Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithms

Sharon Gannot, *Student Member, IEEE,* David Burshtein, *Member, IEEE,* and Ehud Weinstein, *Fellow, IEEE*

***Abstract*—Speech quality and intelligibility might significantly deteriorate in the presence of background noise, especially when the speech signal is subject to subsequent processing. In particular, speech coders and automatic speech recognition (ASR) systems that were designed or trained to act on clean speech signals might be rendered useless in the presence of background noise. Speech enhancement algorithms have therefore attracted a great deal of interest in the past two decades. In this paper, we present a class of Kalman filter-based algorithms with some extensions, modifications, and improvements of previous work. The first algorithm employs the estimate-maximize (EM) method to iteratively estimate the spectral parameters of the speech and noise parameters. The enhanced speech signal is obtained as a byproduct of the parameter estimation algorithm. The second algorithm is a sequential, computationally efficient, gradient descent algorithm. We discuss various topics concerning the practical implementation of these algorithms. Extensive experimental study using real speech and noise signals is provided to compare these algorithms with alternative speech enhancement algorithms, and to compare the performance of the iterative and sequential algorithms.**

I. Introduction

SPEECH quality and intelligibility might significantly deteriorate in the presence of background noise, especially when the speech signal is subject to subsequent processing. In particular, speech coders and automatic speech recognition (ASR) systems that were designed or trained to act on clean speech signals might be rendered useless in the presence of background noise. Speech enhancement algorithms have therefore attracted a great deal of interest in the past two decades [2], [5]–[8], [11], [13], [15]–[18], [20], [21], [23], [25], [26], [29], [30].

Lim and Oppenheim [20] have suggested modeling the speech signal as a stochastic autoregressive (AR) model embedded in additive white Gaussian noise, and use this model for speech enhancement. The proposed algorithm is iterative in nature. It consists of estimating the speech AR parameters by solving the Yule–Walker equations using the current estimate of the speech signal, and then apply the (noncausal) Wiener filter to the observed signal to obtain a hopefully improved estimate of the desired speech signal. It can be shown that the version of the algorithm that uses the covariance of the speech signal estimate, given at the output of the Wiener filter, is in fact the estimate-maximize (EM) algorithm (up to a scale factor) for the problem at hand. As such, it is guaranteed to converge to the maximum likelihood (ML) estimate of the AR parameters, or at least to a local maximum of the likelihood function, and to yield the best linear filtered estimate of the speech signal, computed at the ML parameter estimate. Hansen and Clements [15] proposed to incorporate auditory domain constrains in order to improve the convergence behavior of the Lim and Oppenheim algorithm, and Masgrau *et al*. [21] proposed to incorporate third-order cumulants in the Yule–Walker equations in order to improve the immunity of the AR parameter estimate to additive Gaussian noise.

Weinstein *et al*. [29] presented a time-domain formulation to the problem at hand. Their approach consists of representing the signal model using linear dynamic state equation, and applying the EM method. The resulting algorithm is similar in structure to the Lim and Oppenheim [20] algorithm, only that the noncausal Wiener filter is replaced by the Kalman smoothing equations. In addition to that, sequential speech enhancement algorithms are presented in [29]. These sequential algorithms are characterized by a forward Kalman filter whose parameters are continuously updated. In [30], similar methods were proposed for the related problem of multisensor signal enhancement. Lee *et al*. [17] extended the sequential single sensor algorithm of Weinstein *et al*. by replacing the white Gaussian excitation of the speech signal with a mixed Gaussian term that may account for the presence of an impulse train in the excitation sequence of voiced speech. Lee *et al*. examined the signal-to-noise ratio (SNR) improvement of the algorithm when applied to synthetic speech input. They also provide limited results on real speech signals.

The use of Kalman filtering was previously proposed by Paliwal and Basu [23] for speech enhancement, where experimental results reveal its distinct advantage over the Wiener filter, for the case where the estimated speech parameters are obtained from the clean speech signal (before being corrupted by the noise). Gibson *et al*. [13] proposed to extend the use of the Kalman filter by incorporating a colored noise model in order to improve the enhancement performances for certain class of noise sources. A disadvantage of the above mentioned Kalman filtering algorithms is that they do not address the model parameters estimation problem. Koo and Gibson [16] suggested an algorithm that iterates between Kalman filtering of the given corrupted speech measurements, and estimation of the speech parameters given the enhanced speech waveform. The resulting algorithm is, in fact, an approximated EM algorithm.

In this paper, we present iterative-batch and sequential algorithms with some extensions, modifications, and improvements

Manuscript received June 22, 1996; revised August 27, 1997. The associate editors coordinating the review of this manuscript and approving it for publication was Dr. Jean-Claude Junqua.

The authors are with the Department of Electrical Engineering–Systems, Tel-Aviv University, Tel-Aviv 69978, Israel (e-mail: burstyn@eng.tau.ac.il).

Publisher Item Identifier S 1063-6676(98)04218-7.

of previous work [29], and discuss various topics concerning the practical implementation of these algorithms. This discussion is supported by extensive experimental study using recorded speech signals and actual noise sources. The outcomes consist of the assessment of sound spectrograms, subjective distortion measures such as total output SNR, segmental SNR, and Itakura–Saito, informal subjective tests, speech intelligibility tests, and ASR experiments. The iterative-batch algorithm is compared to various methods, including spectral subtraction [2], the short-time spectral amplitude (STSA) estimator [5], the log spectral amplitude estimator (LSAE) [6], the hidden Markov model (HMM) based filtering algorithms [7], [8], and the Wiener filter approach of [20]. These algorithms may be distinguished by the amount of *a priori* knowledge that is assumed on the statistics of the clean speech signal. For example, the algorithms in [7] and [8] require a training stage in which the clean speech parameters are estimated, prior to the application of the enhancement algorithm, while the other approaches do not require such training stage.

A distinct advantage of the proposed algorithm compared to alternative algorithms is that it enhances the quality and SNR of the speech, while preserving its intelligibility and natural sound. The sequential algorithm is generally inferior to the iterative-batch algorithm. However, at low SNR's the degradation is usually insignificant.

The organization of the paper is as follows. In Section II, we present the signal model. In Section III, we present the iterative-batch algorithm. In Section IV, we show how higher-order statistics might be incorporated in order to improve the performance of the iterative-batch algorithm. The sequential algorithm is presented in Section V. Experimental results are provided in Section VI. In Sections VII and VIII, we discuss and summarize our results.

II. The Signal Model

Let the signal measured by the microphone be given by

$$z(t) = s(t) + v(t) \tag{1}$$

where $s(t)$ represents the sampled speech signal, and $v(t)$ represents additive background noise.

We shall assume the standard LPC modeling for the speech signal over the analysis frame, in which $s(t)$ is modeled as a stochastic AR process, i.e.,

$$s(t) = -\sum_{k=1}^{p} \alpha_k s(t-k) + \sqrt{g_s}u(t) \tag{2}$$

where the excitation $u(t)$ is a normalized (zero mean unit variance) white noise, g_s represents the spectral level, and $\alpha_1, \cdots, \alpha_p$ are the AR coefficients. We may incorporate the more detailed voiced speech model suggested in [3] in which the excitation process is composed of a weighted linear combination of an impulse train and a white noise sequence to represent voiced and unvoiced speech, respectively. However, as indicated in [11], this approach did not yield any significant performance improvements over the standard LPC modeling.

The additive noise $v(t)$ is also assumed to be a realization from a zero-mean, possibly nonwhite stochastic AR process:

$$v(t) = -\sum_{k=1}^{q} \beta_k v(t-k) + \sqrt{g_v}w(t) \tag{3}$$

where $\beta_1, \cdots, \beta_q$ are the AR parameters of the noise process, and g_v represents its power level. Many of the actual noise sources may be closely approximated as low order, all-pole (AR) processes, in which case a significant improvement may be achieved by incorporating the noise model into the estimation process as indicated in [11] and [13].

Following straight-forward algebra manipulations, (1)–(3) may be represented in the following state-space form:

$$\begin{aligned} \boldsymbol{x}(t) &= \Phi \boldsymbol{x}(t-1) + G\boldsymbol{r}(t) \\ \boldsymbol{z}(t) &= \boldsymbol{h}^T \boldsymbol{x}(t) \end{aligned}$$

where the state vector $\boldsymbol{x}(t)$ is defined by:

$$\boldsymbol{x}^T(t) = [\boldsymbol{s}_{p-1}^T(t-1) \quad s(t) \quad \boldsymbol{v}_{q-1}^T(t-1) \quad v(t)]$$

where

$$\begin{aligned} \boldsymbol{s}_p(t) &= [s(t-p+1) \quad s(t-p+2) \quad \cdots \quad s(t)]^T \\ \boldsymbol{v}_q(t) &= [v(t-q+1) \quad v(t-q+2) \quad \cdots \quad v(t)]^T. \end{aligned}$$

