CCIT Report #443
October 2003

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

Israel Cohen

Abstract

The widely-used speech enhancement method of Ephraim and Malah is based on a Gaussian statistical model, presuming spectral components are statistically independent. A major drawback is that the model assumptions conflict with the "decision-directed" *a priori* SNR estimation, which heavily relies on the time-correlation of speech spectra. In this paper, we propose a statistical model for speech enhancement that i) takes into account the time-correlation between successive speech spectral components; ii) admits consistent estimators for the *a priori* SNR and the speech spectral components; iii) retains the simplicity associated with the Ephraim-Malah statistical model; iv) provides insight into the decision-directed approach; and v) enables the extension of existing algorithms to noncausal estimation. In the proposed model, the sequence of speech spectral variances is a random process, which is correlated with the sequence of 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. We show that a special case of the causal estimator degenerates to a "decision-directed" estimator with a *time-varying* weighting factor. Experimental results demonstrate the improved performance of the proposed algorithms.

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].

Two decades ago, Ephraim and Malah proposed a statistical model for speech enhancement [2], [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 spectral components are statistically independent is clearly unfulfilled. However, it 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

The author is with the Department of Electrical Engineering, Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel (email: icohen@ee.technion.ac.il; tel.: +972-4-8294731; fax: +972-4-8295757).

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.

Cappé [20] showed that the dominant factor in the Ephraim-Malah algorithm is the decision-directed estimation approach for the *a priori* 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* SNR's in successive short-term frames are highly correlated.

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 (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.

A major drawback of the above statistical models is that the model assumptions conflict with the decision-directed approach. On the one hand, spectral components are assumed statistically independent when deriving analytical expressions for the speech estimators. On the other hand, the *a priori* SNR, which is the dominant parameter of the speech estimators [20], [22], is obtained by the decision-directed approach, which heavily relies on the strong time-correlation between successive speech spectral components. Quite remarkably, despite this inconsistency, the performance of the LSA algorithm, versus its computational simplicity, is outstanding.

Enhancement schemes based on hidden Morkov models (HMM's) try to circumvent the assumption of specific distributions for the speech and noise processes [23]–[26]. 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 non-diagonal covariance matrices for each subsource, and assuming that a sequence of vectors generated from a given sequence of states is a nonzero order autoregressive process [24], [27]. First order HMM's, for example, with a mixture of Gaussian distributions in each state and minimum mean-square error estimation results in a weighted sum of conditional mean estimators, one for each mixture component in each state, where the weights are the posterior probabilities of the states and mixture components given the noisy signal [28]. Unfortunately, the HMM-based speech enhancement relies on the type of training data [29]. 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 i) takes into account the time-correlation between successive speech spectral components; ii) admits consistent estimators for the *a priori* SNR and the speech spectral components; iii) retains the simplicity associated with the Ephraim-Malah statistical model; iv) provides insight into the decision-directed approach; and v) enables the extension of existing algorithms to noncausal estimation. In the proposed model, the sequence of speech spectral variances is a random process, which is correlated with the sequence of 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, 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* 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 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,

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

where k is the frequency-bin index ($k = 0, 1, \ldots, N-1$), ℓ is the time frame index ($\ell = 0, 1, \ldots$), $h(n)$ 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, an estimate for the clean speech signal is obtained by applying the inverse STFT,

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

where $\tilde{h}(n)$ is a synthesis window that is biorthogonal to the analysis window $h(n)$ [30], and the inverse STFT is efficiently implemented by using the weighted overlap-add method [31].

Let $\mathcal{Y}_0^{\ell'}(k)$ denote a set of spectral measurements $\{Y_0(k), \ldots, Y_{\ell'}(k)\}$, and let $d\left[X_\ell(k), \hat{X}_\ell(k)\right]$ 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

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

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 the spectral components are *not* assumed statistically independent. Therefore, in contrast to existing spectral enhancement techniques (*e.g.*, [1], [2], [4], [12], [32]), the estimation problem is not formulated as that of estimating $X_\ell(k)$ from $Y_\ell(k)$ alone.

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:

1) The squared-error distortion [33]:

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

2) The spectral amplitude distortion [2]:

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

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

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

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

4) The spectral power distortion [21], [28], [32]:

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

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]:

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

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:

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

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 $\beta \ll 1$ [2].

