CCIT Report #460
December 2003

Speech Enhancement Using Supergaussian Speech Models and Noncausal *A Priori* SNR Estimation

Israel Cohen

Abstract

Existing supergaussian speech models in the short-time Fourier transform domain are based on the assumption that distinct spectral components are statistically independent. The corresponding minimum mean-square error (MMSE) spectral estimators require an estimator for the *a priori* SNR. Unfortunately, the latter is often obtained by the decision-directed approach of Ephraim and Malah, which relies on the strong time-correlation between successive speech spectral components. In this paper, we extend the supergaussian speech models by taking into consideration the time-frequency correlation between spectral components. We introduce noncausal *a priori* SNR estimators for Gamma and Laplacian speech models, and derive noncausal estimators for the clean speech spectral components. The noncausal *a priori* SNR estimation consists of two major steps, which follow the rational of Kalman filtering: a "propagation" step and an "update" step. Estimates for the speech spectral variances and the instantaneous power from the previous frame are propagated in time to obtain an estimate for the spectral variance in the current frame. Subsequently, the estimate for the spectral variance is updated by computing the conditional variance of the speech spectral component, based on the underlying speech model. Experimental results demonstrate the improved performance of the proposed algorithms.

I. Introduction

Optimal estimators for speech enhancement in the short-time Fourier transform (STFT) domain are often based on a Gaussian statistical model [1]–[5]. Accordingly, the individual short-term spectral components of the speech and noise signals are modelled as statistically independent Gaussian random variables. Using this model, Ephraim and Malah derived a short-term spectral amplitude (STSA) estimator, which minimizes the mean-square error of the spectral magnitude [1], and a Log-Spectral Amplitude (LSA) estimator, which minimizes the mean-square error of the log-spectra. Wolfe and Godsill [6] derived under the same modeling assumptions three alternative suppression rules, which are based on joint maximum a posteriori (MAP) spectral amplitude and phase estimation, MAP spectral amplitude estimation, and minimum mean-square error (MMSE) spectral power estimation. The resulting suppression

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).

rules are simpler than those of Ephraim and Malah, yet demonstrate similar effect in reducing residual musical noise phenomena. Lotter *et al.* [5] considered a multichannel Gaussian statistical model, where speech spectral amplitudes in different microphones are identical up to a constant channel-dependent factor, while noise components in different microphones are statistically independent Gaussian random variables. They assumed statistical independence across time and frequency in the STFT domain, and generalized the STSA estimator of Ephraim and Malah and the MAP amplitude estimator of Wolfe and Godsill to the multichannel case. Both multichannel estimators provide a significant gain compared to the STSA estimator, when the speech components in different microphones are in phase (nonreverberant environment) and the noise components are sufficiently uncorrelated.

The Gaussian model is motivated by the central limit theorem, as each Fourier expansion coefficient is a weighted sum of random variables resulting from the random sequence [1]. When the span of correlation within the signal is sufficiently short compared to the size of the frames, the probability distribution function of the spectral coefficients asymptotically approaches gaussian as the frame's size increases. The gaussian approximation is in the central region of the gaussian curve near the mean. However, the approximation can be very inaccurate in the tail regions away from the mean [7]. Furthermore, the necessary conditions for the central limit theorem, *e.g.* that a particular few of the member random variables does not dominate the sum [7], are not satisfied for speech signals. In addition, the span of correlation of voiced speech is larger than the typical sizes of frames used in speech enhancement applications [8]. Consequently, statistical models other than the Gaussian model should also be considered [8].

Porter and Boll [9] pointed out that *a priori* speech spectra do not have a Gaussian distribution, but Gamma-like distribution. They proposed to compute the optimal estimator directly from the speech data, rather than from a parametric model of the speech statistics. Martin [8] considered a Gamma speech model, in which the real and imaginary parts of the clean speech spectral components are modeled as independent and identically distributed (IID) Gamma random variables. He assumed that distinct spectral components are statistically independent, and derived MMSE estimators for the complex speech spectral coefficients under Gaussian and Laplacian noise modeling. He showed that under Gaussian noise modeling, the Gamma speech model yields higher improvement in the segmental SNR than the Gaussian speech model. Under Laplacian noise modeling, the Gamma speech model results in lower residual musical noise than the Gaussian speech model. Breithaupt and Martin [10] derived, under the same statistical modeling, MMSE estimators for the magnitude-squared spectral coefficients, and compared their performance to that obtained by using a Gaussian speech model. They showed that improvement in the segmental SNR comes at the expense of additional residual musical noise. Lotter and Vary [11] derived a MAP estimator for the speech spectral amplitude, based on a Gaussian noise model and a supergaussian speech model. They proposed a parametric probability density function (pdf) for the speech spectral amplitude, which approximates, with a proper choice of the parameters, the Gamma and Laplacian densities. Compared with the STSA estimator of Ephraim-Malah, the MAP estimator with Laplacian speech modeling demonstrates improved noise reduction. Martin and Breithaupt [12] showed that modeling the real and imaginary parts of the clean speech spectral components as Laplacian random variables, the MMSE estimators for the complex speech spectral coefficients have similar properties to those estimators derived under Gamma modeling, but are easier to compute and implement.

Unfortunately, in all the above developments the *a priori* SNR, which is the dominant parameter of the spectral estimators, *e.g.*, [6], [13], [14], is obtained by the decision-directed approach of Ephraim and Malah [1]. On the one hand, spectral components in distinct time-frequency bins are assumed statistically independent when deriving analytical expressions for the speech estimators. On the other hand, the decision-directed *a priori* SNR estimator heavily relies on the strong time-correlation between successive speech spectral components. Recently, we introduced a novel Gaussian statistical model for speech enhancement, which takes into account the time-frequency correlation of speech signals [15]. We derived causal and noncausal recursive estimators for the *a priori* SNR, based on the statistical model, and showed their close relation to the decision-directed estimator. The causal estimator degenerates, as a special case, to a "decision-directed" estimator with a *time-varying frequency-dependent* weighting factor. The noncausal estimator employs future spectral measurements to better predict the spectral variances of the clean speech. Under the assumed statistical model, the noncausal *a priori* SNR estimator yields a higher improvement in the segmental SNR, lower log-spectral distortion (LSD), and better Perceptual Evaluation of Speech Quality scores (PESQ, ITU-T P.862), than the decision-directed estimator [16].

In this paper, we extend our recursive estimation approach to Gamma and Laplacian speech models, while the noise model remains gaussian. In contrast with existing supergaussian speech models [8], [10], [12], spectral components are assumed statistically correlated in the STFT domain, and thus their estimation is conditional on the information extracted from measurements in neighboring time-frequency bins. We show that the *a priori* SNR is a more dominant parameter than the *a posteriori* SNR, as is the case with the Ephraim-Malah gain functions [1], [17], which were derived under a Gaussian speech model. However, the MMSE gain functions for Gamma and Laplacian speech models are monotonically increasing as a function of the *a posteriori* SNR, whereas the Ephraim-Malah spectral gains are monotonically *decreasing* functions of the *a posteriori* SNR. The latter behavior is generally preferable, since it introduces a mechanism that counters the musical noise phenomenon [13]. Therefore, when the *a priori* SNR is estimated by the decision-directed method, the MMSE gain functions often produce higher levels of residual musical noise than the Ephraim-Malah gain functions. By contrast, noncausal *a priori* SNR estimators for the Gamma and Laplacian speech models, having a few subsequent spectral measurements at hand, facilitate a distinction between speech onsets and noise irregularities. Local bursts of noise are assigned a lower *a priori* SNR, while speech onsets are assigned a higher *a priori* SNR. Thus, speech onsets are better preserved, while the musical noise effect is reduced.

The proposed noncausal *a priori* SNR estimation consists of two major steps, which follow the rational of Kalman filtering: a "propagation" step and an "update" step. Estimates for the speech spectral variances and the instantaneous power from the previous frame are propagated in time to obtain an estimate for the spectral variance in the current frame. Subsequently, the estimate for the spectral variance is updated by computing the conditional variance of the speech spectral component, based on the underlying speech model. Experimental results show that the noncausal estimator consistently yields a higher segmental SNR and a lower LSD, than the decision-directed method, under all tested environmental conditions and speech models. The performance, in terms of segmental SNR and LSD, is greatest when using a Laplacian speech model and noncausal *a priori* SNR estimator. The performance is worst when using a Gaussian speech model and a decision-directed *a priori* SNR estimator. The Gamma speech model yields a higher segmental SNR and a lower LSD than the other speech models, only when the *a priori* SNR is estimated by the decision-directed method. However, when the *a priori* SNR is estimated by the proposed method, the Laplacian speech model yields a higher segmental SNR and a lower LSD than the other speech models. The differences between the Gaussian, Gamma and Laplacian speech models are smaller when using the noncausal estimators than when using the decision-directed method. Informal listening tests confirm that by using the noncausal estimators, speech components are better preserved, while the residual musical noise is further reduced. The level of residual musical noise is minimal when using a Gaussian speech model and the corresponding noncausal estimator. The residual musical noise is maximal when using a Gamma speech model and the decision-directed method.

The paper is organized as follows. In Section II, we present Gaussian, Gamma and Laplacian speech models, which allow for statistical dependence between speech spectral components in the time-frequency domain. In Section III, we derive noncausal MMSE estimators for the clean speech spectral components, based on the proposed speech models. In Section IV, we introduce noncausal estimators for the *a priori* SNR, and present noncausal recursive speech enhancement algorithms. Finally, in Section V, we evaluate the performance of the proposed algorithms, and show experimental results, which demonstrate their advantage, compared to using the decision-directed approach.

