

Relaxed Statistical Model for Speech Enhancement and *a Priori* SNR Estimation

Israel Cohen, *Senior Member, IEEE*

Abstract—In this paper, we propose a statistical model for speech enhancement that takes into account the time-correlation between successive speech spectral components. It retains the simplicity associated with the Gaussian statistical model, and enables the extension of existing algorithms to noncausal estimation. The sequence of speech spectral variances is a random process, which is generally correlated with the sequence of speech spectral magnitudes. Causal and noncausal estimators for the *a priori* SNR are derived in agreement with the model assumptions and the estimation of the speech spectral components. We show that a special case of the causal estimator degenerates to a “decision-directed” estimator with a time-varying frequency-dependent weighting factor. Experimental results demonstrate the improved performance of the proposed algorithms.

Index Terms—Parameter estimation, sequential estimation, spectral analysis, speech enhancement, time-frequency analysis.

I. INTRODUCTION

ONE OF THE most popular methods for enhancing speech, degraded by uncorrelated additive noise, is the spectral enhancement algorithm of Ephraim and Malah [1], [2]. This algorithm and its derivatives (e.g., [3]–[5]) have been applied to single-channel and multi-channel speech enhancement in speech recognition systems [6], [7], speech coders [8]–[10] digital hearing-aids [11], [12], voice activity detectors [13]–[15], and hands-free mobile communication systems [16]–[18]. The algorithm is based on a Gaussian statistical model [19]. Accordingly, the individual short-term spectral components of the speech and noise signals are modeled as statistically independent Gaussian random variables. The assumption that speech spectral components in the time-frequency domain (i.e., coefficients of the short-time Fourier transform of the speech signal) are statistically independent facilitates a mathematically tractable derivation of useful estimators for various distortion measures. In [2], Ephraim and Malah derived a short-term spectral amplitude (STSA) estimator, which minimizes the mean-square error of the spectral magnitude. In [1], based on the same Gaussian statistical model, they derived a log-spectral amplitude (LSA) estimator, which minimizes the mean-square error of the log-spectra. They found that the LSA estimator is superior to the STSA estimator, since it results in a much lower residual noise level without further affecting the speech itself.

Manuscript received October 1, 2003; revised July 29, 2004. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Peter Vary.

The author is with the Department of Electrical Engineering, The Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: icohen@ee.technion.ac.il; <http://www.ee.technion.ac.il/Sites/People/Israel-Cohen>).

Digital Object Identifier 10.1109/TSA.2005.851940

Cappé [20] showed that the dominant factor in the Ephraim-Malah algorithm is the decision-directed estimation approach for the *a priori* signal-to-noise ratio (SNR). The *a priori* SNR estimate is obtained as a weighted sum of two terms. One representing the *a priori* SNR resulting from the processing of the previous frame. The other term is a maximum likelihood estimate for the *a priori* SNR, based entirely on the current frame. A weighting factor, which represents the importance (weight) of each term, controls the trade-off between the noise reduction and the transient distortion brought into the signal [2], [20]. In practice, the weight of the first term is substantially larger than that of the latter. This indicates that the *a priori* SNRs in successive short-term frames are highly correlated. Unfortunately, the decision-directed estimation approach applies a constraint on the response to speech onsets. The *a priori* SNR estimator cannot respond too fast to an abrupt increase in the instantaneous SNR, since it inevitably yields an increase in the level of musical residual noise.

Martin [12] and Breithaupt and Martin [21] considered a different statistical model, where the clean speech spectral components are gamma distributed, and the noise spectral components are either Gaussian or Laplace distributed. They assumed that distinct spectral components are statistically independent, and derived an estimator for the complex speech spectral coefficients, which minimizes the mean-square error (i.e., a Wiener filter), and a spectral amplitude estimator, which minimizes the mean-square error of the spectral power. However, to estimate the *a priori* SNR they still used the decision-directed approach of Ephraim and Malah.

Enhancement schemes based on hidden Markov models (HMMs) try to circumvent the assumption of specific distributions for the speech and noise processes [22]–[25]. The probability distributions of the two processes are first estimated from long training sequences of clean speech and noise samples, and then used jointly with a given distortion measure to derive an estimator for the speech signal. Normally, vectors generated from a given sequence of states are assumed statistically independent. However, the HMM can be extended to take into account the time-frequency correlation of speech signals by using nondiagonal covariance matrices for each subsurface, and assuming that a sequence of vectors generated from a given sequence of states is a nonzero-order autoregressive process [23], [26]. First-order HMMs, for example, with a mixture of Gaussian distributions in each state and minimum mean-square error estimation result in a weighted sum of conditional mean estimators, one for each mixture component in each state. The weights are the posterior probabilities of the states and mixture components given the noisy signal [27]. Unfortunately, the

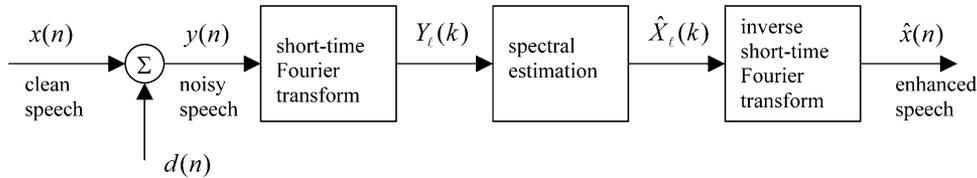


Fig. 1. Spectral estimation approach.

HMM-based speech enhancement relies on the type of training data [28]. It works best with the trained type of noise, but often worse with other type of noise. Furthermore, improved performance generally entails more complex models and higher computational requirements.

In this paper, we propose a statistical model for speech enhancement that takes into account the time-correlation between successive speech spectral components. It retains the simplicity associated with the Ephraim-Malah statistical model, provides insight into the decision-directed approach [41], and most importantly enables the extension of existing algorithms to noncausal estimation [42]. In the proposed model, the sequence of speech spectral variances is a random process, which is correlated with the sequence of the speech spectral components. Causal and noncausal estimators for the *a priori* SNR are derived in agreement with the model assumptions and the estimation of the speech spectral components.

The causal estimator for the *a priori* SNR combines two steps, a “propagation” step and an “update” step, following the rational of Kalman filtering, to recursively predict and update the estimate for the speech spectral variance as new data arrive. The causal *a priori* SNR estimator is closely related to the decision-directed estimator of Ephraim and Malah. A special case of the causal estimator degenerates to a “decision-directed” estimator with a *time-varying frequency-dependent* weighting factor. The weighting factor is monotonically decreasing as a function of the instantaneous SNR, resulting effectively in a larger weighting factor during speech absence, and a smaller weighting factor during speech presence. This reduces both the musical noise and the signal distortion.

The noncausal *a priori* SNR estimator employs future spectral measurements to better predict the spectral variances of the clean speech. A comparison of the causal and noncausal estimators indicates that the differences are primarily noticeable during speech onsets. The *causal a priori* SNR estimator, as well as the decision-directed estimator, cannot respond too fast to an abrupt increase in the instantaneous SNR, since it necessarily implies an increase in the level of musical residual noise. By contrast, the *noncausal* estimator, having a few subsequent spectral measurements at hand, is capable of discriminating between speech onsets and noise irregularities. Experimental results show that the noncausal estimator yields a higher improvement in the segmental SNR and lower log-spectral distortion, than the decision-directed method and the causal estimator. The advantages of the noncausal estimator are particularly perceived during onsets of speech and noise only frames. Onsets of speech are better preserved, while a further reduction of musical noise is achieved.

The paper is organized as follows. In Section II, we formulate the speech enhancement problem. In Section III, a statistical model is proposed that relaxes the independence assumption of spectral components. In Section IV, we derive estimators for the clean speech spectral components and the *a priori* SNR. We present causal and noncausal recursive speech enhancement algorithms, and address their relation to the decision-directed estimation approach. Finally, in Section V, we evaluate the proposed algorithms, and present experimental results, which demonstrate their improved performance.

II. PROBLEM FORMULATION

Let $x(n)$ and $d(n)$ denote speech and uncorrelated additive noise signals, respectively, where n is a discrete-time index. The observed signal $y(n)$, given by $y(n) = x(n) + d(n)$, is transformed into the time-frequency domain by applying the short-time Fourier transform (STFT). Specifically

$$Y_\ell(k) = \sum_{n=0}^{N-1} y(n + \ell M) h(n) e^{-j\frac{2\pi}{N}nk} \quad (1)$$

where k is the frequency-bin index ($k = 0, 1, \dots, N-1$), ℓ is the time frame index ($\ell = 0, 1, \dots$), $h(n)$ is an analysis window of size N (e.g., Hamming window), and M is the framing step (number of samples separating two successive frames). Given an estimate $\hat{X}_\ell(k)$ for the STFT of the clean speech (see Fig. 1), an estimate for the clean speech signal is obtained by applying the inverse STFT

$$\hat{x}(n) = \sum_{\ell} \sum_{k=0}^{N-1} \hat{X}_\ell(k) \tilde{h}(n - \ell M) e^{j\frac{2\pi}{N}k(n - \ell M)} \quad (2)$$

where $\tilde{h}(n)$ is a synthesis window that is biorthogonal to the analysis window $h(n)$ [29], and the inverse STFT is efficiently implemented by using the weighted overlap-add method [30].

