algorithm, in the current context, are the set of signals $\{s(t), w(t), y_1(t), y_2(t)\}$. The complete data are chosen this way if indeed the signals $s(t)$ and $w(t)$, the input to the two channel system of (34), are observed, in addition to the signals $y_1(t)$ and $y_2(t)$, the output of this system, there will be a simple procedure for ML estimation of the parameters of this two-channel system.

Specifically, suppose that these complete data are available. To estimate the parameters, we will maximize its likelihood given by

$$\begin{aligned} L_c(\boldsymbol{\theta}) &= \log f_{y_1,y_2,s,w}(y_1(t), y_2(t), s(t), w(t); \boldsymbol{\theta}) \\ &= \log f_{y_1,y_2,s,w}(y_1(t), y_2(t)/s(t), w(t); \boldsymbol{\theta}) \\ &\quad + \log f_{s,w}(s(t), w(t); \boldsymbol{\theta}). \end{aligned} \tag{36}$$

The signals $y_1(t)$ and $y_2(t)$ are independent, given $s(t)$ and $w(t)$. The signals $s(t)$ and $w(t)$ are independent by

assumption, thus,

$$L_c(\theta) = \underbrace{\log f_{y_1/s,w}(y_1(t)/s(t), w(t); \theta)}_{I} + \underbrace{\log f_{y_2/s,w}(y_2(t)/s(t), w(t); \theta)}_{II} + \underbrace{\log f_s(s(t); \theta)}_{III} + \underbrace{\log f_w(w(t); \theta)}_{IV}. \quad (37)$$

Term *I* depends only on $\{a_k\}$ and $\sigma_{e_1}^2$ and is the log probability of the sequence $e_1(t)$. Similarly, term *II* depends only on $\{b_k\}$ and $\sigma_{e_2}^2$ and is the log probability of the sequence $e_2(t)$. Term *III* is the log probability of the stationary signal $s(t)$ and depends only on its spectral parameters φ_s. Similarly term *IV* is the log probability of the stationary signal $w(t)$ and depends only on its spectral parameters φ_w. Maximizing the likelihood of the complete data with respect to θ is equivalent to maximizing each of the terms *I–IV* separately with respect to the parameters they depend on.

Thus, given the complete data, the parameters φ_s are estimated by

$$\hat{\varphi}_s = \arg\max_{\varphi_s} \log f_s(s(t); \varphi_s) = \arg\min_{\varphi_s} \sum_{\omega_l} \left[\log P_s(\omega_l; \varphi_s) + \frac{|S(\omega_l)|^2}{P_s(\omega_l; \varphi_s)} \right] \quad (38)$$

and φ_w are estimated by

$$\hat{\varphi}_w = \arg\max_{\varphi_w} \log f_w(w(t); \varphi_w) = \arg\min_{\varphi_w} \sum_{\omega_l} \left[\log P_w(\omega_l; \varphi_w) + \frac{|W(\omega_l)|^2}{P_w(\omega_l; \varphi_w)} \right] \quad (39)$$

where $S(\omega)$ and $W(\omega)$ are the Fourier transforms of $s(t)$ and $w(t)$, respectively, i.e.,

$$S(\omega) = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} s(t) e^{-j\omega t}$$

$$W(\omega) = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} w(t) e^{-j\omega t}.$$

The maximization in (38) is sometimes simpler, e.g., when $s(t)$ is assumed to be an AR process, in which case, as shown in the Appendix, maximizing (38) is equivalent to solving the Yule–Walker equation, using the sample autocorrelation of $s(t)$. Similarly, solving (39) is sometimes simpler, and if $w(t)$, the noise source signal, is a white noise signal, it is equivalent to finding the (constant) spectrum level P_w by

$$\hat{P}_w = \frac{1}{N} \sum_{t=0}^{N-1} w^2(t) = \frac{1}{N} \sum_{\omega_l} |W(\omega_l)|^2. \quad (40)$$

Estimating the impulse response coefficients, $\{a_k\}$, and the variance, $\sigma_{e_1}^2$, given the complete data, requires solving a least-squares problem, since

$$\hat{\sigma}_{e_1}^2, \{\hat{a}_k\} = \arg\max_{\sigma_{e_1}^2, \{a_k\}} \log f_{y_1/s,w}(y_1(t)/s(t), w(t); a_0, \cdots, a_{q_a}, \sigma_{e_1}^2) = \arg\min_{\sigma_{e_1}^2, \{a_l\}} \frac{1}{\sigma_{e_1}^2} \sum_{t=0}^{N-1} \left(y_1(t) - s(t) - \sum_{k=0}^{q_a} a_k w(t-k) \right)^2 + N \cdot \log \sigma_{e_1}^2. \quad (41)$$

