Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 47 (2005) 336–350
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Speech enhancement using super-Gaussian speech models and noncausal a priori SNR estimation

Israel Cohen *

Department of Electrical Engineering, Technion—Israel Institute of Technology, Technion City, Haifa 32000, Israel

Received 21 October 2004; received in revised form 23 January 2005; accepted 13 February 2005

Abstract

A priori signal-to-noise ratio (SNR) estimation is of major consequence in speech enhancement applications. Recently, we introduced a noncausal recursive estimator for the a priori SNR based on a Gaussian speech model, and showed its advantage compared to using the decision-directed estimator. In particular, noncausal estimation facilitates a distinction between speech onsets and noise irregularities. In this paper, we extend our noncausal estimation approach to Gamma and Laplacian speech models. We show that the performance of noncausal estimation, when applied to the problem of speech enhancement, is better under a Laplacian model than under Gaussian or Gamma models. Furthermore, the choice of the specific speech model has a smaller effect on the enhanced speech signal when using the noncausal a priori SNR estimator than when using the decision-directed method.

Keywords: Speech enhancement; Noncausal estimation; Spectral enhancement; Super-Gaussian speech modeling

1. Introduction

Optimal estimators for speech enhancement in the short-time Fourier transform (STFT) domain are often based on a Gaussian statistical model (Ephraim and Malah, 1984; Accardi and Cox, 1999; Sohn et al., 1999; Cohen and Berdugo, 2001; Lotter et al., 2003). Accordingly, the individual short-term spectral components of the speech and noise signals are modeled 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 (Ephraim and Malah, 1984), and a log-spectral amplitude (LSA) estimator, which minimizes the mean-square error of the log-spectra. Wolfe and Godsill (2003) derived under the same modeling assumptions three alternative suppression rules, which are based on joint maximum a posteriori

* Tel.: +972 4 8294731; fax: +972 4 8295757.
E-mail address: icohen@ee.technion.ac.il

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.02.011

(MAP) spectral amplitude and phase estimation, MAP spectral amplitude estimation, and minimum mean-square error (MMSE) spectral power estimation. The resulting suppression rules are simpler than those of Ephraim and Malah, yet demonstrate similar effect in reducing residual musical noise phenomena. Lotter et al. (2003) 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 (Ephraim and Malah, 1984). 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 (Davenport, 1970). Porter and Boll (1984) 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 (2002) 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 signal-to-noise ratio (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 (2003) 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 (2003) derived a MAP estimator for the speech spectral amplitude, based on a Gaussian noise model and a super-Gaussian 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 (2003) 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.

In all the above developments the a priori SNR, which is the dominant parameter of the spectral estimators (e.g., Wolfe and Godsill, 2003; Cappé, 1994; Scalart and Vieira-Filho, 1996), is obtained by the decision-directed approach of Ephraim and Malah (1984). Recently, we introduced causal and noncausal recursive estimators for the a priori SNR, which take into account the time-frequency correlation of speech signals (Cohen, 2004a,b, in press). We showed their close relation to the decision-directed estimator of Ephraim and Malah. 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 a few future spectral measurements (fixed lag) to better predict the

spectral variances of the clean speech. In some applications, e.g., digital voice recording, surveillance, and speaker identification, a delay of a few short-term frames between the enhanced speech and the noisy observation is tolerable. In such cases, 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 (Cohen, 2004a).

In this paper, we extend our noncausal estimation approach to Gamma and Laplacian speech models, while the noise model remains Gaussian. Spectral components in the STFT domain are assumed statistically correlated along the frequency axis, as well as along time-trajectories, due to the finite length of the analysis frame in the STFT and the overlap between successive frames (Cohen, in press). Hence, the noncausal 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 (Ephraim and Malah, 1984, 1985), 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 (Cappé, 1994). 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. Experimental results confirm that the noncausal estimators consistently yield 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. Furthermore, 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 indicate that 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 2, we review MMSE estimators for clean speech spectral components, based on Gaussian, Gamma and Laplacian speech models. In Section 3, we introduce noncausal a priori SNR estimators for Gamma and Laplacian speech models. In Section 4, we evaluate the performance of noncausal estimation under various speech models, and show experimental results, which demonstrate its advantage compared to using the decision-directed approach.

2. MMSE signal estimation

In this section, we review MMSE estimators in the STFT domain under Gaussian, Gamma and Laplacian speech models. 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 ($\ell = 0, 1, \ldots$). We assume that $X(k,\ell)$ and $D(k,\ell)$ are zero-mean random variables, and denote by $\lambda_X(k,\ell) \triangleq E\{|X(k,\ell)|^2\}$ and $\lambda_D(k,\ell) \triangleq E\{|D(k,\ell)|^2\}$ the speech and noise spectral variances, respectively. Then, the noise spectral components $\{D(k,\ell)\}$ are often assumed statistically independent zero-mean complex Gaussian random variables, given their variances $\lambda_D(k,\ell)$. However, the conditional pdf $X(k,\ell)$ of given the spectral variance $\lambda_X(k,\ell)$ is assumed either Gaussian (Ephraim and Malah, 1984, 1985), Gamma (Martin, 2002) or Laplacian (Breithaupt and Martin, 2003; Martin and Breithaupt, 2003). Let X_R and X_I denote, respectively, the real and imaginary parts of a clean speech spectral component X. Let $p(X_\rho|\lambda_X)$ denote the conditional pdf of X_ρ ($\rho \in \{R, I\}$) given the spectral variance λ_X. Then, for a Gaussian speech model