Let $\mathcal{Y}_0^{\ell'}(k)$ denote a set of spectral measurements $\{Y_0(k), \dots, Y_{\ell'}(k)\}$, and let $d[X_\ell(k), \hat{X}_\ell(k)]$ be a given distortion measure between $X_\ell(k)$ and $\hat{X}_\ell(k)$. Our objective is to find an estimator $\hat{X}_\ell(k)$, which minimizes the conditional expected value of the distortion measure, given the set of spectral noisy measurements

$$\hat{X}_\ell(k) = \arg \min_{\hat{X}} E \left\{ d[X_\ell(k), \hat{X}] \middle| \mathcal{Y}_0^{\ell'}(k) \right\}. \quad (3)$$

We consider a causal estimation of $X_\ell(k)$ (in which case $\ell' \leq \ell$), as well as a noncausal estimation (in which case $\ell' > \ell$)¹, while

¹Note that causality is defined with respect to the spectral components, rather than with respect to the samples in the time domain.

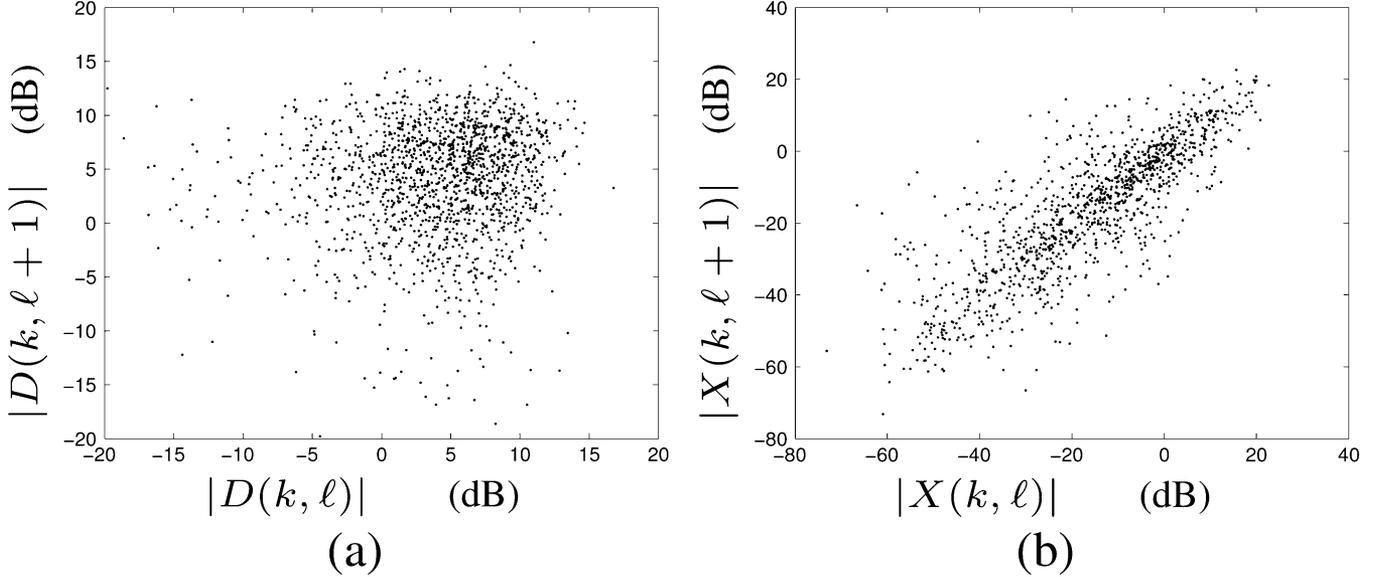


Fig. 2. Scatter plots for successive spectral magnitudes of (a) white Gaussian noise signal and (b) speech signal at center frequency 500 Hz ($k = 17$). The overlap between successive frames is 50%.

the spectral components are *not* assumed statistically independent.

Let $A_\ell(k)$ and $\varphi_\ell(k)$ denote respectively the magnitude and phase of $X_\ell(k)$. Then, distortion measures that are of particular interest for speech enhancement applications are as follows.

- 1) The squared-error distortion [31]

$$d_{SE}[X_\ell(k), \hat{X}_\ell(k)] \triangleq |X_\ell(k) - \hat{X}_\ell(k)|^2. \quad (4)$$

- 2) The spectral amplitude distortion [2]

$$d_{SA}[X_\ell(k), \hat{X}_\ell(k)] \triangleq [A_\ell(k) - \hat{A}_\ell(k)]^2. \quad (5)$$

- 3) The log-spectral amplitude distortion [1]

$$d_{LSA}[X_\ell(k), \hat{X}_\ell(k)] \triangleq [\log A_\ell(k) - \log \hat{A}_\ell(k)]^2. \quad (6)$$

- 4) The spectral power distortion [21], [27], [32]

$$d_{SP}[X_\ell(k), \hat{X}_\ell(k)] \triangleq [A_\ell^2(k) - \hat{A}_\ell^2(k)]^2. \quad (7)$$

The last three distortion measures are insensitive to the estimation error of $\hat{\varphi}_\ell(k)$. Therefore, it is constructive to combine them with the following constrained optimization problem [2]:

$$\min_{\hat{\varphi}_\ell(k)} E \left\{ \left| e^{j\varphi_\ell(k)} - e^{j\hat{\varphi}_\ell(k)} \right|^2 \right\} \quad \text{subject to} \quad \left| e^{j\hat{\varphi}_\ell(k)} \right| = 1. \quad (8)$$

This yields an estimator for the complex exponential of the phase, constrained to not affecting the spectral magnitude estimate. Alternatively, an estimate for the spectral phase $\hat{\varphi}_\ell(k)$ is obtained by minimizing the expected value of the following distortion measure [2]:

$$d_\varphi[\varphi_\ell(k), \hat{\varphi}_\ell(k)] \triangleq 1 - \cos[\varphi_\ell(k) - \hat{\varphi}_\ell(k)]. \quad (9)$$

This measure is invariant under modulo 2π transformation of the estimation error $\varphi_\ell(k) - \hat{\varphi}_\ell(k)$, and for small estimation errors it closely resembles the squared-error distortion measure, since $1 - \cos \beta \approx \beta^2/2$ for $\beta \ll 1$.

III. SPEECH SPECTRAL MODEL

In this section, we propose a statistical model that takes into account the time-correlation between successive spectral components of the speech signal. To see graphically the relation between successive spectral components of a speech signal, in comparison with a noise signal, we present scatter plots for successive spectral magnitudes, and investigate the sample autocorrelation coefficient sequences (ACS) of the STFT coefficients along time-trajectories (the frequency-bin index k is held fixed). We consider a speech signal that is constructed from six different utterances, without intervening pauses. The utterances, half from male speakers and half from female speakers, are taken from the TIMIT database [33]. The speech signal is sampled at 16 kHz, and transformed into the STFT domain using Hamming analysis windows of 512 samples (32 ms) length, and 256 samples framing step (50% overlap between successive frames).

Fig. 2 shows an example of scatter plots for successive spectral magnitudes of white Gaussian noise (WGN) and speech signals. It implies that 50% overlap between successive frames does not yield a significant correlation between the spectral magnitudes of the WGN signal. However, successive spectral magnitudes of the speech signal are highly correlated. Fig. 3 shows the ACSs of the speech spectral components along time-trajectories, for various frequency-bins and framing steps. The 95 percent confidence limits (e.g., [34]) are depicted as horizontal dotted lines. In order to prevent an upward bias of the autocovariance estimates due to irrelevant (nonspeech) spectral components, the ACSs are computed from spectral components whose magnitudes are within 30 dB of the maximal magnitude.