The state transition matrix Φ is given by:

$$\Phi = \begin{bmatrix} \Phi_s & \mathbf{0} \\ \mathbf{0} & \Phi_v \end{bmatrix}$$

where

$$\Phi_s = \begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \\ -\alpha_p & -\alpha_{p-1} & \cdots & \cdots & -\alpha_2 & -\alpha_1 \end{bmatrix}$$

$$\Phi_v = \begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \\ -\beta_q & -\beta_{q-1} & \cdots & \cdots & -\beta_2 & -\beta_1 \end{bmatrix}$$

$$r(t) = [u(t) \quad w(t)]^T,$$

$$G = \begin{bmatrix} \boldsymbol{g}_s & \mathbf{0} \\ \mathbf{0} & \boldsymbol{g}_v \end{bmatrix}$$

where $\boldsymbol{g}_s$ and $\boldsymbol{g}_v$ are the following p and q dimensional vectors:

$$\boldsymbol{g}_s^T = [0 \quad \cdots \quad 0 \quad \sqrt{g_s}] \qquad \boldsymbol{g}_v^T = [0 \quad \cdots \quad 0 \quad \sqrt{g_v}]$$

and

$$\boldsymbol{h}^T = [\boldsymbol{h}_s^T \quad \boldsymbol{h}_v^T]$$

where $\boldsymbol{h}_s$ and $\boldsymbol{h}_v$ are the following p and q dimensional vectors:

$$\boldsymbol{h}_s^T = [0 \quad \cdots \quad 0 \quad 1] \qquad \boldsymbol{h}_v^T = [0 \quad \cdots \quad 0 \quad 1].$$

Assuming that all the signal and noise parameters are known, which implies that $\Phi, \boldsymbol{h}$, and G are known, the optimal minimum mean square error (MMSE) linear state estimate, which includes the desired speech signal $s(t)$, is obtained using the Kalman smoothing equations. However, since the signal and noise parameters are not known *a priori*, they must also be estimated within the algorithm.

III. EM-Based Algorithm

Applying the EM method to the problem at hand, and following the considerations in [27] and [29] (see also [30] that considers the two channel case), we obtain in Appendix A the following algorithm that iterates between state and parameter estimation.

Let $\boldsymbol{\theta}$ be the vector of unknown parameters in the extended model

$$\begin{aligned}\boldsymbol{\theta} &= [\boldsymbol{\alpha}^T \quad g_s \quad \boldsymbol{\beta}^T \quad g_v]^T \\ \boldsymbol{\alpha} &= [\alpha_p \quad \alpha_{p-1} \quad \cdots \quad \alpha_1]^T \\ \boldsymbol{\beta} &= [\beta_q \quad \beta_{q-1} \quad \cdots \quad \beta_1]^T \end{aligned} \tag{4}$$

and let $\hat{\boldsymbol{\theta}}^{(l)}$ be the estimate of $\boldsymbol{\theta}$ after l iterations of the algorithms. Finally, let

$$\boldsymbol{z} = [z(1) \quad z(2) \quad \cdots \quad z(N)]^T \tag{5}$$

be the vector of observed data in the current analysis frame. We will use the notation

$$\widehat{(\cdot)} = E_{\hat{\boldsymbol{\theta}}^{(l)}}\{\cdot|\boldsymbol{z}\}. \tag{6}$$

To obtain $\hat{\boldsymbol{\theta}}^{(l+1)}$, we use the following two-stage iteration.

1) State Estimation (E-step): Define by

$$\begin{aligned}\boldsymbol{\mu}(t|N) &= \widehat{\boldsymbol{x}(t)} \\ P(t|N) &= \widehat{\boldsymbol{x}(t)\boldsymbol{x}^T(t)} - \widehat{\boldsymbol{x}(t)}\,\widehat{\boldsymbol{x}(t)}^T\end{aligned}$$

i.e., $\boldsymbol{\mu}(t|N)$ represents the current state estimate based on $z(t) t = 1, \cdots, N$, and $P(t|N)$ represents the associated error covariance matrix.

Then $\boldsymbol{\mu}(t|N)$ and $P(t|N)$ are computed using a forward, Kalman filtering recursion, followed by a backward Kalman smoothing recursion, as follows.

- *Forward (filtering) recursion*:

For $t = 1, 2, \cdots, N$

Propagation Equations

$$\boldsymbol{\mu}(t|t-1) = \hat{\Phi}\boldsymbol{\mu}(t-1|t-1) \tag{7}$$

$$P(t|t-1) = \hat{\Phi}P(t-1|t-1)\hat{\Phi}^T + \hat{G}\hat{G}^T \tag{8}$$

Updating Equations

$$\boldsymbol{\mu}(t|t) = \boldsymbol{\mu}(t|t-1) + \boldsymbol{k}(t)[z(t) - \boldsymbol{h}^T\boldsymbol{\mu}(t|t-1)] \tag{9}$$

$$P(t|t) = P(t|t-1) - \boldsymbol{k}(t)\boldsymbol{h}^T P(t|t-1) \tag{10}$$

where

$$\boldsymbol{k}(t) = \frac{P(t|t-1)\boldsymbol{h}}{\boldsymbol{h}^T P(t|t-1)\boldsymbol{h}}.$$

- *Backward (smoothing) recursion:* For $t = N, N-1, \cdots, 1$

$$\begin{aligned}\boldsymbol{\mu}(t-1|N) = {} & \boldsymbol{\mu}(t-1|t-1) + S(t-1) \\ & \cdot (\boldsymbol{\mu}[t-1|N] - \hat{\Phi}\boldsymbol{\mu}(t-1|t-1))\end{aligned} \tag{11}$$

$$\begin{aligned}P(t-1|N) = {} & P(t-1|t-1) - S(t-1) \\ & \cdot (P(t|N) - P(t|t-1))S(t-1)^T\end{aligned} \tag{12}$$

where

$$S(t-1) = P(t-1|t-1)\hat{\Phi}^T(P(t|t-1))^{-1}$$

where $\hat{\Phi}$ and $\hat{G}$ are the matrices Φ and G, respectively, computed using the current parameter estimates.

2) Parameter Estimation (M-step):

$$\hat{\boldsymbol{\alpha}}^{(l+1)} = -\left[\sum_{t=1}^{N} \widehat{\boldsymbol{s}_p(t-1)\boldsymbol{s}_p^T(t-1)}\right]^{-1} \sum_{t=1}^{N} \widehat{\boldsymbol{s}_p(t-1)s(t)} \tag{13}$$

$$\hat{g}_s^{(l+1)} = \frac{1}{N}\sum_{t=1}^{N}[\widehat{s^2(t)} + (\hat{\boldsymbol{\alpha}}^{(l+1)})^T \widehat{\boldsymbol{s}_p(t-1)s(t)}] \tag{14}$$

$$\hat{\boldsymbol{\beta}}^{(l+1)} = -\left[\sum_{t=1}^{N} \widehat{\boldsymbol{v}_q(t-1)\boldsymbol{v}_q^T(t-1)}\right]^{-1} \sum_{t=1}^{N} \widehat{\boldsymbol{v}_q(t-1)v(t)} \tag{15}$$

$$\hat{g}_v^{(l+1)} = \frac{1}{N}\sum_{t=1}^{N}[\widehat{v^2(t)} + (\hat{\boldsymbol{\beta}}^{(l+1)})^T \widehat{\boldsymbol{v}_q(t-1)v(t)}]. \tag{16}$$

We note that $\widehat{\boldsymbol{s}_p(t-1)\boldsymbol{s}_p^T(t-1)}$ is the upper left $p \times p$ submatrix of $\widehat{\boldsymbol{x}(t)\boldsymbol{x}^T(t)}$. $\widehat{\boldsymbol{s}_p(t-1)s(t)}$, $\widehat{s^2(t)}$, $\widehat{\boldsymbol{v}_q(t-1)\boldsymbol{v}_q^T(t-1)}$, $\widehat{\boldsymbol{v}_q(t-1)v(t)}$, and $\widehat{v^2(t)}$ may similarly be extracted from $\widehat{\boldsymbol{x}(t)\boldsymbol{x}^T(t)}$.

Note that (13) and (15) are similar to the standard Yule–Walker (YW) solution for estimating the coefficients of an AR process, except that the covariance is replaced by its *a posteriori* value.

Since the algorithm is based on the EM method, it is guaranteed to converge monotonically to the ML estimate of all unknown parameters (under Gaussian assumptions), or at least to a local maximum of the likelihood function, where each iteration increases the likelihood of the estimate of the parameters. As a byproduct, it yields the optimal linear state (signal) estimate, computed using the estimated parameters.

This algorithm is an extension of the algorithm presented in [29] for the case in which the additive noise is modeled more generally as a colored AR process. Since the signal and the noise parameter estimates are computed separately within the algorithm, the increase in computational complexity is quite moderate. However, the realizable improvement in the enhancement performance may be quite significant, as indicated in [11] and [13].

In order to reduce the computations involved, we suggest to replace the full smoothing operation with fixed-lag smooth-

ing (delayed Kalman filter estimate) [23] or even just by filtering. That is, instead of using $\hat{s}(t|N)$, the pth entry of $\boldsymbol{\mu}(t|N)$, as the enhanced signal estimate, it is proposed to use $\hat{s}(t-p+1|t)$ (fixed lag smoothing) or $\hat{s}(t|t)$ (filtering), that are the first and the pth entries of $\boldsymbol{\mu}(t|t)$, respectively. Similar observations apply to the enhanced noise estimate, used by the EM algorithm. With these modifications, we do not need to apply the smoothing (11) and (12) that are computationally expensive. As indicated in [11], the resulting algorithm still maintains its nice monotonic convergence behavior.