II. Statistical Model

Let x and d denote speech and uncorrelated additive noise signals, and let $y = x + d$ represent the observed signal. Applying the STFT to the observed signal, we have in the time-frequency domain

$$Y(k,\ell) = X(k,\ell) + D(k,\ell) \tag{1}$$

where k is the frequency-bin index ($k = 0, 1, \ldots, K-1$) and ℓ is the time frame index ($\ell = 0, 1, \ldots$). A statistical model, which takes into account the time-frequency correlation of speech signals, was recently proposed in [18]. Accordingly,

1) The noise spectral components $D(k,\ell)$ are statistically independent zero-mean complex Gaussian random variables. The real and imaginary parts of $D(k,\ell)$ are IID.
2) For fixed k and ℓ, a speech spectral component $X(k,\ell)$ is a zero-mean complex random variable. Its real and imaginary parts are IID.
3) The sequence of speech spectral variances $\{\lambda_X(k,\ell) \mid \ell = 0, 1, \ldots\}$, where $\lambda_X(k,\ell) \triangleq E\left\{|X(k,\ell)|^2\right\}$, is a random process. The spectral variances $\lambda_X(k,\ell)$ are generally correlated with the speech spectral magnitudes $|X(k',\ell')|$. However, given $\lambda_X(k,\ell)$, $X(k,\ell)$ is statistically independent of $X(k',\ell')$ for all $(k',\ell') \neq (k,\ell)$.

The conditional pdf of $X(k,\ell)$ given the spectral variance $\lambda_X(k,\ell)$ is determined by the specific speech model. Let $X_R = \Re\{X\}$ and $X_I = \Im\{X\}$ denote, respectively, the real and imaginary parts of a clean speech spectral component X. Let $p\left(X_\rho \,|\, \lambda_X\right)$ denote the conditional pdf of X_ρ ($\rho \in \{R, I\}$) given the spectral variance λ_X. Then, for a Gaussian speech model [1]

$$p\left(X_\rho \,|\, \lambda_X\right) = \frac{1}{\sqrt{\pi\,\lambda_X}} \exp\left(-\frac{X_\rho^2}{\lambda_X}\right) , \tag{2}$$

for a Gamma speech model [8]

$$p\left(X_\rho \,|\, \lambda_X\right) = \frac{1}{2\sqrt{\pi}} \left(\frac{3}{2\,\lambda_X}\right)^{1/4} |X_\rho|^{-1/2} \exp\left(-\sqrt{\frac{3}{2\,\lambda_X}}\,|X_\rho|\right) , \tag{3}$$

and for a Laplacian speech model [10], [12]

$$p\left(X_\rho \,|\, \lambda_X\right) = \frac{1}{\sqrt{\lambda_X}} \exp\left(-\frac{2\,|X_\rho|}{\sqrt{\lambda_X}}\right) . \tag{4}$$

In contrast with existing supergaussian speech models [8], [10], [12], successive spectral components are correlated, as the random processes $\{X(k,\ell) \,|\, \ell = 0, 1, \ldots\}$ and $\{\lambda_X(k,\ell) \,|\, \ell = 0, 1, \ldots\}$ are not independent. Therefore, the speech enhancement problem cannot be formulated as that of estimating $X(k,\ell)$ from $Y(k,\ell)$ alone.

III. MMSE Signal Estimation

In this section, we derive a noncausal MMSE estimator for $X(k,\ell)$, for Gaussian, Gamma and Laplacian speech models. We assume knowledge of the noise spectrum, which in practice can be estimated by using the *Minima Controlled Recursive Averaging* approach [19]. For notational simplicity, we often omit the arguments k and ℓ when there is no confusion.

Let $\mathcal{Y}_0^{\ell+L} = \{Y(k,\ell') \,|\, 0 \le k \le K-1,\ 0 \le \ell' \le \ell + L\}$ represent the set of spectral measurements up to frame $\ell + L$, where L ($L \ge 0$) denotes an admissible time delay in frames between the noisy speech signal and the enhanced signal. Let $p\left(X_\rho \,|\, \mathcal{Y}_0^{\ell+L},\, \lambda_X\right)$ denote the conditional pdf of X_ρ ($\rho \in \{R, I\}$) given the spectral variance λ_X and the noisy measurements $\mathcal{Y}_0^{\ell+L}$. Let $p\left(\lambda_X \,|\, \mathcal{Y}_0^{\ell+L}\right)$ denote the conditional pdf of λ_X given $\mathcal{Y}_0^{\ell+L}$. Then, a noncausal MMSE estimator $\hat{X}_\rho$ for X_ρ is obtained by

$$\hat{X}_\rho = E\left\{X_\rho \,|\, \mathcal{Y}_0^{\ell+L}\right\} = \iint X_\rho\, p\left(X_\rho \,|\, \mathcal{Y}_0^{\ell+L},\, \lambda_X\right) p\left(\lambda_X \,|\, \mathcal{Y}_0^{\ell+L}\right)\, dX_\rho\, d\lambda_X\, . \tag{5}$$

Applying Bayes' rule, we have

$$p\left(X_\rho \,|\, \mathcal{Y}_0^{\ell+L},\, \lambda_X\right) = \frac{p\left(Y_\rho \,|\, X_\rho,\, \mathcal{Y}_0^{\ell+L}\backslash\{Y_\rho\},\, \lambda_X\right) p\left(X_\rho \,|\, \mathcal{Y}_0^{\ell+L}\backslash\{Y_\rho\},\, \lambda_X\right)}{\int p\left(Y_\rho \,|\, X_\rho,\, \mathcal{Y}_0^{\ell+L}\backslash\{Y_\rho\},\, \lambda_X\right) p\left(X_\rho \,|\, \mathcal{Y}_0^{\ell+L}\backslash\{Y_\rho\},\, \lambda_X\right)\, dX_\rho}\, . \tag{6}$$

The model assumptions imply

$$p\left(Y_\rho \,|\, X_\rho,\, \mathcal{Y}_0^{\ell+L}\backslash\{Y_\rho\},\, \lambda_X\right) = p\left(Y_\rho \,|\, X_\rho\right) , \tag{7}$$

$$p\left(X_\rho \,|\, \mathcal{Y}_0^{\ell+L}\backslash\{Y_\rho\},\, \lambda_X\right) = p\left(X_\rho \,|\, \lambda_X\right) . \tag{8}$$

Approximating the conditional pdf of λ_X given the noisy observations $\mathcal{Y}_0^{\ell+L}$ by a Dirac delta function at position $\lambda_{X|\ell+L} \triangleq E\left\{|X|^2 \,|\, \mathcal{Y}_0^{\ell+L}\right\}$, and substituting (7) and (8) into (6), the spectral estimator $\hat{X}_\rho$ is given by

$$\begin{aligned} \hat{X}_\rho &= \iint X_\rho \, p\left(X_\rho \,|\, Y_\rho,\, \lambda_X\right) \delta\left(\lambda_X - \lambda_{X|\ell+L}\right) dX_\rho \, d\lambda_X \\ &= \int X_\rho \, p\left(X_\rho \,|\, Y_\rho,\, \lambda_{X|\ell+L}\right) dX_\rho = E\left\{X_\rho \,|\, Y_\rho,\, \lambda_{X|\ell+L}\right\} . \end{aligned} \tag{9}$$

That is, given the set of noisy measurements $\mathcal{Y}_0^{\ell+L}$, we first derive an estimate for the clean speech spectral variance $\lambda_{X|\ell+L}$. Subsequently, the estimation problem for the speech spectral component X_ρ reduces to that of estimating X_ρ from Y_ρ alone, assuming knowledge of the variance of X_ρ. The latter problem, when the *a priori* SNR is defined appropriately, can be solved similar to the MMSE estimation problem under the assumption that speech spectral components $X(k,\ell)$ and $X(k',\ell')$ are independent for $(k,\ell) \neq (k',\ell')$ [8], [12], [20]. Accordingly, an estimate for X is obtained by applying spectral gains to the real and imaginary parts of Y

$$\hat{X} = G\left(\xi,\, \gamma_R\right) Y_R + j\, G\left(\xi,\, \gamma_I\right) Y_I \tag{10}$$

where the *a priori* SNR ξ is defined by

$$\xi(k,\ell) \triangleq \frac{\lambda_{X|\ell+L}(k,\ell)}{\lambda_D(k,\ell)}\,, \tag{11}$$

the *a posteriori* SNR's γ_R and γ_I, corresponding to the real and imaginary parts of Y, are defined by

$$\gamma_R(k,\ell) \triangleq \frac{Y_R^2(k,\ell)}{\lambda_D(k,\ell)}\,, \quad \gamma_I \triangleq \frac{Y_I^2(k,\ell)}{\lambda_D(k,\ell)}\,, \tag{12}$$

and $\lambda_D(k,\ell) \triangleq E\left\{|D(k,\ell)|^2\right\}$ denotes the noise spectral variance.