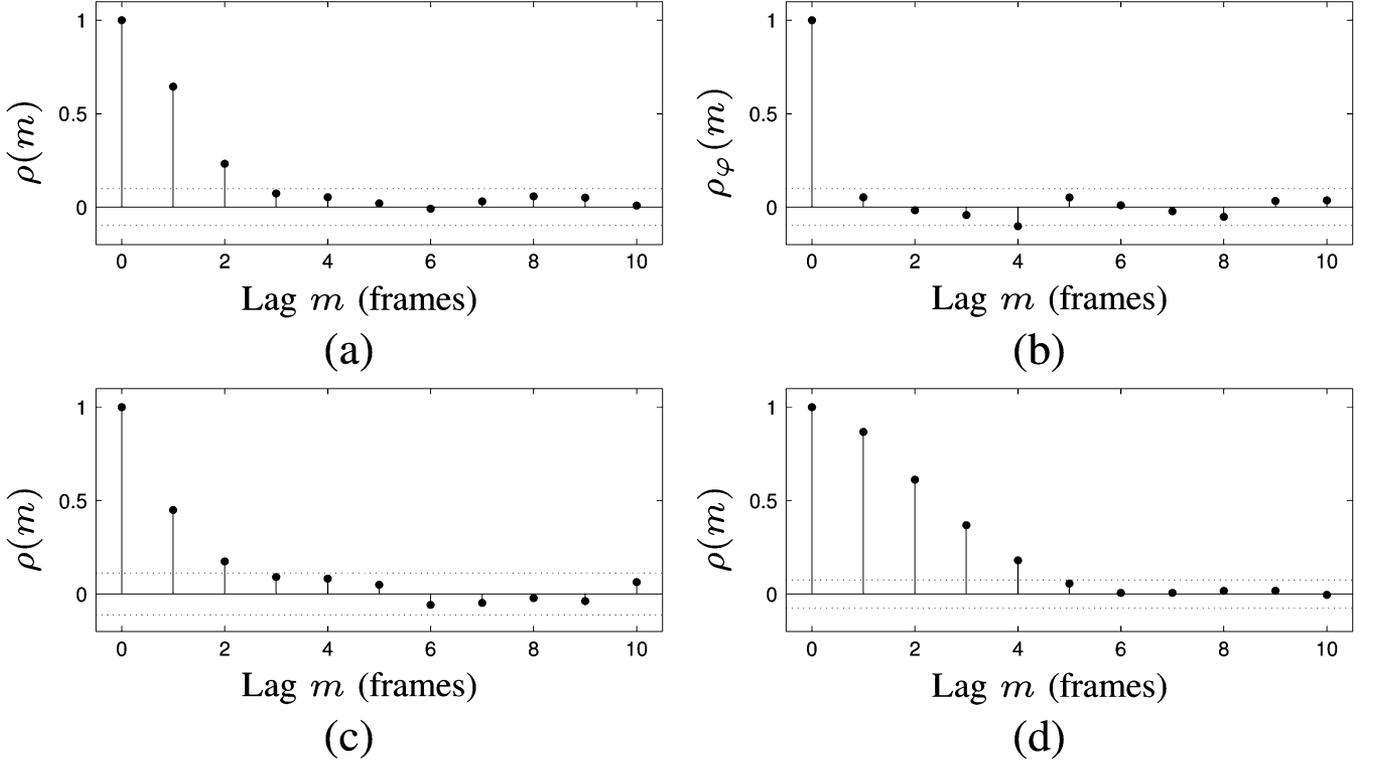


Fig. 3. Sample autocorrelation coefficient sequences (ACSs) of clean speech STFT coefficients along time-trajectories, for various frequency-bins and framing steps. The dotted lines represents 95 percent confidence limits. (a) ACS of the spectral magnitude at frequency-bin $k = 17$ (center frequency 500 Hz), framing step $M = N/2$ (50% overlap between frames); (b) ACS of the spectral phase, $k = 17$, $M = N/2$; (c) ACS of the spectral magnitude, $k = 65$ (center frequency 2 kHz), $M = N/2$; and (d) ACS of the spectral magnitude, $k = 17$, $M = N/4$ (75% overlap between frames).

Specifically, the sample autocorrelation coefficients of the spectral magnitudes are calculated by

$$\rho(m) = \frac{\sum_{\ell \in \mathcal{L}} [A_{\ell}(k) - \bar{A}(k)][A_{\ell+m}(k) - \bar{A}(k)]}{\sum_{\ell \in \mathcal{L}} [A_{\ell}(k) - \bar{A}(k)]^2} \quad (10)$$

where

$$\bar{A}(k) = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} A_{\ell}(k)$$

denotes the sample mean, m is the lag in frames, and \mathcal{L} represents the set of relevant spectral components

$$\mathcal{L} = \{\ell | A_{\ell}(k) \geq 10^{-30/20} \max_{\ell} \{A_{\ell}(k)\}\}.$$

The corresponding sample autocorrelation coefficients of the spectral phases are obtained by

$$\rho_{\varphi}(m) = \frac{\sum_{\ell \in \mathcal{L}} \varphi_{\ell}(k) \varphi_{\ell+m}(k)}{\sum_{\ell \in \mathcal{L}} \varphi_{\ell}^2(k)}. \quad (11)$$

Fig. 4 shows the variation of the correlation between successive spectral magnitudes on frequency and on overlap between successive frames. Figs. 3 and 4 demonstrate that for speech signals, successive spectral magnitudes are highly correlated, while the correlation is generally larger at lower frequencies, and it increases as the overlap between successive frames increases.

Fig. 5 shows, for a realization of WGN, the variation of the correlation between successive spectral magnitudes on the overlap between frames. A comparison of Figs. 5 and 4 reveals that for a sufficiently large framing step ($M \geq N/2$,

i.e., overlap between frames $\leq 50\%$), successive spectral components of the *noise* signal, but clearly not of the *speech* signal, can be assumed uncorrelated. For smaller framing steps, the correlation between successive spectral noise components has also to be taken into consideration. Furthermore, since the length of the analysis window cannot be too large (its typical length is 20–40 ms [2]), for a given frame ℓ adjacent Fourier expansion coefficients of the noise signal, $D_{\ell}(k)$ and $D_{\ell}(k+1)$, as well as adjacent coefficients of the speech signal, $X_{\ell}(k)$ and $X_{\ell}(k+1)$, are also correlated to a certain degree. Nevertheless, our primary goal is to propose a valid and consistent statistical model for both the spectral enhancement and the *a priori* SNR estimation, while keeping the resulting algorithms simple. Therefore, we continue with the statistical independence assumption for distinct frequency-bins ($X_{\ell}(k)$ and $X_{\ell'}(k')$ are assumed statistically independent if $k \neq k'$), as implied in the estimation problem (3).

In conclusion of the above discussion, we propose the following statistical model for the speech and noise spectral components.

- 1) The noise spectral components $\{D_{\ell}(k)\}$ are statistically independent zero-mean complex Gaussian random variables. The real and imaginary parts of $D_{\ell}(k)$ are independent and identically distributed (iid).
- 2) The speech spectral phases $\{\varphi_{\ell}(k)\}$ are iid uniform random variables on $[-\pi, \pi]$.
- 3) The random processes $\{A_{\ell}(k) | \ell = 0, 1, \dots\}$ and $\{A_{\ell'}(k') | \ell = 0, 1, \dots\}$ are statistically independent for $k \neq k'$.

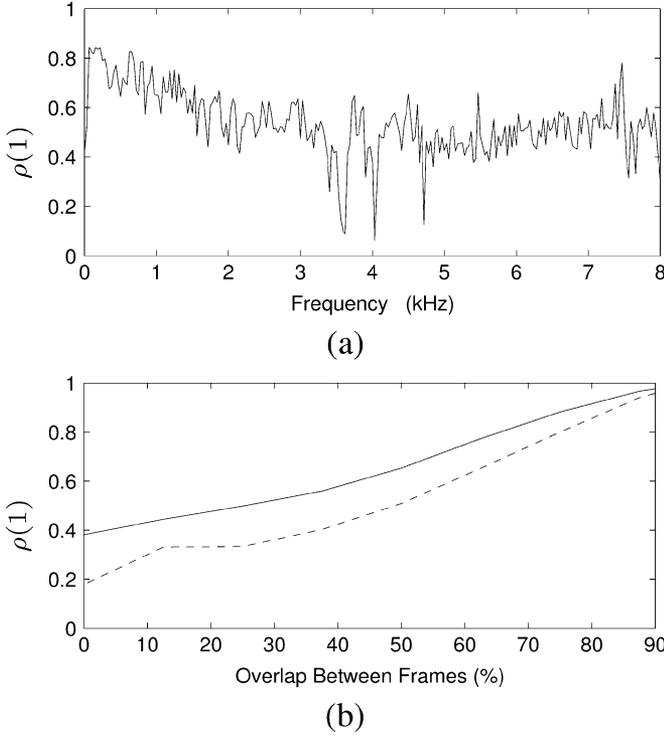


Fig. 4. Variation of the correlation between successive spectral magnitudes of the speech signal. (a) Variation of $\rho(1)$ on frequency for $M = N/2$ (50% overlap between frames) and (b) Variation of $\rho(1)$ on overlap between frames for $k = 33$ (center frequency 1 kHz; solid line) and $k = 65$ (center frequency 2 kHz; dashed line).

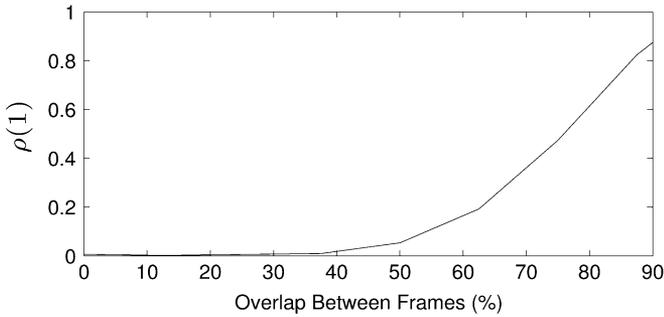


Fig. 5. Variation of the correlation between successive spectral magnitudes on the overlap between frames for a realization of white Gaussian noise.

- 4) For fixed k and ℓ , a speech spectral component $X_\ell(k)$ is conditionally a zero-mean complex Gaussian random variable, given its variance $\lambda_{X_\ell}(k) = E\{A_\ell^2(k)\}$.
- 5) The sequence of speech spectral variances $\{\lambda_{X_\ell}(k) | \ell = 0, 1, \dots\}$ is a random process, generally correlated with the sequence of speech spectral magnitudes $\{A_{\ell'}(k) | \ell' = 0, 1, \dots\}$. However, given $\lambda_{X_\ell}(k)$, $A_\ell(k)$ is statistically independent of $A_{\ell'}(k)$ for all $\ell' \neq \ell$.

Clearly, the first assumption does not hold when the overlap between successive frames is too large (see Fig. 5). Therefore, we assume that the STFT is implemented in accordance with this assumption (e.g., the overlap is not greater than 50%). The third assumption allows to formulate independent estimation problems for each frequency bin k , which greatly simplifies the re-

sulting algorithms. The fifth assumption allows to further simplify the algorithms, as shown in the next section, by splitting the estimation problem into two parts: the first part is spectral variance estimation from a given set of noisy measurements; the second part is spectral *amplitude* estimation from a *single* spectral noisy measurement, using the spectral variance estimate obtained in the first part. Note that successive spectral components are still assumed correlated, since the random processes $\{X_\ell(k) | \ell = 0, 1, \dots\}$ and $\{\lambda_{X_\ell}(k) | \ell = 0, 1, \dots\}$ are not independent.