A simplified EM algorithm may be obtained by iteratively estimating the speech parameters using the enhanced speech signal (by employing the ordinary YW equation set), and then using these parameters to improve the estimate of the enhanced signal (the noise parameters are estimated, using signal segments at which voice activity is assumed not to be present). This simplified EM algorithm was suggested by Koo *et al.* [16]. We found that unlike the EM algorithm, which is guaranteed to be stable and to monotonically increase the likelihood function, the simplified EM algorithm does not possess such properties. The simplified EM algorithm results in performance degradation, which is very significant at the lower SNR range. Similar behavior was noticed by Lim and Oppenheim [20] in the context of an iterative Wiener filter algorithm for the enhancement of speech in the presence of white Gaussian noise.

IV. Parameter Estimation Using Higher-Order Statistics

To obtain a reliable estimate of the speech signal, it is essential to have a powerful initialization algorithm for the speech and noise parameters. Otherwise, the estimation algorithm might converge to a local minimum of the likelihood function. When the SNR is high, an initial estimate of the speech parameters may be obtained using standard LPC processing, and an initial estimate of the noise parameters may be obtained by employing a voice activity detector, so that the noise statistics are accumulated during silence periods.

Unfortunately, this initialization procedure breaks down at low SNR conditions, below 5 dB in our experiments. However, if the additive noise $v(t)$ is assumed to be Gaussian, then higher-order statistics (HOS) may be incorporated in order to improve the initial estimate of the speech parameters. In that case, the quality of the enhanced speech signal is significantly improved compared to the standard initialization method that was indicated above.

It can be shown, by invoking the basic cumulant properties in [22, Sect. II-B3], and recalling (2), that

$$\begin{aligned}&\operatorname{cum}(s(t), s(t-l_1), s(t-l_2), \cdots, s(t-l_M))\\ &\quad = -\sum_{k=1}^{p} \alpha_k \operatorname{cum}(s(t-k), s(t-l_1), s(t-l_2), \cdots,\\ &\qquad s(t-l_M)) \end{aligned} \tag{17}$$

whenever $M \geq 1$, where $\operatorname{cum}(\cdot,\cdot,\cdots)$ denotes the joint cumulant of the bracketed variables. We note that

$$\begin{aligned}&\operatorname{cum}(z(t)) = E\{z(t)\}\\ &\operatorname{cum}(z(t), z(t-l_1)) = E\{z(t)z(t-l_1)\}\\ &\operatorname{cum}(z(t), z(t-l_1), z(t-l_2))\\ &\qquad = E\{z(t)z(t-l_1)z(t-l_2)\}\\ &\operatorname{cum}(z(t), z(t-l_1), z(t-l_2), z(t-l_3))\\ &\qquad = E\{z(t)z(t-l_1)z(t-l_2)z(t-l_3)\}\\ &\qquad\quad - E\{(z(t)z(t-l_1)\}\cdot E\{z(t-l_2)z(t-l_3)\}\\ &\qquad\quad - E\{(z(t)z(t-l_2)\}\cdot E\{z(t)z(t-l_3)\}\\ &\qquad\quad - E\{(z(t)z(t-l_3)\}\cdot E\{z(t-l_1)z(t-l_2)\}.\end{aligned}$$

A general formula for expressing cumulants in terms of moments can be found in [22].

Now, under the assumption that $v(t)$ is Gaussian, it can be shown, by invoking the same basic cumulant properties in [22] and recalling (1) and the statistical independence of $v(t)$ and $s(t))$, that

$$\begin{aligned}&\operatorname{cum}(z(t), z(t-l_1), z(t-l_2), \cdots, z(t-l_M))\\ &\quad = -\sum_{k=1}^{p} \alpha_k \operatorname{cum}(z(t-k), z(t-l_1), z(t-l_2), \cdots,\\ &\qquad z(t-l_M))\end{aligned}$$

whenever $M \geq 2$. For $M = 1$, we obtain the standard Yule–Walker equations based on second-order statistics. However, in this case the equations do not hold because of the contribution of the additive noise, and this is why the parameter initialization breaks down at low SNR. For $M \geq 2$, we obtain additional Yule–Walker type equations that are insensitive to the presence of additive Gaussian noise. These equations appear to be very useful if the additive noise is "more Gaussian" than the speech signal in the sense that its higher-order cumulants are relatively small in magnitude.

In practice the cumulants are approximated by substituting the unavailable ensemble averages with sample averages, thus obtaining a set of linear equations that may be used to compute the AR parameters $\alpha_1, \cdots, \alpha_p$ directly from the observed signal $z(t)$. The spectral level g_s of the excitation signal may be computed by applying a whitening filter using the estimated AR parameters.

Since the equations are satisfied for all $M \geq 2$ and any combinations of lags $l_1, l_2, \cdots, l_M$, we have an overdetermined set of equations that may be used to improve numerical and statistical stability of the resulting parameters estimates.

Experimental results using actual speech signal in several typical noise environments indicated that at low SNR conditions, below 5 dB, using fourth-order cumulants $(M = 3)$, one typically obtains a better and more robust initial estimate of the speech parameters as compared with the conventional LPC approach based on second-order statistics. The use of third-order cumulants $(M = 2)$, as suggested in [24], was not that effective.

We also tried to incorporate HOS into the iterative algorithm, and not merely as an initialization tool. For that purpose, consider (17) for $M \geq 1$. By using the cumulants

of the enhanced speech signal, $\hat{s}(t)$, in (17) we obtain an iterative algorithm that employs HOS to iteratively estimate the speech AR parameters. However, experiments showed that the resulting algorithm produces low-quality enhanced speech, with reduced bandwidth formants. Similar observation was noted by Masgrau *et al*. [21], when incorporating third-order statistics into the approximated EM, Wiener filter-based algorithm of [20].

V. Sequential Algorithm

The iterative-batch EM algorithm requires the use of an analysis window over which the signal and noise statistics are assumed to be stationary. To avoid this assumption, we now suggest a sequential speech enhancement algorithm that is no longer an EM algorithm. The resulting sequential algorithm is more computationally efficient than the iterative-batch algorithm. Another benefit of the sequential algorithm is that it is delayless, unlike the iterative-batch algorithm that has an inherent delay of one processing window frame.

Following the considerations in [29] (see also [30], which considers the two-channel case) we obtain in Appendix B the following sequential speech enhancement algorithm.

This algorithm consists of a forward Kalman filter, given by (7)–(10), whose parameters are continuously up-dated according to

$$\hat{\boldsymbol{\alpha}}(t+1) = \hat{\boldsymbol{\alpha}}(t) - \frac{\rho_s}{g_s}[Q_{12}^s(t) + Q_{11}^s(t)\hat{\boldsymbol{\alpha}}(t)] \tag{18}$$

$$\hat{g}_s(t+1) = \frac{1-\lambda_s}{1-\lambda_s^t}[Q_{22}^s(t) + \hat{\boldsymbol{\alpha}}^T(t)Q_{12}^s(t)] \tag{19}$$

$$\hat{\boldsymbol{\beta}}(t+1) = \hat{\boldsymbol{\beta}}(t) - \frac{\rho_v}{g_v}[Q_{12}^v(t) + Q_{11}^v(t)\hat{\boldsymbol{\beta}}(t)] \tag{20}$$

$$\hat{g}_v(t+1) = \frac{1-\lambda_v}{1-\lambda_v^t}[Q_{22}^v(t) + \hat{\boldsymbol{\beta}}^T(t)Q_{12}^v(t)] \tag{21}$$

where $Q^s(t), Q^v(t)$ are defined by

$$\begin{aligned} Q^s(t) &= \begin{bmatrix} Q_{11}^s(t) & Q_{12}^s(t) \\ Q_{21}^s(t) & Q_{22}^s(t) \end{bmatrix} = \sum_{\tau=1}^{t} \lambda_s^{t-\tau} \widehat{\boldsymbol{s}_{p+1}(\tau)\boldsymbol{s}_{p+1}^T(\tau)} \\ &= \lambda_s Q^s(t-1) + \widehat{\boldsymbol{s}_{p+1}(t)\boldsymbol{s}_{p+1}^T(t)} \\ Q^v(t) &= \begin{bmatrix} Q_{11}^v(t) & Q_{12}^v(t) \\ Q_{21}^v(t) & Q_{22}^v(t) \end{bmatrix} = \sum_{\tau=1}^{t} \lambda_v^{t-\tau} \widehat{\boldsymbol{v}_{q+1}(\tau)\boldsymbol{v}_{q+1}^T(\tau)} \\ &= \lambda_v Q^v(t-1) + \widehat{\boldsymbol{v}_{q+1}(t)\boldsymbol{v}_{q+1}^T(t)}. \end{aligned}$$

$Q_{11}^s(t)$ is a $p \times p$ matrix, $Q_{12}^s(t) = (Q_{21}^s(t))^T$ is a $p \times 1$ matrix, and $Q_{22}^s(t)$ is a scalar value. Similarly, $Q_{11}^v(t)$ is a $q \times q$ matrix, $Q_{12}^v(t) = (Q_{21}^v(t))^T$ is a $q \times 1$ matrix, and $Q_{22}^v(t)$ is a scalar value.

$0 \le \lambda_s, \lambda_v \le 1$ are exponential weighting factors and ρ_s, ρ_v are the update stepsizes.