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

for a Gamma speech model

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

and for a Laplacian speech model

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

An MMSE estimator $\widehat{X}$ for X is obtained by

$$\widehat{X} = E\{X|Y,\lambda_X\} = \widehat{X}_R + j\widehat{X}_I, \tag{5}$$

where $\widehat{X}_\rho$ ($\rho \in \{R, I\}$) is an MMSE estimator for X_ρ given by

$$\widehat{X}_\rho = E\{X_\rho|Y_\rho,\lambda_X\} = \int X_\rho p(X_\rho|Y_\rho,\lambda_X)\,dX_\rho. \tag{6}$$

The expression for $\widehat{X}_\rho$ can be written as

$$\widehat{X}_\rho = G(\xi,\gamma_\rho)Y_\rho, \tag{7}$$

where

$$\xi(k,\ell) \triangleq \frac{\lambda_X(k,\ell)}{\lambda_D(k,\ell)}, \quad \gamma_\rho \triangleq \frac{Y_\rho^2(k,\ell)}{\lambda_D(k,\ell)}, \tag{8}$$

represent the a priori and a posteriori SNRs, respectively. The specific expression for the spectral gain function $G(\xi,\gamma_\rho)$ depends on the particular choice of a speech model. For a Gaussian speech model, the gain function is independent of the a posteriori SNR. It is often referred to as Wiener filter, given by (Lim and Oppenheim, 1979)

$$G(\xi) = \frac{\xi}{1+\xi}. \tag{9}$$

For a Gamma speech model,the gain function is given by (Martin, 2002; see also Appendix A)

$$G(\xi,\gamma_\rho) = \frac{1}{C_{\rho+} - C_{\rho-}} \times \frac{\exp(C_{\rho-}^2/4)D_{-1.5}(c_{\rho-}) - \exp(C_{\rho+}^2/4)D_{-1.5}(C_{\rho+})}{\exp(c_{\rho-}^2/4)D_{-0.5}(C_{\rho-}) + \exp(c_{\rho+}^2/4)D_{-0.5}(C_{\rho+})}, \tag{10}$$

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

$$C_{\rho\pm} \triangleq \frac{\sqrt{3}}{2\sqrt{\xi}} \pm \sqrt{2}\gamma_\rho \tag{11}$$

and $D_p(z)$ denotes the parabolic cylinder function (Gradshteyn and Ryzhik, 1980, Eq. (9.240)). For a Laplacian speech model, the gain function is given by (Martin and Breithaupt, 2003; see also Appendix B)

$$G(\xi,\gamma_\rho) = \frac{2}{L_{\rho+} - L_{\rho-}} \times \frac{L_{\rho+}\mathrm{erfcx}(L_{\rho+}) - L_{\rho-}\mathrm{erfcx}(L_{\rho-})}{\mathrm{erfcx}(L_{\rho+}) + \mathrm{erfcx}(L_{\rho-})}, \tag{12}$$

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

$$L_{\rho\pm} \triangleq \frac{1}{\sqrt{\xi}} \pm \sqrt{\gamma}_\rho, \tag{13}$$

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

$$\mathrm{erfcx}(x) \triangleq e^{x^2}\frac{2}{\sqrt{\pi}}\int_X^\infty e^{-t^2}\,dt. \tag{14}$$

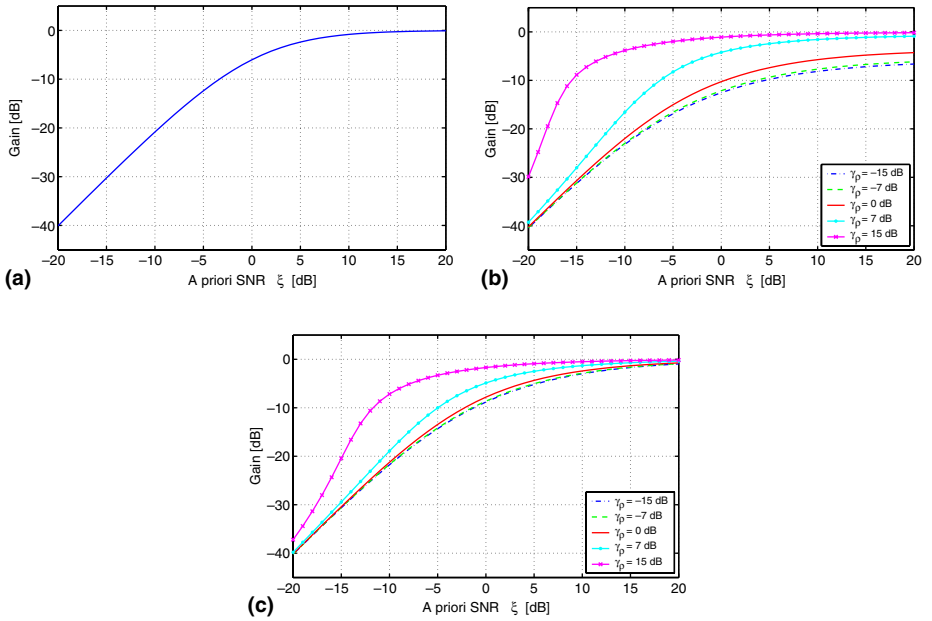

Fig. 1. Parametric gain curves describing the MMSE gain function $G(\xi, \gamma_\rho)$ for different speech models: (a) Gain for Gaussian speech model, obtained by (9); (b) Gain curves for Gamma speech model, obtained by (10); (c) Gain curves for Laplacian speech model, obtained by (12).