IV. SIGNAL ESTIMATION

In this section, we derive estimators for $X_\ell(k)$ based on the proposed statistical model and the various distortion measures specified in Section II. We show that similar to conventional spectral estimators, $\hat{X}_\ell(k)$ is obtained by applying a real-valued gain function to the corresponding spectral measurement $Y_\ell(k)$. The spectral gain depends on two parameters: the *a priori* and *a posteriori* SNRs. However, rather than evaluating the *a priori* SNR by the decision-directed approach, the *a priori* SNR estimation relies on the statistical model. For notational simplicity, the frequency-bin index k is henceforth omitted, since according to the statistical model, an estimate $\hat{X}_\ell(k)$ can be found independently for each k . Furthermore, we assume knowledge of the noise power spectral density (PSD), which in practice can be estimated by using the *Minima Controlled Recursive Averaging* approach [35].

A. Spectral Enhancement

Let $p(X_\ell | \mathcal{Y}_0^{\ell'}, \lambda_{X_\ell})$ denote the conditional probability density function (pdf) of a speech spectral component X_ℓ given its variance λ_{X_ℓ} and the noisy measurements $\mathcal{Y}_0^{\ell'}$. Let $p(\lambda_{X_\ell} | \mathcal{Y}_0^{\ell'})$ denote the conditional pdf of the clean speech spectral variance at frame ℓ given $\mathcal{Y}_0^{\ell'}$. Then, the spectral estimator \hat{X}_ℓ is obtained from

$$\begin{aligned} \min_{\hat{X}_\ell} E \left\{ d(X_\ell, \hat{X}_\ell) \middle| \mathcal{Y}_0^{\ell'} \right\} \\ = \min_{\hat{X}_\ell} \iint d(X_\ell, \hat{X}_\ell) p(X_\ell | \mathcal{Y}_0^{\ell'}, \lambda_{X_\ell}) \\ \times p(\lambda_{X_\ell} | \mathcal{Y}_0^{\ell'}) dX_\ell d\lambda_{X_\ell}. \end{aligned} \quad (12)$$

The proposed statistical model implies

$$p(X_\ell | \mathcal{Y}_0^{\ell'}, \lambda_{X_\ell}) = p(X_\ell | Y_\ell, \lambda_{X_\ell}). \quad (13)$$

To simplify the algorithm, we first derive a minimum mean-squared error (MMSE) estimate for λ_{X_ℓ} from the set of noisy measurements $\mathcal{Y}_0^{\ell'}$

$$\hat{\lambda}_{X_\ell} = E \left\{ A_\ell^2(k) \middle| \mathcal{Y}_0^{\ell'} \right\} \triangleq \lambda_{X_\ell | \ell'}. \quad (14)$$

Then, given $\lambda_{X_\ell | \ell'}$, we derive a conditional estimate for X_ℓ which minimizes the expected value of the distortion measure

$$\hat{X}_\ell = \arg \min_{\hat{X}_\ell} \int d(X_\ell, \hat{X}_\ell) p(X_\ell | \mathcal{Y}_0^{\ell'}, \lambda_{X_\ell | \ell'}) dX_\ell. \quad (15)$$

Substituting (13) into (15), the spectral estimate \hat{X}_ℓ is obtained from

$$\begin{aligned} \min_{\hat{X}_\ell} \int d(X_\ell, \hat{X}_\ell) p(X_\ell | Y_\ell, \lambda_{X_\ell | \ell'}) dX_\ell \\ = \min_{\hat{X}_\ell} E\{d(X_\ell, \hat{X}_\ell) | Y_\ell, \lambda_{X_\ell | \ell'}\}. \end{aligned} \quad (16)$$

The latter problem, when the *a priori* SNR is defined appropriately, reduces to the classical spectral enhancement problem as formulated by Ephraim and Malah [1], [2]. Accordingly, it is unnecessary to assume that speech spectral components are statistically independent (e.g., [3], [12], [15], [21], [32]). It is sufficient to assume that *given a speech spectral variance* λ_{X_ℓ} , X_ℓ is statistically independent of $X_{\ell'}$ for all $\ell' \neq \ell$ (see model assumptions no. 5 and 2). This enables to derive estimators for X_ℓ and λ_{X_ℓ} under consistent model assumptions.

An estimate for X_ℓ is obtained by applying a spectral gain function to each noisy spectral component of the speech signal

$$\hat{X}_\ell = G(\xi_{\ell | \ell'}, \gamma_\ell) Y_\ell \quad (17)$$

where the *a priori* and *a posteriori* SNRs are defined respectively by²

$$\xi_{\ell | \ell'} \triangleq \frac{\lambda_{X_\ell | \ell'}}{\lambda_{D_\ell}} \quad (18)$$

$$\gamma_\ell \triangleq \frac{|Y_\ell|^2}{\lambda_{D_\ell}} \quad (19)$$

and where $\lambda_{D_\ell} \triangleq E\{|D_\ell|^2\}$ denotes the noise spectral variance. The specific expression for the spectral gain function $G(\xi_{\ell | \ell'}, \gamma_\ell)$ depends on the particular choice of a distortion measure $d(X_\ell, \hat{X}_\ell)$. For squared-error distortion (see (4)), the gain function is given by [31]

$$G_{SE}(\xi_{\ell | \ell'}) = \frac{\xi_{\ell | \ell'}}{1 + \xi_{\ell | \ell'}}. \quad (20)$$

In case of combining the spectral amplitude, the log-spectral amplitude, or the spectral power distortion measures (see (5)–(7)) with the constrained optimization problem (8), the gain functions can respectively be written as [1], [2], [27], [32]

$$\begin{aligned} G_{SA}(\xi_{\ell | \ell'}, \gamma_\ell) &= \frac{\sqrt{\pi v_\ell}}{2\gamma_\ell} \left[(1 + v_\ell) I_0\left(\frac{v_\ell}{2}\right) \right. \\ &\quad \left. + v_\ell I_1\left(\frac{v_\ell}{2}\right) \right] \exp\left(-\frac{v_\ell}{2}\right) \end{aligned} \quad (21)$$

$$G_{LSA}(\xi_{\ell | \ell'}, \gamma_\ell) = \frac{\xi_{\ell | \ell'}}{1 + \xi_{\ell | \ell'}} \exp\left(\frac{1}{2} \int_{v_\ell}^{\infty} \frac{e^{-t}}{t} dt\right) \quad (22)$$

$$G_{SP}(\xi_{\ell | \ell'}, \gamma_\ell) = \sqrt{\frac{\xi_{\ell | \ell'}}{1 + \xi_{\ell | \ell'}} \left(\frac{1}{\gamma_\ell} + \frac{\xi_{\ell | \ell'}}{1 + \xi_{\ell | \ell'}} \right)} \quad (23)$$

where $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of zero and first order, respectively, and v_ℓ is defined by $v_\ell \triangleq$

²Note that in [2], the *a priori* SNR is defined by $\xi_\ell = \lambda_{X_\ell} / \lambda_{D_\ell}$, where the variance λ_{X_ℓ} is a parameter of the prior pdf of X_ℓ .

$\xi_{\ell | \ell'} \gamma_\ell / (1 + \xi_{\ell | \ell'})$. It still remains to estimate the *a priori* SNR $\xi_{\ell | \ell'}$, as defined in (18) and (14), based on the statistical model.

B. Causal Recursive Estimation

In this section, we propose a causal conditional estimator $\hat{\xi}_{\ell | \ell}$ for the *a priori* SNR given the noisy measurements up to frame ℓ . The estimator combines two steps, a “propagation” step and an “update” step, following the rational of Kalman filtering, to recursively predict and update the estimate for λ_{X_ℓ} as new data arrive.

Suppose we are given an estimate $\hat{\lambda}_{X_\ell | \ell-1}$, which is conditioned on the noisy measurements up to frame $\ell - 1$, and a new noisy spectral component Y_ℓ is observed. Then, the estimate for λ_{X_ℓ} can be updated by computing the conditional variance of X_ℓ given Y_ℓ and $\hat{\lambda}_{X_\ell | \ell-1}$

$$\hat{\lambda}_{X_\ell | \ell} = E\left\{A_\ell^2 | \hat{\lambda}_{X_\ell | \ell-1}, Y_\ell\right\}. \quad (24)$$

This is obtained by applying the gain function $G_{SP}(\hat{\xi}_{\ell | \ell-1}, \gamma_\ell)$ to Y_ℓ , and computing the squared absolute value of the result³

$$\begin{aligned} \hat{\lambda}_{X_\ell | \ell} &= G_{SP}^2(\hat{\xi}_{\ell | \ell-1}, \gamma_\ell) |Y_\ell|^2 \\ &= \frac{\hat{\xi}_{\ell | \ell-1}}{1 + \hat{\xi}_{\ell | \ell-1}} \left(\frac{1}{\gamma_\ell} + \frac{\hat{\xi}_{\ell | \ell-1}}{1 + \hat{\xi}_{\ell | \ell-1}} \right) |Y_\ell|^2. \end{aligned} \quad (25)$$

Dividing both sides of (25) by λ_{D_ℓ} , we have

$$\hat{\xi}_{\ell | \ell} = \frac{\hat{\xi}_{\ell | \ell-1}}{1 + \hat{\xi}_{\ell | \ell-1}} \left(1 + \frac{\hat{\xi}_{\ell | \ell-1} \gamma_\ell}{1 + \hat{\xi}_{\ell | \ell-1}} \right). \quad (26)$$

We call (26) the “update” step.