Similarly, estimating $\{b_k\}$ and $\sigma_{e_2}^2$ given the complete data requires solving the following least squares problem:

$$\hat{\sigma}_{e_2}^2, \{\hat{b}_k\} = \arg\max_{\sigma_{e_2}^2, \{b_k\}} \log f_{y_2/s,w}(y_2(t)/s(t), w(t); b_0, \cdots, b_{q_b}, \sigma_{e_2}^2) = \arg\min_{\sigma_{e_2}^2, \{b_k\}} \frac{1}{\sigma_{e_2}^2} \sum_{t=0}^{N-1} \left(y_2(t) - w(t) - \sum_{k=0}^{q_b} b_k s(t-k) \right)^2 + N \cdot \log \sigma_{e_2}^2. \quad (42)$$

The explicit solution of the least-squares problems implied by (41) and (42) is achieved by solving the following "normal" linear equations:

$$\mathcal{R}_w \cdot \boldsymbol{a} = \boldsymbol{r}_{wy_1} - \boldsymbol{r}_{ws} \quad (43)$$

$$\mathcal{R}_s \cdot \boldsymbol{b} = \boldsymbol{r}_{sy_2} - \boldsymbol{r}_{sw} \quad (44)$$

where $\mathcal{R}_w$ is the correlation matrix of $w(t)$ of order q_a, i.e.,

$$\mathcal{R}_w = \begin{bmatrix} r_w(0) & r_w(1) & r_w(2) & \cdots & \cdots & r_w(q_a) \\ r_w(1) & r_w(0) & & & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ & & & & & r_w(1) \\ r_w(q_a) & \cdots & \cdots & r_w(2) & r_w(1) & r_w(0) \end{bmatrix}$$

where $r_w(k) = \frac{1}{N} \sum_{t=0}^{N-1} w(t)\, w(t-k)$. (45)

$\mathcal{R}_s$ is the order q_b correlation matrix of $s(t)$

$$\mathcal{R}_s = \begin{bmatrix} r_s(0) & r_s(1) & r_s(2) & \cdots & \cdots & r_s(q_b) \\ r_s(1) & r_s(0) & & & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ & & & & & r_s(1) \\ r_s(q_b) & \cdots & \cdots & r_s(2) & r_s(1) & r_s(0) \end{bmatrix} \quad \text{where } r_s(k) = \frac{1}{N}\sum_{t=0}^{N-1} s(t)\, s(t-k). \tag{46}$$

The vectors $\boldsymbol{r}_{wy_1}$, $\boldsymbol{r}_{ws}$, $\boldsymbol{r}_{sy_2}$, $\boldsymbol{r}_{sw}$ represent the appropriate cross correlation between the signals, e.g.,

$$\boldsymbol{r}_{ws} = \begin{bmatrix} r_{ws}(0) \\ \vdots \\ r_{ws}(q_a) \end{bmatrix}$$

$$\text{where } r_{ws}(k) = \frac{1}{N}\sum_{t=0}^{N-1} s(t)\, w(t-k) \tag{47}$$

and the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are the unknown impulse response coefficients of the systems A and B.

Observing the required procedures for maximizing the likelihood of the complete data, i.e., equations (38), (39) and (43), (44), we see that the sufficient statistics of the complete data contain quadratic terms, which are the sample autocorrelation (or the sample spectrum) and the sample cross correlation (or cross spectrum) of the various signals, in addition to the linear terms (i.e., the signals themselves). Thus, the E step of the algorithm (with the above choice of complete data) requires the expectations

$$\hat{s}(t) = E\left\{s(t)/y_1(t), y_2(t); \boldsymbol{\theta}^{(n)}\right\} \tag{48a}$$

$$\hat{w}(t) = E\left\{s(t)/y_1(t), y_2(t); \boldsymbol{\theta}^{(n)}\right\} \tag{48b}$$

and the quadratic terms

$$\hat{r}_s(k) = E\left\{r_s(k)/y_1(t), y_2(t); \boldsymbol{\theta}^{(n)}\right\} \tag{49a}$$

$$\hat{r}_w(k) = E\left\{r_w(k)/y_1(t), y_2(t); \boldsymbol{\theta}^{(n)}\right\} \tag{49b}$$

$$\hat{r}_{sw}(k) = \hat{r}_{ws}(-k) = E\left\{r_{sw}(k)/y_1(t), y_2(t); \boldsymbol{\theta}^{(n)}\right\}. \tag{49c}$$