Fig. 1 displays gain curves $G(\xi, \gamma_\rho)$for several values of γ_ρ, which result from (9), (10) and (12). 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 (Ephraim and Malah, 1984, 1985), 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 a priori SNR. Such a behavior is related to the useful mechanism that counters the musical noise phenomenon (Cappé, 1994). 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 (Ephraim and Malah, 1984, 1985; Porter and Boll, 1984). 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, 1984, 1985). 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.

3. Noncausal estimation of speech spectral variance

A noncausal estimator for the a priori SNR was recently developed under a Gaussian speech model (Cohen, 2004a,b, in press). The noncausal 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. In this section, we extend the derivation of the noncausal a priori SNR estimator to Gamma and Laplacian speech models.

Let $\mathscr{Y}_0^{\ell+L} = \{Y(k,\ell')|0 \leqslant k \leqslant K-1, 0 \leqslant \ell' \leqslant \ell+L\}$ represent the set of spectral measurements up to frame $\ell+L$, where L ($L \geqslant 0$) denotes an admissible time delay in frames between the noisy speech signal and the enhanced signal. Let $\lambda'_{X|\ell+L}(k,\ell) \triangleq E\left\{|X(k,\ell)|^2|\mathscr{Y}_0^{\ell+L}\backslash\{Y(k,\ell)\}\right\}$ denote the conditional variance of X given $\mathscr{Y}_0^{\ell+L}$ excluding the noisy measurement Y. Let $\lambda'_{x|[\ell,\ell+L]}(k,\ell) \triangleq E\{|X(k,\ell)|^2|\mathscr{Y}_\ell^{\ell+L}\backslash\{Y(k,\ell)\}\}$ denote the conditional variance of X given the noisy measurements $\mathscr{Y}_\ell^{\ell+L}\backslash\{Y\}$. Then, the estimate for $\lambda_{x|\ell+L}$ is "updated", when the noisy measurement Y is obtained, by computing the conditional variance of X given Y and $\hat{\lambda}'_{X|\ell+L}$:

$$\begin{aligned}\hat{\lambda}_{X|\ell+L} &= E\left\{|X|^2|\hat{\lambda}'_{X|\ell+L}, Y\right\} \\ &= E\left\{X_{\mathrm{R}}^2|\hat{\lambda}'_{X|\ell+L}, Y_{\mathrm{R}}\right\} + E\left\{X_{\mathrm{I}}^2|\hat{\lambda}'_{X|\ell+L}, Y_{\mathrm{I}}\right\}.\end{aligned} \tag{15}$$

Since X_{R} and X_{I} are IID, as well as the noise components D_{R} and D_{I}, we can write for $Y_\rho \neq 0$ ($\rho \in \{\mathrm{R},\mathrm{I}\}$)

$$E\left\{X_\rho^2|\hat{\lambda}'_{X|\ell+L}, Y_\rho\right\} = H(\xi',\gamma_\rho)Y_\rho^2, \tag{16}$$

where ξ' is an a priori SNR defined by

$$\xi'(k,\ell) = \frac{\lambda'_{X|\ell+L}(k,\ell)}{\lambda_D(k,\ell)}, \tag{17}$$

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

$$H(\xi',\gamma_\rho) = \frac{\xi'}{1+\xi'}\left(\frac{1}{2\gamma_\rho} + \frac{\xi'}{1+\xi'}\right). \tag{18}$$

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

$$\begin{aligned}H(\xi',\gamma_\rho) = &\frac{3}{(C_{\rho+}-C_{\rho-})^2} \\ &\times \frac{\exp(C_{\rho-}^2/4)D_{-2.5}(C_{\rho-}) + \exp(C_{\rho+}^2/4)D_{-2.5}(C_{\rho+})}{\exp(C_{\rho-}^2/4)D_{-0.5}(C_{\rho-}) + \exp(C_{\rho+}^2 4)D_{-0.5}(C_{\rho+})},\end{aligned} \tag{19}$$

where $C_{\rho\pm}$ are obtained from (11) by substituting ξ with ξ'. For a Laplacian speech model, the spectral power gain function is given by (see Appendix B)

$$H(\xi',\gamma_\rho) = \frac{4}{(L_{\rho+}-L_{\rho-})^2}\frac{(L_{\rho+}^2+0.5)\mathrm{erfcx}(L_{\rho+}) + (L_{\rho-}^2+0.5)\mathrm{erfcx}(L_{\rho-}) - (L_{\rho+}+L_{\rho-})/\sqrt{\pi}}{\mathrm{erfcx}(L_{\rho+}) + \mathrm{erfcx}(L_{\rho-})}, \tag{20}$$

[1] Note that (19) is a much simpler expression than the one derived in (Breithaupt and Martin, 2003, Sec. 3.2). In particular, confluent hypergeometric functions are not involved, and the same expression holds for $C_{\rho-} \geqslant 0$ and $C_{\rho-} < 0$.

where $L_{\rho\pm}$ are obtained from (13) by substituting ξ with ξ'.