An improvement in the convergence behavior of the algorithm is obtained by normalizing the stepsizes, i.e., using

$$\begin{aligned} \hat{\boldsymbol{\alpha}}(t+1) &= \hat{\boldsymbol{\alpha}}(t) - \rho_s \frac{Q_{12}^s(t) + Q_{11}^s(t)\hat{\boldsymbol{\alpha}}(t)}{||Q_{12}^s(t) + Q_{11}^s(t)\hat{\boldsymbol{\alpha}}(t)||} \\ \hat{\boldsymbol{\beta}}(t+1) &= \hat{\boldsymbol{\beta}}(t) - \rho_v \frac{Q_{12}^v(t) + Q_{11}^v(t)\hat{\boldsymbol{\beta}}(t)}{||Q_{12}^v(t) + Q_{11}^v(t)\hat{\boldsymbol{\beta}}(t)||}. \end{aligned}$$

VI. Experiments

In order to evaluate the performances of the proposed Kalman-EM-iterative (KEMI) algorithm and Kalman-gradient descent-sequential (KGDS) algorithm, both objective and subjective tests were conducted. The performances of these algorithms were compared with the following algorithms:

1) the log spectral amplitude estimator (LSAE), suggested by Ephraim and Malah [6], which is an improvement of the short-time spectral amplitude (STSA) estimator [5];
2) the HMM-based speech enhancement algorithms suggested by Ephraim *et al*. [7], [8];
3) the spectral subtraction algorithm suggested by Boll [2];
4) the Wiener-EM (WEM) algorithm of Lim and Oppenheim [20].

In the experiments that we describe below, five iterations were required for the KEMI algorithm to converge. The AR order used to model the speech signal was ten, and the AR order used to model the noise signal was four. The frame size was 16 ms, although small changes in this value did not degrade the performance. The analysis frames were nonoverlapping (overlapping frames did not yield improved performance). The LSAE algorithm was implemented with a decision directed *a priori* SNR estimation. The HMM algorithm was based on the minimum mean square error (MMSE) criterion, since it was reported to be superior over the alternative maximum *a posteriori* (MAP) criterion [8]. The WEM algorithm was the RLMAP variant in [20], since it yielded the best results.

Both Hebrew and English sentences uttered by male and female speakers were used in our experiments. The sampling rate was 8 kHz. The speech signal was degraded by additive noise, at various SNR's. Various recorded noise sources, including refrigerator, air conditioner and computer fan, were considered. All were found to obey the Gaussian assumption with a good degree of approximation. The computer fan noise is typical of an office environment. Its slowly time-varying short-term spectrum can be modeled adequately by an AR process of order $q = 4$. In Fig. 1, we assess the validity of the Gaussian approximation, by plotting the empirical cumulative distribution function (CDF) of both the computer fan noise and speech samples over a segment of 50 ms. The vertical axis employs a nonlinear division of the interval [0, 1], such that the vertical coordinate of a sample with CDF c, is $y = G(c)$, where G is the CDF of a Gaussian random variable, whose mean and variance are the empirical mean and variance of the given signal segment. Hence, a Gaussian random variable corresponds to a straight line (presented by a dashed line in the figure). As can be seen, the noise segment shown (which is typical of that noise signal) is very close to an ideal Gaussian curve. On the other hand, the voiced segment shown (unvoiced speech segments possess similar Gaussian curves) deviates significantly from the Gaussian curve. More precisely, for the noise segment shown, the deviation from the Gaussian curve is significant only for data samples with cumulative probability value higher than 99% or lower than 3%. On the other hand, for the speech segment shown, the deviation from the Gaussian curve is already significant for

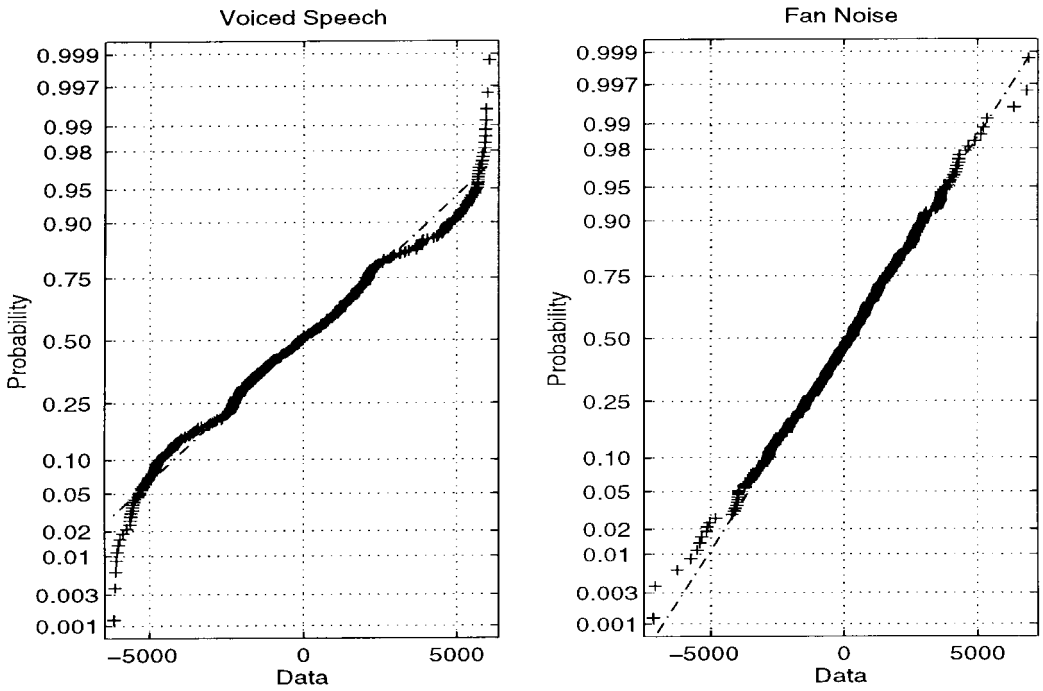

Fig. 1. Gaussian curves for typical speech and computer fan noise segments.

data samples whose cumulative probability is higher than 80% or lower than 20%.

Our informal speech quality test involved ten listeners. Each considered 40 Hebrew and English sentences. Some of the English sentences were taken from the TIMIT data base [12]. The rest were recorded in a silent environment. The speech signal was corrupted by several noise signals at various SNR's. Each listener had to characterize the quality of the enhanced speech and to compare it with the quality of the corrupted one. Each listener examined each enhanced speech signal, without knowing which signal corresponds to which algorithm. All listeners indicated that the quality of the speech processed by the KEMI algorithm is superior to the quality of the corrupted speech at the entire SNR range examined (between -10 and $+15$ dB). They also indicated a significant reduction in the noise level without any severe distortion of the speech signal.

At SNR values below 5 dB, the speech quality of the KGDS algorithm is only slightly inferior to that of the KEMI. Above 5 dB the KGDS algorithm was sometimes unstable, with a time varying signal level. We attribute this phenomenon to the fact that at high SNR's the estimated noise model parameters might be very inaccurate (since the noise is masked by the signal). A possible solution is to replace the sequential update equation of the noise parameters by an estimator that, based on a voice activity detector, considers only signal segments where speech activity is not detected.

A comparison between the filtered output and the fixed-lag smoothed output showed a slight advantage to the former (although the fixed-lag smoothed output was sometimes characterized as being slightly muffled).

The spectral subtraction algorithm significantly reduced the noise level, but generated an annoying "musical noise" effect, i.e., the enhanced speech contained tones with fast shifting frequencies. The algorithm collapses at SNR values below -5 dB.

Both HMM-MMSE and LSAE algorithms showed the most significant noise reduction, but were also characterized by a noticeable distortion of the natural sound of the speech signal. However, the speech distortion of the HMM-MMSE algorithm is much more noticeable, especially at SNR values lower than 0 dB. This observation was also noted in [8]. At SNR values lower than -10 dB, the intelligibility of the speech, processed by the HMM-MMSE algorithm, is severely damaged.

All these observations were valid both for a synthetic white Gaussian noise and for a recorded fan noise signal. The WEM algorithm was designed under the assumption of white Gaussian noise, hence was not tested for colored-noise environments. Our listening tests indicate some advantage to the KEMI algorithm over the WEM algorithm.

Fig. 2 shows the spectrograms of some clean speech segment (upper left), and the corresponding noisy segment (upper right), enhanced KEMI (fixed lag smoothing version) signal (lower left), and enhanced LSAE signal (lower right). As can be seen, the LSAE algorithm shows better noise reduction, at the expense of larger distortion of the speech signal, which is expressed by formant widening.