The specific expression for the spectral gain function $G\left(\xi,\, \gamma_\rho\right)$ $(\rho \in \{R, I\})$ depends on the particular choice of a speech model. For a Gaussian speech model, the gain function is independent of the *a posteriori* SNR's. It is often referred to as Wiener filter, given by [20]

$$G\left(\xi\right) = \frac{\xi}{1+\xi}\,. \tag{13}$$

For a Gamma speech model, the gain function is given by [8] (see also Appendix I)

$$G\left(\xi,\, \gamma_\rho\right) = \frac{1}{C_{\rho+} - C_{\rho-}} \, \frac{\exp\left(C_{\rho-}^2/4\right) D_{-1.5}\left(C_{\rho-}\right) - \exp\left(C_{\rho+}^2/4\right) D_{-1.5}\left(C_{\rho+}\right)}{\exp\left(C_{\rho-}^2/4\right) D_{-0.5}\left(C_{\rho-}\right) + \exp\left(C_{\rho+}^2/4\right) D_{-0.5}\left(C_{\rho+}\right)} \tag{14}$$

where $C_{\rho+}$ and $C_{\rho-}$ are defined by

$$C_{\rho\pm} \triangleq \frac{\sqrt{3}}{2\sqrt{\xi}} \pm \sqrt{2\gamma_\rho}\,, \tag{15}$$

and $D_p(z)$ denotes the parabolic cylinder function [21, eq. 9.240]. For a Laplacian speech model, the gain function is given by [12] (see also Appendix II)

$$G\left(\xi,\, \gamma_\rho\right) = \frac{2}{L_{\rho+} - L_{\rho-}} \, \frac{L_{\rho+}\,\mathrm{erfcx}(L_{\rho+}) - L_{\rho-}\,\mathrm{erfcx}(L_{\rho-})}{\mathrm{erfcx}(L_{\rho+}) + \mathrm{erfcx}(L_{\rho-})} \tag{16}$$

where $L_{\rho+}$ and $L_{\rho-}$ are defined by

$$L_{\rho\pm} \triangleq \frac{1}{\sqrt{\xi}} \pm \sqrt{\gamma_\rho}\,, \tag{17}$$

and $\mathrm{erfcx}(x)$ is the scaled complementary error function, defined by

$$\mathrm{erfcx}(x) \triangleq e^{x^2} \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2}\, dt\,. \tag{18}$$

Figure 1 displays gain curves $G\left(\xi,\, \gamma_\rho\right)$ for several values of γ_ρ, which result from (13), (14) and (16). It shows that generally the *a priori* SNR is a more dominant parameter than the *a posteriori* SNR. The influence of the *a posteriori* SNR on the spectral gain is largest for a Gamma model, while it has no effect on the gain for a Gaussian model. Furthermore, the spectral gains for Gamma and Laplacian speech models are monotonically increasing functions of the *a posteriori* SNR, when the *a priori* SNR is kept constant.

It is worth making a comparison between the above MMSE gain functions and the Ephraim-Malah gain functions [1], [17], which were derived under a Gaussian speech model for minimizing the mean-square error distortion of the spectral or log-spectral amplitude. The *a priori* SNR is likewise a more dominant parameter than the *a posteriori* SNR. However, the Ephraim-Malah spectral gains are monotonically *decreasing* functions of the *a posteriori* SNR, for a fixed value of the the *a priori* SNR. Such a behavior is related to the useful mechanism that counters the musical noise phenomenon [13]. Local bursts of the *a posteriori* SNR, during noise-only frames, are “pulled down” to the average noise level, thus avoiding local buildup of noise whenever it exceeds its average characteristics. Unfortunately, the MMSE gain function for a Gaussian speech model is independent of the *a posteriori* SNR, while the MMSE gain functions for Gamma and Laplacian speech models are adversely increasing as a function of the *a posteriori* SNR. Therefore, in case the *a priori* SNR is estimated by the decision-directed method, the MMSE gain functions are expected to produce higher levels of residual musical noise, when compared with the Ephraim-Malah gain functions.

In speech enhancement applications, estimators which minimize the mean-square error distortion of the spectral amplitude or log-spectral amplitude have been found advantageous to MMSE estimators [1], [9], [17]. Hence, it would be constructive to derive such estimators for Gamma and Laplacian speech models, and compare their performances to those obtained under Gaussian modeling (*i.e.*, compare with the STSA and LSA estimators of Ephraim and Malah [1], [17]). However, this will not be pursued in this paper. Rather, we present in the next section noncausal estimators for the *a priori* SNR. These estimators employ future spectral measurements, for discriminating between speech onsets and noise irregularities. Local bursts of noise are assigned a lower *a priori* SNR, while speech onsets are assigned a higher *a priori* SNR. Thus, speech onsets are better preserved, while the musical noise effect is reduced.

IV. Noncausal *A Priori* SNR Estimation

In this section, we derive noncausal estimators for the *a priori* SNR for Gaussian, Gamma and Laplacian speech models. The noncausal *a priori* SNR estimation consists of two major steps, which follow the rational of Kalman filtering: a “propagation” step and an “update” step. Estimates for the speech spectral variances and the instantaneous power from the previous frame are propagated in time to obtain an estimate for the spectral variance in the current frame. Subsequently, the estimate for the spectral variance is updated by computing the conditional variance of the speech spectral component, based on the underlying speech model.

Let $\lambda'_{X|\ell+L}(k,\ell) \triangleq E\left\{|X(k,\ell)|^2 \,|\, \mathcal{Y}_0^{\ell+L} \backslash \{Y(k,\ell)\}\right\}$ denote the conditional variance of X given $\mathcal{Y}_0^{\ell+L}$ excluding the noisy measurement Y. Let $\lambda'_{X\,|\,[\ell,\ell+L]}(k,\ell) \triangleq E\left\{|X(k,\ell)|^2 \,|\, \mathcal{Y}_\ell^{\ell+L} \backslash \{Y(k,\ell)\}\right\}$ denote the conditional variance of X given the noisy measurements $\mathcal{Y}_\ell^{\ell+L} \backslash \{Y\}$. Then, an estimate for $\lambda_{X|\ell+L}$ can be “updated”, when the noisy measurement Y is obtained, by computing the conditional variance of X given Y and $\hat{\lambda}'_{X|\ell+L}$:

$$\hat{\lambda}_{X|\ell+L} = E\left\{|X|^2 \,|\, \hat{\lambda}'_{X|\ell+L}\,, Y\right\} = E\left\{X_R^2 \,|\, \hat{\lambda}'_{X|\ell+L}\,, Y_R\right\} + E\left\{X_I^2 \,|\, \hat{\lambda}'_{X|\ell+L}\,, Y_I\right\} \,. \tag{19}$$

Since X_R and X_I are IID, as well as the noise components D_R and D_I, we can write for $Y_\rho \neq 0$ ($\rho \in \{R, I\}$)

$$E\left\{X_\rho^2 \,|\, \hat{\lambda}'_{X|\ell+L}\,, Y_\rho\right\} = H\left(\xi', \gamma_\rho\right) Y_\rho^2 \tag{20}$$

where ξ' is an *a priori* SNR defined by

$$\xi'(k,\ell) = \frac{\lambda'_{X|\ell+L}(k,\ell)}{\lambda_D(k,\ell)} \tag{21}$$

and $H\left(\xi', \gamma_\rho\right)$ is a MMSE gain function in the spectral power domain. The specific expression for $H\left(\xi', \gamma_\rho\right)$ depends on the particular choice of a speech model. For a Gaussian speech model, the spectral power gain function is given by [15]

$$H\left(\xi',\, \gamma_\rho\right) = \frac{\xi'}{1+\xi'}\left(\frac{1}{\gamma_\rho} + \frac{\xi'}{1+\xi'}\right) \,. \tag{22}$$

For a Gamma speech model, the spectral power gain function is given by[1] (see Appendix I)

$$H\left(\xi',\, \gamma_\rho\right) = \frac{3}{\left(C_{\rho+} - C_{\rho-}\right)^2} \, \frac{\exp\left(C_{\rho-}^2/4\right) D_{-2.5}\left(C_{\rho-}\right) + \exp\left(C_{\rho+}^2/4\right) D_{-2.5}\left(C_{\rho+}\right)}{\exp\left(C_{\rho-}^2/4\right) D_{-0.5}\left(C_{\rho-}\right) + \exp\left(C_{\rho+}^2/4\right) D_{-0.5}\left(C_{\rho+}\right)} \tag{23}$$

where $C_{\rho\pm}$ are obtained from (15) by substituting ξ with ξ'. For a Laplacian speech model, the spectral power gain function is given by (see Appendix II)

$$H\left(\xi',\, \gamma_\rho\right) = \frac{4}{\left(L_{\rho+} - L_{\rho-}\right)^2} \, \frac{(L_{\rho+}^2 + 0.5)\mathrm{erfcx}(L_{\rho+}) + (L_{\rho-}^2 + 0.5)\mathrm{erfcx}(L_{\rho-}) - (L_{\rho+} + L_{\rho-})/\sqrt{\pi}}{\mathrm{erfcx}(L_{\rho+}) + \mathrm{erfcx}(L_{\rho-})} \tag{24}$$

where $L_{\rho\pm}$ are obtained from (17) by substituting ξ with ξ'.

[1] Note that (23) is a much simpler expression than the one derived in [10, sec. 3.2]. In particular, confluent hypergeometric functions are not involved, and the same expression holds for $C_{\rho-} \geq 0$ and $C_{\rho-} < 0$.

