ON THE DECISION-DIRECTED ESTIMATION APPROACH OF EPHRAIM AND MALAH

Israel Cohen
Department of Electrical Engineering, Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
icohen@ee.technion.ac.il

ABSTRACT

The decision-directed approach of Ephraim and Malah is widely used for a priori SNR estimation and speech enhancement. However, it conflicts with common model assumptions. In this paper, we propose recursive estimators for the a priori SNR and the speech spectral components. We introduce a novel statistical model that takes into account the time-correlation between successive speech spectral components, while keeping the resulting algorithms simple. This model provides new insight into the decision-directed approach, and enables the extension of existing speech enhancement algorithms to noncausal estimation. The causal a priori SNR estimator degenerates, as a special case, to a “decision-directed” estimator with a time-varying frequency-dependent weighting factor. The noncausal estimator is capable of discriminating between speech onsets and noise irregularities, achieving lower levels of both musical noise and speech distortion.

1. INTRODUCTION

Two decades ago, Ephraim and Malah proposed a popular statistical model for speech enhancement [1]. Accordingly, the individual short-term spectral components of the speech and noise signals are modeled as statistically independent Gaussian random variables. The assumption of statistical independency is clearly unfulfilled. However, it facilitates a mathematically tractable derivation of useful estimators for various distortion measures [1–3]. Cappé [4] showed that the dominant factor in the Ephraim-Malah algorithm is the decision-directed estimation approach for the a priori SNR. Unfortunately, this approach conflicts with the model assumptions. 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 estimate heavily relies on the strong time-correlation between successive speech spectral components.

In this paper, we propose estimators for the a priori SNR and the speech spectral components. We introduce a novel statistical model that takes into account the time-correlation between successive speech spectral components, while keeping the resulting algorithms simple. This model provides new insight into the decision-directed approach, and 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 the causal a priori SNR estimator degenerates, as a special case, to a “decision-directed” estimator with a time-varying frequency-dependent weighting factor. Furthermore, the noncausal estimator, having a few subsequent spectral measurements at hand, is capable of discriminating between speech onsets and noise irregularities. This yields lower levels of both musical noise and speech distortion.

In Sec. 2, we formulate the speech enhancement problem. In Sec. 3, we present the statistical model. In Sec. 4, we derive estimators for the speech spectral components and the a priori SNR. In Sec. 5, we address the relation to the decision-directed estimation approach. Finally, in Section 6, we discuss the advantages of the proposed estimation approach.

2. PROBLEM FORMULATION

Let \( x(n) \) and \( d(n) \) denote speech and uncorrelated additive noise signals, respectively, where \( n \) is a discrete-time index. Applying the short-time Fourier transform (STFT) to the observed signal \( y(n) \), we have in the time-frequency domain

\[
Y(k) = X(k) + D(k)
\]

where \( k \) is the frequency-bin index and \( \ell \) is the time frame index. Let \( Y(k) = \{Y_0(k), \ldots, Y_{\ell'}(k)\} \) denote a set of spectral measurements, and let \( d \{X(k), \hat{X}(k)\} \) be a given distortion measure between \( X(k) \) and \( \hat{X}(k) \). Our objective is to find an estimator \( \hat{X}(k) \), which minimizes the conditional expected value of the distortion measure, given the set of spectral noisy measurements

\[
\hat{X}(k) = \arg \min \mathbb{E} \left\{ d \left[ X(k), \hat{X} \right] \mid Y_{\ell}(k) \right\}.
\]

We consider a causal estimation of \( X(k) \) (in which case \( \ell' = \ell \)), as well as a noncausal estimation (in which case \( \ell' > \ell \)).

3. SPEECH SPECTRAL MODEL

The relation between successive spectral components of a speech signal, in comparison with a noise signal, can be investigated by evaluating the autocorrelation sequences of the STFT coefficients along time-trajectories (the frequency-bin index \( k \) is held

\footnote{Note that causality is defined with respect to the spectral components, rather than with respect to the samples in the time domain.}
fixed). The time-correlation between successive spectral magnitudes of speech signals is shown to be much higher than that of white Gaussian noise [5]. To enable recursive estimators for the speech spectral components and the a priori SNR, while keeping the resulting algorithms simple, we propose the following statistical model:

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 statistically independent (IID).