We will implement the E step in the frequency domain, since for stationary processes with large observation time, the DFT coefficients at each frequency are statistically independent and can be processed separately. In each frequency ω_l, the observation may be written as

$$\begin{bmatrix} Y_1(\omega_l) \\ Y_2(\omega_l) \end{bmatrix} = \begin{bmatrix} 1 & A(\omega_l) \\ B(\omega_l) & 1 \end{bmatrix} \cdot \begin{bmatrix} \hat{S}(\omega_l) \\ W(\omega_l) \end{bmatrix}. \tag{50}$$

The E step requires the conditional expectation of $S(\omega_l)$, $W(\omega_l)$, $|S(\omega_l)|^2$, $|W(\omega_l)|^2$, and $S(\omega_l)\, W^*(\omega_l)$.

At each step of the algorithm, the current values of the parameters are used. We will denote by $A^{(n)}(\omega)$ and $B^{(n)}(\omega)$ the current estimate of the frequency responses of the unknown systems A and B, and by $P_s^{(n)}(\omega)$ and $P_w^{(n)}(\omega)$ the current estimate of the power spectra of $s(t)$ and $w(t)$. Let $H(\omega_l)$ denote the matrix

$$H(\omega_l) = \begin{bmatrix} 1 & A^{(n)}(\omega_l) \\ B^{(n)}(\omega_l) & 1 \end{bmatrix} \tag{51}$$

and let $\Phi(\omega_l)$ and Σ denote the power spectra matrices

$$\Phi(\omega_l) = \begin{bmatrix} P_s^{(n)}(\omega_l) & 0 \\ 0 & P_w^{(n)}(\omega_l) \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \sigma_{e1}^2 & 0 \\ 0 & \sigma_{e2}^2 \end{bmatrix}. \tag{52}$$

The required conditional expectations are readily available, using linear estimation theory. These conditional estimates may be interpreted, in this case, as performing a two-channel Wiener filter (see Appendix) and calculating its error covariance matrix. Thus, the estimate of the linear terms is given by

$$\begin{bmatrix} \hat{S}(\omega_l) \\ \hat{W}(\omega_l) \end{bmatrix} = K(\omega_l) \cdot \begin{bmatrix} Y_1(\omega_l) \\ Y_2(\omega_l) \end{bmatrix} \tag{53}$$

where $K(\omega_l)$ is the matrix

$$K(\omega_l) = \Phi(\omega_l) \cdot H(\omega_l)^{\dagger}\left(H(\omega_l) \cdot \Phi(\omega_l) \cdot H(\omega_l)^{\dagger} + \Sigma\right)^{-1}. \tag{54}$$

For the quadratic terms, we have to calculate the error covariance matrix of this Wiener filter, i.e.,

$$\begin{aligned} \hat{\Phi}(\omega_l) &= \left(\Phi^{-1}(\omega_l) + H(\omega_l) \cdot \Sigma^{-1} \cdot H(\omega_l)^{\dagger}\right)^{-1} \\ &= \Phi(\omega_l) - \Phi(\omega_l)\, H(\omega_l)^{\dagger}\left(H(\omega_l) \cdot \Phi(\omega_l) \cdot H(\omega_l)^{\dagger} + \Sigma\right)^{-1} H(\omega_l)\, \Phi(\omega_l) \end{aligned} \tag{55}$$

and the quadratic terms are obtained by

$$\begin{aligned} M_S(\omega_l) &= E\left\{|S(\omega_l)|^2/Y_1(\omega_l), Y_2(\omega_l)\right\} \\ &= |\hat{S}(\omega_l)|^2 + \hat{\Phi}_{11}(\omega_l) \end{aligned} \tag{56}$$

$$\begin{aligned} M_W(\omega_l) &= E\left\{|W(\omega_l)|^2/Y_1(\omega_l), Y_2(\omega_l)\right\} \\ &= |\hat{W}(\omega_l)|^2 + \hat{\Phi}_{22}(\omega_l) \end{aligned} \tag{57}$$

$$\begin{aligned} M_{SW}(\omega_l) &= E\left\{S(\omega_l)\, W^*(\omega_l)/Y_1(\omega_l), Y_2(\omega_l)\right\} \\ &= \hat{S}(\omega_l)\, \hat{W}^*(\omega_l) + \hat{\Phi}_{12}(\omega_l). \end{aligned} \tag{58}$$

The E and M steps of the EM algorithm for maximizing the likelihood of the observations [given by (5) and (35)]

for this more general case may now be stated explicitly. Here, for simplicity, we will assume that we do not have to estimate $\sigma_{e_1}^2$ and $\sigma_{e_2}^2$.