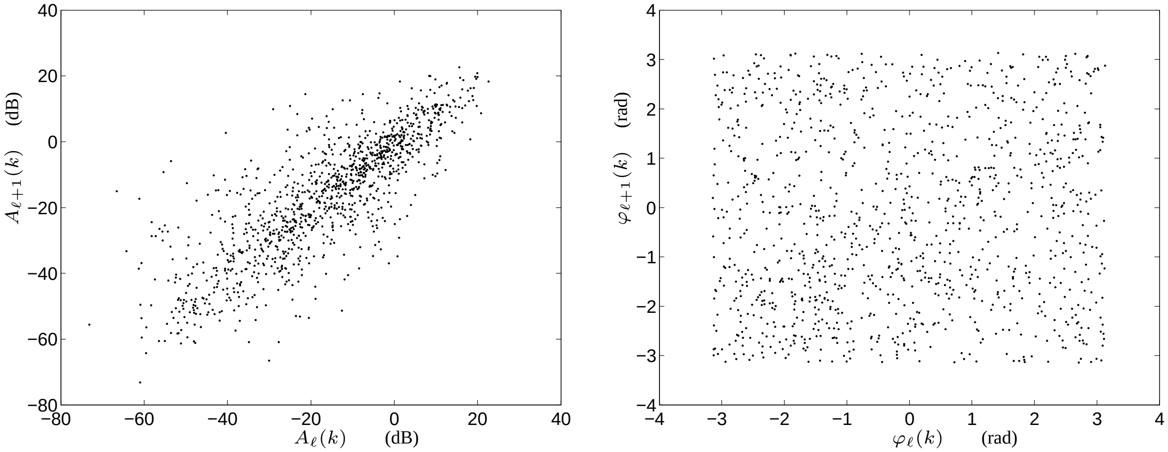
III. Speech Spectral Model

In this section, we propose a statistical model that takes into account the time-correlation between successive spectral components of the speech signal. In particular, the Gaussian statistical model of Ephraim and Malah [2] is relaxed by assuming that $\{X_0(k), X_1(k), \ldots\}$ are statistically dependent.

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 phases, and investigate the autocorrelation sequences (ACS's) of 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 [34]. The speech signal is sampled at 16 kHz, and transformed into the STFT domain using Hamming analysis windows of 512 samples length, and 256 samples framing step (50% overlap between successive frames).

Figure 1 shows scatter plots for successive spectral magnitudes and phases of the speech signal, at center frequency 500 Hz ($k = 17$). Similar plots are obtained for other frequency-bins, whatever speech signals are taken. Figure 2 shows scatter plots for successive spectral magnitudes and phases of a *white Gaussian noise* (WGN) signal. These figures imply that successive spectral magnitudes of speech signals are highly correlated, whereas successive spectral phases are much less correlated. In contrast, successive spectral magnitudes of a WGN signal are weakly correlated.

Figure 3 shows the ACS's of the speech spectral components along time-trajectories, for various frequency-bins and framing steps. The 95 percent confidence limits (*e.g.,* [35]) are depicted as horizontal dotted lines. In order to prevent an upward bias of the autocorrelation estimates due to irrelevant (non-speech) spectral components, the



(a)

(b)

Fig. 1. Scatter plots for successive spectral components of a *speech* signal, at center frequency 500 Hz ($k = 17$). (a) Scatter plot for successive spectral magnitudes; (b) Scatter plot for successive spectral phases.

ACS's are computed from spectral components whose magnitudes are within 30 dB of the maximal magnitude. Specifically, the autocorrelation coefficients of the spectral magnitudes are estimated by

$$\rho(m) \triangleq \frac{E\left\{A_\ell(k)A_{\ell+m}(k)\right\}}{E\left\{A_\ell^2(k)\right\}} \approx \frac{\sum_{\ell\in\mathcal{L}} A_\ell(k)A_{\ell+m}(k)}{\sum_{\ell\in\mathcal{L}} A_\ell^2(k)} \tag{10}$$

where m is the lag in frames, and $\mathcal{L}$ represents the set of relevant spectral components

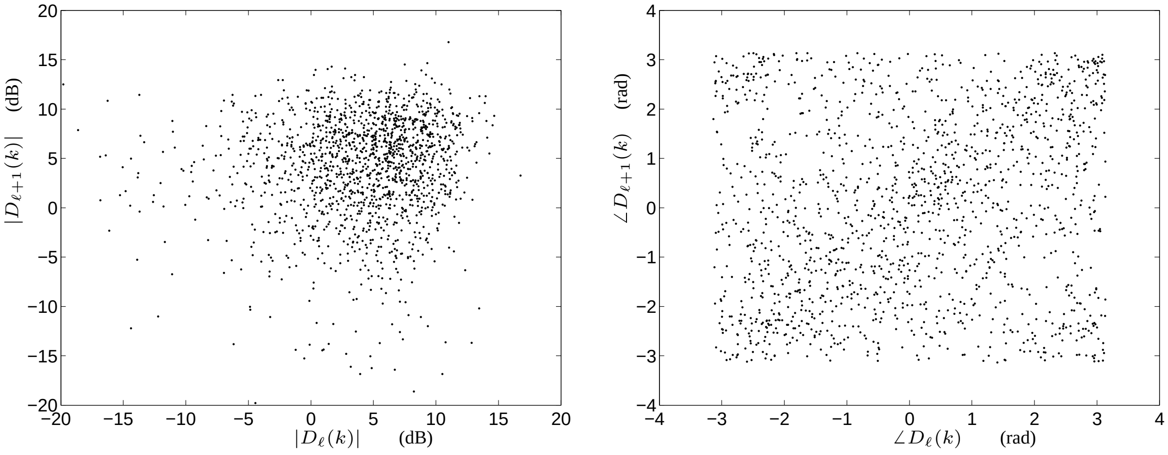
$$\mathcal{L} = \left\{ \ell \;\middle|\; A_\ell(k) \geq 10^{-30/20} \max_{\ell}\{A_\ell(k)\} \right\} .$$

The corresponding autocorrelation coefficients of the spectral phases are obtained by

$$\rho(m) \triangleq \frac{E\left\{\varphi_\ell(k)\,\varphi_{\ell+m}(k)\right\}}{E\left\{\varphi_\ell^2(k)\right\}} \approx \frac{\sum_{\ell\in\mathcal{L}} \varphi_\ell(k)\,\varphi_{\ell+m}(k)}{\sum_{\ell\in\mathcal{L}} \varphi_\ell^2(k)} . \tag{11}$$

Figure 4 shows the variation of the correlation between successive spectral magnitudes on frequency and on overlap between successive frames. Figures 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.