2. The speech spectral phases \( \angle X_{\ell}(k) \) 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 random process, where \( A_{\ell}(k) \sim \mathcal{N}(0, \sigma^2_{\ell}) \). 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. Let \( \chi_{\ell}(k) \) denote the speech spectral variance. Then given \( \chi_{\ell}(k) \), a spectral component \( X_{\ell}(k) \) is a zero-mean complex Gaussian random variable with IID real and imaginary parts.

5. The sequence of speech spectral variances \( \{ \chi_{\ell}(k) \mid \ell = 0, 1, \ldots \} \) is a random process. For fixed \( k \) and \( \ell \), \( X_{\ell}(k) \) is correlated with the sequence of speech spectral magnitudes \( \{ A_{\ell'}(k) \mid \ell' = 0, 1, \ldots \} \). However, given \( \chi_{\ell}(k) \), \( X_{\ell}(k) \) is statistically independent of \( X_{\ell'}(k) \) for \( \ell' \neq \ell \).

We note that the fundamental difference between the proposed statistical model and that of Ephraim and Malah stems from the last assumption.

4. SIGNAL ESTIMATION

In this section, we derive an estimator for \( X_{\ell}(k) \) based on the proposed statistical model. We assume knowledge of the noise PSD, which in practice can be estimated by using the Minima Controlled Recursive Averaging approach [6]. For notational simplicity, the frequency-bin index \( k \) is henceforth omitted.

4.1. Spectral Enhancement

Let \( p \left( X_{\ell} \mid Y_{\ell}^{\ell} \right) \) denote the conditional pdf of a speech spectral component \( X_{\ell}(k) \) given the noisy measurements \( Y_{\ell}^{\ell} \). Then, the spectral estimator \( \hat{X}_{\ell} \) is obtained by minimizing

\[
E \left\{ d \left( X_{\ell}, \hat{X}_{\ell} \right) \mid Y_{\ell}^{\ell} \right\} = \int d \left( X, \hat{X} \right) p \left( X_{\ell} \mid Y_{\ell}^{\ell} \right) dX_{\ell}. \tag{3}
\]

Since the statistical model is generally nonlinear, and there exists no simple solution for the spectral estimation, we first derive an estimate for \( \chi_{\ell} \), from the noisy measurements \( Y_{\ell}^{\ell} \), \( \hat{\chi}_{\ell} \), and subsequently obtain the estimator \( \hat{X}_{\ell} \). The proposed statistical model implies

\[
p \left( X_{\ell} \mid Y_{\ell}^{\ell}, \chi_{\ell} \right) = p \left( X_{\ell} \mid Y_{\ell}, \chi_{\ell} \right) \tag{4}
\]

for \( \ell' \geq \ell \). Hence, given \( \hat{\chi}_{\ell} \), we can compute a spectral estimate from

\[
\min_{\hat{X}_{\ell}} \int d \left( X_{\ell}, \hat{X}_{\ell} \right) p \left( X_{\ell} \mid Y_{\ell}, \hat{\chi}_{\ell} \right) dX_{\ell} = \min_{\hat{X}_{\ell}} E \left\{ d \left( X_{\ell}, \hat{X}_{\ell} \right) \mid Y_{\ell}, \hat{\chi}_{\ell} \right\}. \tag{5}
\]

This 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 \( X_{\ell} \) is obtained by applying a spectral gain function to each noisy spectral component of the speech signal:

\[
\hat{X}_{\ell} = G \left( \xi_{\ell}(\ell', \gamma_{\ell}) \right) Y_{\ell} \tag{6}
\]

where the a priori and a posteriori SNR’s are defined respectively by

\[
\xi_{\ell}(\ell', \gamma_{\ell}) = \frac{\lambda_{\ell} \gamma_{\ell}}{\lambda_{\ell} \gamma_{\ell} + \lambda_{\ell'}}, \quad \gamma_{\ell} = \frac{|Y_{\ell}|^2}{\lambda_{\ell}} \tag{7}
\]

and where \( \lambda_{\ell} \) is correlated with the sequence \( X_{\ell}(k) \) for \( \ell' \neq \ell \). The specific expression for the spectral gain function \( G(\xi_{\ell}(\ell', \gamma_{\ell})) \) depends on the particular choice of a distortion measure \( d \left( X_{\ell}, \hat{X}_{\ell} \right) \) [3].

4.2. Causal Recursive A Priori SNR Estimation

In this subsection, we propose a causal recursive estimator \( \hat{\xi}_{\ell} \) for the a priori SNR. The estimator combines a “propagation” step and an “update” step, following the rational of Kalman filtering, to recursively predict and update the estimate for \( \chi_{\ell} \) as new data arrive.