- *The E Step, the nth Iteration:*
  - Calculate the conditional expectations $\hat{S}(\omega_l)$ and $\hat{W}(\omega_l)$ by (53).
  - Calculate $M_S(\omega_l)$, $M_W(\omega_l)$, and $M_{SW}(\omega_l)$ by (56)–(58).
  - The signal estimates $\hat{s}(t)$ and $\hat{w}(t)$, and the correlation estimates $\hat{r}_s(k)$, $\hat{r}_w(k)$, and $\hat{r}_{sw}(k)$, are achieved by inverse Fourier transforming $\hat{S}(\omega)$, $\hat{W}(\omega)$, $M_S(\omega)$, $M_W(\omega)$, and $M_{SW}(\omega)$, respectively.
- *The M Step, the nth Iteration:*
  - Solve the linear equations of (43) and (44) for $\boldsymbol{a}$ and $\boldsymbol{b}$, using the estimates $\hat{r}_s(k)$, $\hat{r}_w(k)$, and $\hat{r}_{sw}(k)$ from the E step, and with

$$\hat{r}_{wy_1}(k) = \frac{1}{N}\sum_{t=0}^{N-1} \hat{w}(t)\, y_1(t-k) \tag{59a}$$

$$\hat{r}_{sy_2}(k) = \frac{1}{N}\sum_{t=0}^{N-1} \hat{s}(t)\, y_2(t-k). \tag{59b}$$

The result is the updated coefficients $\boldsymbol{a}^{(n+1)}$ and $\boldsymbol{b}^{(n+1)}$ of the systems A and B.

  - Update the spectral parameter estimate, by solving (38) and (39), using $M_S(\omega_l)$ and $M_W(\omega_l)$ instead of $|S(\omega_l)|^2$ and $|W(\omega_l)|^2$. For LPC parameters of the speech signal $s(t)$, solve the Yule–Walker equations, using $\hat{r}_s(k)$.

The EM algorithm above iterates, until some convergence criterion is met. This algorithm is summarized in Fig. 7.

Further Research—Sequential Algorithms: The procedures suggested in this section, and also in the previous section, are implemented in each iteration on the entire data. We, however, may also be interested in adaptive and sequential procedures, where in each iteration new measurements are processed and an updated segment of enhanced signal is produced. Examining the suggested batch procedure illustrated in Fig. 7, a sequential algorithm comes in mind. The Wiener filter of the E step will be replaced by the sequential Kalman filter, and the linear least-squares problems of the M step will be solved via a sequential RLS-type algorithm. This and other adaptive algorithms could potentially be an alternative to the LMS and RLS algorithms suggested for solving the least-squares problem that arises in Widrow's approach in [2]. This possibility remains to be carefully explored. The details, the analysis, and experiments with this adaptive version are now under investigation and are the subject of further research; some initial results may be found in [17].

VI. Experimental Results

The EM algorithm for both the simplified scenario of Section IV and the more general scenario of Section V has been implemented. In this section we discuss the results.

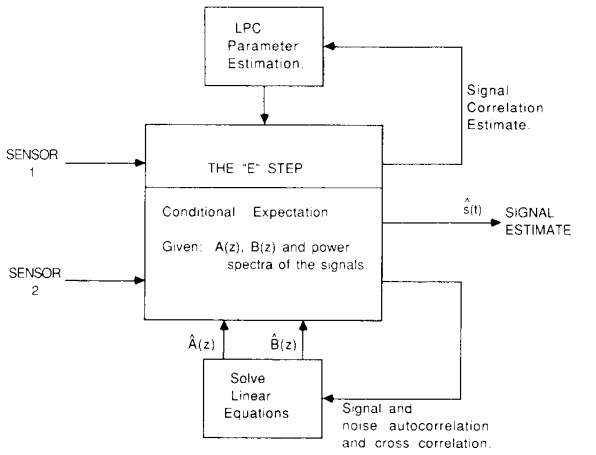


Fig. 7. The EM algorithm; more general scenario.

A. The Simplified Scenario

The EM algorithm developed in Section IV has been implemented with $s(t)$, a speech signal, and $y_2(t)$, a band-limited noise signal, with a flat spectrum from 0 to 3 kHz. The FIR filter $A(z)$ was of order 10. $y_1(t)$ was generated according to Fig. 4, and the SNR in $y_1(t)$ was approximately -20 dB. The level of the independent noise source $e(t)$ was 20 dB below the level of $w(t)$. The results were compared to a "batch" version of the least-squares algorithm, corresponding to estimating the $\{a_k\}$'s via the least-square problem

$$\min_{\{a_k\}} \sum_t \left(y_1(t) - \sum_{k=1}^{q} a_k y_2(t-k) \right)^2$$

and then cancelling the noise and estimating the signal as

$$s(t) = y_1(t) - \sum_{k=1}^{q} a_k y_2(t-k).$$

Both algorithms produced good enhancement of the speech signal, and although there were perceptible differences, the overall quality of both was similar.