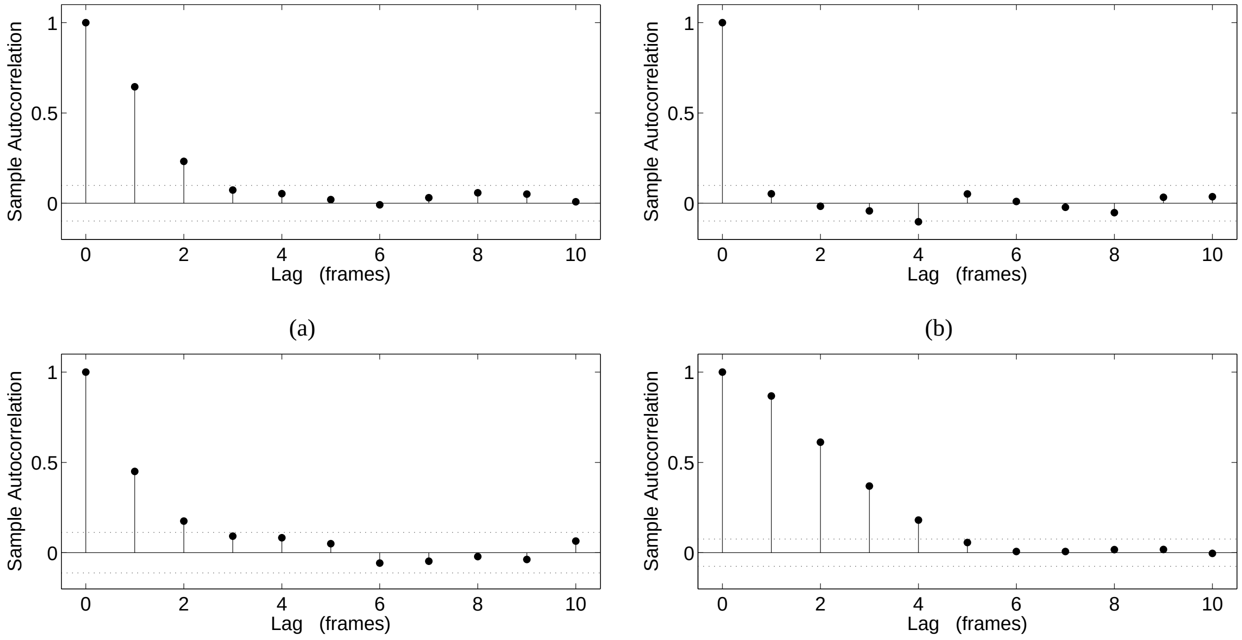
Figure 5 shows the ACS's of WGN spectral magnitude along time-trajectories, for various framing steps. Figure 6 demonstrates, for a realization of WGN, the variation of the correlation between successive spectral magnitudes on the overlap between frames. A comparison of Figs. 6 and 4 reveals that for 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



(a)

(b)

Fig. 2. Scatter plots for successive spectral components of a *white Gaussian noise* signal ($k = 17$). (a) Scatter plot for successive spectral magnitudes; (b) Scatter plot for successive spectral phases.



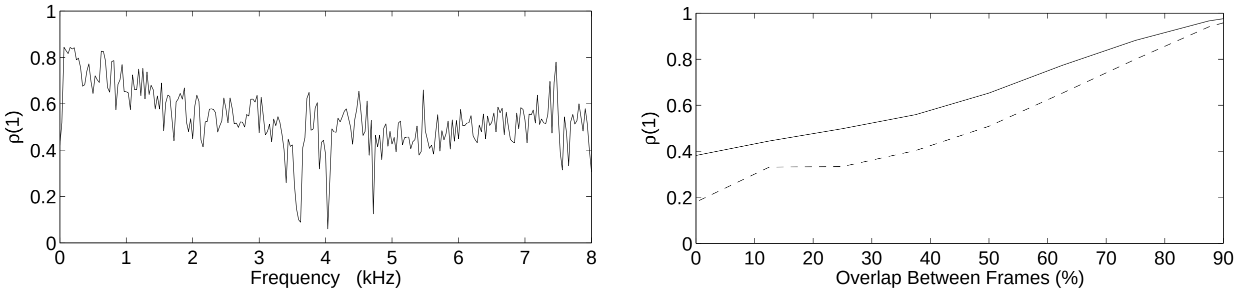
(a)

(b)

(c)