Computation of the update step requires the estimate

$$\hat{\xi}_{\ell | \ell-1} \triangleq \frac{\hat{\lambda}_{X_\ell | \ell-1}}{\lambda_{D_{\ell-1}}} \quad (27)$$

for the *a priori* SNR given $\mathcal{Y}_0^{\ell-1}$. Note that in (27), $\hat{\lambda}_{X_\ell | \ell-1}$ is divided by $\lambda_{D_{\ell-1}}$ rather than by λ_{D_ℓ} , since given the measurements up to frame $\ell - 1$ the noise variance estimate at frame ℓ is given by $\lambda_{D_{\ell-1}}$. Assume we are given at frame $\ell - 1$ estimates for the spectral amplitude $A_{\ell-1}$ and the spectral variance $\lambda_{X_{\ell-1}}$, conditioned on $\mathcal{Y}_0^{\ell-1}$. Then, these estimates can be “propagated” in time to obtain an estimate for λ_{X_ℓ} . Since λ_{X_ℓ} is correlated with both $\lambda_{X_{\ell-1}}$ and $A_{\ell-1}$, we propose to use an estimate of the form

$$\hat{\lambda}_{X_\ell | \ell-1} = \max\left\{(1 - \alpha)\hat{\lambda}_{X_{\ell-1} | \ell-1} + \alpha\hat{A}_{\ell-1}^2, \lambda_{\min}\right\} \quad (28)$$

where α ($0 \leq \alpha \leq 1$) is related to the degree of nonstationarity of the random process $\{\lambda_{X_\ell} | \ell = 0, 1, \dots\}$, and λ_{\min} is a lower bound on the variance of X_ℓ . In case of a pseudo-stationary process, α is set to a small value, since $\hat{\lambda}_{X_\ell | \ell-1} \approx \hat{\lambda}_{X_{\ell-1} | \ell-1}$. In case of a nonstationary process, α is set to a larger value, since the variances at successive frames are less correlated, and

³Recall that G_{SP} minimizes the expected spectral power distortion, yielding the square root of the conditional expected spectral power. That is, $G_{SP}(\xi_\ell, \gamma_\ell) |Y_\ell| = [E\{A_\ell^2 | \xi_\ell, Y_\ell\}]^{1/2}$.

TABLE I
SUMMARY OF THE CAUSAL RECURSIVE SPEECH ENHANCEMENT ALGORITHM

Initialization:	$\hat{A}_{-1} = 0, \hat{\xi}_{-1 -1} = \xi_{\min}$.
For all short-time frames $\ell = 0, 1, \dots$	
“Propagation” step:	
	$\hat{\xi}_{\ell \ell-1} = \max \left\{ (1 - \alpha)\hat{\xi}_{\ell-1 \ell-1} + \alpha \frac{\hat{A}_{\ell-1}^2}{\lambda_{D\ell-1}}, \xi_{\min} \right\}$
“Update” step:	
	$\hat{\xi}_{\ell \ell} = \frac{\hat{\xi}_{\ell \ell-1}}{1 + \hat{\xi}_{\ell \ell-1}} \left(1 + \frac{\hat{\xi}_{\ell \ell-1} \gamma_{\ell}}{1 + \hat{\xi}_{\ell \ell-1}} \right)$
Spectral enhancement:	
	$\hat{X}_{\ell} = G(\hat{\xi}_{\ell \ell}, \gamma_{\ell}) Y_{\ell}$

the relative importance of $\hat{\lambda}_{X_{\ell-1}|\ell-1}$ to predict $\hat{\lambda}_{X_{\ell}|\ell-1}$ decreases. Dividing both sides of (28) by $\lambda_{D\ell-1}$, we obtain the “propagation” step

$$\hat{\xi}_{\ell|\ell-1} = \max \left\{ (1 - \alpha)\hat{\xi}_{\ell-1|\ell-1} + \alpha \frac{\hat{A}_{\ell-1}^2}{\lambda_{D\ell-1}}, \xi_{\min} \right\} \quad (29)$$

where ξ_{\min} is a lower bound on the *a priori* SNR. The steps of the causal recursive spectral enhancement algorithm are summarized in Table I. The algorithm is initialized at frame $\ell = -1$ with $\hat{A}_{-1} = 0$ and $\hat{\xi}_{-1|-1} = \xi_{\min}$. Then, for $\ell = 0, 1, \dots$, the propagation and update steps are iterated to obtain estimates for the nonstationary *a priori* SNR. The gain function $G(\hat{\xi}_{\ell|\ell}, \gamma_{\ell})$ employed for the spectral enhancement step is determined by the particular choice of the distortion measure.

C. Relation to “Decision-Directed” Estimation

The proposed causal conditional estimator $\hat{\xi}_{\ell|\ell}$ for the *a priori* SNR is closely related to the decision-directed estimator of Ephraim and Malah [2]. The decision-directed estimator is given by

$$\hat{\xi}_{\ell|\ell}^{\text{DD}} = \mu \frac{\hat{A}_{\ell-1}^2}{\lambda_{D\ell-1}} + (1 - \mu) \max\{\gamma_{\ell} - 1, 0\} \quad (30)$$

where μ ($0 \leq \mu \leq 1$) is a weighting factor that controls the trade-off between the noise reduction and the transient distortion introduced into the signal [2], [20]. A larger value of μ results in a greater reduction of the musical noise phenomena, but at the expense of attenuated speech onsets and audible modifications of transient components. As a compromise, a value 0.98 of μ was determined by simulations and informal listening tests [2].

The update step (26) of the causal conditional estimator can be written as (see the Appendix)

$$\hat{\xi}_{\ell|\ell} = \alpha_{\ell} \hat{\xi}_{\ell|\ell-1} + (1 - \alpha_{\ell})(\gamma_{\ell} - 1) \quad (31)$$

where α_{ℓ} is defined by

$$\alpha_{\ell} \triangleq 1 - \frac{\hat{\xi}_{\ell|\ell-1}^2}{(1 + \hat{\xi}_{\ell|\ell-1})^2}. \quad (32)$$

Substituting (29) into (31) and (32) with $\alpha \equiv 1$, and applying the lower bound constraint to $\hat{\xi}_{\ell|\ell}$ rather than $\hat{\xi}_{\ell|\ell-1}$, we have

$$\hat{\xi}_{\ell|\ell} = \max \left\{ \alpha_{\ell} \frac{\hat{A}_{\ell-1}^2}{\lambda_{D\ell-1}} + (1 - \alpha_{\ell})(\gamma_{\ell} - 1), \xi_{\min} \right\}, \quad (33)$$

$$\alpha_{\ell} = 1 - \frac{\hat{A}_{\ell-1}^4}{(\lambda_{D\ell-1} + \hat{A}_{\ell-1}^2)^2}. \quad (34)$$

The expression (33) with $\alpha_{\ell} \equiv \mu$ is actually a practical form of the decision-directed estimator

$$\hat{\xi}_{\ell|\ell}^{\text{DD}} = \max \left\{ \mu \frac{\hat{A}_{\ell-1}^2}{\lambda_{D\ell-1}} + (1 - \mu)(\gamma_{\ell} - 1), \xi_{\min} \right\} \quad (35)$$

that includes a lower bound constraint to further reduce the level of residual musical noise [20]. Accordingly, a special case of the causal recursive estimator with $\alpha \equiv 1$ degenerates to a “decision-directed” estimator with a *time-varying frequency-dependent* weighting factor α_{ℓ} .

It is interesting to note that the weighting factor α_{ℓ} , given by (34), is monotonically decreasing as a function of the instantaneous SNR, $\hat{A}_{\ell-1}^2/\lambda_{D\ell-1}$. A decision-directed estimator with a larger weighting factor is indeed preferable during speech absence (to reduce musical noise phenomena), while a smaller weighting factor is more advantageous during speech presence (to reduce signal distortion) [20]. The above special case of the causal recursive estimator conforms to such a desirable behavior. Moreover, the general form of the causal recursive estimator provides an additional degree of freedom for adjusting the value of α in (29) to the degree of spectral nonstationarity. This may produce even further improvement in the performance.

The different behaviors of the causal recursive estimator $\hat{\xi}_{\ell|\ell}^{\text{RE}}$ (Table I) and the decision-directed estimator $\hat{\xi}_{\ell|\ell}^{\text{DD}}$ (35) are illustrated in the example of Fig. 6. The analyzed signal contains only white Gaussian noise during the first and last 20 frames, and in between it contains an additional sinusoidal component at the displayed frequency with 0 dB SNR.⁴ The signal is transformed into the STFT domain using half overlapping Hamming windows. The *a priori* SNR estimates, $\hat{\xi}_{\ell|\ell}^{\text{RE}}$ and $\hat{\xi}_{\ell|\ell}^{\text{DD}}$, are obtained by using the parameters $\xi_{\min} = -25$ dB, $\alpha = 0.9$, $\mu = 0.98$. Employing d_{LSA} as the distortion measure (see (6)), the spectral amplitude estimate \hat{A}_{ℓ} is recursively obtained by applying $G_{\text{LSA}}(\hat{\xi}_{\ell|\ell}, \gamma_{\ell})$ to the noisy spectral measurements [see (22) and (17)].

Fig. 6 shows that when the *a posteriori* SNR γ_{ℓ} is sufficiently low, the proposed *a priori* SNR estimate is smoother than the decision-directed estimate, which helps reducing the level of musical noise. When γ_{ℓ} increases, the response of the *a priori* SNR $\hat{\xi}_{\ell|\ell}^{\text{RE}}$ is initially slower than $\hat{\xi}_{\ell|\ell}^{\text{DD}}$, but it then builds up faster to the *a posteriori* SNR. When γ_{ℓ} is sufficiently high, $\hat{\xi}_{\ell|\ell}^{\text{DD}}$ follows the *a posteriori* SNR with a delay of 1 frame, whereas $\hat{\xi}_{\ell|\ell}^{\text{RE}}$ follows the *a posteriori* SNR instantaneously. When γ_{ℓ} decreases, the response of $\hat{\xi}_{\ell|\ell}^{\text{RE}}$ is immediate, while that of $\hat{\xi}_{\ell|\ell}^{\text{DD}}$ is delayed by

⁴Note that the SNR is computed in the time domain, whereas the *a priori* and *a posteriori* SNRs are computed in the time-frequency domain. Therefore, the latter SNRs may increase at the displayed frequency well above the average SNR.