The direct least-squares approach assumed that $y_2(t)$ and $s(t)$ are uncorrelated, and this assumption is critical. Our algorithm does not require this assumption. In a second experiment, $y_2(t)$ included a delayed version of the speech signal, as illustrated in Fig. 8. (Note that this scheme is different than the scheme considered in the more general scenario, since we have a direct measurement of the input to the system $A(z)$.)

In a second experiment, the levels of $w(t)$ and $e(t)$ were as before, so that the SNR in $y_1(t)$ was again approximately -20 dB. The direct least-squares approach cancelled part of the signal, together with the noise, resulting in poor quality. In comparison, the performance of our algorithm was still good.

B. The More General Scenario

The scenario assumed in Section V was simulated, where again $s(t)$ was a speech signal and $w(t)$ was a white noise signal. In order to simulate a realistic scenario, we

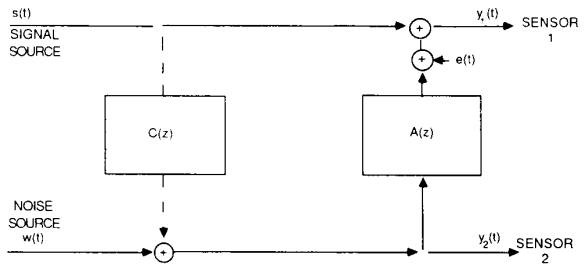


Fig. 8. Correlation between the reference and desired signals. $C(z) = 0.1z^{-5}$.

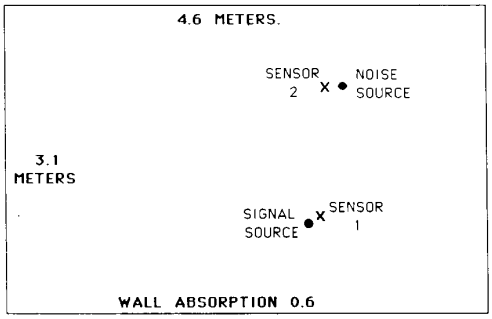


Fig. 9. The living room acoustic environment.

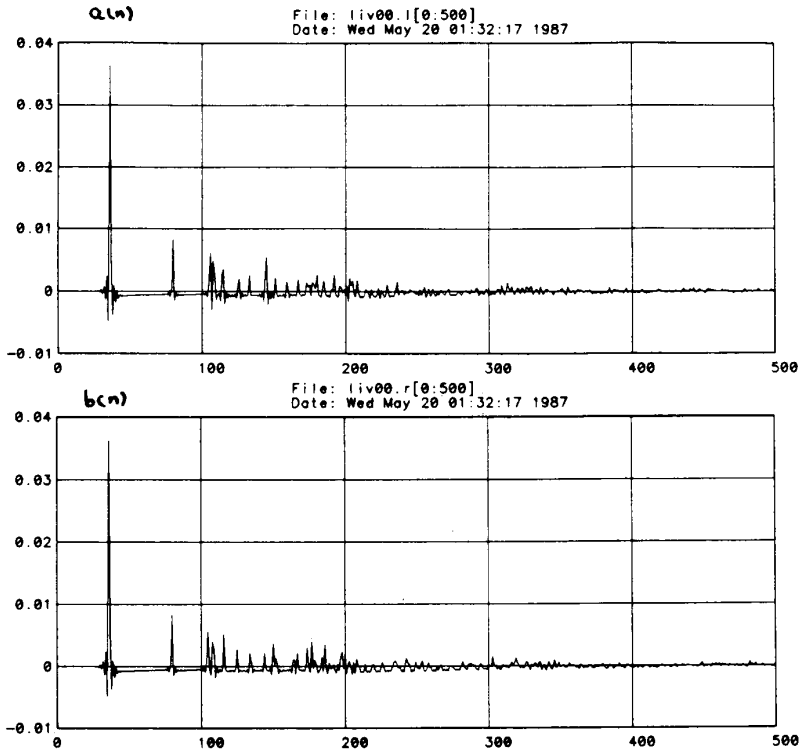


Fig. 10. Simulated "room acoustics" impulse responses.

assumed a living room environment with signal and noise source located as illustrated in Fig. 9. We used a simulation program developed by Peterson [18], and we generated FIR impulse responses having 2000 coefficients each, for the systems A and B. The first 500 coefficients of these impulse responses are plotted in Fig. 10. Monitoring the level of the noise source, we have considered the cases of $+20$, 0, and -20 dB SNR in $y_1(t)$. In all these experiments, the levels of $e_1(t)$ and $e_2(t)$ were 20 dB below the level of $w(t)$.