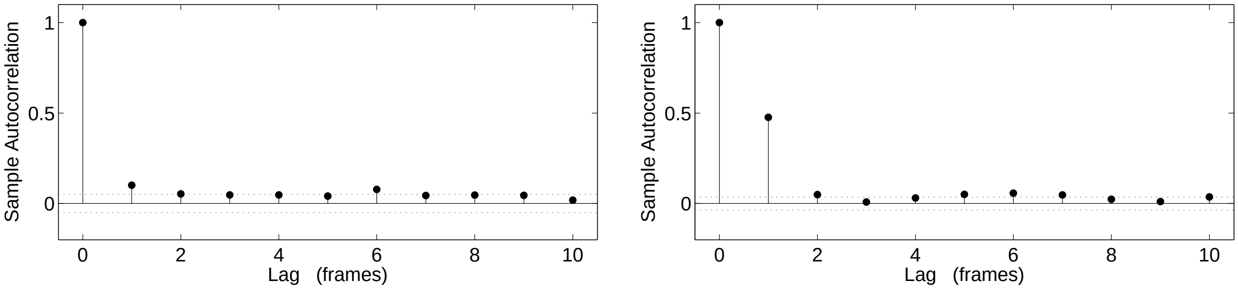
(d)

Fig. 3. Autocorrelation sequences (ACS's) 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), $M = N/2$; (d) ACS of the spectral magnitude, $k = 17$, $M = N/4$ (75% overlap between frames).



(a) (b)

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); (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).



(a) (b)

Fig. 5. Autocorrelation sequences of white Gaussian noise spectral magnitude (along time-trajectories) for various framing steps. The dotted lines represents 95 percent confidence limits. (a) $M = N/2$ (50% overlap between frames); (b) $M = N/4$ (75% overlap between frames).

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 manifested 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) For a fixed frequency-bin index k, the sequence of speech spectral magnitudes $\{A_\ell(k) \mid \ell = 0, 1, \ldots\}$ is a

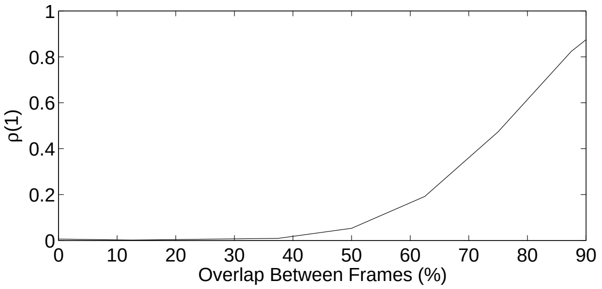


Fig. 6. Variation of the correlation between successive spectral magnitudes on the overlap between frames for a realization of white Gaussian noise.

random process. For $k \neq k'$, the two random processes $\{A_\ell(k) \mid \ell = 0, 1, \ldots\}$ and $\{A_\ell(k') \mid \ell = 0, 1, \ldots\}$ are statistically independent.

4) For fixed k and ℓ, a speech spectral component $X_\ell(k)$ is a zero-mean complex Gaussian random variable. Its real and imaginary parts are IID.

5) The sequence of speech spectral variances $\{\lambda_{X_\ell}(k) \mid \ell = 0, 1, \ldots\}$, where $\lambda_{X_\ell}(k) \overset{\triangle}{=} E\left\{A_\ell^2(k)\right\}$, is a random process. For fixed k and ℓ, $\lambda_{X_\ell}(k)$ is correlated with the sequence of speech spectral magnitudes $\{A_{\ell'}(k) \mid \ell' = 0, 1, \ldots\}$. However, given $\lambda_{X_\ell}(k)$, $A_\ell(k)$ is statistically independent of $A_{\ell'}(k)$ for $\ell' \neq \ell$.

Note that the fundamental difference between the proposed statistical model and that of Ephraim and Malah originates from the last assumption. Here, the variance sequence of $X_\ell(k)$ is a random process, rather than a sequence of parameters. Furthermore, successive spectral components are correlated, as the random processes $\{X_\ell(k) \mid \ell = 0, 1, \ldots\}$ and $\{\lambda_{X_\ell}(k) \mid \ell = 0, 1, \ldots\}$ are not independent.

IV. Signal Estimation

In this section, we derive estimators for $X_\ell(k)$, as formulated in (3), 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* SNR's. 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 PSD, which in practice can be estimated by using the *Minima Controlled Recursive Averaging* approach [36].

A. *Spectral Enhancement*

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

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

Applying Bayes' rule to the conditional pdf of X_ℓ, we have

$$p\left(X_\ell \,|\, \mathcal{Y}_0^{\ell'},\, \lambda_{X_\ell}\right) = \frac{p\left(Y_\ell \,|\, X_\ell,\, \mathcal{Y}_0^{\ell-1},\, \mathcal{Y}_{\ell+1}^{\ell'},\, \lambda_{X_\ell}\right) p\left(X_\ell \,|\, \mathcal{Y}_0^{\ell-1},\, \mathcal{Y}_{\ell+1}^{\ell'},\, \lambda_{X_\ell}\right)}{\int p\left(Y_\ell \,|\, X_\ell,\, \mathcal{Y}_0^{\ell-1},\, \mathcal{Y}_{\ell+1}^{\ell'},\, \lambda_{X_\ell}\right) p\left(X_\ell \,|\, \mathcal{Y}_0^{\ell-1},\, \mathcal{Y}_{\ell+1}^{\ell'},\, \lambda_{X_\ell}\right)\, dX_\ell}\, . \tag{13}$$