Intelligibility tests were also conducted for the KEMI algorithm. The clean speech database was the high quality connected digits recorded at TI, (TIDIGITS) [19] by 225 adult female and male speakers. Using the TIDIGITS data base, we created two new databases, each consisting of 200 utterances of isolated digits. The first database consisted of digit utterances that were corrupted by additive computer fan noise, at SNR level of -15 dB. The second data base was created by enhancing each noisy utterance, using the KEMI

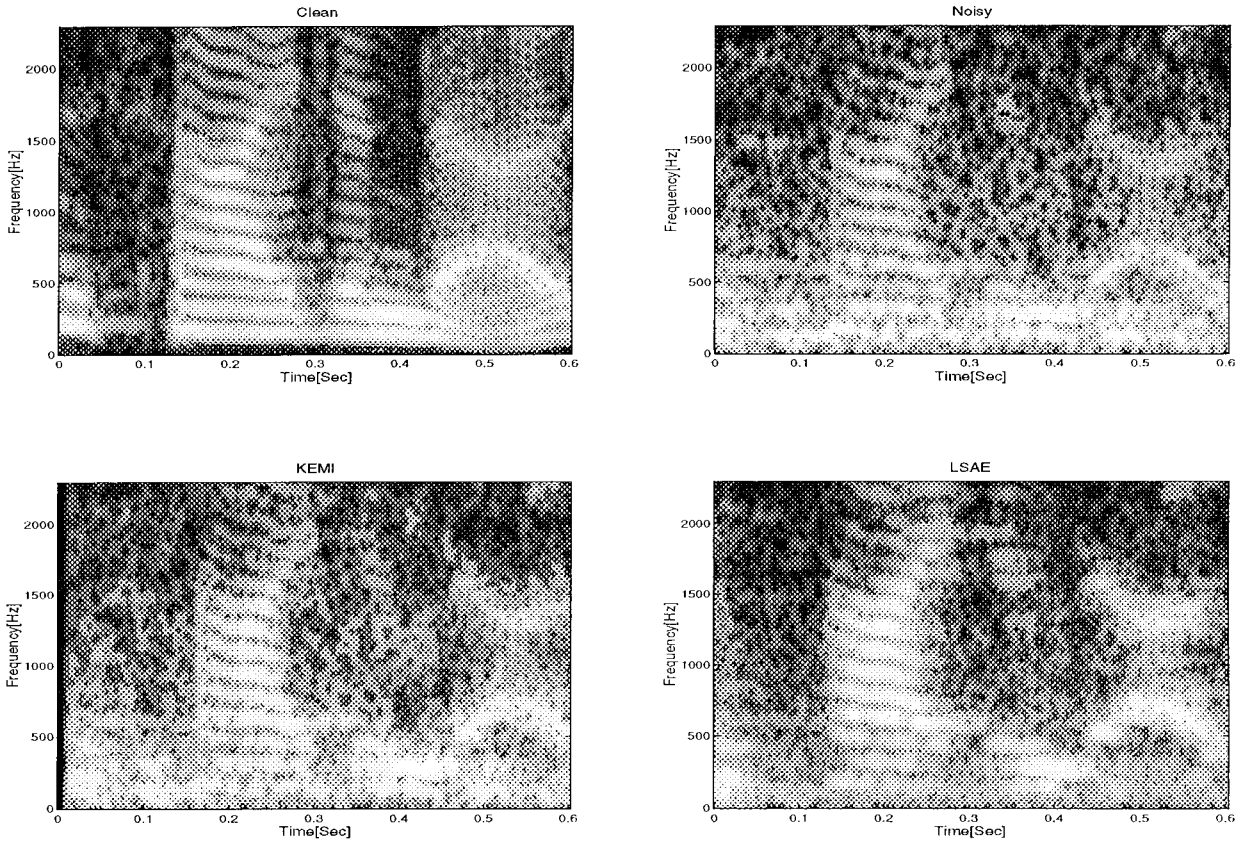

Fig. 2. Clean, noisy and enhanced (KEMI and LSAE) sound spectrograms.

algorithm. Twelve listeners participated in the experiment. The recognition rate of each listener was evaluated by considering 20 noisy utterances and 20 enhanced utterances. All 40 utterances were randomly selected, and were presented to the listener at a random order. The overall word error rate was slightly decreased from 31% (74 wrong decisions) without speech enhancement, to 26% (63 wrong decisions) when preprocessing the speech signal, using the KEMI algorithm.

Our objective set of experiments consisted of total output SNR measurements, segmental SNR and Itakura–Saito (IS) distortion measurements. These distortion measures are known to be correlated with the subjective perception of speech quality [14]. The IS distortion measure and the segmental SNR possess the highest correlation, with a small advantage to the former. Let $s(t)$ and $\hat{s}(t)$ denote the clean and enhanced speech signals, respectively. The total output SNR is defined by

$$\mathrm{SNR} = \frac{\sum_t s^2(t)}{\sum_t (s(t) - \hat{s}(t))^2} \tag{22}$$

where the time summations are over the entire duration of the signals. To obtain the segmental output SNR we used the median value of the individual SNR measurements [using (22)] of all the frames of the signal. Median averaging eliminates outliers, and is therefore superior to the common definition of segmental SNR that involves simple averaging. Similarly, to obtain a representative IS distance measure for the enhanced signal, we calculated the median of the IS measurements of all the frames of the signal. The data base consisted of four sentences uttered by two female (1, 3) and two male speakers (2, 4) both in Hebrew (1, 3) and in English (2, 4). The duration of each of the first three sentences was 25 s. The duration of the last sentence was 5 s.

Fig. 3 shows the total SNR of the various enhanced signals for all sentences combined. The noise source was the fan noise signal. The algorithms that were examined were the proposed KEMI algorithm (the fixed lag variant is presented. The filtered variant was slightly inferior), the KGDS algorithm, the LSAE algorithm, and the HMM-MMSE algorithm. Fig. 4 shows the same data for white Gaussian noise, and the KEMI, HMM-MMSE, and WEM algorithms. We have also evaluated the segmental SNR of the various enhanced signals. However, the conclusions from the segmental SNR experiments are very similar to the conclusions from the total SNR experiments. Hence, only total SNR results are provided. Fig. 5 presents the median IS distance of the various enhanced signals for each sentence contaminated by fan noise. Fig. 3 demonstrates that above 0 dB input SNR (fan noise), the LSAE algorithm is superior to the KEMI algorithm by between 1–2 dB enhancement. Fig. 4 shows that above 0 dB input SNR (white noise), the HMM-MMSE algorithm improves the KEMI algorithm by up to 2.5 dB. In this region, the KEMI algorithm improves the WEM algorithm by up to 2.5 dB. Our total SNR results for the HMM-MMSE algorithm agree with the results that are reported in [8] and [25]. At the lower input SNR range (below

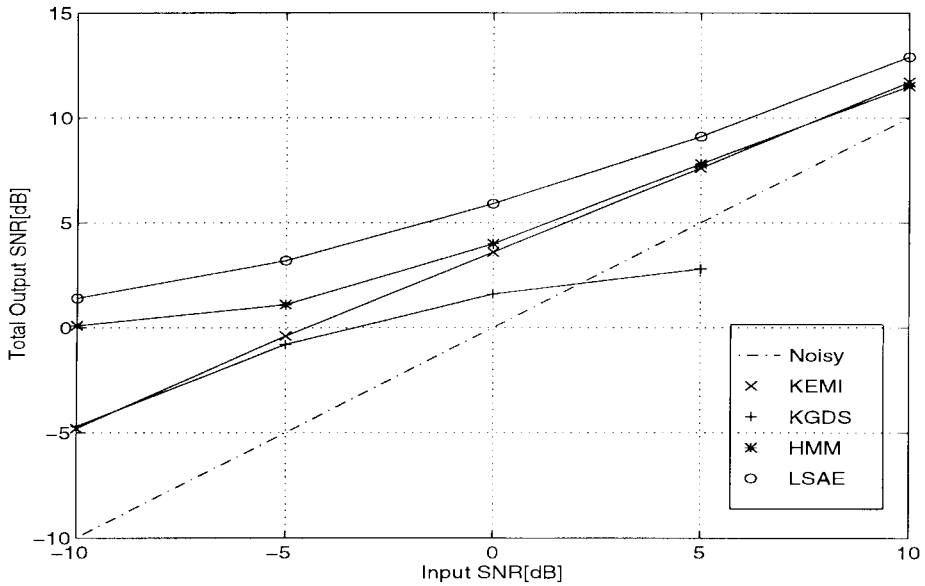

Fig. 3 Total SNR level averaged over the four sentences (fan noise).

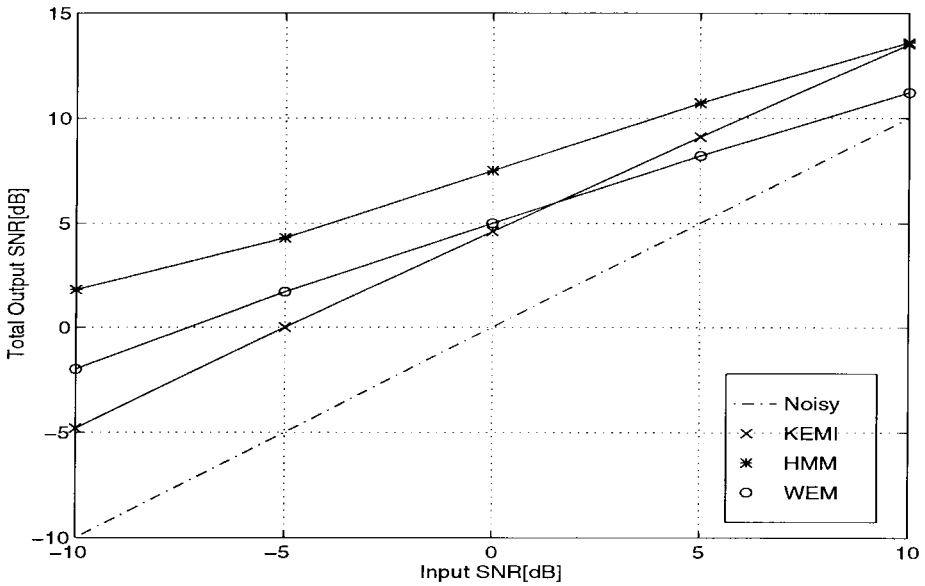

Fig. 4. Total SNR level averaged over the four sentences (white noise).