We have implemented the EM algorithm described in Section V, where we assumed that the level of the white

noise signals $w(t)$, $e_1(t)$, and $e_2(t)$ are known. The results were compared to the least-squares method, by informal listening. Both algorithms estimated up to 500 coefficients of the impulse response. In all SNR levels, our algorithm performed better, and its output, unlike the least squares output, was reverberation free.

At high SNR (+20 dB), the output of the least-squares method output sounded worse than the unprocessed measurement signal, due to the signal cancelling effect. The output of our method sounded better than the original measurement signal.

At 0 dB, the least-squares output sounded better than the measurement signal. However, it sounded much worse than the output of our algorithm, which at this SNR level generated an almost clean signal.

At −20 dB SNR, the output of the ML method sounded better then the least-squares method. However, the distinction between the two was not as significant as in the case of 0 dB SNR. This is perhaps a result of the fact that in order to generate a low SNR, we increased the level of the noise source, which resulted in a high noise-to-signal ratio in the reference microphone, which in turn resulted in a lower signal cancellation since the situation becomes closer to that assumed by the least-squares method.

Appendix A
Maximum Likelihood Parameter Estimation of an AR Process with Long Observation Time

Let $s(t)$ be a sample function from a Gaussian stationary AR process. Suppose that $s(t)$ has been observed in the time window $0 \le t \le N-1$. We want to estimate the AR parameters using the ML criterion. If the observation window is long enough, maximizing the likelihood of the observation is equivalent to minimizing (25), where the power spectrum $P_s(\omega)$ of the process is given by (8), i.e., we have to minimize

$$\sum_{\omega_l}\left[\log G - \log\left|\sum_{i=0}^{p} h_i e^{j\omega_l i}\right|^2 + \frac{\left|S(\omega_l)\right|^2 \cdot \left|\sum_{i=0}^{p} h_i e^{j\omega_l i}\right|^2}{G}\right] \tag{60}$$

where $h_0 = -1$.

We take now the derivatives of (60) with respect to $\{h_k\}_{k=1}^{p}$. Then, setting the derivatives equal to zero, we get

$$\mathrm{Re}\left\{\sum_{\omega_l}\left[\frac{e^{j\omega_l k}}{\sum_{i=0}^{p} h_i e^{j\omega_l i}} + \frac{1}{G}\left|S(\omega_l)\right|^2 \sum_{i=0}^{p} h_i e^{j\omega_l (i-k)}\right]\right\} = 0 \qquad k = 1, \cdots, p. \tag{61}$$

For large N, however,

$$\sum_{\omega_l} \frac{e^{j\omega_l k}}{\sum_{i=0}^{p} h_i e^{j\omega_l i}} \approx \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{e^{j\omega k}}{\sum_{i=0}^{p} h_i e^{j\omega i}}\, d\omega = 0 \qquad \text{for } k > 0 \tag{62}$$

where the equality to zero can be shown following, e.g., the technique in [19, eqs. (14) and (15)]. Thus, (61) is equivalent to

$$r_k - \sum_{i=1}^{p} h_i r_{i-k} = 0 \qquad k = 1, \cdots, p \tag{63}$$

where r_k is the kth sample correlation coefficient, which is achieved by inverse Fourier transforming the sample power spectrum $|S(\omega)|^2$, i.e.,

$$r_k = \frac{1}{N}\sum_{\omega_l} \left|S(\omega_l)\right|^2 e^{j\omega_l k}. \tag{64}$$

Similarly, taking the derivative with respect to G, we get

$$\sum_{\omega_l}\left[\frac{1}{G} - \frac{\left|S(\omega_l)\right|^2 \sum_{i=0}^{p}\sum_{n=0}^{p} h_i h_n e^{j\omega_l (i-n)}}{G^2}\right] = \frac{1}{G^2}\sum_{\omega_l}\left[G - \sum_{i=0}^{p}\sum_{n=0}^{p} h_i h_n \left|S(\omega_l)\right|^2 e^{j\omega_l (i-n)}\right] = 0 \tag{65}$$

which is equivalent to solving

$$G = \sum_{i=0}^{p}\sum_{n=0}^{p} h_i h_n r_{i-n}. \tag{66}$$

Since $h_0 = -1$ and $\{h_k\}_{k=1}^{p}$ satisfy (63), we get

$$G = r_0 - \sum_{k=1}^{p} h_k r_k. \tag{67}$$

Thus, the AR parameters are estimated by (63) and (67) which are the Yule–Walker equations, using the sample correlation coefficients $\{r_k\}$.