TABLE II
SUMMARY OF THE NONCAUSAL RECURSIVE SPEECH ENHANCEMENT ALGORITHM

Initialization: $\hat{A}_{-1} = 0$, $\hat{\xi}_{-1|L-1} = \xi_{\min}$.

For all short-time frames $\ell = 0, 1, \dots$

”Backward estimation”:

$$\hat{\xi}_{\ell|[\ell+1, \ell+L]} = \begin{cases} \frac{1}{L} \sum_{n=1}^L \gamma_{\ell+n} - \beta, & \text{if nonnegative,} \\ 0, & \text{otherwise.} \end{cases}$$

”Backward-forward propagation”:

$$\hat{\xi}'_{\ell|\ell+L} = \max \left\{ \alpha \frac{\hat{A}_{\ell-1}^2}{\lambda_{D_{\ell-1}}} + (1 - \alpha) \left[\alpha' \hat{\xi}_{\ell-1|\ell+L-1} + (1 - \alpha') \hat{\xi}_{\ell|[\ell+1, \ell+L]} \right], \xi_{\min} \right\}$$

”Update” step:

$$\hat{\xi}_{\ell|\ell+L} = \frac{\hat{\xi}_{\ell|\ell+L}}{1 + \hat{\xi}'_{\ell|\ell+L}} \left(1 + \frac{\hat{\xi}'_{\ell|\ell+L} \gamma_{\ell}}{1 + \hat{\xi}'_{\ell|\ell+L}} \right)$$

Spectral enhancement:

$$\hat{X}_{\ell} = G \left(\hat{\xi}_{\ell|\ell+L}, \gamma_{\ell} \right) Y_{\ell}$$

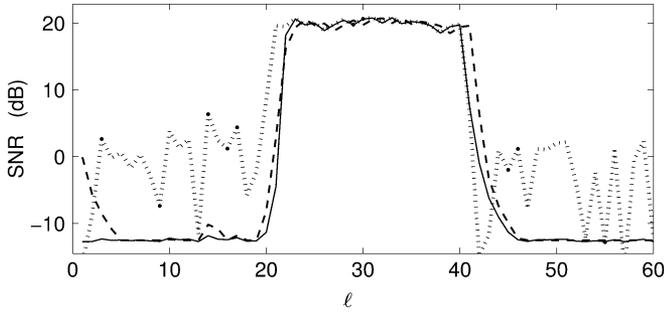


Fig. 6. SNRs in successive short-time frames: A posteriori SNR γ_{ℓ} (dotted line), decision-directed a priori SNR $\hat{\xi}_{\ell|\ell}^{\text{DD}}$ (dashed line), and causal recursive a priori SNR estimate $\hat{\xi}_{\ell|\ell}^{\text{RE}}$ (solid line).

1 frame. As a consequence, we expect that the causal recursive estimator, in comparison with the decision-directed estimator, may produce a lower level of musical noise while not increasing the audible distortion in the enhanced signal.

D. Noncausal Recursive Estimation

In this section, we propose a noncausal conditional estimator $\hat{\xi}_{\ell|\ell+L}$ for the a priori SNR, given the noisy measurements up to frame $\ell + L$, where $L > 0$ denotes the admissible time delay in frames. Similar to the causal estimator, the noncausal estimator combines update and propagation steps to recursively estimate $\lambda_{X_{\ell}}$ as new data arrive. However, future spectral measurements are also employed in the process to better predict the spectral variances of the clean speech.

Let $\lambda'_{X_{\ell}|\ell+L} \triangleq E\{A_{\ell}^2 | \mathcal{Y}_0^{\ell-1}, \mathcal{Y}_{\ell+1}^{\ell+L}\}$ denote the conditional spectral variance of X_{ℓ} given $\mathcal{Y}_0^{\ell+L}$ excluding the noisy measurement at frame ℓ . Let $\lambda_{\ell|[\ell+1, \ell+L]} \triangleq E\{A_{\ell}^2 | \mathcal{Y}_{\ell+1}^{\ell+L}\}$ denote the conditional spectral variance of X_{ℓ} given the subsequent noisy measurements $\mathcal{Y}_{\ell+1}^{\ell+L}$. Then, similar to (25), the estimate for λ_{ℓ} given $\hat{\lambda}'_{X_{\ell}|\ell+L}$ and Y_{ℓ} can be updated by

$$\begin{aligned} \hat{\lambda}_{X_{\ell}|\ell+L} &= E \left\{ A_{\ell}^2 | \hat{\lambda}'_{X_{\ell}|\ell+L}, Y_{\ell} \right\} \\ &= \frac{\hat{\xi}'_{\ell|\ell+L}}{1 + \hat{\xi}'_{\ell|\ell+L}} \left(\frac{1}{\gamma_{\ell}} + \frac{\hat{\xi}'_{\ell|\ell+L}}{1 + \hat{\xi}'_{\ell|\ell+L}} \right) |Y_{\ell}|^2 \end{aligned} \quad (36)$$

where $\hat{\xi}'_{\ell|\ell+L} \triangleq \hat{\lambda}'_{X_{\ell}|\ell+L} / \lambda_{D_{\ell-1}}$ is the a priori SNR estimate given $\mathcal{Y}_0^{\ell-1}$ and $\mathcal{Y}_{\ell+1}^{\ell+L}$. Dividing both sides of (36) by $\lambda_{D_{\ell}}$, we have the “update” step

$$\hat{\xi}_{\ell|\ell+L} = \frac{\hat{\xi}'_{\ell|\ell+L}}{1 + \hat{\xi}'_{\ell|\ell+L}} \left(1 + \frac{\hat{\xi}'_{\ell|\ell+L} \gamma_{\ell}}{1 + \hat{\xi}'_{\ell|\ell+L}} \right). \quad (37)$$

To obtain an estimate for $\lambda'_{X_{\ell}|\ell+L}$, we employ the estimates $\hat{A}_{\ell-1}$ and $\hat{\lambda}_{\ell-1|\ell+L-1}$ from the previous frame, and derive an estimate for $\lambda_{X_{\ell}}$ from the measurements $\mathcal{Y}_{\ell+1}^{\ell+L}$. Suppose an estimate $\hat{\lambda}_{\ell|[\ell+1, \ell+L]}$ is given, we propose to propagate the estimates from frame $\ell - 1$ to frame ℓ by

$$\hat{\lambda}'_{\ell|\ell+L} = \max \left\{ \alpha \hat{A}_{\ell-1}^2 + (1 - \alpha) \left[\alpha' \hat{\lambda}_{\ell-1|\ell+L-1} + (1 - \alpha') \hat{\lambda}_{\ell|[\ell+1, \ell+L]} \right], \xi_{\min} \right\} \quad (38)$$

where α ($0 \leq \alpha \leq 1$) is related to the stationarity of the random process $\{\lambda_{X_{\ell}} | \ell = 0, 1, \dots\}$, and α' ($0 \leq \alpha' \leq 1$) is associated with the reliability of the estimate $\hat{\lambda}_{\ell|[\ell+1, \ell+L]}$ in comparison with that of $\hat{\lambda}_{\ell-1|\ell+L-1}$. Dividing both sides of (38) by $\lambda_{D_{\ell-1}}$, we have the following “backward-forward propagation” step

$$\hat{\xi}'_{\ell|\ell+L} = \max \left\{ \alpha \frac{\hat{A}_{\ell-1}^2}{\lambda_{D_{\ell-1}}} + (1 - \alpha) \left[\alpha' \hat{\xi}_{\ell-1|\ell+L-1} + (1 - \alpha') \hat{\xi}_{\ell|[\ell+1, \ell+L]} \right], \xi_{\min} \right\}. \quad (39)$$

An estimate for the a priori SNR ξ_{ℓ} given the measurements $\mathcal{Y}_{\ell+1}^{\ell+L}$ is obtained by

$$\hat{\xi}_{\ell|[\ell+1, \ell+L]} = \begin{cases} \frac{1}{L} \sum_{n=1}^L \gamma_{\ell+n} - \beta, & \text{if nonnegative,} \\ 0, & \text{otherwise} \end{cases} \quad (40)$$

where β ($\beta \geq 1$) is an over-subtraction factor to compensate for a sudden increase in the noise level. This estimator is an anticausal version of the maximum-likelihood a priori SNR estimator suggested in [2].

The steps of the noncausal recursive spectral enhancement algorithm are summarized in Table II. The algorithm is initialized

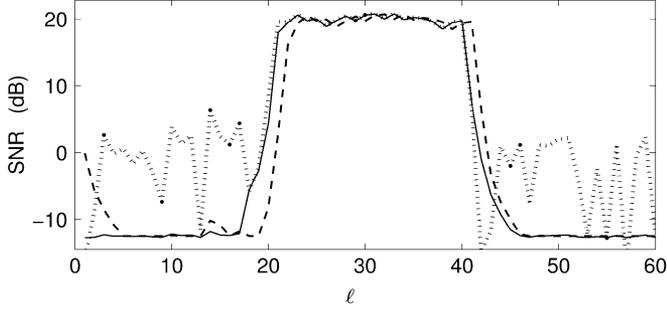


Fig. 7. SNRs in successive short-time frames: *A posteriori* SNR γ_ℓ (dotted line), decision-directed *a priori* SNR $\hat{\xi}_{\ell|\ell}^{\text{DD}}$ (dashed line), and noncausal recursive *a priori* SNR estimate $\hat{\xi}_{\ell|\ell+3}^{\text{RE}}$ with 3 frames delay (solid line).

at frame $\ell = -1$ with $\hat{A}_{-1} = 0$ and $\hat{\xi}_{-1|L-1} = \xi_{\min}$. Then, for $\ell = 0, 1, \dots$, the propagation and update steps are iterated to obtain estimates for the *a priori* SNR and the speech spectral components.