The proposed statistical model (particularly the first and last model assumptions) implies

$$p\left(Y_\ell \,|\, X_\ell,\, \mathcal{Y}_0^{\ell-1},\, \mathcal{Y}_{\ell+1}^{\ell'},\, \lambda_{X_\ell}\right) \ = \ p\left(Y_\ell \,|\, X_\ell\right)\, , \tag{14}$$

$$p\left(X_\ell \,|\, \mathcal{Y}_0^{\ell-1},\, \mathcal{Y}_{\ell+1}^{\ell'},\, \lambda_{X_\ell}\right) \ = \ p\left(X_\ell \,|\, \lambda_{X_\ell}\right)\, . \tag{15}$$

Approximating the conditional pdf of λ_{X_ℓ} given the noisy observations $\mathcal{Y}_0^{\ell'}$ by a Dirac delta function at position $\lambda_{X_\ell|\ell'} \triangleq E\left\{A_\ell^2(k) \,|\, \mathcal{Y}_0^{\ell'}\right\}$, and substituting (14) and (15) into (13), the spectral estimator $\hat{X}_\ell(k)$ is obtained from

$$\begin{aligned} &\min_{\hat{X}_\ell} \iint d\left(X_\ell, \hat{X}_\ell\right) p\left(X_\ell \,|\, Y_\ell,\, \lambda_{X_\ell}\right) \delta\left(\lambda_{X_\ell} - \lambda_{X_\ell|\ell'}\right)\, dX_\ell\, d\lambda_{X_\ell} \\ &= \min_{\hat{X}_\ell} \int d\left(X_\ell, \hat{X}_\ell\right) p\left(X_\ell \,|\, Y_\ell,\, \lambda_{X_\ell|\ell'}\right)\, dX_\ell\, . \end{aligned} \tag{16}$$

That is, given the set of noisy measurements $\mathcal{Y}_0^{\ell'}$, we first derive an estimate for the clean speech spectral variance $\lambda_{X_\ell|\ell'}$ at frame ℓ. Subsequently, the estimation problem for the speech spectral component X_ℓ reduces to that of estimating X_ℓ from Y_ℓ alone, assuming knowledge of the variance of X_ℓ. The latter problem, when the *a priori* SNR is defined appropriately, is essentially the classical spectral enhancement problem as formulated by Ephraim and Malah [1], [2]. As a result, an estimate for X_ℓ is obtained by applying a spectral gain function to each noisy spectral component of the speech signal:

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

where the *a priori* and *a posteriori* SNR's are defined respectively by[2]

$$\xi_{\ell|\ell'} \ \triangleq \ \frac{\lambda_{X_\ell|\ell'}}{\lambda_{D_\ell}} \tag{18}$$

$$\gamma_\ell \ \triangleq \ \frac{|Y_\ell|^2}{\lambda_{D_\ell}} \tag{19}$$

[2] Note 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.

and where $\lambda_{D_\ell} \triangleq E\left\{|D_\ell|^2\right\}$ denotes the noise spectral variance. The specific expression for the spectral gain function $G\left(\xi_{\ell|\ell'}, \gamma_\ell\right)$ depends on the particular choice of a distortion measure $d\left(X_\ell, \hat{X}_\ell\right)$. For squared-error distortion, the gain function is given by [33]

$$G_{\mathrm{SE}}\left(\xi_{\ell|\ell'}\right) = \frac{\xi_{\ell|\ell'}}{1+\xi_{\ell|\ell'}} \,. \tag{20}$$

In case of combining the spectral amplitude, the log-spectral amplitude, or the spectral power distortion measures with the constrained optimization problem (8), the gain functions can respectively be written as [1], [2], [28], [32]

$$\begin{aligned} G_{\mathrm{SA}}\left(\xi_{\ell|\ell'}, \gamma_\ell\right) &= \frac{\sqrt{\pi\, \upsilon_\ell}}{2\gamma_\ell}\left[(1+\upsilon_\ell) I_0\left(\frac{\upsilon_\ell}{2}\right) + \upsilon_\ell\, I_1\left(\frac{\upsilon_\ell}{2}\right)\right]\exp\left(-\frac{\upsilon_\ell}{2}\right) & (21)\\ G_{\mathrm{LSA}}\left(\xi_{\ell|\ell'}, \gamma_\ell\right) &= \frac{\xi_{\ell|\ell'}}{1+\xi_{\ell|\ell'}}\exp\left(\frac{1}{2}\int_{\upsilon_\ell}^{\infty}\frac{e^{-t}}{t}dt\right) & (22)\\ G_{\mathrm{SP}}\left(\xi_{\ell|\ell'}, \gamma_\ell\right) &= \sqrt{\frac{\xi_{\ell|\ell'}}{1+\xi_{\ell|\ell'}}\left(\frac{1}{\gamma_\ell}+\frac{\xi_{\ell|\ell'}}{1+\xi_{\ell|\ell'}}\right)} & (23) \end{aligned}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of zero and first order, respectively, and υ_ℓ is defined by $\upsilon_\ell \triangleq \xi_{\ell|\ell'}\,\gamma_\ell/(1+\xi_{\ell|\ell'})$. It still remains to estimate the *a priori* SNR $\xi_{\ell|\ell'}$, as defined in (18), based on the statistical model.

B. Causal Recursive Estimation

In this subsection, 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, 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 $\hat{\lambda}_{X_\ell|\ell-1}$:

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

This is obtained by applying the gain function $G_{\mathrm{SP}}\left(\hat{\xi}_{\ell|\ell-1}, \gamma_\ell\right)$ to Y_ℓ, and computing the squared absolute value of the result[3]

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

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

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

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

We call (26) the "update" step.