Suppose we are given an estimate \( \hat{\chi}_{\ell} \), and a new noisy spectral component \( Y_{\ell} \) is observed. Then, the estimate for \( \chi_{\ell} \) can be updated by computing the conditional variance of \( X_{\ell} \) given \( Y_{\ell} \) and \( \chi_{\ell} \):

\[
\chi_{\ell} \left| Y_{\ell} \right| = \frac{\lambda_{\ell} \gamma_{\ell}}{\lambda_{\ell} \gamma_{\ell} + \lambda_{\ell'}}, \quad \gamma_{\ell} = \frac{|Y_{\ell}|^2}{\lambda_{\ell}}, \tag{8}
\]

Since we assume that \( X_{\ell} \mid \chi_{\ell} \) and \( D_{\ell} \) are statistically independent Gaussian complex variables, the conditional distribution of \( X_{\ell} \mid \chi_{\ell} \) given \( \chi_{\ell} \) is Gaussian with mean and variance

\[
E \left\{ X_{\ell} \mid \chi_{\ell} \right\} = \frac{\chi_{\ell} \gamma_{\ell}}{\lambda_{\ell} \gamma_{\ell} + \lambda_{\ell'}} \tag{9}
\]

\[
\text{var} \left\{ X_{\ell} \mid \chi_{\ell} \right\} = \frac{X_{\ell} \gamma_{\ell}}{\lambda_{\ell} \gamma_{\ell} + \lambda_{\ell'}} \tag{10}
\]

Substituting (9) and (10) into (8), we have

\[
\hat{\chi}_{\ell} = \frac{\lambda_{\ell} \gamma_{\ell}}{\lambda_{\ell} \gamma_{\ell} + \lambda_{\ell'}} \frac{X_{\ell} \gamma_{\ell}}{\lambda_{\ell} \gamma_{\ell} + \lambda_{\ell'}} \tag{11}
\]

Using (7) and dividing both sides of (11) by \( \lambda_{\ell} \), we have

\[
\hat{\xi}_{\ell} = \frac{\lambda_{\ell} \gamma_{\ell}}{\lambda_{\ell} \gamma_{\ell} + \lambda_{\ell'}} \left( 1 + \frac{\hat{\xi}_{\ell} \gamma_{\ell}}{1 + \hat{\xi}_{\ell} \gamma_{\ell}} \right). \tag{12}
\]

2Note that in [1], the a priori SNR is defined by \( \xi_{\ell} = \chi_{\ell} \), where the variance \( \chi_{\ell} \) is a parameter of the prior pdf of \( X_{\ell} \).
Table 1. Summary of the Causal Recursive Speech Enhancement Algorithm.

| Initialization: \( A_{-1} = 0 \), \( \xi_{-1} = \min \). 
For all short-time frames \( \ell = 0, 1 \ldots \) 
Obtain \( \xi_{\ell-1} \) by using the “propagation” step (14). 
Obtain \( \xi_{\ell} \) by using the “update” step (12). 
Estimate the speech spectral component \( X_\ell \) by (6) and (7). |

Table 2. Summary of the Noncausal Recursive Speech Enhancement Algorithm.

| Initialization: \( A_{-1} = 0 \), \( \xi_{-1} = \min \). 
For all short-time frames \( \ell = 0, 1 \ldots \) 
Obtain \( \xi_{\ell \mid \ell+1} \) by “backward estimation” (18). 
Obtain \( \xi_{\ell \mid \ell+1} \) by “backward-forward propagation” (17). 
Obtain \( \xi_{\ell+1} \) by the “update” step (15). 
Estimate the speech spectral component \( X_\ell \) by (6) and (7). |

We call (12) the “update” step. Assume we are given at frame \( \ell - 1 \) estimates for \( A_{\ell-1} \) and \( \lambda_{X_{\ell-1}} \), conditioned on \( Y_{\ell-1} \). Then, these estimates can be “propagated” in time to obtain an estimate for \( \lambda_{X_\ell} \). Since \( \lambda_{X_\ell} \) is correlated with both \( \lambda_{X_{\ell-1}} \) and \( A_{\ell-1} \), we propose to use a first-order predictor of the form

\[
\lambda_{X_\ell \mid \ell-1} = \max \left\{ (1 - \alpha) \lambda_{X_{\ell-1} \mid \ell-1} + \alpha \hat{A}_{\ell-1}^2, \lambda_{\min} \right\}
\]

where \( \alpha (0 \leq \alpha \leq 1) \) 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 \( X_\ell \). Dividing both sides of (13) by \( \lambda_{D_{\ell-1}} \), we obtain the “propagation” step

\[
\xi_{\ell \mid \ell-1} = \max \left\{ (1 - \alpha) \xi_{\ell-1} + \alpha \hat{A}_{\ell-1}^2, \xi_{\min} \right\}
\]

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 1.