Fig. 7 demonstrates the behavior of the noncausal recursive estimator in the same example of Fig. 6. The noncausal *a priori* SNR estimate $\hat{\xi}_{\ell|\ell+3}^{\text{RE}}$ is obtained with the parameters $\xi_{\min} = -25$ dB, $\alpha = \alpha' = 0.9$, $\beta = 2$, and $L = 3$ frames delay. A comparison of Figs. 6 and 7 indicates that the differences between the causal and noncausal recursive estimators are primarily noticeable during onsets of signal components. Clearly, the *causal a priori* SNR estimator, as well as the decision-directed estimator, cannot respond too fast to an abrupt increase in γ_ℓ , since it necessarily implies an increase in the level of musical residual noise. By contrast, the *noncausal* estimator, having a few subsequent spectral measurements at hand, is capable of discriminating between speech onsets and irregularities in γ_ℓ corresponding to noise only. Therefore, in comparison with the decision-directed estimator, the noncausal *a priori* SNR estimator is expected to produce even lower levels of musical noise and signal distortion.

V. EXPERIMENTAL RESULTS

In this section, the performance of the causal and noncausal recursive estimators are evaluated, and compared to that of the decision-directed estimator. The evaluation includes two objective quality measures, and informal listening tests. The first quality measure is the segmental SNR, in dB, defined by [36] (41), shown at the bottom of the page, where J represents the number of frames in the signal, $N = 512$ is the number of samples per frame (corresponding to 32 ms half overlapping

frames), and \mathcal{T} confines the SNR at each frame to perceptually meaningful range between 35 dB and -10 dB ($\mathcal{T}x \triangleq \min[\max(x, -10), 35]$). The operator \mathcal{T} prevents the segmental SNR measure from being biased in either a positive or negative direction due to a few silence or unusually high SNR frames, that do not contribute significantly to the overall speech quality [37], [38]. The second quality measure is log-spectral distortion (LSD), in dB, which is defined by (42), shown at the bottom of the page, where $\mathcal{C}X_\ell(k) \triangleq \max\{|X_\ell(k)|^2, \delta\}$ is the spectral power, clipped such that the log-spectrum dynamic range is confined to about 50 dB (that is, $\delta = 10^{-50/10} \cdot \max_{k,\ell}\{|X_\ell(k)|^2\}$).

The noise signals used in our evaluation are taken from the Noisex92 database [39]. They include white Gaussian noise, car interior noise, F16 cockpit noise, and babble noise. The speech signal is constructed from six different utterances, without intervening pauses. The utterances, half from male speakers and half from female speakers, are taken from the TIMIT database [33]. The speech signal is sampled at 16 kHz and degraded by the various noise types with segmental SNRs in the range $[-5, 10]$ dB.

The noisy signals are transformed into the STFT domain using half overlapping Hamming analysis windows of 512 samples length. The causal recursive estimation algorithm (Table I) is applied to the noisy speech signals, with parameters $\xi_{\min} = -20$ dB and $\alpha = 0.9$. The noncausal recursive estimation algorithm (Table II) is applied to the noisy signals, with parameters $\xi_{\min} = -20$ dB, $\alpha = \alpha' = 0.9$, $\beta = 2$, and $L = 3$ frames delay. Alternatively, the *a priori* SNR is estimated by the decision-directed method (30), with parameters $\xi_{\min} = -20$ dB and $\mu = 0.98$ (this value of μ was determined in [1], [2] by simulations and informal listening tests).

The spectral gain function used in our evaluation is G_{LSA} (see (22)). The PSD of the noise is estimated by recursively averaging past spectral power values of the noise signal:

$$\hat{\lambda}_{D_\ell} = 0.85\hat{\lambda}_{D_{\ell-1}} + 0.15|D_\ell|^2.$$

In practice, the periodogram of the noise $|D_\ell|^2$ is unknown, and λ_{D_ℓ} can be estimated by using the *Minima Controlled Recursive Averaging* approach [35]. However, to isolate the influence of the *a priori* SNR estimator and to show its importance, a practical noise PSD estimator is not employed to produce the results. In fact, including a practical noise estimator in the speech enhancement algorithms emphasizes the distinction between the proposed and the decision-directed methods, since the noise estimator interacts with the speech estimator and causes the inferior algorithm to be even worse.

$$\text{SegSNR} = \frac{1}{J} \sum_{\ell=0}^{J-1} \mathcal{T} \left\{ 10 \log_{10} \frac{\sum_{n=0}^{N-1} x^2(n + \ell N/2)}{\sum_{n=0}^{N-1} [x(n + \ell N/2) - \hat{x}(n + \ell N/2)]^2} \right\} \quad (41)$$

$$\text{LSD} = \frac{1}{J} \sum_{\ell=0}^{J-1} \left\{ \frac{1}{N/2 + 1} \sum_{k=0}^{N/2} [10 \log_{10} \mathcal{C}X_\ell(k) - 10 \log_{10} \mathcal{C}\hat{X}_\ell(k)]^2 \right\}^{\frac{1}{2}} \quad (42)$$

TABLE III
SEGMENTAL SNR IMPROVEMENT FOR VARIOUS NOISE TYPES AND LEVELS, OBTAINED BY USING THE DECISION-DIRECTED APPROACH (DD), CAUSAL RECURSIVE ESTIMATION (CRE), AND NONCAUSAL RECURSIVE ESTIMATION WITH 3 FRAMES DELAY (NCRE)

Input SegSNR [dB]	Stationary WGN			Car interior noise			F16 cockpit noise			Babble noise		
	DD	CRE	NCRE	DD	CRE	NCRE	DD	CRE	NCRE	DD	CRE	NCRE
-5	7.37	7.37	7.89	7.71	7.76	8.22	6.20	6.23	6.86	6.10	6.17	6.71
0	5.85	5.88	6.53	6.62	6.68	7.18	4.75	4.80	5.52	4.76	4.85	5.51
5	4.39	4.47	5.18	5.46	5.54	6.13	3.41	3.49	4.26	3.50	3.61	4.28
10	3.05	3.20	3.94	4.33	4.45	5.05	2.24	2.36	3.14	2.34	2.46	3.15

TABLE IV
LOG-SPECTRAL DISTANCE FOR VARIOUS NOISE TYPES AND LEVELS, OBTAINED BY USING THE DECISION-DIRECTED APPROACH (DD), CAUSAL RECURSIVE ESTIMATION (CRE), AND NONCAUSAL RECURSIVE ESTIMATION WITH 3 FRAMES DELAY (NCRE)

Input SegSNR [dB]	Stationary WGN			Car interior noise			F16 cockpit noise			Babble noise		
	DD	CRE	NCRE	DD	CRE	NCRE	DD	CRE	NCRE	DD	CRE	NCRE
-5	2.77	2.79	2.65	4.46	4.46	4.37	3.36	3.38	3.19	2.94	2.95	2.77
0	2.16	2.17	1.97	3.61	3.60	3.53	2.40	2.41	2.20	2.08	2.08	1.91
5	1.59	1.58	1.35	2.84	2.83	2.78	1.64	1.63	1.43	1.42	1.41	1.26
10	1.06	1.03	0.82	2.16	2.15	2.11	1.06	1.04	0.88	0.95	0.94	0.82

Table III presents the results of the segmental SNR improvement achieved by the causal and noncausal recursive estimators and by the decision-directed method for various noise types and levels. The noncausal recursive estimator consistently yields a higher improvement in the segmental SNR, than the decision-directed method and the causal recursive estimator, under all tested environmental conditions. The results of the log-spectral distance are summarized in Table IV. It shows that the noncausal recursive estimator obtains lower LSD than the decision-directed method and the causal recursive estimator. A subjective study of speech spectrograms and informal listening tests confirm that the advantages of the noncausal recursive estimator are particularly perceived during onsets of speech and noise only frames. Onsets of speech are better preserved, while a further reduction of noise irregularities (musical noise) is achieved. We note that the results of the segmental SNR and the LSD obtained by using the *causal* recursive estimator are very similar to those obtained by using the decision-directed method. Therefore, in case the delay between the enhanced speech and the noisy observation needs to be minimized, the decision-directed method is perhaps preferable due to its computational simplicity. However, in applications where a few frames delay is tolerable, the *noncausal* recursive estimation approach is definitely more advantageous than the decision-directed approach.

VI. CONCLUSION

We have introduced a statistical model for speech enhancement and *a priori* SNR estimation, which realizes the significance of the statistical dependence between successive speech spectral components. Moreover, it enables consistent derivation of estimators for the speech spectral components and the *a priori* SNR, while keeping the resulting algorithms simple.

We proposed causal and noncausal recursive estimators for the *a priori* SNR. The causal estimator is closely related to the decision-directed estimator of Ephraim and Malah. It degenerates, as a special case, to a “decision-directed” estimator with a *time-varying frequency-dependent* weighting factor, which is monotonically decreasing as a function of the instantaneous SNR. A larger weighting factor is engaged during speech absence, to reduce musical noise phenomena, and a smaller weighting factor evolves during speech presence to reduce signal distortion. The general form of the causal recursive estimator provides an additional degree of freedom, which is adjustable to the degree of spectral nonstationarity. The noncausal recursive estimator, when compared with the causal estimator, is particularly useful during speech onsets. The causal estimator, alike the decision-directed estimator, cannot respond too fast to an abrupt increase in the instantaneous SNR, since it inevitably increases the level of musical residual noise. By contrast, the noncausal estimator, having a few subsequent spectral measurements at hand, is capable of discriminating between speech onsets and noise irregularities. In comparison with the decision-directed estimator, the noncausal estimator produces lower levels of musical noise and signal distortion.