Computation of the update step requires the estimate

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

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 elementary nonlinear predictor of the form

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

where α ($0 \leq \alpha \leq 1$) is related to the degree of nonstationarity of the random process $\{\lambda_{X_\ell} \mid \ell = 0, 1, \ldots\}$, and $\lambda_{\min}$ 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 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

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

where $\xi_{\min}$ 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 $\ell = 0, 1, \ldots$, the propagation and update steps are iterated to obtain estimates for the nonstationary *a priori* SNR. The gain function $G\left(\hat{\xi}_{\ell|\ell}, \, \gamma_\ell\right)$ 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

$$\hat{\xi}^{\mathrm{DD}}_{\ell|\ell} = \mu\frac{\hat{A}^2_{\ell-1}}{\lambda_{D_{\ell-1}}} + (1-\mu)\max\left\{\gamma_\ell - 1, 0\right\} \tag{30}$$

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

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 $\ell = 0, 1, \ldots$
    “Propagation” step:
        $\hat{\xi}_{\ell|\ell-1} = \max\left\{(1-\alpha)\hat{\xi}_{\ell-1|\ell-1} + \alpha\frac{\hat{A}^2_{\ell-1}}{\lambda_{D_{\ell-1}}}\,,\ \xi_{\min}\right\}$
    “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:
        $\hat{X}_\ell = G\left(\hat{\xi}_{\ell|\ell},\,\gamma_\ell\right)\,Y_\ell$
```

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

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

where α_ℓ is defined by

$$\alpha_\ell \overset{\triangle}{=} 1 - \frac{\hat{\xi}^2_{\ell|\ell-1}}{\left(1+\hat{\xi}_{\ell|\ell-1}\right)^2}\,. \tag{32}$$

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

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

$$\alpha_\ell \quad = \quad 1 - \frac{\hat{A}^4_{\ell-1}}{\left(\lambda_{D_{\ell-1}} + \hat{A}^2_{\ell-1}\right)^2}\,. \tag{34}$$

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

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

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* 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, $\hat{A}^2_{\ell-1}/\,\lambda_{D_{\ell-1}}$. A decision-directed estimator with a larger weighting factor is indeed preferable

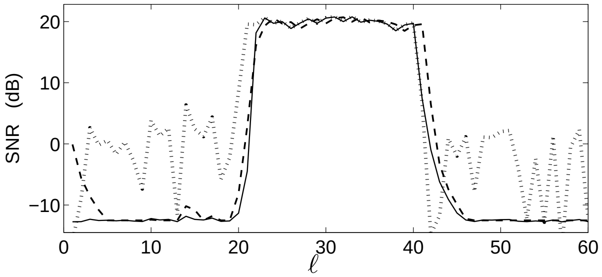


Fig. 7. SNR's in successive short-time frames: *A posteriori* SNR γ_ℓ (dotted line), decision-directed *a priori* SNR $\hat{\xi}^{\mathrm{DD}}_{\ell|\ell}$ (dashed line), and causal recursive *a priori* SNR estimate $\hat{\xi}^{\mathrm{RE}}_{\ell|\ell}$ (solid line).

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 $\hat{\xi}^{\mathrm{RE}}_{\ell|\ell}$ (Table I) and the decision-directed estimator $\hat{\xi}^{\mathrm{DD}}_{\ell|\ell}$ (35) are illustrated in the example of Fig. 7. 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. The signal is transformed to the STFT domain using Hamming windows with 50% overlap between successive frames. The *a priori* SNR estimates, $\hat{\xi}^{\mathrm{RE}}_{\ell|\ell}$ and $\hat{\xi}^{\mathrm{DD}}_{\ell|\ell}$, are obtained by using the G_{LSA} spectral gain function (22), and the parameters $\xi_{\min} = -25$ dB, $\alpha = 0.9$, $\mu = 0.98$. It 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 $\hat{\xi}^{\mathrm{RE}}_{\ell|\ell}$ is initially slower than $\hat{\xi}^{\mathrm{DD}}_{\ell|\ell}$, but it then builds up faster to the *a posteriori* SNR. When γ_ℓ is sufficiently high, $\hat{\xi}^{\mathrm{DD}}_{\ell|\ell}$ follows the *a posteriori* SNR with a delay of 1 frame, whereas $\hat{\xi}^{\mathrm{RE}}_{\ell|\ell}$ follows the *a posteriori* SNR instantaneously. When γ_ℓ decreases, the response of $\hat{\xi}^{\mathrm{RE}}_{\ell|\ell}$ is immediate, while that of $\hat{\xi}^{\mathrm{DD}}_{\ell|\ell}$ is delayed by 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 subsection, 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 λ_{X_ℓ} 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 $\lambda'_{X_\ell|\ell+L} \triangleq E\left\{A_\ell^2 \mid \mathcal{Y}_0^{\ell-1}, \mathcal{Y}_{\ell+1}^{\ell+L}\right\}$ denote the conditional spectral variance of X_ℓ given $\mathcal{Y}_0^{\ell+L}$ excluding the noisy measurement at frame ℓ. Let $\lambda_{\ell\,|\,[\ell+1,\ell+L]} \triangleq E\left\{A_\ell^2 \mid \mathcal{Y}_{\ell+1}^{\ell+L}\right\}$ 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

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

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 λ_{D_ℓ}, we have the "update" step

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

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 λ_{X_ℓ} 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

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

where α $(0 \leq \alpha \leq 1)$ is related to the stationarity of the random process $\{\lambda_{X_\ell} \mid \ell = 0, 1, \ldots\}$, 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:

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

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

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

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