Equation (20) does not hold in the case $Y_\rho \to 0$, since it yields $H\left(\xi',\, \gamma_\rho\right) \to \infty$, and as a consequence the conditional variance of X_ρ is generally not zero. For $Y_\rho = 0$ (or practically for Y_ρ smaller than a predetermined threshold) we use the following expressions: For a Gaussian speech model

$$E\left\{X_\rho^2 \mid \hat{\lambda}'_{X|\ell+L}\,, Y_\rho = 0\right\} = \frac{\xi'}{1+\xi'}\,\lambda_D\,, \tag{25}$$

for a Gamma speech model we have (see Appendix I)

$$E\left\{X_\rho^2 \mid \hat{\lambda}'_{X|\ell+L}\,, Y_\rho = 0\right\} = \frac{3\,D_{-2.5}\left(\frac{\sqrt{3}}{2\sqrt{\xi'}}\right)}{8\,D_{-0.5}\left(\frac{\sqrt{3}}{2\sqrt{\xi'}}\right)}\,\lambda_D \tag{26}$$

and for a Laplacian speech model we have (see Appendix II)

$$E\left\{X_\rho^2 \mid \hat{\lambda}'_{X|\ell+L}\,, Y_\rho = 0\right\} = \sqrt{\frac{2}{\pi}}\,\frac{\exp\left(\frac{1}{2\xi'}\right) D_{-3}\left(\sqrt{\frac{2}{\xi'}}\right)}{\operatorname{erfcx}\left(\frac{1}{\sqrt{\xi'}}\right)}\,\lambda_D \tag{27}$$

Figure 2 shows parametric gain curves describing the spectral power gain functions $H(\xi', \gamma_\rho)$ for several values of γ_ρ, which result from (22), (23) and (24). In contrast with the gain functions $G(\xi, \gamma_\rho)$, which minimize the MSE between X_ρ and $\hat{X}_\rho$, the gain functions $H(\xi', \gamma_\rho)$ minimize the MSE between X_ρ^2 and $\widehat{X_\rho^2}$, and are not monotonically increasing functions of the *a posteriori* SNR. On the contrary, for a Gaussian speech model $H(\xi', \gamma_\rho)$ is a decreasing function of γ_ρ, and for Gamma and Laplacian speech models $H(\xi', \gamma_\rho)$ is a decreasing function of γ_ρ when γ_ρ is sufficiently small (depending on the *a priori* SNR ξ'). Therefore, local bursts of noise, which are associated with moderate values of γ_ρ and small values of ξ', are assigned lower values of $H(\xi', \gamma_\rho)$. This implies lower values of $\hat{\lambda}_{X|\ell+L}$, lower values of the *a priori* SNR estimate $\hat{\xi}$, and eventually lower spectral gains $G(\xi, \gamma_\rho)$. Such a behavior avoids the local buildup of noise, and thus counters the musical noise phenomenon.

To obtain an estimate for $\lambda'_{X|\ell+L}(k, \ell)$, we "propagate" in time the estimates $\hat{X}(k, \ell{-}1)$ and $\left\{\hat{\lambda}_{X|\ell+L-1}(k, \ell-1)\right\}_{k=0}^{K-1}$ from the previous frame, and employ the measurements $\mathcal{Y}_\ell^{\ell+L} \backslash \{Y(k, \ell)\}$. Suppose an estimate $\hat{\lambda}'_{X\,|\,[\ell,\ell+L]}(k, \ell)$ for λ_X is given, based on the measurements $\mathcal{Y}_\ell^{\ell+L} \backslash \{Y\}$. Let b denote a normalized window function of length $2w+1$, *i.e.*, $\sum_{i=-w}^{w} b(i) = 1$. Then, a useful estimator for $\lambda'_{X|\ell+L}(k, \ell)$, which combines the information from past and future frames, is given by [16]

$$\begin{aligned}\hat{\lambda}'_{X|\ell+L}(k, \ell) \quad = \quad \max\Bigg\{ & \mu|\hat{X}(k, \ell-1)|^2 + (1-\mu)\left[\mu' \sum_{i=-w}^{w} b(i)\, \hat{\lambda}_{X|\ell+L-1}(k-i, \ell-1)\right. \\ & \left. +(1-\mu')\hat{\lambda}'_{X\,|\,[\ell,\ell+L]}(k, \ell)\right]\,,\ \lambda_{\min}\Bigg\}\end{aligned} \tag{28}$$

where μ $(0 \le \mu \le 1)$ is related to the degree of nonstationarity of the random process $\{\lambda_X(k, \ell) \mid \ell = 0, 1, \ldots\}$, b is related to the correlation between frequency bins of λ_X, μ' $(0 \le \mu' \le 1)$ is associated with the reliability of the estimate $\hat{\lambda}'_{X\,|\,[\ell,\ell+L]}$ in comparison with that of $\hat{\lambda}_{X|\ell+L-1}$, and $\lambda_{\min}$ is a lower bound on the variance of X. An estimate for $\lambda'_{X\,|\,[\ell,\ell+L]}(k,\ell)$ given the measurements $\mathcal{Y}_{\ell}^{\ell+L}\backslash\{Y\}$ can be obtained by local averaging. Specifically,

$$\hat{\lambda}'_{X\,|\,[\ell,\ell+L]}(k,\ell) = \begin{cases} \frac{\sum_{(n,i)\in\Gamma} b(i)\,|Y(k-i,\ell+n)|^2}{\sum_{(n,i)\in\Gamma} b(i)} - \beta\,\lambda_D\,, & \text{if nonnegative,} \\ 0\,, & \text{otherwise,} \end{cases} \tag{29}$$

where $\Gamma \overset{\triangle}{=} \{(n,i)\,|\,0 \leq n \leq L,\ -w \leq i \leq w,\ (n,i) \neq (0,0)\}$ designates the time-frequency indices of the measurements, and β ($\beta \geq 1$) is an over-subtraction factor to compensate for a sudden increase in the noise level. The steps of the noncausal spectral enhancement algorithm for Gaussian, Gamma and Laplacian speech models are summarized in Table I.

For comparison, using the decision-directed approach of Ephraim and Malah [1], [13], an estimate for the *a priori* SNR $\xi(k,\ell)$, as defined in (11), can be obtained by

$$\hat{\xi}^{\mathrm{DD}}(k,\ell) = \max\left\{\alpha\frac{|\hat{X}(k,\ell-1)|^2}{\lambda_D} + (1-\alpha)\left[\gamma_R(k,\ell) + \gamma_I(k,\ell) - 1\right]\,,\ \xi_{\min}\right\}\,, \tag{30}$$

where α ($0 \leq \alpha \leq 1$) is a weighting factor that controls the trade-off between noise reduction and transient distortion introduced into the signal, and $\xi_{\min}$ is a lower bound on the *a priori* SNR. In the next section we present experimental results that show the improved performance of the noncausal *a priori* SNR estimator, compared with the decision-directed estimator, for MMSE estimation and Gaussian, Gamma and Laplacian speech models.

V. Experimental Results

In this section, the performance of the noncausal *a priori* SNR estimator is evaluated for different speech models, and compared to that of the decision-directed estimator. Figure 3 demonstrates the different behaviors of the noncausal and the decision-directed estimators for Gaussian, Gamma and Laplacian speech priors. The analyzed signal is sampled at 16 kHz, and transformed into the STFT domain using half overlapping Hamming windows of 512 samples length (32 ms). It contains only white Gaussian noise (WGN) during the first and last 20 frames, and in between it contains an additional sinusoidal component at the displayed frequency with 0 dB SNR[2]. The noncausal *a priori* SNR estimate $\hat{\xi}$ is obtained by using the algorithm in Table I, with the parameters $\mu = 0.8$, $\mu' = 0.5$, $b = \begin{bmatrix}0.25 & 0.5 & 0.25\end{bmatrix}$, $L = 2$, $\beta = 2$, $\lambda_{\min} = \xi_{\min}\lambda_D$, and $\xi_{\min} = -25$ dB. The decision-directed estimator $\hat{\xi}^{\mathrm{DD}}$ is obtained by (30) with the parameters $\alpha = 0.95$ and $\xi_{\min} = -25$ dB. Figure 3 shows that when the *a posteriori* SNR's γ_R and γ_I are sufficiently low, the noncausal *a priori* SNR estimate is smoother than the decision-directed estimate for all tested speech models. When γ_R or γ_I increases, the noncausal estimator, having a few subsequent spectral measurements at hand, is capable of discriminating between speech onsets and irregularities in the *a posteriori* SNR's corresponding to noise. It responds quickly to speech onsets, but remains close to its lower bound in case of speech irregularities. On the other hand, the decision-directed estimator cannot respond too fast to

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

an abrupt increase in γ_R or γ_I, since it necessarily implies an increase in the level of musical noise. When γ_R and γ_I decrease, the response of $\hat{\xi}$ is immediate, while that of $\hat{\xi}^{\mathrm{DD}}$ is delayed by 1 frame. Consequently, in comparison with the decision-directed estimator, the noncausal *a priori* SNR estimator entails lower levels of musical noise and signal distortion. Furthermore, the suppression of the musical noise phenomenon is more significant under a Gaussian speech model than under Gamma or Laplacian speech models. This is attributable to characteristics of the gain curves in Figs. 1 and 2. Under a Gaussian speech model, the spectral power gain function $H(\xi', \gamma_\rho)$ decreases as a function of γ_ρ, while the spectral gain $G(\xi, \gamma_\rho)$ is independent of γ_ρ. Thus, abrupt bursts of γ_ρ during noise-only frames are suppressed. On the other hand, under Gamma or Laplacian speech models, $H(\xi', \gamma_\rho)$ decreases as a function of γ_ρ only for sufficiently small γ_ρ, while $G(\xi, \gamma_\rho)$ increases as a function of γ_ρ. Thus, the mechanism, which counters the musical noise phenomenon, is not as much effective.

An experimental evaluation of the noncausal *a priori* SNR estimator is performed by enhancing noisy speech signals under various noise conditions and speech models, and comparing the results to those obtained by using 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 [22]

$$\mathrm{SegSNR} = \frac{1}{J}\sum_{\ell=0}^{J-1} \mathcal{T}\left\{10\,\log_{10}\frac{\sum_{n=0}^{N-1} x^2(n+\ell N/2)}{\sum_{n=0}^{N-1}\left[x(n+\ell N/2)-\hat{x}(n+\ell N/2)\right]^2}\right\} \quad [\mathrm{dB}] \tag{31}$$

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 [23], [24]. The second quality measure is log-spectral distortion, which is defined by

$$\mathrm{LSD} = \frac{1}{J}\sum_{\ell=0}^{J-1}\left\{\frac{1}{N/2+1}\sum_{k=0}^{N/2}\left[10\,\log_{10}\mathcal{C}X(k,\ell) - 10\,\log_{10}\mathcal{C}\hat{X}(k,\ell)\right]^2\right\}^{\frac{1}{2}} \quad [\mathrm{dB}] \tag{32}$$

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, $\delta = 10^{-50/10}\max\limits_{k,\ell}\left\{|X(k,\ell)|^2\right\}$).