Eq. (16) does not hold in the case $Y_\rho \to 0$, since it yields $H(\xi', \gamma_\rho) \to \infty$, 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

$$E\left\{X_\rho^2 | \hat{\lambda}'_{X|\ell+L}, Y_\rho = 0\right\} = \frac{\xi'}{1+\xi'}\lambda_D, \tag{21}$$

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

$$E\left\{X_\rho^2 | \hat{\lambda}'_{X|\ell+L}, Y_\rho = 0\right\} = \frac{3D_{-2.5}\left(\frac{\sqrt{3}}{2\sqrt{\xi'}}\right)}{8D_{-0.5}\left(\frac{\sqrt{3}}{2\sqrt{\xi'}}\right)}\lambda_D, \tag{22}$$

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

$$E\left\{X_\rho^2 | \hat{\lambda}'_{X|\ell+L}, Y_\rho = 0\right\} = \sqrt{\frac{2}{\pi}}\frac{\exp\left(\frac{1}{2\xi'}\right)D_{-3}\left(\sqrt{\frac{2}{\xi'}}\right)}{\operatorname{erfcx}\left(\frac{1}{\sqrt{\xi'}}\right)}\lambda_D. \tag{23}$$

Fig. 2 shows parametric gain curves describing the spectral power gain functions $H(\xi', \gamma_\rho)$ for several values of γ_ρ, which result from (18)–(20). In contrast with the gain functions $G(\xi, \gamma_\rho)$, which minimize the MSE between X_ρ and $\widehat{X}_\rho$, 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').

In the present work, $G(\xi, \gamma_\rho)$ is used for estimating the clean speech spectral component X_ρ (see (7)), whereas $H(\xi', \gamma_\rho)$ is used for estimating the speech spectral variance (see (15) and (16)). This combination yields a desirable effect on the residual musical noise. Local bursts of noise, which are associated with higher (but 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.

The estimate for $\lambda'_{X|\ell+L}(k,\ell)$ is obtained by "propagating" in time the estimates $\widehat{X}(k,\ell-1)$ and $\{\hat{\lambda}_{X|\ell+L-1}(k,\ell-1)\}_{k=0}^{K-1}$ from the previous frame, and employing the measurements $\mathscr{Y}_\ell^{\ell+L} \setminus \{Y(k,\ell)\}$ (Cohen, 2004a,b, in press). Specifically, the estimator for $\lambda'_{X|\ell+L}(k,\ell)$, which combines the information from past and future frames, is given by

$$\hat{\lambda}'_{X|\ell+L}(k,\ell) = \max\left\{\mu|\widehat{X}(k,\ell-1)|^2 + (1-\mu)\left[\mu'\sum_{i=-\omega}^{\omega} b(i)\hat{\lambda}_{X|\ell+L-1}(k-i,\ell-1) + (1-\mu')\hat{\lambda}'_{X|[\ell,\ell+L]}(k,\ell)\right], \lambda_{\min}\right\}, \tag{24}$$

where μ ($0 \leqslant \mu \leqslant 1$) is related to the degree of nonstationarity of the random process $\{\lambda_X(k,\ell)|\ell = 0,1,\ldots\}$, b denotes a normalized window function of length $2\omega+1$ (i.e., $\sum_{i=-\omega}^{\omega} b(i) = 1$) which is related to the correlation between frequency bins of λ_X, μ' ($0 \leqslant \mu' \leqslant 1$) 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. The estimate for $\lambda'_{X|[\ell,\ell+L]}(k,\ell)$ is obtained by local averaging (Cohen, 2004a):

$$\hat{\lambda}'_{X|[\ell,\ell+L]}(k,\ell) = \begin{cases} \dfrac{\sum_{(n,i)\in\Gamma} b(i)|Y(k-i,\ell+n)|^2}{\sum_{(n,i)\in\Gamma} b(i)} - \beta\lambda_D, & \text{if nonnegative,} \\ 0, & \text{otherwise,} \end{cases} \tag{25}$$