Appendix B
Multichannel Wiener Filter

Suppose that a k-component signal vector $\boldsymbol{s}(t)$ is passing through a given multiinput, multioutput linear system $\mathcal{H}$, generates an m-component output signal which is measured with additive noise, i.e., we measure

$$\boldsymbol{y}(t) = \mathcal{H}\{\boldsymbol{s}(t)\} + \boldsymbol{n}(t). \tag{68}$$

In the frequency domain, we may write

$$\boldsymbol{Y}(\omega) = H(\omega) \cdot \boldsymbol{S}(\omega) + \boldsymbol{N}(\omega) \tag{69}$$

where $\boldsymbol{S}(\omega)$ is the $(1 \times k)$ Fourier transform of $\boldsymbol{s}(t)$, $\boldsymbol{N}(\omega)$ is the $(1 \times m)$ Fourier transform of $\boldsymbol{n}(t)$, $\boldsymbol{Y}(\omega)$ is the $(1 \times m)$ Fourier transform of $\boldsymbol{y}(t)$, and $H(\omega)$ is the $(k \times m)$ frequency response of $\mathcal{H}$.

Suppose the signals are observed throughout the time axis, $-\infty \le n \le \infty$. Also assume that $\boldsymbol{s}(t)$ and $\boldsymbol{n}(t)$ are sample signals from zero-mean stationary Gaussian processes with the power spectra matrices

$$\Phi_s(\omega) = E\{\boldsymbol{S}(\omega)\,\boldsymbol{S}(\omega)^{\dagger}\}, \qquad \Phi_n(\omega) = E\{\boldsymbol{N}(\omega)\,\boldsymbol{N}(\omega)^{\dagger}\}. \tag{70}$$

The minimum mean square estimate of the signal vector $\boldsymbol{s}(t)$ is the noncausal Wiener filter. This Wiener filter is expressed in the frequency domain, and is given by (see

[20, ch. 5])

$$\hat{S}(\omega) = E\{S(\omega)/Y(\omega)\} = \Phi_s(\omega) \cdot H(\omega)^{\dagger}\big(H(\omega)\,\Phi_s(\omega)\,H(\omega)^{\dagger} + \Phi_n(\omega)\big)^{-1} \cdot Y(\omega). \quad (71)$$

Note that for the scalar case and $H(\omega) = 1$, equation (71) reduces to the familiar Wiener filter form

$$\hat{S}(\omega) = \frac{\Phi_s(\omega)}{\Phi_s(\omega) + \Phi_n(\omega)} \cdot Y(\omega).$$

The error covariance matrix is given by (again from [20])

$$\mathcal{P}(\omega) = E\{(S(\omega) - \hat{S}(\omega))(S(\omega) - \hat{S}(\omega))^{\dagger}\} = \big(\Phi^{-1}(\omega) + H(\omega) \cdot \Sigma^{-1} \cdot H(\omega)^{\dagger}\big)^{-1} = \Phi(\omega) - \Phi(\omega) \cdot H(\omega)^{\dagger}\big(H(\omega) \cdot \Phi(\omega) \cdot H(\omega)^{\dagger} + \Sigma\big)^{-1} H(\omega) \cdot \Phi(\omega) \quad (72)$$

which in the scalar case and $H(\omega) = 1$ reduces to the familiar form

$$\mathcal{P}(\omega) = E\{(S(\omega) - \hat{S}(\omega))(S(\omega) - \hat{S}(\omega))^{*}\} = \frac{\Phi_s(\omega) \cdot \Phi_n(\omega)}{\Phi_s(\omega) + \Phi_n(\omega)}.$$

For the two-channel case considered in Section V, with the power spectra and $H(\omega)$ defined as in (52) and (51), the multichannel Wiener filter (71) and (72) reduces to (53) and (55).

References

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

[2] B. Widrow *et al.*, "Adaptive noise cancelling: Principles and applications," *Proc. IEEE*, vol. 63, pp. 1692-1716, 1975.

[3] S. F. Boll and D. C. Pulsipher, "Suppression of acoustic noise in speech using two microphone adaptive noise cancellation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 752-753, 1980.

[4] W. A. Harrison, J. S. Lim, and E. Singer, "A new application of adaptive noise cancellation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 21-27, 1986.

[5] D. Falconer and L. Ljung, "Application of fast Kalman estimation to adaptive equalization," *IEEE Trans. Commun.*, vol. COM-26, pp. 1439-1446, 1978.