The noise signals used in our evaluation are taken from the Noisex92 database [25]. 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 [26]. 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 half overlapping Hamming analysis windows of 512 samples length.

The noncausal speech enhancement algorithm (Table I) is applied to the noisy speech signals, using the same parameters as in the example of Fig. 3. Alternatively, the *a priori* SNR ξ is estimated by the decision-directed method (30), with the parameters $\xi_{\min} = -25$ dB and $\alpha = 0.98$ (this value of α was determined in [1], [17] by simulations and informal listening tests), and the spectral estimate $\hat{X}(k, \ell)$ is computed via (10) by using the appropriate spectral gain function (13), (14) or (16), according to the speech model. The noise spectral variance is estimated by recursively averaging past spectral power values of the noise signal: $\hat{\lambda}_D(k, \ell) = 0.95\, \hat{\lambda}_D(k, \ell - 1) + 0.05\, |D(k, \ell)|^2$. In practice, the periodogram of the noise $|D(k, \ell)|^2$ is unknown, and $\lambda_D(k, \ell)$ can be estimated by using the *Minima Controlled Recursive Averaging* approach [19].

Figure 4 shows the results of the segmental SNR improvement achieved by the noncausal and the decision-directed *a priori* SNR estimators for different speech models. The results of the log-spectral distance are displayed in Figure 5. The noncausal estimator consistently yields a higher segmental SNR and a lower LSD, than the decision-directed method, under all tested environmental conditions and speech models. The performance, in terms of segmental SNR and LSD, is greatest when using a Laplacian speech model and noncausal *a priori* SNR estimator. The performance is worst when using a Gaussian speech model and a decision-directed *a priori* SNR estimator. The Gamma speech model yields a higher segmental SNR and a lower LSD than the other speech models, only when the *a priori* SNR is estimated by the decision-directed method. However, when the *a priori* SNR is estimated by the proposed method, the Laplacian speech model yields a higher segmental SNR and a lower LSD than the other speech models. Informal listening tests confirm that by using the noncausal estimator, speech components are better preserved, while the residual musical noise is further reduced. The level of residual musical noise is minimal when using a Gaussian speech model and the noncausal estimator. The residual musical noise is maximal when using a Gamma speech model and the decision-directed method. Additionally, the differences between the Gaussian, Gamma and Laplacian speech models, in terms of segmental SNR, LSD and residual musical noise, are smaller when using the noncausal estimator than when using the decision-directed method.

VI. Conclusion

We have proposed noncausal recursive algorithms for MMSE estimation of speech signals for Gaussian, Gamma and Laplacian speech models. Noncausal estimation of the *a priori* SNR is accomplished by propagating across time and frequency spectral variance estimates from past and future frames, and updating the result by computing the conditional variance of the speech spectral component, based on the underlying speech model. We show that the noncausal *a priori* SNR estimator yields a higher segmental SNR, a lower LSD, and lower musical noise than the decision-directed estimator, under all tested environmental conditions and speech models. It should be noted that the heuristic estimator (28) is not relying on a model for the speech spectral variance process (*e.g.*, a Markovian), from which the estimator of the signal evolves [27], [28]. The parameters in (28) are related to the nonstationarity of the variance process, the correlation between frequency bins, and the reliability of the variance estimate from future noisy measurements. Furthermore, the assumption about the Dirac distribution of the speech spectral variance in (9) allows for the variance estimate to be substituted into the spectral estimate, which significantly simplifies the resultant algorithm.

We have shown that the spectral gains for Gamma and Laplacian speech models are monotonically increasing functions of the *a posteriori* SNR, when the *a priori* SNR is kept constant. Such a behavior is adverse to the useful mechanism that counters the musical noise phenomenon, since local bursts of noise are assigned higher gain values and further emphasized relative to the average noise characteristics. Using the noncausal *a priori* SNR estimator instead of the decision-directed estimator, local bursts of noise are assigned a lower *a priori* SNR, while speech onsets are assigned a higher *a priori* SNR. Thus, speech onsets are better preserved, while the musical noise effect is reduced. Experimental results show that the performance of the noncausal *a priori* SNR estimator, when combined with MMSE signal estimation, is best in terms of segmental SNR and LSD improvement under a Laplacian speech prior. However, the level of the residual musical noise is slightly higher than the level obtained under a Gaussian speech prior. Additionally, the differences between the Gaussian, Gamma and Laplacian speech models are smaller when using the noncausal *a priori* SNR estimator than when using the decision-directed method. Therefore, by taking into account the uncertainty of speech presence in the noisy measurements [1], [4], [29], [30], the Laplacian speech model should be very attractive. A Bernoulli-Laplacian speech model may lead to further suppression of the residual musical noise during speech absence, while preserving the same segmental SNR and LSD during speech presence. Another deserving study is related to the distortion measure, which is employed for the spectral enhancement. Estimators which minimize the mean-square error distortion of the spectral amplitude or log-spectral amplitude are more suitable for speech enhancement than MMSE estimators [1], [9], [17]. Hence, it may prove beneficial to utilize such estimators derived under Gamma or Laplacian speech modeling. These subjects are currently under investigation.

Appendix I
Conditional Moments $E\{X_\rho^n \,|\, \lambda_X, Y_\rho\}$ for a Gamma Speech Model

The conditional moments $E\{X_\rho^n \,|\, \lambda_X, Y_\rho\}$ for $n = 1, 2, \ldots$ and $\rho \in \{R, I\}$ are obtained by

$$E\{X_\rho^n \,|\, \lambda_X, Y_\rho\} = \frac{\int_{-\infty}^{\infty} X_\rho^n \, p\,(Y_\rho \,|\, X_\rho, \lambda_X)\, p\,(X_\rho \,|\, \lambda_X)\; dX_\rho}{\int_{-\infty}^{\infty} p\,(Y_\rho \,|\, X_\rho, \lambda_X)\, p\,(X_\rho \,|\, \lambda_X)\; dX_\rho} \tag{33}$$

Assuming a Gamma speech model and a Gaussian noise, we have

$$\begin{aligned} E\{X_\rho^n \,|\, \lambda_X, Y_\rho\} &= \frac{\int_{-\infty}^{\infty} X_\rho^n \, |X_\rho|^{-1/2} \, \exp\left(-\frac{(Y_\rho - X_\rho)^2}{\lambda_D} - \sqrt{\frac{3}{2\,\lambda_X}}|X_\rho|\right) dX_\rho}{\int_{-\infty}^{\infty} |X_\rho|^{-1/2} \, \exp\left(-\frac{(Y_\rho - X_\rho)^2}{\lambda_D} - \sqrt{\frac{3}{2\,\lambda_X}}|X_\rho|\right) dX_\rho} & (34) \\ &= \frac{\int_0^{\infty} X_\rho^{n-\frac{1}{2}} \left[\exp\left(-\frac{X_\rho^2}{\lambda_D} - \frac{G_{\rho-}}{\sqrt{\lambda_D}} X_\rho\right) + (-1)^n \exp\left(-\frac{X_\rho^2}{\lambda_D} - \frac{G_{\rho+}}{\sqrt{\lambda_D}} X_\rho\right)\right] dX_\rho}{\int_0^{\infty} X_\rho^{-\frac{1}{2}} \left[\exp\left(-\frac{X_\rho^2}{\lambda_D} - \frac{G_{\rho-}}{\sqrt{\lambda_D}} X_\rho\right) + \exp\left(-\frac{X_\rho^2}{\lambda_D} - \frac{G_{\rho+}}{\sqrt{\lambda_D}} X_\rho\right)\right] dX_\rho} & (35) \end{aligned}$$

where $G_{\rho\pm}$ are defined by

$$G_{\rho\pm} \triangleq \frac{\sqrt{3}}{2\sqrt{\xi}} \pm \frac{\sqrt{2}\,Y_\rho}{\sqrt{\lambda_D}}\,. \tag{36}$$

By using [21, eqs. 3.462.1, 8.339.2, 8.338.2], we obtain

$$E\{X_\rho^n \,|\, \lambda_X, Y_\rho\} = \frac{(2n-1)!!}{2^n}\left(\frac{\lambda_D}{2}\right)^{\frac{n}{2}} \frac{\exp\left(G_{\rho-}^2/4\right) D_{-n-0.5}\left(G_{\rho-}\right) + (-1)^n \exp\left(G_{\rho+}^2/4\right) D_{-n-0.5}\left(G_{\rho+}\right)}{\exp\left(G_{\rho-}^2/4\right) D_{-0.5}\left(G_{\rho-}\right) + \exp\left(G_{\rho+}^2/4\right) D_{-0.5}\left(G_{\rho+}\right)} \tag{37}$$

where $(2n-1)!! \triangleq 1 \cdot 3 \ldots (2n-1)$. Since $C_{\rho\pm}$, as defined by (36), are related to $G_{\rho\pm}$ by

$$G_{\rho\pm} = \begin{cases} C_{\rho\pm}\,, & \text{if } Y_\rho \geq 0\,, \\ C_{\rho\mp}\,, & \text{otherwise,} \end{cases} \tag{38}$$

we can rewrite (37) for $Y_\rho \neq 0$ as

$$E\{X_\rho^n \,|\, \lambda_X, Y_\rho\} = \frac{(2n-1)!!}{(C_{\rho+} - C_{\rho-})^n} \frac{\exp\left(C_{\rho-}^2/4\right) D_{-n-0.5}\left(C_{\rho-}\right) + (-1)^n \exp\left(C_{\rho+}^2/4\right) D_{-n-0.5}\left(C_{\rho+}\right)}{\exp\left(C_{\rho-}^2/4\right) D_{-0.5}\left(C_{\rho-}\right) + \exp\left(C_{\rho+}^2/4\right) D_{-0.5}\left(C_{\rho+}\right)}\, Y_\rho^n\,. \tag{39}$$