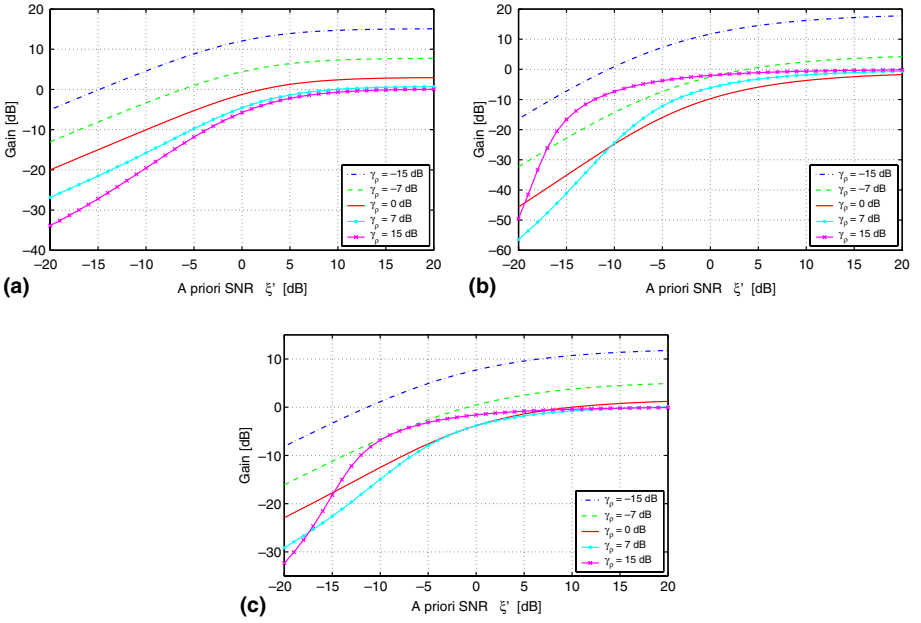

Fig. 2. Parametric gain curves describing the MMSE spectral power gain function $H(\xi', \gamma_\rho)$ for different speech models: (a) Gain curves for Gaussian speech model, obtained by (18); (b) Gain curves for Gamma speech model, obtained by (19); (c) Gain curves for Laplacian speech model, obtained by (20).

where $\Gamma \triangleq \{(n,i)|0 \leqslant n \leqslant L, -\omega \leqslant i \leqslant \omega, (n,i) \neq (0,0)\}$ designates the time-frequency indices of the measurements, and β ($\beta \geqslant 1$) is an over-subtraction factor to compensate for a sudden increase in the noise level. The steps of the noncausal spectral enhancement algorithm under Gaussian, Gamma or Laplacian speech models is summarized in Table 1.

4. Experimental results

In this section, the performance of the noncausal a priori SNR estimator is evaluated under different speech models, and compared to that of the decision-directed approach of Ephraim and Malah (1984) (Cappé, 1994). The decision directed estimator for the a priori SNR is given by

Table 1
Summary of the noncausal speech enhancement algorithm for Gaussian, Gamma and Laplacian speech models

Initialization at the first frame for all frequency bins k:
  $\widehat{X}(k,-1) = 0$, $\hat{\lambda}_{X|L-1}(k,-1) = \lambda_{\min}$
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 (25)
    Compute the spectral variance estimate $\hat{\lambda}'_{X|\ell+L]}(k,\ell)$ by using (24)
    Compute the a priori SNR $\xi'(k,\ell)$ by using (17), and the a posteriori SNRs $\gamma_\rho(k,\ell)$ ($\rho \in \{\mathrm{R},\mathrm{I}\}$) by using (8)
    Compute the MMSE spectral-power gains $H(\xi', \gamma_\rho)$ ($\rho \in \{\mathrm{R},\mathrm{I}\}$) by using (18), (19) or (20), according to the speech model
    Update the spectral variance estimate $\hat{\lambda}_{X|\ell+L}(k,\ell)$ by using (15) and (16), and update the a priori SNR $\xi(k,\ell)$ by using (8)
    Compute the MMSE spectral gains $G(\xi, \gamma_\rho)$ ($\rho \in \{\mathrm{R},\mathrm{I}\}$) by using (9), (10) or (12), according to the speech model
    Compute the speech spectral estimate $\widehat{X}(k,\ell)$ by using (5) and (7)

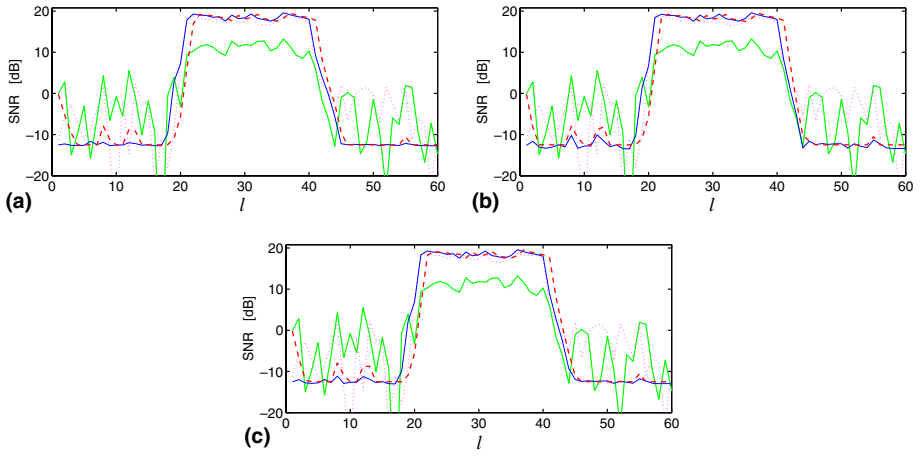

Fig. 3. SNRs in successive short-time frames for (a) Gaussian, (b) Gamma, and (c) Laplacian speech models: A posteriori SNRs γ_R (solid thin line) and γ_I (dotted line), decision-directed a priori SNR estimate $\hat{\xi}^{DD}$ (dashed line), and noncausal a priori SNR estimate $\hat{\xi}$ (solid heavy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

$$\hat{\xi}^{DD}(k,\ell)=\max\left\{\alpha\frac{|\widehat{X}(k,\ell-1)|^2}{\lambda_D}+(1-\alpha)[\gamma_R(k,\ell)+\gamma_I(k,\ell)-1],\xi_{min}\right\}, \tag{26}$$

where α ($0\leqslant\alpha\leqslant 1$) is a weighting factor that controls the trade-off between noise reduction and transient distortion introduced into the signal, and ξ_{min} is a lower bound on the a priori SNR.