4.3 Noncausal Recursive A Priori SNR Estimation

Now we propose a noncausal recursive estimator \( \hat{\xi}_{\ell \mid \ell+L} \) for the a priori SNR, given the noisy measurements up to frame \( \ell + L \), where \( L > 0 \) denotes the admissible delay in frames.

Let \( \lambda_{X_{\ell \mid \ell+L}} \equiv E \{ A_\ell^2 \mid Y_0^{\ell-1}, Y_{\ell+L}^L \} \) and \( \lambda_{\ell \mid \ell+1} \equiv E \{ A_\ell^2 \mid Y_{\ell+1}^L \} \) denote conditional spectral variances of \( X_\ell \). An estimate for \( \lambda_{\ell \mid \ell+1} \) given \( \lambda_{X_{\ell \mid \ell+L}} \) and \( Y_\ell \) can be updated similarly to (11) by using the “update” step

\[
\hat{\xi}_{\ell \mid \ell+L} = \frac{\hat{\xi}_{\ell \mid \ell+L}}{1 + \hat{\xi}_{\ell \mid \ell+L} \gamma_\ell} + \frac{\hat{\xi}_{\ell \mid \ell+L} \gamma_\ell}{1 + \hat{\xi}_{\ell \mid \ell+L}} \hat{\xi}_{\ell \mid \ell+L}.
\]

(15)

To obtain an estimate for \( \lambda_{X_{\ell \mid \ell+L}} \), we employ the estimates \( \hat{A}_{\ell-1} \) and \( \lambda_{X_{\ell-1} \mid \ell+L-1} \) from the previous frame, and derive an estimate for \( \lambda_{X_\ell} \) from the measurements \( Y_{\ell+1}^L \). Suppose an estimate \( \hat{\lambda}_{\ell \mid \ell+1} \) is given, we propose to propagate the estimates from frame \( \ell - 1 \) to frame \( \ell \) by

\[
\hat{\lambda}_{\ell \mid \ell+L} = \max \left\{ \alpha \hat{A}_{\ell-1}^2 + (1 - \alpha) \left[ \alpha' \hat{\lambda}_{\ell-1} \mid \ell+L-1 \right], \lambda_{\min} \right\}
\]

where \( \alpha (0 \leq \alpha \leq 1) \) is related to the degree of nonstationarity of the random process \( \{ \lambda_{X_\ell} \mid \ell = 0, 1, \ldots \} \), and \( \alpha' (0 \leq \alpha' \leq 1) \) is associated with the reliability of the estimate \( \hat{\lambda}_{\ell \mid \ell+1-L} \) in comparison with that of \( \hat{\lambda}_{\ell-1-L} \). Dividing both sides of (16) by \( \lambda_{D_{\ell-1}} \), we have the following “backward-forward propagation” step:

\[
\hat{\xi}_{\ell \mid \ell+L} = \max \left\{ \alpha A_{\ell-1}^2 + (1 - \alpha) \left[ \alpha \hat{\xi}_{\ell \mid \ell+L-1} \right], \lambda_{\min} \right\}.
\]

(17)

An estimate for the a priori SNR \( \hat{\xi}_{\ell} \) given the measurements \( Y_{\ell+1}^L \) is obtained by

\[
\hat{\xi}_{\ell \mid \ell+1} = \frac{1}{\sum_{n=1}^{L+1} \gamma_{\ell+n} - \beta}, \quad \lambda_{\min} \quad \text{if nonnegative, otherwise,}
\]

(18)

where \( \beta (\beta \geq 1) \) is an over-subtraction factor to compensate for a sudden increase in the noise level. This estimator is an anti-causal version of the maximum-likelihood a priori SNR estimator suggested in [1]. The steps of the noncausal recursive spectral enhancement algorithm are summarized in Table 2.