0 dB), SNR improvement figures might be misleading (e.g., an algorithm that produces a constant zero signal, improves SNR). In fact, below 0 dB, the distortion of the HMM-MMSE algorithm is very significant. The distortion of the LSAE algorithm in that region is less severe, but the enhanced signal sounds unnatural. On the other hand, the KEMI algorithm improves SNR over the full range of input SNR values, without affecting the intelligibility and natural sound of the speech. The IS distance measure results in Fig. 5 show an advantage to the KEMI algorithm over the full range of input SNR's. The IS values of the HMM algorithm were much worse (i.e., higher) compared to the other algorithms, and are therefore not presented. At SNR's below -5 dB, the performances of the KEMI and KGDS algorithms are essentially identical. Note that some of the measurements of the KGDS algorithm at input SNR equal to ten are missing. This is due to the possible unstable behavior of the KGDS algorithm at high SNR's.

ASR experiments were conducted using a continuous density HMM-based speech recognition system. The acoustic front end is comprised of eight cepstral values and their time derivatives, computed by using standard LPC analysis. The resulting 16-dimensional feature vector is modeled by a mixture of three diagonal covariance Gaussians. Each word in the vocabulary was modeled by one, five-state, left-to-right HMM. Training was performed using the Baum–Welch algorithm. The decoder was a conventional Viterbi algorithm. The speech data base was the speaker-independent, high-quality connected digits recorded at TI [19]. This data base is divided into training and testing digit strings uttered by 225 adult talkers. The single digits (word) recognition rate of the

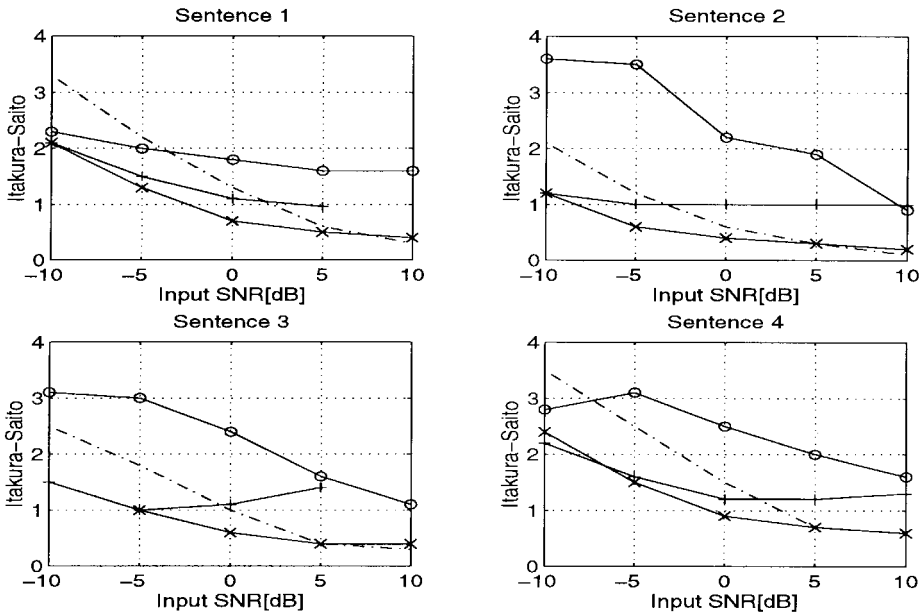

Fig. 5. Median Itakura–Saito distortion measure (the same legend of Fig. 3, except for HMM).

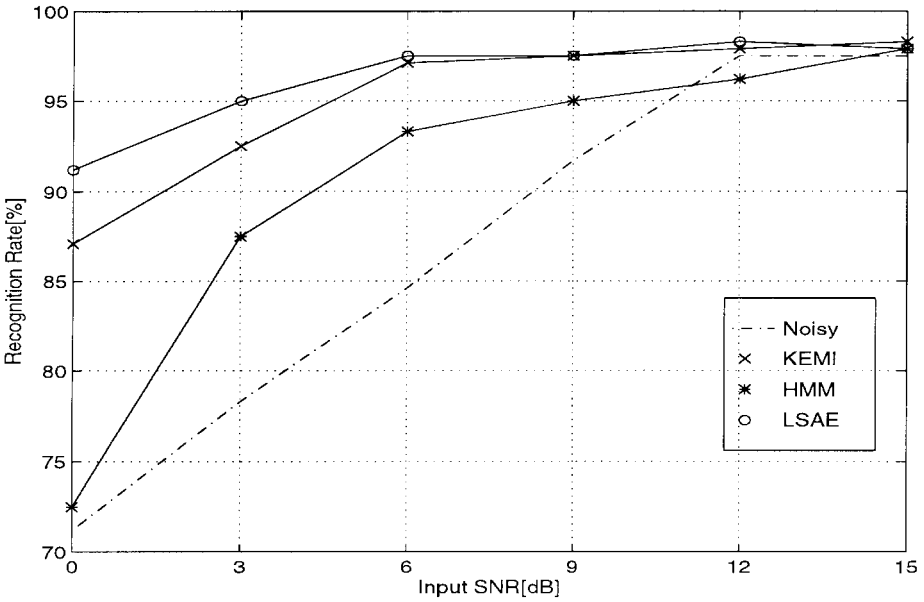

Fig. 6. Single digits recognition rate with preprocessing (KEMI, HMM, and LSAE) and without.

system, when tested on the clean isolated digits sentences, was 99.1%.

The single digits recognition rate of the system when subject to speech signals contaminated by computer fan noise at various SNR's (The SNR was measured in the frequency region of interest, between 200 and 3200 Hz) is summarized in Fig. 6. We also show the corresponding recognition rate, when the noisy speech is preprocessed by the KEMI algorithm (fixed-lag smoothing version; the filtered version was slightly inferior), by the LSAE algorithm, and by the HMM-MMSE algorithm. The main purpose of this comparison is to compare the performance of the enhancement algorithms using an additional task. Alternative noise adaptation algorithms [10], [28] that adapt the parameters of the recognizer, instead of preprocessing the input speech, cannot produce an enhanced speech signal. Hence, the performance of these algorithms was not evaluated. As can be seen, the KEMI algorithm improves the performance by about 6 dB (i.e., when speech enhancement is not employed, the SNR needs to be increased by 6 dB in order to obtain the same recognition rate). The KEMI algorithm shows superior performance compared to the HMM-MMSE algorithm, and is comparable to the LSAE at input SNR's higher than 6 dB. Below 6 dB input SNR, the LSAE is superior to the KEMI algorithm.

VII. Discussion

The proposed Kalman filter-based enhancement algorithms use parallel AR models to describe the dynamics of the speech

and noise signals. The combined model is used to decompose the noisy signal into speech and noise components. HMM-based enhancement algorithms also use parallel models, one set for each of the components into which the signal is to be decomposed (although usually the noise model degenerates to a single state HMM). The main difference between the two approaches is that the HMM-based methods constrain the estimated speech (and noise) parameters to some codebook of possible spectra that is obtained from some clean speech data base. This codebook is in fact a detailed model to the speech signal. The success of the HMM-based methods, depends on the accuracy of this model. A mismatch between the data base used to construct the speech codebook and the actual speech signal that needs to be enhanced might deteriorate the quality of the enhanced signal. For example, speaker-dependent applications might be more successful than speaker-independent applications when using the HMM-based methods. The LSAE algorithm shows improved performance compared to HMM-MMSE, especially at the low SNR range. We attribute this phenomenon to the fact that LSAE (also STSA) tends to attenuate the noisy speech signal less than Wiener filtering does, especially at the lower SNR range (in addition to that, recall that HMM-MMSE is a weighted combination of Wiener filters).

HMM-based signal decomposition has also been used to enhance the robustness of HMM-based speech recognition systems [10], [28]. In this case, parallel HMM model combination is used to transform the probability distribution of the clean speech signal into the probability distribution of the noisy signal.

The Kalman filter-based algorithms employ the MMSE criterion, although other criteria such as minimax (H_∞) are also possible in the enhancement stage of the algorithm. In fact, in [26], a Kalman filter based on the minimax criterion is shown to be superior over a standard (MMSE) Kalman filter. It should be noted, however, that the implementation of the minimax criterion for the parameter estimation stage of the algorithm seems to be much more complicated.

The proposed algorithms use Kalman filtering instead of the Wiener filter approach of Lim and Oppenheim [20]. The advantage of Kalman filtering compared to Wiener filtering is attributed to the fact that the Kalman filter approach enables accurate modeling of the nonstationary transitions at frame boundaries. The advantage of the Kalman filter compared to the Wiener filter was previously noted in [23], for the case where the estimated speech parameters are obtained from the clean speech signal (e.g., 4.5 dB improvement in segmental SNR for speech contaminated by white noise at 0 dB input SNR, and 5.5 dB improvement when using fixed lag Kalman smoothing).