Fig. 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 1, with the parameters $\mu=0.8$, $\mu'=0.5$, $b=[0.25\ 0.5\ 0.25]$, $L=2$, $\beta=2$, $\lambda_{min}=\xi_{min}\lambda_D$, and $\xi_{min}=-25$ dB. The decision-directed estimator $\hat{\xi}^{DD}$ is obtained by (26) with the parameters $\alpha=0.95$ and $\xi_{min}=-25$ dB. Fig. 3 shows that when the a posteriori SNRs 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 SNRs 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 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 $\hat{\xi}^{DD}$ 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

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

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, in dB, defined by (Quackenbush et al., 1988)

$$\text{SegSNR} = \frac{1}{J}\sum_{\ell=0}^{J-1} \mathscr{T}\left\{10\log_{10}\frac{\sum_{n=0}^{N-1} x^2(n+\ell N/2)}{\sum_{n=0}^{N-1}[x(n+\ell N/2)-\hat{x}(n+\ell N/2)]^2}\right\}, \tag{27}$$

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 $\mathscr{T}$ confines the SNR at each frame to perceptually meaningful range between 35 dB and -10 dB ($\mathscr{T}x \triangleq \min[\max(x,-10),35]$). The operator $\mathscr{T}$ 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 (Deller et al., 2000; Papamichalis, 1987). The second quality measure is log-spectral distortion, in dB, which is defined by

$$\text{LSD} = \frac{1}{J}\sum_{\ell=0}^{J-1}\left\{\frac{1}{N/2+1}\sum_{k=0}^{N/2}[10\log_{10}\mathscr{C}X(k,\ell) - 10\log_{10}\mathscr{C}\widehat{X}(k,\ell)]^2\right\}^{\frac{1}{2}}, \tag{28}$$

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

The noise signals used in our evaluation are taken from the Noisex92 database (Varga and Steeneken, 1993). 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 (Garofolo et al., 1988). The speech signal is sampled at 16 kHz and degraded by the various noise types with segmental SNRs in the range $[-5, 10]$ dB. The noisy signals are transformed into the STFT domain using half overlapping Hamming analysis windows of 512 samples length.

The noncausal speech enhancement algorithm (Table 1) 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 (26), with the parameters $\xi_{\min} = -25$ dB and $\alpha = 0.98$ [this value of α was determined in (Ephraim and Malah, 1984, 1985) by simulations and informal listening tests], and the spectral estimate $\widehat{X}(k,\ell)$ is computed via (7) by using the appropriate spectral gain function (9), (10) or (12), 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 (Cohen, 2003).

Fig. 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 Fig. 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,

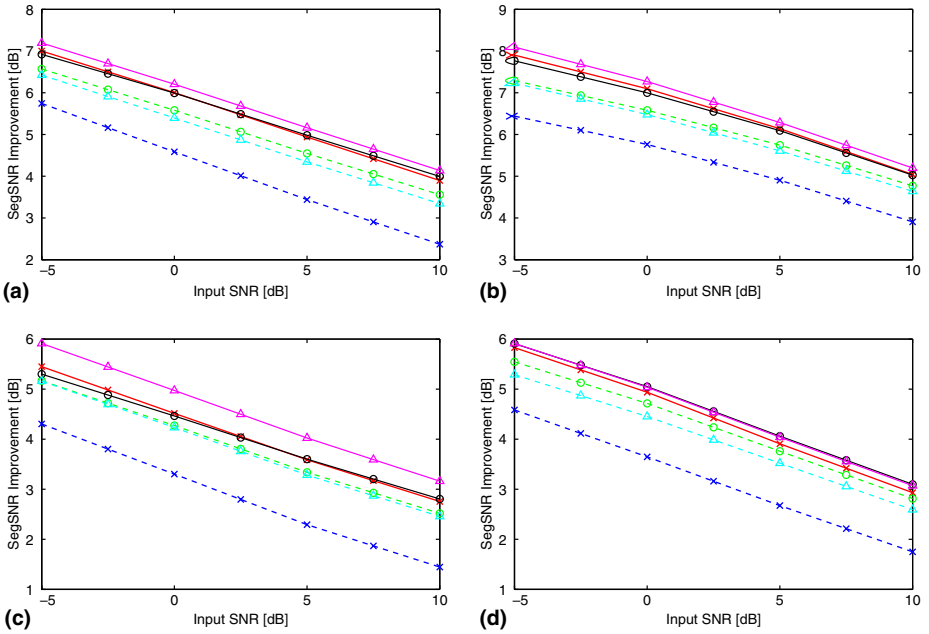

Fig. 4. Segmental SNR improvement for various noise types and levels, obtained by using Gaussian (×), Gamma (○) and Laplacian (△) speech models. The a prior SNR is obtained by either noncausal recursive estimation (solid lines) or by the decision-directed approach (dashed lines). (a) White Gaussian noise;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