In particular, for $n = 1$ we have $E\{X_\rho \,|\, \lambda_X, Y_\rho\} = G\left(\xi,\, \gamma_\rho\right) Y_\rho$, where $G\left(\xi,\, \gamma_\rho\right)$ is defined by (14), and for $n = 2$ we have $E\{X_\rho^2 \,|\, \lambda_X, Y_\rho\} = H\left(\xi,\, \gamma_\rho\right) Y_\rho^2$, where $H\left(\xi,\, \gamma_\rho\right)$ is defined by (23). Note that for $Y_\rho = 0$, (37) reduces to

$$E\{X_\rho^n \,|\, \lambda_X, Y_\rho = 0\} = \frac{1 + (-1)^n}{2} \frac{(2n-1)!!}{2^n}\left(\frac{\lambda_D}{2}\right)^{\frac{n}{2}} \frac{D_{-n-0.5}\left(\sqrt{\frac{3\lambda_D}{4\lambda_X}}\right)}{D_{-0.5}\left(\sqrt{\frac{3\lambda_D}{4\lambda_X}}\right)}\,, \tag{40}$$

which is not zero in case n is an even number.

Appendix II

Conditional Moments $E\{X_\rho^n \,|\, \lambda_X, Y_\rho\}$ for a Laplacian Speech Model

Assuming a Laplacian speech model and a Gaussian noise, the conditional moments $E\{X_\rho^n \,|\, \lambda_X, Y_\rho\}$ for $n = 1, 2, \ldots$ and $\rho \in \{R, I\}$ are given by

$$\begin{aligned} E\{X_\rho^n \,|\, \lambda_X, Y_\rho\} &= \frac{\int_{-\infty}^{\infty} X_\rho^n \exp\left(-\frac{(Y_\rho - X_\rho)^2}{\lambda_D} - \frac{2}{\sqrt{\lambda_X}}|X_\rho|\right) dX_\rho}{\int_{-\infty}^{\infty} \exp\left(-\frac{(Y_\rho - X_\rho)^2}{\lambda_D} - \frac{2}{\sqrt{\lambda_X}}|X_\rho|\right) dX_\rho} & (41) \\ &= \frac{\int_0^{\infty} X_\rho^n \left[\exp\left(-\frac{X_\rho^2}{\lambda_D} - \frac{2F_{\rho-}}{\sqrt{\lambda_D}}X_\rho\right) + (-1)^n \exp\left(-\frac{X_\rho^2}{\lambda_D} - \frac{2F_{\rho+}}{\sqrt{\lambda_D}}X_\rho\right)\right] dX_\rho}{\int_0^{\infty} \left[\exp\left(-\frac{X_\rho^2}{\lambda_D} - \frac{2F_{\rho-}}{\sqrt{\lambda_D}}X_\rho\right) + \exp\left(-\frac{X_\rho^2}{\lambda_D} - \frac{2F_{\rho+}}{\sqrt{\lambda_D}}X_\rho\right)\right] dX_\rho} & (42) \end{aligned}$$

where $F_{\rho\pm}$ are defined by

$$F_{\rho\pm} \triangleq \frac{1}{\sqrt{\xi}} \pm \frac{Y_\rho}{\sqrt{\lambda_D}}\,. \tag{43}$$

By using [21, eqs. 3.462.1, 3.322.2], we obtain

$$E\{X_\rho^n \,|\, \lambda_X, Y_\rho\} = n!\sqrt{\frac{2}{\pi}}\left(\frac{\lambda_D}{2}\right)^{\frac{n}{2}} \frac{\exp\left(F_{\rho-}^2/2\right) D_{-n-1}\left(\sqrt{2}F_{\rho-}\right) + (-1)^n \exp\left(F_{\rho+}^2/2\right) D_{-n-1}\left(\sqrt{2}F_{\rho+}\right)}{\text{erfcx}(F_{\rho+}) + \text{erfcx}(F_{\rho-})} \tag{44}$$

The relation between $L_{\rho\pm}$, which are defined by (17), and $F_{\rho\pm}$ is given by

$$F_{\rho\pm} = \begin{cases} L_{\rho\pm}\,, & \text{if } Y_\rho \geq 0\,, \\ L_{\rho\mp}\,, & \text{otherwise,} \end{cases} \tag{45}$$

Hence, we can rewrite (44) for $Y_\rho \neq 0$ as

$$E\{X_\rho^n \,|\, \lambda_X, Y_\rho\} = \frac{n!\sqrt{2^{n+1}/\pi}}{(L_{\rho+} - L_{\rho-})^n} \frac{\exp\left(L_{\rho-}^2/2\right) D_{-n-1}\left(\sqrt{2}L_{\rho-}\right) + (-1)^n \exp\left(L_{\rho+}^2/2\right) D_{-n-1}\left(\sqrt{2}L_{\rho+}\right)}{\mathrm{erfcx}(L_{\rho+}) + \mathrm{erfcx}(L_{\rho-})} Y_\rho^n \tag{46}$$

In particular, for $n = 1$ we have $E\{X_\rho \,|\, \lambda_X, Y_\rho\} = G\,(\xi,\, \gamma_\rho)\, Y_\rho$, where $G\,(\xi,\, \gamma_\rho)$ is obtained from (46) by using [21, eq. 9.254.2], and is given by (16). For $n = 2$, we have $E\{X_\rho^2 \,|\, \lambda_X, Y_\rho\} = H\,(\xi,\, \gamma_\rho)\, Y_\rho^2$, where $H\,(\xi,\, \gamma_\rho)$ is obtained from (46) by using [21, eqs. 9.247.1, 9.254.1, 9.254.2], and is given by (24). Note that for $Y_\rho = 0$, (44) reduces to

$$E\{X_\rho^n \,|\, \lambda_X, Y_\rho = 0\} = \frac{1 + (-1)^n}{2} n! \sqrt{\frac{2}{\pi}} \left(\frac{\lambda_D}{2}\right)^{\frac{n}{2}} \frac{\exp\left(\frac{1}{2\xi}\right) D_{-3}\left(\sqrt{\frac{2}{\xi}}\right)}{\mathrm{erfcx}(\frac{1}{\sqrt{\xi}})}\,, \tag{47}$$

which is not zero in case n is an even number.

ACKNOWLEDGEMENT

The author thanks Prof. David Malah for his helpful comments.

REFERENCES

[1] Y. Ephraim and D. Malah, "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.

[2] A. J. Accardi and R. V. Cox, "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.

[3] 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.

[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] T. Lotter, C. Benien, and P. Vary, "Multichannel speech enhancement using bayesian spectral amplitude estimation," in *Proc. 28th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-03*, Hong Kong, 6–10 April 2003, pp. I_832–I_835.

[6] 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.

[7] J. W. B. Davenport, *Probability and Random Processes: an Introduction for Applied Scientists and Engineers*. New York: McGraw-Hill, 1970.

[8] 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.

[9] J. Porter and S. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. IEEE Internat. Conf. Acoust. Speech, Signal Process. (ICASSP)*, San Diego, California, 19–21 March 1984, pp. 18A.2.1–18A.2.4.

[10] 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.

[11] T. Lotter and P. Vary, “Noise reduction by maximum a posteriori spectral amplitude estimation with supergaussian speech modeling,” in *Proc. 8th Internat. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, 8–11 September 2003, pp. 83–86.

[12] R. Martin and C. Breithaupt, “Speech enhancement in the DFT domain using Laplacian speech priors,” in *Proc. 8th Internat. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, 8–11 September 2003, pp. 87–90.

[13] 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.

[14] 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.

[15] I. Cohen, “Relaxed statistical model for speech enhancement and *a priori* SNR estimation,” Technion - Israel Institute of Technology, Haifa, Israel, Technical Report, EE PUB 1384, October 2003.

[16] ——, “Speech enhancement using a noncausal *a priori* SNR estimator,” *to appear in IEEE Signal Processing Letters*.

[17] 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.

[18] I. Cohen, “Recursive estimation of speech spectral components based on a statistical model,” submitted.

[19] ——, “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.

[20] 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.

[21] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 4th ed. Academic Press, 1980.

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

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

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

[25] 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.

[26] 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).

[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, April 1992.

[28] ——, “Statistical-model-based speech enhancement systems,” *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, October 1992.