5. RELATION TO “DECISION-DIRECTED” ESTIMATION

The proposed causal recursive estimator \( \hat{\xi}_{\ell} \) for the a priori SNR is closely related to the decision-directed estimator of Ephraim and Malah [1]. The decision-directed estimator is given by

\[
\hat{\xi}_{DD} = \min \left\{ \frac{\hat{\xi}_{\ell-1}^2}{\lambda_{D_{\ell-1}}} + (1 - w) \max \{ \gamma_{\ell} - 1, 0 \} \right\}
\]

(19)

where \( w (0 \leq w \leq 1) \) is a weighting factor that controls the trade-off between the noise reduction and the transient distortion introduced into the signal [1, 4]. The update step (12) of the causal recursive estimator can be written as

\[
\hat{\xi}_{\ell} = \alpha_{\ell} \hat{\xi}_{\ell-1} + (1 - \alpha_{\ell}) (\gamma_{\ell} - 1)
\]

(20)

where \( \alpha_{\ell} \) is defined by

\[
\alpha_{\ell} \equiv 1 - \frac{\hat{\xi}_{\ell-1}^2}{\left( 1 + \hat{\xi}_{\ell-1} \right)^2}.
\]

(21)

Substituting (14) into (20) and (21) with the parameter \( \alpha \) set to 1, and applying the lower bound constraint to \( \hat{\xi}_{\ell} \) rather than \( \hat{\xi}_{\ell-1} \),
we have
\[
\hat{\xi}_{\ell|\ell} = \max \left\{ \alpha \frac{A_{\ell-1}^2}{\lambda^{\text{DD}}_{\ell-1}} + (1 - \alpha \ell)(\gamma_{\ell} - 1), \xi_{\min} \right\} \tag{22}
\]
\[
\alpha_{\ell} = 1 - \frac{A_{\ell-1}^4}{\left(\lambda^{\text{DD}}_{\ell-1} + A_{\ell-1}^2\right)^4}. \tag{23}
\]

The expression (22) with \( \alpha_{\ell} \equiv w \) is actually a practical form of the decision-directed estimator,
\[
\hat{\xi}_{\ell|\ell}^{\text{DD}} = \max \left\{ w \frac{A_{\ell-1}^2}{\lambda^{\text{DD}}_{\ell-1}} + (1 - w)(\gamma_{\ell} - 1), \xi_{\min} \right\}, \tag{24}
\]
that includes a lower bound constraint to further reduce the level of residual musical noise [4]. 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 \( \alpha_{\ell} \).

6. DISCUSSION

The weighting factor \( \alpha_{\ell} \) in (23) is monotonically decreasing as a function of the instantaneous SNR, \( A_{\ell-1}^2 / \lambda^{\text{DD}}_{\ell-1} \). 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) [4]. The 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 \( \alpha \) in (14) to the degree of spectral nonstationarity. This may further improve the performance.

The different behaviors of the causal recursive estimator and the decision-directed estimator are illustrated in the example of Fig. 1. 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 by using half overlapping Hamming windows. The a priori SNR estimates are obtained by using the spectral gain function, which minimizes the mean squared error of the log-spectral amplitude [2], and the parameters \( \xi_{\min} = -25 \text{ dB}, \alpha = 0.9, w = 0.98 \). It shows that when the a posteriori SNR \( \gamma_{\ell} \) 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 \( \gamma_{\ell} \) increases, the response of \( \hat{\xi}_{\ell|\ell} \) is initially slower than \( \hat{\xi}_{\ell|\ell}^{\text{DD}} \), but it then builds up faster to the a posteriori SNR. When \( \gamma_{\ell} \) is sufficiently high, \( \hat{\xi}_{\ell|\ell}^{\text{DD}} \) follows the a posteriori SNR with a delay of 1 frame, whereas \( \hat{\xi}_{\ell|\ell} \) follows the a posteriori SNR instantaneously. When \( \gamma_{\ell} \) decreases, the response of \( \hat{\xi}_{\ell|\ell} \) is immediate, while that of \( \hat{\xi}_{\ell|\ell}^{\text{DD}} \) is delayed by 1 frame.

Figure 1(b) demonstrates the behavior of the noncausal recursive estimator. The noncausal a priori SNR estimate \( \hat{\xi}_{\ell|\ell+3} \) is obtained with the parameters \( \xi_{\min} = -25 \text{ dB}, \alpha = \alpha' = 0.9, \beta = 2 \), and \( L = 3 \) frames delay. 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 \( \gamma_{\ell} \), 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 \( \gamma_{\ell} \) corresponding to noise only. Experimental results [5] 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 musical noise is achieved.

7. REFERENCES