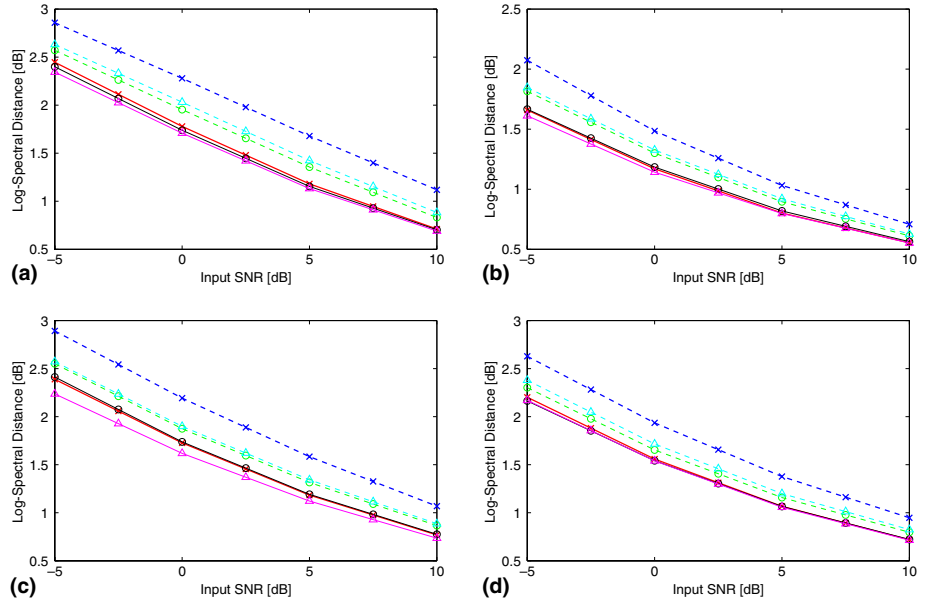

Fig. 5. Log-spectral distance for various noise types and levels, obtained by using Gaussian (×), Gamma (○) and Laplacian (△) speech models. The a prior SNR is obtained by either noncausal recursive estimation (solid lines) or by the decision-directed approach (dashed lines). (a) White Gaussian noise;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

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, conducted by four experienced listeners, 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.

5. Conclusion

We have proposed and evaluated the performance of noncausal recursive estimators for the a priori SNR under Gamma and Laplacian speech modeling. The noncausal estimation is accomplished by propagating spectral variance estimates across the time and frequency axes (see (24) and (25)), 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 (24) is not relying on a model for the speech spectral variance process (e.g., a Markovian), from which the estimator of the signal evolves (Ephraim, 1992a,b). The parameters in (24) 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 (Cohen, in press).

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 (Ephraim and Malah, 1984; Cohen and Berdugo, 2001; McAulay and Malpass, 1980; Malah et al., 1999), 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 (Ephraim and Malah, 1984, 1985; Porter and Boll, 1984). Hence, it may prove beneficial to utilize such estimators derived under Gamma or Laplacian speech modeling.

Acknowledgements

The author thanks Prof. David Malah and the anonymous reviewers for their helpful comments.

Appendix A. 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 $\rho \in \{\mathrm{R}, \mathrm{I}\}$ are obtained by

$$E\{X_\rho^n|\lambda_X, Y_\rho\} = \frac{\int_{-\infty}^{\infty} X_\rho^n p(Y_\rho|X_\rho, \lambda_X) p(X_\rho|\lambda_X)\,\mathrm{d}X_\rho}{\int_{-\infty}^{\infty} p(Y_\rho|X_\rho, \lambda_X) p(X_\rho|\lambda_X)\,\mathrm{d}X_\rho}. \tag{29}$$

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

$$E\{X_\rho^n|\lambda_X, Y_\rho\} = \frac{\int_{-\infty}^{\infty} X_\rho^n |X_\rho|^{-1/2} \exp\left(-\frac{(Y_\rho - X_\rho)^2}{\lambda_D} - \sqrt{\frac{3}{2\lambda_X}}|X_\rho|\right)\mathrm{d}X_\rho}{\int_{-\infty}^{\infty} |x_\rho|^{-1/2} \exp\left(-\frac{(Y_\rho - X_\rho)^2}{\lambda_D} - \sqrt{\frac{3}{2\lambda_X}}|X_\rho|\right)\mathrm{d}X_\rho} \tag{30}$$

$$= \frac{\int_0^{\infty} X_\rho^{n-\frac{1}{2}} \left[\exp\left(-\frac{X_\rho^2}{\lambda_D} - \frac{G_{\rho-}}{\sqrt{\lambda_D}}X_\rho\right) + (-1)^n \exp\left(-\frac{X_\rho^2}{\lambda_D} - \frac{G_{\rho+}}{\sqrt{\lambda_D}}X_\rho\right)\right]\mathrm{d}X_\rho}{\int_0^{\infty} X_\rho^{-\frac{1}{2}} \left[\exp\left(-\frac{X_\rho^2}{\lambda_D} - \frac{G_{\rho-}}{\sqrt{\lambda_D}}X_\rho\right) + \exp\left(-\frac{X_\rho^2}{\lambda_D} - \frac{G_{\rho+}}{\sqrt{\lambda_D}}X_\rho\right)\right]\mathrm{d}X_\rho}, \tag{31}$$