Figure 8 demonstrates the behavior of the noncausal recursive estimator in the same example of Fig. 7. The noncausal *a priori* SNR estimate $\hat{\xi}^{\mathrm{RE}}_{\ell|\ell+3}$ is obtained with the parameters $\xi_{\min} = -25$ dB, $\alpha = \alpha' = 0.9$, $\beta = 2$, and $L = 3$ frames delay. A comparison of Figs. 7 and 8 indicates that the differences between the causal and noncausal recursive estimators are primarily noticeable during onsets of signal components. Clearly, the *causal a*

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 $\ell = 0, 1, \ldots$

”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”:

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

“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:

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

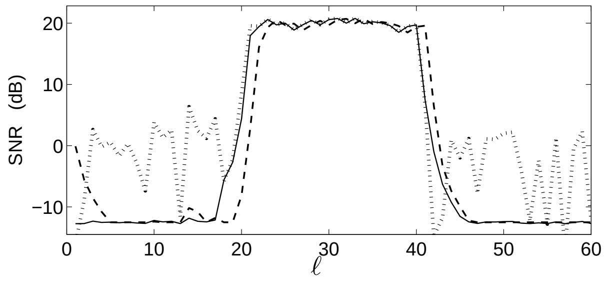


Fig. 8. SNR’s in successive short-time frames: *A posteriori* SNR γ_ℓ (dotted line), decision-directed *a priori* SNR $\hat{\xi}^{\mathrm{DD}}_{\ell|\ell}$ (dashed line), and noncausal recursive *a priori* SNR estimate $\hat{\xi}^{\mathrm{RE}}_{\ell|\ell+3}$ with 3 frames delay (solid line).

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 defined by [37]

$$\begin{aligned} \text{SegSNR} &= \frac{1}{J}\sum_{\ell=0}^{J-1} \text{SNR}_\ell \\ &= \frac{1}{J}\sum_{\ell=0}^{J-1} 10 \cdot \log \frac{\sum_{n=0}^{N-1} x^2(n+\ell N/2)}{\sum_{n=0}^{N-1} \left[x(n+\ell N/2) - \hat{x}(n+\ell N/2)\right]^2} \quad [\text{dB}] \qquad (41) \end{aligned}$$

where J represents the number of frames in the signal, $N = 512$ is the number of samples per frame (corresponding to 32 ms frames), and the overlap between successive frames is 50%. The SNR at each frame, SNR_ℓ, is limited to perceptually meaningful range between 35 dB and -10 dB. This 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 [38], [39]. The second quality measure is log-spectral distance (LSD), which is defined by

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

where $\mathcal{C}X(k,\ell) \triangleq \max\left\{ |X(k,\ell)|^2, \delta \right\}$ is the spectral power, clipped such that the log-spectrum dynamic range is confined to about 50 dB (that is, $\delta = 10^{-50/10} \cdot \max\limits_{k,\ell}\left\{ |X(k,\ell)|^2 \right\}$).

The noise signals used in our evaluation are taken from the Noisex92 database [40]. 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 [34]. The speech signal is sampled at 16 kHz and degraded by the various noise types with segmental SNR's in the range $[-5, 10]$ dB.

The noisy signals are transformed into the STFT domain using Hamming analysis windows of 512 samples length, and 256 samples framing step (50% overlap between successive frames). 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} (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 [36]. 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

TABLE III

Segmental SNR Improvement for Various Noise Types and Levels, Obtained Using the Decision-Directed Approach (DD), Causal Recursive Estimation (CRE), and Noncausal Recursive Estimation with 3 Frames Delay (NCRE).

Input SegSNR	Stationary WGN			Car interior noise			F16 cockpit noise			Babble noise		
[dB]	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 Using the Decision-Directed Approach (DD), Causal Recursive Estimation (CRE), and Noncausal Recursive Estimation with 3 Frames Delay (NCRE).

Input SegSNR	Stationary WGN			Car interior noise			F16 cockpit noise			Babble noise		
[dB]	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**

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.

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 generates consistent estimators for the speech spectral components and the *a priori* SNR, while keeping the resulting algorithms simple. It extends existing speech enhancement algorithms, which were developed under the assumption that speech spectral components are statistically independent. It also explains the notable performance of the log-spectral amplitude estimator when combined with the decision-directed estimation approach.

The main differences between the proposed statistical model and that of Ephraim and Malah are that the sequence of speech spectral variances is referred to as a random process, rather than a sequence of parameters, and that successive spectral components are correlated through the statistical dependence between the spectral variances and spectral components. Estimators for the speech spectral components are derived based on the proposed model for various distortion measures. We show that similar to conventional spectral estimators, spectral enhancement is obtained by applying a real-valued gain function to the spectral noisy measurements. However, the *a priori* SNR estimation relies on the statistical model, rather than the decision-directed approach.

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* 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], [41]. In this case, we need to find also an estimator for the speech presence probability, that is consistent with the model assumptions and the *a priori* SNR estimation. This subject is currently under investigation.

ACKNOWLEDGEMENT

The author thanks Prof. David Malah for proofreading the manuscript and for his helpful comments.

REFERENCES

[1] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-33, no. 2, pp. 443–445, April 1985.

[2] ——, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[3] D. Malah, R. V. Cox, and A. J. Accardi, “Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments,” in *Proc. 24th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-99*, Phoenix, Arizona, 15–19 March 1999, pp. 789–792.