The proposed model can be extended to take into account the statistical dependence between spectral components in distinct frequency-bins. A simple strategy is to “propagate” the spectral variances from frame $\ell-1$ to frame ℓ by considering the spectral variances from all frequency bins, and weighting them in accordance with the time-frequency correlation in the speech signal. A further improvement of the speech enhancement results can be achieved by utilizing the uncertainty of speech presence in the noisy measurements [2]–[4], [40]. In this case, one needs to find also an estimator for the speech presence probability, that is consistent with the model assumptions and the *a priori* SNR estimation.

APPENDIX

Equation (26) can be written as

$$\begin{aligned}\hat{\xi}_{\ell|\ell} &= \frac{\hat{\xi}_{\ell|\ell-1}}{1 + \hat{\xi}_{\ell|\ell-1}} + \left(\frac{\hat{\xi}_{\ell|\ell-1}}{1 + \hat{\xi}_{\ell|\ell-1}} \right)^2 (\gamma_{\ell} - 1) \\ &\quad + \left(\frac{\hat{\xi}_{\ell|\ell-1}}{1 + \hat{\xi}_{\ell|\ell-1}} \right)^2 \\ &= \frac{(2\hat{\xi}_{\ell|\ell-1} + 1)\hat{\xi}_{\ell|\ell-1}}{(1 + \hat{\xi}_{\ell|\ell-1})^2} + \frac{\hat{\xi}_{\ell|\ell-1}^2}{(1 + \hat{\xi}_{\ell|\ell-1})^2} (\gamma_{\ell} - 1).\end{aligned}\quad (43)$$

Define α_{ℓ} by

$$\alpha_{\ell} \triangleq 1 - \frac{\hat{\xi}_{\ell|\ell-1}^2}{(1 + \hat{\xi}_{\ell|\ell-1})^2} = \frac{2\hat{\xi}_{\ell|\ell-1} + 1}{(1 + \hat{\xi}_{\ell|\ell-1})^2}.\quad (44)$$

Then, (43) can be written as (31).

ACKNOWLEDGMENT

The author thanks Prof. D. Malah and Prof. Y. Ephraim for their constructive comments and helpful suggestions. He also thanks the anonymous reviewers for their valuable comments.

REFERENCES

- [1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.
- [2] —, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
- [3] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments," in *Proc. 24th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-99*, Phoenix, AZ, Mar. 15–19, 1999, pp. 789–792.
- [4] I. Cohen and B. Berdugo, "Speech enhancement for nonstationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, Nov. 2001.
- [5] N. S. Kim and J.-H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Processing Lett.*, vol. 7, no. 5, pp. 108–110, May 2000.
- [6] M. Kleinschmidt, J. Tchorz, and B. Kollmeier, "Combining speech enhancement and auditory feature extraction for robust speech recognition," *Speech Commun.*, vol. 34, no. 1–2, pp. 75–91, Apr. 2001.
- [7] A. J. Accardi and R. V. Cox, "Robust digit recognition in noise: An evaluation using the AURORA corpus," in *Proc. 24th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-99*, Phoenix, AZ, Mar. 15–19, 1999, pp. 201–204.
- [8] —, "A modular approach to speech enhancement with an application to speech coding," in *Proc. 24th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-99*, Phoenix, AZ, Mar. 15–19, 1999, pp. 201–204.
- [9] J. Thyssen, Y. Gao, A. Benyassine, E. Shlomot, C. C. Murgia, H.-Y. Su, K. Mano, Y. Hiwasaki, H. Ehara, K. Yasunaga, C. Lamblin, B. Kovcsi, J. Stegmann, and H.-G. Kang, "A candidate for the ITU-T 4 kbit/s speech coding standard," in *Proc. 26th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-OI*, Salt Lake City, UT, May 7–11, 2001, pp. 681–684.
- [10] T. Agarwal and P. Kabal, "Pre-processing of noisy speech for voice coders," in *Proc. IEEE Workshop on Speech Coding*, Tsukuba, Japan, Oct. 6–9, 2002, pp. 169–171.
- [11] T. Fillon and J. Prado, "Evaluation of an ERB frequency scale noise reduction for hearing aids: A comparative study," *Speech Commun.*, vol. 39, no. 1–2, pp. 23–32, Jan. 2003.
- [12] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proc. 27th IEEE Internal. Conf. Acoust. Speech Signal Process., ICASSP-02*, Orlando, FL, May 13–17, 2002, pp. I-253–I-256.
- [13] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detector," *IEEE Signal Processing Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
- [14] Y. D. Cho and A. Kondo, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Processing Lett.*, vol. 8, no. 10, pp. 276–278, Oct. 2001.
- [15] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 498–505, Sep. 2003.
- [16] M. Matassoni, M. Omologo, A. Santarelli, and P. Svaizer, "On the joint use of noise reduction and MLLR adaptation for in-car hands-free speech recognition," in *Proc. 27th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-02*, Orlando, FL, May 13–15, 2002, pp. I-289–I-292.
- [17] I. Cohen, S. Gannot, and B. Berdugo, "An integrated real-time beamforming and postfiltering system for nonstationary noise environments," *Special Issue of EURASIP JASP on Signal Processing for Acoustic Communication System*, vol. 2003, no. II, pp. 1064–1073, Oct. 2003.
- [18] A. Gilloire, P. Scalart, C. Lamblin, C. Mokbel, and S. Proust, "Innovative speech processing for mobile terminals: An annotated bibliography," *Signal Process.*, vol. 80, no. 7, pp. 1149–1166, July 2000.
- [19] Y. Ephraim and I. Cohen, "Recent advancements in speech enhancement," in *The Electrical Engineering Handbook*, 3rd ed. Boca Raton, FL: CRC Press, to be published [Online]. Available: <http://ece.gmu.edu/~yephraim/ephraim.html>.
- [20] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 2, no. 2, pp. 345–349, Apr. 1994.
- [21] C. Breithaupt and R. Martin, "MMSE estimation of magnitude-squared DFT coefficients with supergaussian priors," in *Proc. 28th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-03*, Hong Kong, Apr. 6–10, 2003, pp. I-896–I-899.
- [22] B. H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 6, pp. 1404–1413, Dec. 1985.
- [23] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.
- [24] H. Sheikhzadeh and L. Deng, "Waveform-based speech recognition using hidden filter models: Parameter selection and sensitivity to power normalization," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 80–91, Jan. 1994.
- [25] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1518–1568, June 2002.
- [26] C. J. Wellekens, "Explicit time correlations in hidden Markov models for speech recognition," in *Proc. 12th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-87*, Dallas, TX, Apr. 6–9, 1987, pp. 384–386.
- [27] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 40, no. 4, pp. 725–735, Apr. 1992.
- [28] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 5, pp. 445–455, Sep. 1998.
- [29] J. Wexler and S. Raz, "Discrete Gabor expansions," *Speech Process.*, vol. 21, no. 3, pp. 207–220, Nov. 1990.
- [30] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1983.
- [31] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.
- [32] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP J. Appl. Signal Process., Special Issue on Digital Audio for Multimedia Communications*, vol. 2003, no. 10, pp. 1043–1051, Sep. 2003.
- [33] J. S. Garofolo, "Getting Started with the DARPA TIM1T CD-ROM: An Acoustic Phonetic Continuous Speech Database," National Institute of Standards and Technology (NIST), Gaithersburg, MD, Tech. Rep., Dec. 1988. prototype as of.
- [34] A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics*, 6th ed. London, U.K.: Edward Arnold, 1994, vol. 1.
- [35] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept. 2003.
- [36] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
- [37] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, 2nd ed. New York: IEEE Press, 2000.

- [38] P. E. Papamichalis, *Practical Approaches to Speech Coding*. Englewood Cliffs, NJ: Prentice-Hall, 1987.
- [39] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.
- [40] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.
- [41] I. Cohen, "On the decision-directed estimation approach of Ephraim and Malah," in *Proc. 29th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-2004*, vol. I, Montreal, QC, Canada, May 17–21, 2004, pp. 293–296.
- [42] I. Cohen, "Speech enhancement using a noncausal *A Priori* SNR estimator," *IEEE Signal Process. Lett.*, vol. 11, no. 9, pp. 725–728, Sep. 2004.



Israel Cohen (M'01–SM'03) received the B.Sc. (summa cum laude), M.Sc., and Ph.D. degrees in electrical engineering in 1990, 1993, and 1998, respectively, all from The Technion—Israel Institute of Technology, Haifa.

From 1990 to 1998, he was a Research Scientist at RAFAEL Research Laboratories, Haifa, Israel Ministry of Defense. From 1998 to 2001, he was a Post-doctoral Research Associate at the Computer Science Department, Yale University, New Haven, CT. Since 2001, he has been a Senior Lecturer with the Elec-

trical Engineering Department, The Technion. His research interests are statistical signal processing, analysis, and modeling of acoustic signals, speech enhancement, noise estimation, microphone arrays, source localization, blind source separation, system identification, and adaptive filtering.

Dr. Cohen serves as Associate Editor for the *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING* and *IEEE SIGNAL PROCESSING LETTERS*, and as guest editor for a special issue of the *EURASIP Journal on Applied Signal Processing on Advances in Multimicrophone Speech Processing*.