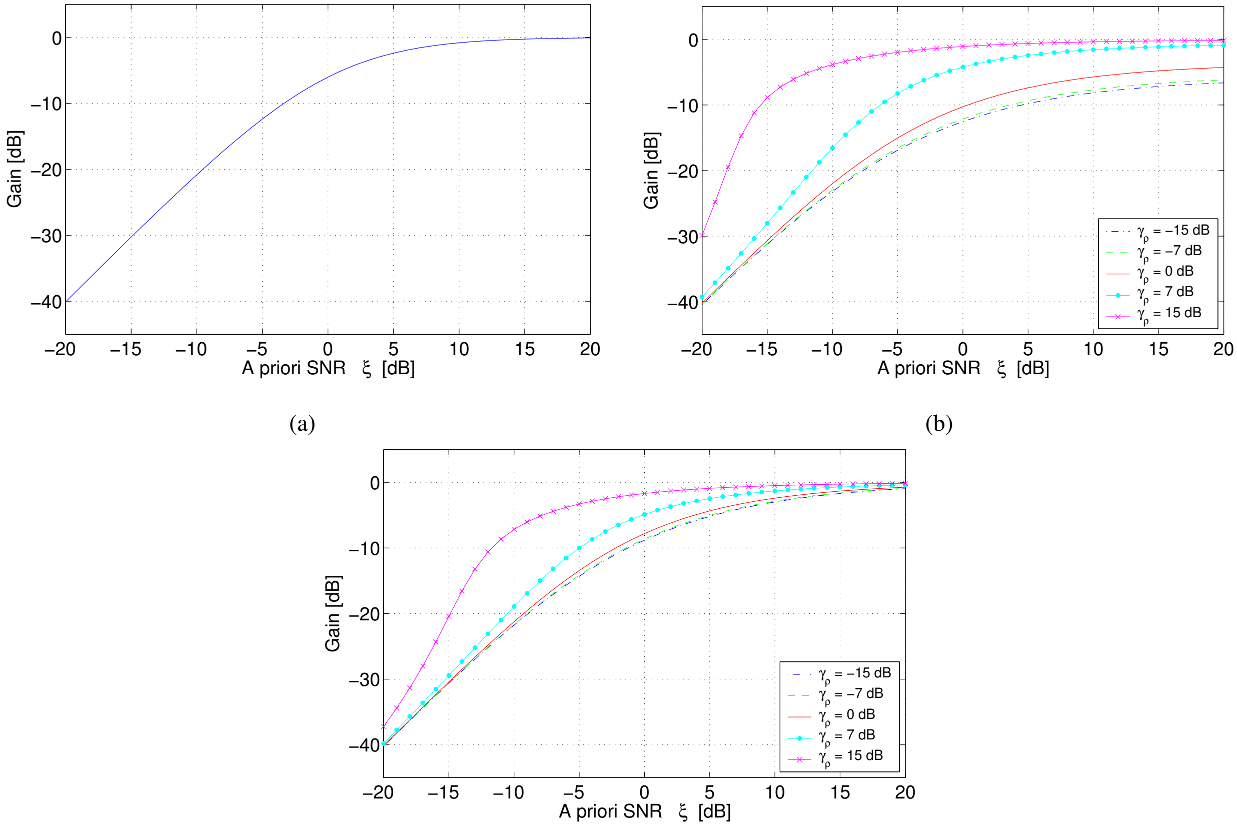
[29] R. J. McAulay and M. L. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-28, no. 2, pp. 137–145, April 1980.

[30] 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.

TABLE I

Summary of the Noncausal Speech Enhancement Algorithm for Gaussian, Gamma and Laplacian Speech Models.

```
Initialization at the first frame for all frequency bins $k$:
    $\hat{X}(k,-1) = 0\,,\ \hat{\lambda}_{X|L-1}(k,-1) = \lambda_{\min}\,.$
For all short-time frames $\ell = 0, 1, \ldots$
    For all frequency bins $k = 0, \ldots, K-1$
        Compute the spectral variance estimate $\hat{\lambda}'_{X|\,[\ell,\ell+L]}(k,\ell)$ by using (29).
        Compute the spectral variance estimate $\hat{\lambda}'_{X|\ell+L}(k,\ell)$ by using (28).
        Compute the a priori SNR $\xi'(k,\ell)$ by using (21), and the a posteriori SNR's $\gamma_\rho(k,\ell)$ $(\rho \in \{R, I\})$ by using (12).
        Compute the MMSE spectral-power gains $H\left(\xi', \gamma_\rho\right)$ $(\rho \in \{R, I\})$ by using (22), (23) or (24), according to the speech model.
        Update the spectral variance estimate $\hat{\lambda}_{X|\ell+L}(k,\ell)$ by using (19) and (20), and update the a priori SNR $\xi(k,\ell)$ by using (11).
        Compute the MMSE spectral gains $G\left(\xi, \gamma_\rho\right)$ $(\rho \in \{R, I\})$ by using (13), (14) or (16), according to the speech model.
        Compute the speech spectral estimate $\hat{X}(k,\ell)$ by using (10).
```

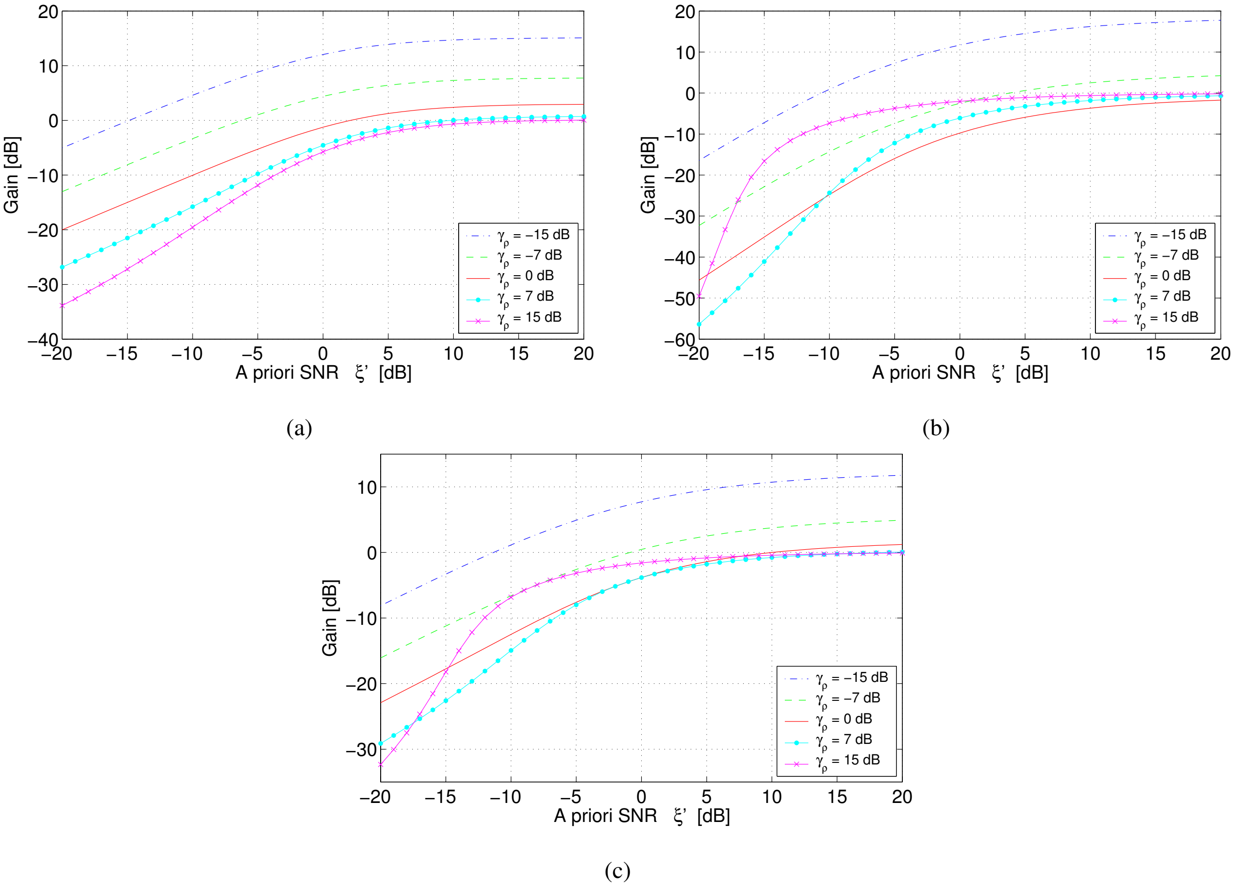


(a)

(b)

(c)

Fig. 1. Parametric gain curves describing the MMSE gain function $G(\xi, \gamma_\rho)$ for different speech models: (a) Gain for Gaussian speech model, obtained by (13); (b) Gain curves for Gamma speech model, obtained by (14); (c) Gain curves for Laplacian speech model, obtained by (16).

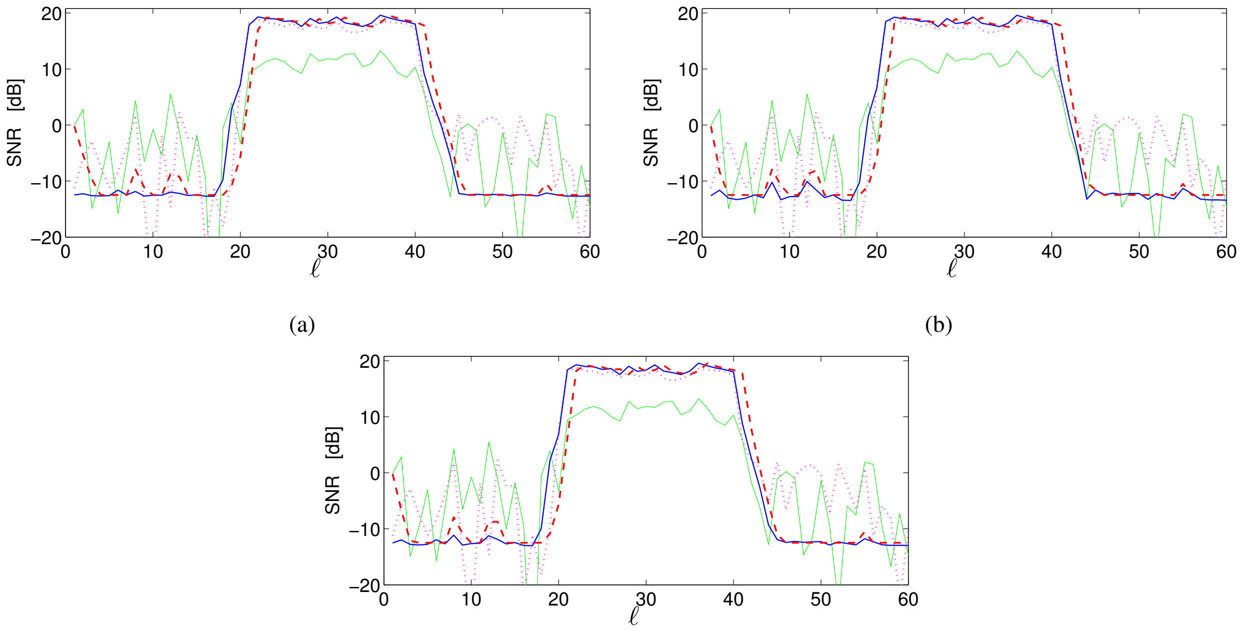


(a)

(b)

(c)

Fig. 2. Parametric gain curves describing the MMSE spectral power gain function $H(\xi', \gamma_\rho)$ for different speech models: (a) Gain curves for Gaussian speech model, obtained by (22); (b) Gain curves for Gamma speech model, obtained by (23); (c) Gain curves for Laplacian speech model, obtained by (24).