In [18], the HMM-based enhancement in [7] and [8] that employs parallel Wiener filters was modified by replacing Wiener filtering by Kalman filtering.

An important feature of the proposed algorithm is that, unlike the alternative algorithms that were examined, a voice activity detector (VAD) is not required. In fact, the implementation of these alternative algorithms in Section VI, assumed an ideal VAD that produces the true noise spectrum. In that sense, the comparison was biased in favor of these competing algorithms. This advantage of the proposed algorithm is especially significant for noise sources with fast changing spectrum, since a VAD is less useful in that case. It is also significant for the lower SNR region, where it is more difficult to construct reliable VAD's.

VIII. Summary

We presented iterative-batch and sequential speech enhancement algorithms in the presence of colored background noise, and compared the performance of these algorithms with alternative speech enhancement algorithms. The iterative-batch algorithm employs the EM method to estimate the spectral parameters of the speech signal and noise process. Each iteration of the algorithm is composed of an estimation (E) step and a maximization (M) step. The E-step is implemented by using the Kalman filtering equations. The M-step is implemented by using a nonstandard YW equation set, in which correlations are replaced by their *a posteriori* values, that are calculated by using the Kalman filtering equations. The enhanced speech is obtained as a byproduct of the E-step. The performance of this algorithm was compared to that of alternative speech enhancement algorithms. A distinct advantage of the proposed algorithm compared to alternative algorithms is that it enhances the quality and SNR of the speech, while preserving its intelligibility and natural sound. Another advantage of the algorithm is that a VAD is not required.

Our development assumes a colored, rather than white, Gaussian noise model. The incremental computational price that is paid for this extension is moderate. However, the realizable improvement in the enhancement performance may be quite significant, as indicated in [11] and [13].

Fixed-lag Kalman smoothing was superior to Kalman filtering in terms of the objective distance measures (total and segmental SNR and Itakura–Saito distance) and in terms of the ASR performance. However, our informal speech quality tests suggest the opposite conclusion (i.e., that filtering is slightly superior to fixed-lag smoothing).

Fourth-order cumulant based equations were shown to provide a reliable initialization to the EM algorithm. Alternative initialization methods that we tried, such as third-order statistics based equations, were not as effective.

In order to reduce the computational load and to eliminate the delay of the iterative-batch algorithm, the sequential algorithm may be used. Although in general, the performance of the iterative-batch algorithm is superior, at low SNR's the differences in performance are small.

Appendix A

We provide a derivation of the EM algorithm presented in Section III.

Let $\boldsymbol{z}$ defined by (5) be the vector of corrupted speech samples (observed data) possessing the probability distribution function (PDF) $f_{\boldsymbol{Z}}(\boldsymbol{z};\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is defined by (4).

The ML estimate of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \log f_{\boldsymbol{Z}}(\boldsymbol{z};\boldsymbol{\theta}). \tag{23}$$

Our objective is to estimate the clean speech samples $s(t)$ from the observed data $z(t)$. Such an estimate will be obtained as a byproduct of the ML parameter estimation algorithm. The solution to (23) is obtained by using the EM algorithm [4], which is a general iterative procedure for obtaining the solution to the ML optimization problem. To apply the EM algorithm we need to define a "complete data" vector $\boldsymbol{y}$ that is related to the observed data vector ("incomplete data") through a (generally noninvertible) transformation, $F(\cdot)$, i.e.,

$$\boldsymbol{z} = F(\boldsymbol{y}). \tag{24}$$

The general lth iteration of the EM algorithm consists of the following estimation (E) step and maximization (M) step.

E-Step:

$$Q\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(l)}\right) = E_{\hat{\boldsymbol{\theta}}^{(l)}}\{\log f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})|\boldsymbol{z}\}. \tag{25}$$

M-Step:

$$\hat{\boldsymbol{\theta}}^{(l+1)} = \arg\max_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(l)}\right).$$

$\hat{\boldsymbol{\theta}}^{(l)}$ is the estimate of $\boldsymbol{\theta}$ after l iterations of the algorithm. Intuitively, the E-step yields an estimate of the *a posteriori* "complete data" statistics given the "incomplete data." The crucial point in any implementation of the EM algorithm is how to define the "complete data" such that the implementation of the maximization required by the M-step is simpler than the maximization required by the original ML criterion (23).

Now consider our noisy speech parameter estimation problem. The observed data vector ("incomplete data") in the current analysis frame is

$$\boldsymbol{z} = [z(1) \quad z(2) \quad \cdots \quad z(N)]^T$$

The corresponding vectors of speech and noise samples are

$$\boldsymbol{s} = [s(-p+1) \quad s(-p+2) \quad \cdots \quad s(N)]^T$$
$$\boldsymbol{v} = [v(-q+1) \quad v(-q+2) \quad \cdots \quad v(N)]^T$$

(N is the frame length. p and q are the speech and noise AR orders). The "complete data" vector, $\boldsymbol{y}$, is defined to be a concatenation of the clean speech samples $\boldsymbol{s}$, and the noise samples $\boldsymbol{v}$, i.e.,

$$\boldsymbol{y}^T = [\boldsymbol{s}^T \quad \boldsymbol{v}^T].$$

Invoking Bayes's rule

$$\log f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta}) = \log f_{\boldsymbol{S}}(\boldsymbol{s};\boldsymbol{\theta}) + \log f_{\boldsymbol{V}}(\boldsymbol{v};\boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ is the vector of unknown parameters defined in (4).

Under the assumption that both the speech innovation sequence, $u(t)$, and the noise innovation sequence, $w(t)$ are Gaussian (hence, $s(t)$ and $v(t)$ are also assumed Gaussian), and recalling (2) and (3), one obtains

$$\begin{aligned}\log f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta}) = {} & C + \log f(\boldsymbol{s}_p(0)) + \log f(\boldsymbol{v}_q(0)) \\ & - \frac{N}{2}\log g_s - \frac{N}{2}\log g_v \\ & - \frac{1}{2g_s}\sum_{t=1}^{N}[s(t) + \boldsymbol{\alpha}^T\boldsymbol{s}_p(t-1)]^2 \\ & - \frac{1}{2g_v}\sum_{t=1}^{N}[v(t) + \boldsymbol{\beta}^T\boldsymbol{v}_q(t-1)]^2 \end{aligned} \tag{26}$$

where C is a constant, independent of the parameter vector $\boldsymbol{\theta}$. Under the assumption that $N \gg p, q$, the contributions of $\log f(\boldsymbol{s}_p(0))$ and $\log f(\boldsymbol{v}_q(0))$ in (26) are negligible. Hence, taking the conditional expectation given the corrupted measurements $\boldsymbol{z}$ at $\hat{\boldsymbol{\theta}}^{(l)}$ yields

$$\begin{aligned} Q\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(l)}\right) &= E_{\hat{\boldsymbol{\theta}}^{(l)}}\{\log f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})|\boldsymbol{z}\} \\ &= C - \frac{N}{2}\log g_s - \frac{N}{2}\log g_v \\ &\quad - \frac{1}{2g_s}\sum_{t=1}^{N}\widehat{s^2(t)} + 2\boldsymbol{\alpha}^T\widehat{\boldsymbol{s}_p(t-1)s(t)} \\ &\quad + \boldsymbol{\alpha}^T\widehat{\boldsymbol{s}_p(t-1)\boldsymbol{s}_p^T(t-1)}\boldsymbol{\alpha} \\ &\quad - \frac{1}{2g_v}\sum_{t=1}^{N}\widehat{v^2(t)} + 2\boldsymbol{\beta}^T\widehat{\boldsymbol{v}_q(t-1)v(t)} \\ &\quad + \boldsymbol{\beta}^T\widehat{\boldsymbol{v}_q(t-1)\boldsymbol{v}_q^T(t-1)}\boldsymbol{\beta} \end{aligned} \tag{27}$$

where the notation (6) has been used.

Equation (27) implies that the maximization of $Q\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(l)}\right)$ with respect to $\boldsymbol{\theta}$ (M-step) is completely decoupled to two separate optimization problems, one with respect to the speech parameters, and the other with respect to the noise parameters. That is a very desirable property of the algorithm. Equations (13)–(16) are obtained by straightforward differentiation of (27).

APPENDIX B

We provide a derivation of the sequential algorithm presented in Section III.

The suggested recursive algorithm is based on the following gradient descent algorithm for solving the ML optimization problem, (23):

$$\hat{\theta}_i^{(l+1)} = \hat{\theta}_i^{(l)} + \frac{\rho_i}{N} \cdot \frac{\partial \log f_{\boldsymbol{Z}}(\boldsymbol{z};\boldsymbol{\theta})}{\partial \theta_i}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(l)}} \tag{28}$$

where $\hat{\theta}_i^{(l)}$ is the estimate of θ_i after l iteration cycles. The constants ρ_i are the update stepsizes. For sufficiently small stepsizes, this algorithm converges to a local maxima of the likelihood function. To compute the partial derivatives in (28), we suggest using Fisher's identity [9]. Let the vector $\boldsymbol{y}$ ("complete data") be some vector, that is related to the

measurements vector $\boldsymbol{z}$ by the transformation (24). Then Fisher's identity asserts the following:

$$\frac{\partial}{\partial\theta_i}\log f_{\boldsymbol{Z}}(\boldsymbol{z};\boldsymbol{\theta})_{|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(l)}} = \frac{\partial}{\partial\theta_i}E_{\hat{\boldsymbol{\theta}}^{(l)}}\{\log f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})|\boldsymbol{z}\} = \frac{\partial}{\partial\theta_i}Q\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(l)}\right)_{|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(l)}} \tag{29}$$

where we have used the definition (25). In order to make this identity useful, $\boldsymbol{y}$ should be chosen such that the differentiation of $Q\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(l)}\right)$ is simpler to implement than the direct differentiation of $\log f_{\boldsymbol{Z}}(\boldsymbol{z};\boldsymbol{\theta})$.