[4] I. Cohen and B. Berdugo, “Speech enhancement for non-stationary noise environments,” *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, October 2001.

[5] N. S. Kim and J.-H. Chang, “Spectral enhancement based on global soft decision,” *IEEE Signal Processing Letters*, 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 Communication*, vol. 34, no. 1-2, pp. 75–91, April 2001.

[7] A. J. Accardi and R. V. Cox, “Robust digit recognition in noise: An evaluation using the AURORA corpus,” in *Proc. 24th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-99*, Phoenix, Arizona, 15–19 March 1999, pp. 201–204.

[8] ——, “A modular approach to speech enhancement with an application to speech coding,” in *Proc. 24th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-99*, Phoenix, Arizona, 15–19 March 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. Kovesi, J. Stegmann, and H.-G. Kang, “A candidate for the ITU-T 4 kbit/s speech coding standard,” in *Proc. 26th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-01*, Salt Lake City, Utah, 7–11 May 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*, Tsukaba, Japan, 6–9 October 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 Communication*, vol. 39, no. 1-2, pp. 23–32, January 2003.

[12] R. Martin, “Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors,” in *Proc. 27th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-02*, Orlando, Florida, 13–17 May 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 Letters*, vol. 6, no. 1, pp. 1–3, January 1999.

[14] Y. D. Cho and A. Kondoz, “Analysis and improvement of a statistical model-based voice activity detector,” *IEEE Signal Processing Letters*, vol. 8, no. 10, pp. 276–278, October 2001.

[15] S. Gazor and W. Zhang, “A soft voice activity detector based on a laplacian-gaussian model,” *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 5, pp. 498–505, September 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 Internat. Conf. Acoust. Speech Signal Process., ICASSP-02*, Orlando, Florida, 13–15 May 2002, pp. I–289–I–292.

[17] I. Cohen, S. Gannot, and B. Berdugo, "An integrated real-time beamforming and postfiltering system for non-stationary noise environments," *to appear in special issue of EURASIP JASP on Signal Processing for Acoustic Communication System*, 2003.

[18] A. Gilloire, P. Scalart, C. Lamblin, C. Mokbel, and S. Proust, "Innovative speech processing for mobile terminals: an annotated bibliography," *Signal Processing*, vol. 80, no. 7, pp. 1149–1166, July 2000.

[19] Y. Ephraim and D. Malah, "Speech enhancement using optimal nonlinear spectral amplitude estimation," in *Proc. 8th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-83*, Boston, Massachusetts, 14–16 April 1983, pp. 1118–1121.

[20] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 2, no. 2, pp. 345–349, April 1994.

[21] C. Breithaupt and R. Martin, "MMSE estimation of magnitude-squared DFT coefficients with supergaussian priors," in *Proc. 28th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-03*, Hong Kong, 6–10 April 2003, pp. I_896–I_899.

[22] P. Scalart and J. Vieira-Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. 21th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-96*, Atlanta, Georgia, 7–10 May 1996, pp. 629–632.

[23] B. H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-33, no. 6, pp. 1404–1413, December 1985.

[24] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, October 1992.

[25] H. Sheikhzadeh and L. Deng, "Waveform-based speech recognition using hidden filter models: Parameter selection and sensitivity to power normalization," *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 80–91, January 1994.

[26] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Information Theory*, vol. 48, no. 6, pp. 1518–1568, June 2002.

[27] C. J. Wellekens, "Explicit time correlations in hidden Markov models for speech recognition," in *Proc. 12th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-87*, Dallas, Texas, 6–9 April 1987, pp. 384–386.

[28] Y. Ephraim, "A bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 40, no. 4, pp. 725–735, April 1992.

[29] 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 and Audio Processing*, vol. 6, no. 5, pp. 445–455, September 1998.

[30] J. Wexler and S. Raz, "Discrete Gabor expansions," *Speech Processing*, vol. 21, no. 3, pp. 207–220, November 1990.

[31] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*. Englewood Cliffs, New Jersey: Prentice-Hall, 1983.

[32] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *to appear in special issue of EURASIP JASP on Digital Audio for Multimedia Communications*, 2003.

[33] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, December 1979.

[34] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Tech. Rep., (prototype as of December 1988).

[35] A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics*, 6th ed. London, UK: Edward Arnold, 1994, vol. 1.

[36] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, September 2003.

[37] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1988.

[38] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, 2nd ed. New York: IEEE Press, 2000.

[39] P. E. Papamichalis, *Practical Approaches to Speech Coding*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1987.

[40] 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 Communication*, vol. 12, no. 3, pp. 247–251, July 1993.

[41] R. J. McAulay and M. L. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-28, no. 2, pp. 137–145, April 1980.