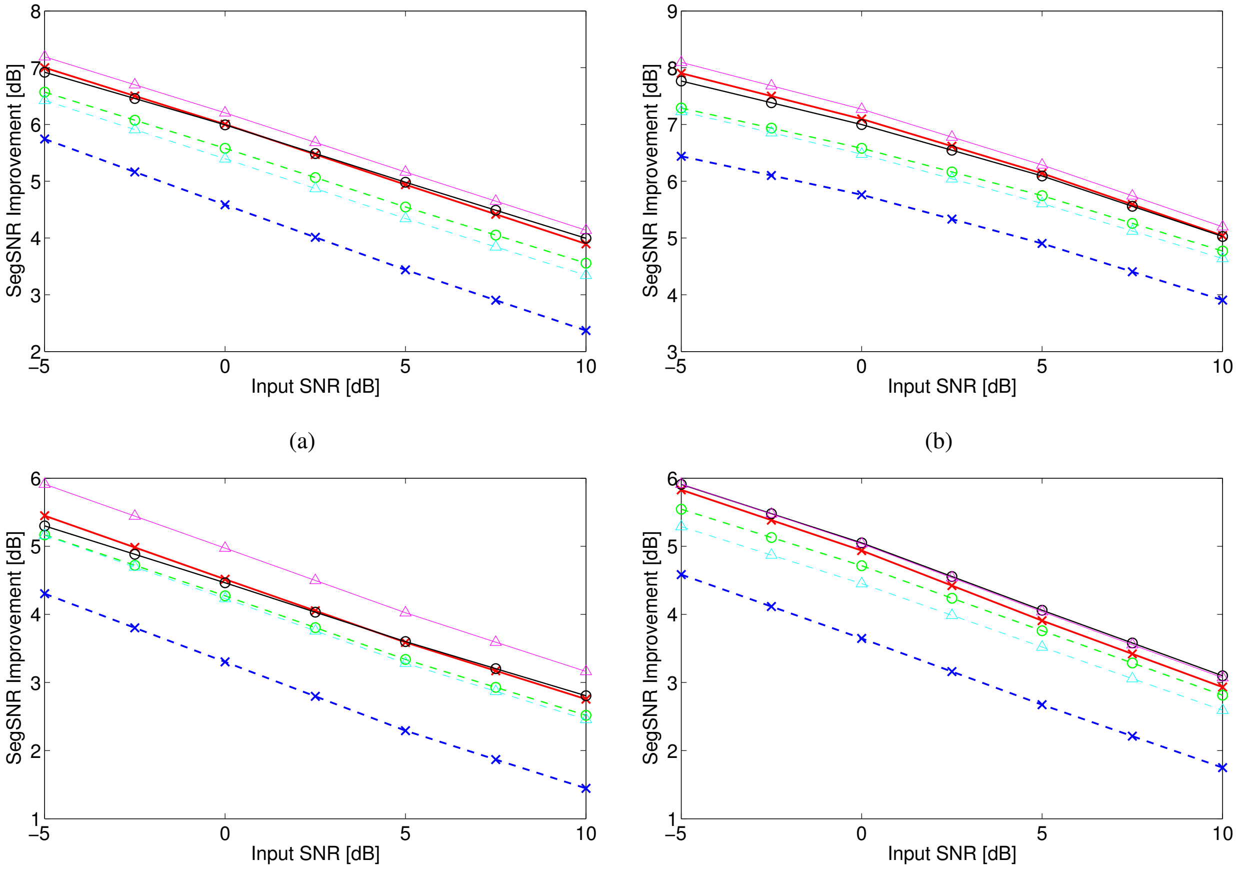


(a)

(b)

(c)

Fig. 3. SNR's in successive short-time frames for (a) Gaussian, (b) Gamma, and (c) Laplacian speech models: *A posteriori* SNR's γ_R (solid thin line) and γ_I (dotted line), decision-directed *a priori* SNR estimate $\hat{\xi}^{\mathrm{DD}}$ (dashed line), and noncausal *a priori* SNR estimate $\hat{\xi}$ (solid heavy line).



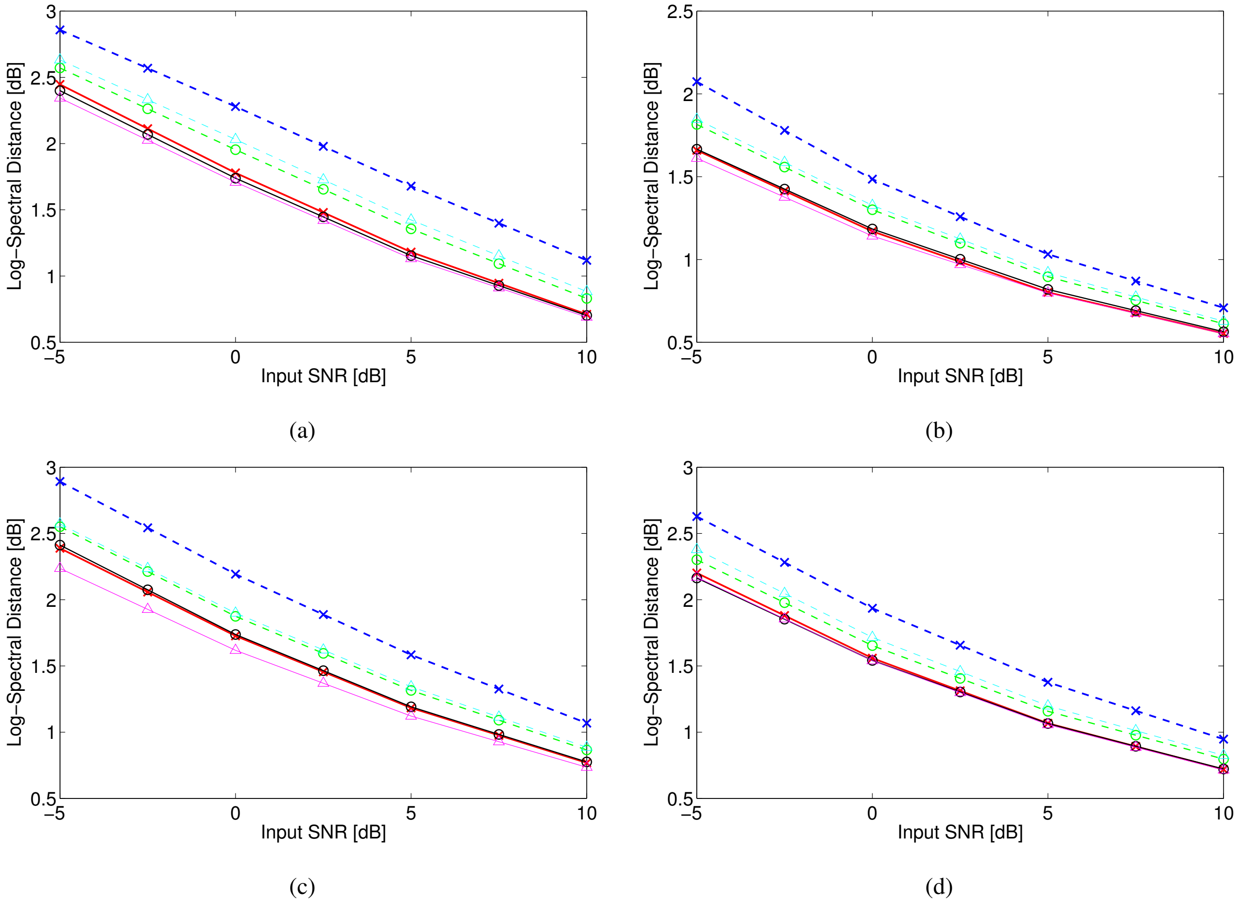
(a)

(b)

(c)

(d)

Fig. 4. Segmental SNR improvement for various noise types and levels, obtained by using Gaussian (×), Gamma (○) and Laplacian (△) speech models. The *a prior* SNR is obtained by either noncausal recursive estimation (solid lines) or by the decision-directed approach (dashed lines). (a) White Gaussian noise; (b) Car interior noise; (c) F16 cockpit noise; (c) Babble noise.



(a)

(b)

(c)

(d)

Fig. 5. Log-spectral distance for various noise types and levels, obtained by using Gaussian (×), Gamma (○) and Laplacian (△) speech models. The *a prior* SNR is obtained by either noncausal recursive estimation (solid lines) or by the decision-directed approach (dashed lines). (a) White Gaussian noise; (b) Car interior noise; (c) F16 cockpit noise; (c) Babble noise.