[6] R. A. Monzingo and T. W. Miller, *Introduction to Adaptive Arrays*. New York: Wiley, 1980.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc.*, Ser. 39, pp. 1-38, 1977.

[8] J. S. Lim and A. V. Oppenheim, "All pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 197-210, 1978.

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

[10] L. J. Griffiths and C. W. Jim, "Linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagat.*, vol. AP-30, pp. 27-34, 1982.

[11] H. L. Van Trees, *Detection Estimation and Modulation Theory*. New York: Wiley, 1968.

[12] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 1965.

[13] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Stat.*, vol. 11, pp. 95-103, 1983.

[14] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.*, vol. 41, pp. 164-171, 1970.

[15] H. O. Hartley and R. R. Hocking, "The analysis of incomplete data," *Biometrics*, vol. 27, pp. 783-808, 1971.

[16] T. Orchard and M. A. Woodsbury, "A missing information principle: Theory and applications," in *Proc. 6th Berkeley Symp. Math. Stat. Prob.*, 1972, pp. 697-715.

[17] M. Feder, E. Weinstein, and A. V. Oppenheim, "A new class of sequential and adaptive algorithms with application to noise cancellation," in *Proc. 1988 Int. Conf. Acoust., Speech, Signal Processing*, 1988.

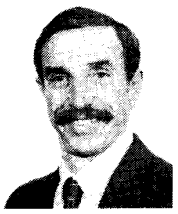
[18] P. M. Peterson, "Simulating the response of multiple microphone to a single acoustic source in a reverberant room," *J. Acoust. Soc. Amer.*, vol. 76, Nov. 1986.

[19] J. A. Edward and M. M. Fitelson, "Notes on maximum-entropy processing," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 232-234, 1973.

[20] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Englewood Cliffs, NJ: Prentice-Hall, 1979.

Meir Feder (S'81-M'87) received the B.Sc. and M.Sc. degrees (summa cum laude) from Tel-Aviv University, Tel-Aviv, Israel, and the Sc.D. degree from the Massachusetts Institute of Technology, Cambridge, and the Woods Hole Oceanographic Institution, Woods Hole, MA, all in electrical engineering, in 1980, 1984, and 1987, respectively.

He is currently a Research Associate at the Department of Electrical Engineering and Computer Science at M.I.T. His research interests lie in the broad area of signal processing, stochastic estimation, and information theory. He is interested in investigating new statistical estimation techniques and using them in signal processing applications, e.g., speech, image, and array processing problems.



Alan V. Oppenheim (S'57-M'65-SM'71-F'77) received the S.B. and S.M. degrees in 1961 and the Sc.D. degree in 1964, all in electrical engineering, from the Massachusetts Institute of Technology, Cambridge.

In 1964 he joined the Faculty at M.I.T. where he is currently Professor of Electrical Engineering and Computer Science. From 1978 to 1980 he was Associate Head of the Data Systems Division at M.I.T. Lincoln Laboratory. Since 1977 he has also been a Guest Investigator at the Woods Hole Oceanographic Institution, Woods Hole, MA. His research interests are in the general area of signal processing and its application to speech, image, and seismic data processing. A current area of emphasis is knowledge-based signal processing. He is coauthor of a text on digital signal processing, coauthor of a text on signals and systems, Editor of a book on applications of digital signal processing, Co-editor of a book entitled *Advanced Topics in Signal Processing*, and Editor of a reprint book entitled *Papers on Digital Signal Processing*. He has been Editor of the Prentice-Hall *Series on Signal Processing* since 1975. He is also author of two video tape lecture series and study guides on signal processing.

Dr. Oppenheim has been a Guggenheim Fellow, a Sackler Fellow, and has held the Cecil H. Green Distinguished Chair in Electrical Engineering and Computer Science. He has also received a number of awards for outstanding research and teaching, including the 1988 IEEE Education Medal, and is an elected member of the National Academy of Engineering. He is also a member of Tau Beta Pi, Eta Kappa Nu, and Sigma Xi.

Ehud Weinstein (M'82-SM'86) received the B.Sc. degree from the Technion-Israel Institute of Technology, and the Ph.D. degree from Yale University, New Haven, CT, both in electrical engineering, in 1975 and 1978, respectively.

He is currently associated with the Department of Electronic Systems, Faculty of Engineering, Tel-Aviv University, Israel, and with the Department of Ocean Engineering, Woods Hole Oceanographic Institute, MA. His research interests are in the general area of parameter estimation and its applications to signal and array processing.

Dr. Weinstein received the IEEE Acoustics, Speech, and Signal Processing Society 1983 Senior Award.