Differentiating (26) with respect to $\boldsymbol{\alpha}$ and invoking (28), (29) yields

$$\hat{\boldsymbol{\alpha}}^{(l+1)} = \hat{\boldsymbol{\alpha}}^{(l)} - \frac{\rho_s}{Ng_s}\sum_{\tau=1}^{N}[\widehat{\boldsymbol{s}_p(\tau-1)\boldsymbol{s}_p^T(\tau-1)}\hat{\boldsymbol{\alpha}}^{(l)} + \widehat{\boldsymbol{s}_p(\tau-1)s(\tau)}].$$

To obtain our sequential algorithm, the iteration index is replaced by the time index. We also incorporate a forgetting factor for calculating the covariance terms. For convenience, we define

$$\begin{aligned} Q^s(t) &\triangleq \begin{bmatrix} Q_{11}^s(t) & Q_{12}^s(t) \\ Q_{21}^s(t) & Q_{22}^s(t) \end{bmatrix} \\ &\triangleq \sum_{\tau=1}^{t}\lambda_s^{t-\tau}\,\widehat{\boldsymbol{s}_{p+1}(\tau)\boldsymbol{s}_{p+1}^T(\tau)} \\ &= \lambda_s Q^s(t-1) + \widehat{\boldsymbol{s}_{p+1}(t)\boldsymbol{s}_{p+1}^T(t)}. \end{aligned}$$

$Q_{11}^s(t)$ is a $p\times p$ matrix, $Q_{12}^s(t) = (Q_{21}^s(t))^T$ is a $p\times 1$ matrix, and $Q_{22}^s(t)$ is a scalar value. λ_s and λ_v are forgetting factors for the speech and noise, respectively, that satisfy

$$0 \le \lambda_s, \quad \lambda_v \le 1$$

and control the update rate. Then our sequential update of $\hat{\boldsymbol{\alpha}}(t)$ [see (18)] is

$$\begin{aligned} \hat{\boldsymbol{\alpha}}(t+1) &= \hat{\boldsymbol{\alpha}}(t) - \frac{\rho_s}{g_s}\sum_{\tau=1}^{t}\lambda_s^{t-\tau}[\widehat{\boldsymbol{s}_p(\tau-1)\boldsymbol{s}_p^T(\tau-1)}\hat{\boldsymbol{\alpha}}(t) \\ &\quad + \widehat{\boldsymbol{s}_p(\tau-1)s(\tau)}] \\ &= \hat{\boldsymbol{\alpha}}(t) - \frac{\rho_s}{g_s}[Q_{12}^s(t) + Q_{11}^s(t)\hat{\boldsymbol{\alpha}}(t)]. \end{aligned}$$

g_s may be obtained similarly. Alternatively, we may use the sequential variant of (14) [see (19)]. Similar update equations apply to the noise parameters [see (20) and (21)].

ACKNOWLEDGMENT

The authors would like to thank J. Goldberger of Tel-Aviv University for providing the speech recognition software, A. Erell of DSPC-Israel for providing the HMM-MMSE program, and O. Bahat and U. Sharony for helping with efficient programming of the algorithms.

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Englewood Cliffs, NJ: Prentice-Hall, 1979.
[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113–120, 1979.
[3] D. Burshtein, "Joint modeling and maximum likelihood estimation of pitch and linear prediction coefficient parameters," *J. Acoust. Soc. Amer.*, vol. 91, pp. 1531–1537, 1992.
[4] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc.*, pp. 1–38, 1977.
[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 1109–1121, 1984.
[6] ——, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443–445, 1985.
[7] Y. Ephraim, D. Malah, and B. H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1846–1856, 1989.
[8] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 40, pp. 725–735, 1992.
[9] R. A. Fisher, "Theory of statistical estimation," in *Proc. Cambridge Phil. Soc.*, vol. 22, pp. 700–725, 1925.
[10] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 352–359, 1996.
[11] S. Gannot, "Algorithms for single microphone speech enhancement," M.Sc. thesis, Tel-Aviv Univ., Israel, Apr. 1995.
[12] Nat. Inst. Standards Technol. "The DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," NIST, 1990.
[13] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 39, pp. 1732–1742, 1991.
[14] R. M. Gray, A. Buzo, A. H. Gray and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 367–376, 1980.
[15] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 795–805, 1991.
[16] B. Koo and J. D. Gibson, "Filtering of colored noise for speech enhancement and coding," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1989, pp. 349–352.
[17] B. G. Lee, K. Y. Lee, and S. Ann, "An EM-based approach for parameter enhancement with an application to speech signals," *Signal Process.*, vol. 46, pp. 1–14, 1995,
[18] K. Y. Lee and K. Shirai, "Efficient recursive estimation for speech enhancement in colored noise," *IEEE Signal Processing Lett.*, vol. 3, pp. 196–199, 1996.
[19] R. G. Leonard, "A database for speaker independent digit recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1984, pp. 42.11.1–42.11.4.
[20] 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.
[21] E. Masgrau, J. Salavedra, A. Moreno, and A. Ardanuy, "Speech enhancement by adaptive Wiener filtering based on cumulant AR modeling," *Speech Processing in Adverse Conditions*, M. Grenie and J. C. Junqua, Eds., 1992, pp. 143–146.
[22] C. L. Nikias and A. P. Petropulu, *Higher-Order Spectra Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[23] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1987, pp. 177–180.
[24] K. K. Paliwal and M. M. Sondhi, "Recognition of noisy speech using cumulant based linear prediction analysis," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1991, pp. 429–432.
[25] H. Sameti, H. Sheikhzadeh, L. Deng, and R. Brennan, "Comparative performance of spectral subtraction and HMM-based speech enhance-

ment strategies with application to hearing aid design," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 13–16, 1994.
[26] X. Shen, L. Deng, and A. Yasmin, "H-infinity filtering for speech enhancement," in *Proc. Int. Conf. Spoken Language Processing*, 1996, pp. 873–876.
[27] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *J. Time Series Anal.*, vol. 3, pp. 253–264, 1982.
[28] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1990, pp. 845–848.
[29] E. Weinstein, A. V. Oppenheim, and M. Feder, "Signal enhancement using single and multi-sensor measurements," RLE Tech. Rep. 560, Mass. Inst. Technol., Cambridge, MA, 1990.
[30] E. Weinstein, A. V. Oppenheim, M. Feder, and J. R. Buck, "Iterative and sequential algorithms for multisensor signal enhancement," *IEEE Trans. Signal Processing*, vol. 42, pp. 846–859, 1994.

Sharon Gannot (S'95) received the B.Sc. degree (summa cum laude) from the Technion—Israel Institute of Technology, Haifa, in 1986, and the M.Sc. degree (cum laude) from Tel-Aviv University, Israel, in 1995, both in electrical engineering. Currently, he is a Ph.D. student in the Department of Electrical Engineering-Systems, Tel-Aviv University.

From 1986 to 1993, he was head of research and development section in the Israeli Defense Forces. His research interests include parameter estimation, statistical signal processing, and speech processing using either single or multimicrophone arrays.

David Burshtein (M'92) received the B.Sc. (summa cum laude) and Ph.D. degrees in electrical engineering from Tel-Aviv University, Israel, in 1982 and 1987, respectively.

From 1982 to 1987, he was a Senior Research Engineer in the Research Laboratories, Israel Ministry of Defense, and was involved in research and development of digital signal processing systems. During 1988 and 1989, he was a Research Staff Member in the Speech Recognition Group, IBM T. J. Watson Research Center. In October 1989, he joined the Department of Electrical Engineering-Systems, Tel-Aviv University, where he is currently a faculty member. In the past six years, he has also been acting as a consultant to the Israeli Ministry of Defense and to DSPC-Israel. His research interests include parameter estimation, speech processing, speech recognition, statistical pattern recognition, and neural networks.

Ehud Weinstein (M'82–SM'86–F'94) was born in Tel-Aviv, Israel, on May 9, 1950. He received the B.Sc. degree from the Technion—Israel Institute of Technology, Haifa, and the Ph.D. degree from Yale University, New Haven, CT, both in electrical engineering, in 1975 and 1978, respectively.

In 1980, he joined the Department of Electrical Engineering-Systems, Faculty of Engineering, Tel-Aviv University, Israel, where he is currently a Professor. Since 1978, he has been affiliated with the Woods Hole Oceanographic Institute. He is also a Research Affiliate in the Research Laboratory of Electronics at the Massachusetts Institute of Technology, Cambridge, since 1990. His research interests are in the general areas of estimation theory, statistical signal processing, array processing, and digital communications.

Dr. Weinstein is a co-recipient of the 1983 Senior Award of the IEEE Acoustics, Speech, and Signal Processing Society.