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

$$G_{\rho\pm} \triangleq \frac{\sqrt{3}}{2\sqrt{\xi}} \pm \frac{\sqrt{2}Y_\rho}{\sqrt{\lambda_D}}. \tag{32}$$

By using (Gradshteyn and Ryzhik, 1980, Eqs. 3.462.1, 8.339.2, 8.338.2), we obtain

$$E\{X_\rho^n|\lambda_X, Y_\rho\} = \frac{(2n-1)!!}{2n}\left(\frac{\lambda_D}{2}\right)^{\frac{n}{2}} \frac{\exp(G_{\rho-}^2/4) D_{-n-0.5}(G_{\rho-}) + (-1)^n \exp(G_{\rho+}^2/4) D_{-n-0.5}(G_{\rho+})}{\exp(G_{\rho-}^2/4) D_{-0.5}(G_{\rho-}) + \exp(G_{\rho+}^2/4) D_{-0.5}(G_{\rho+})}, \tag{33}$$

where $(2n-1)!! \triangleq 1 \cdot 3 \cdots (2n-1)$. Since $C_{\rho\pm}$, as defined by (32), are related to $G_{\rho\pm}$ by

$$G_{\rho\pm} = \begin{cases} C_{\rho\pm}, & \text{if } Y_\rho \geqslant 0, \\ C_{\rho\mp}, & \text{otherwise}, \end{cases} \tag{34}$$

we can rewrite (33) for $Y_\rho \neq 0$ as

$$E\{X_\rho^n|\lambda_X, Y_\rho\} = \frac{(2n-1)!!}{(C_{\rho+} - C_{\rho-})^n} \frac{\exp(C_{\rho-}^2/4) D_{-n-0.5}(C_{\rho-}) + (-1)^n \exp(C_{\rho+}^2/4) D_{-n-0.5}(C_{\rho+})}{\exp(C_{\rho-}^2/4) D_{-0.5}(C_{\rho-}) + \exp(C_{\rho+}^2/4) D_{-0.5}(C_{\rho+})} Y_\rho^n. \tag{35}$$

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 defined by (10), and 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 defined by (19). Note that for $Y_\rho = 0$, (33) reduces to

$$E\{X_\rho^n|\lambda_X, Y_\rho = 0\} = \frac{1 + (-1)^n}{2} \frac{(2n-1)!!}{2^n} \left(\frac{\lambda_D}{2}\right)^{\frac{n}{2}} \times \frac{D_{-n-0.5}\left(\sqrt{\frac{3\lambda_D}{4\lambda_X}}\right)}{D_{-0.5}\left(\sqrt{\frac{3\lambda_D}{4\lambda_X}}\right)}, \tag{36}$$

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

Appendix B. 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 $\rho \in \{\mathrm{R}, \mathrm{I}\}$ are given by

$$E\{X_\rho^n|\lambda_X,Y_\rho\}=\frac{\int_{-\infty}^{\infty}X_\rho^n\exp\left(-\frac{(Y_\rho-X_\rho)^2}{\lambda_D}-\frac{2}{\sqrt{\lambda_X}}|X_\rho|\right)\mathrm{d}X_\rho}{\int_{-\infty}^{\infty}\exp\left(-\frac{(Y_\rho-X_\rho)^2}{\lambda_D}-\frac{2}{\sqrt{\lambda_X}}|X_\rho|\right)\mathrm{d}X_\rho} \tag{37}$$

$$=\frac{\int_0^{\infty}X_\rho^n\left[\exp\left(-\frac{X_\rho^2}{\lambda_D}-\frac{2F_{\rho-}}{\sqrt{\lambda_D}}X_\rho\right)+(-1)^n\exp\left(-\frac{X_\rho^2}{\lambda_D}-\frac{2F_{\rho+}}{\sqrt{\lambda_D}}X_\rho\right)\right]\mathrm{d}X_\rho}{\int_0^{\infty}\left[\exp(-\frac{X_\rho^2}{\lambda_D}-\frac{2F_{\rho-}}{\sqrt{\lambda_D}}X_\rho)+\exp\left(-\frac{X_\rho^2}{\lambda_D}-\frac{2F_{\rho+}}{\sqrt{\lambda_D}}X_\rho\right)\right]\mathrm{d}X_\rho}, \tag{38}$$

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

$$F_{\rho\pm}\triangleq\frac{1}{\sqrt{\xi}}\pm\frac{Y_\rho}{\sqrt{\lambda_D}}. \tag{39}$$

By using (Gradshteyn and Ryzhik, 1980, Eqs. 3.462.1, 3.322.2), we obtain

$$E\{X_\rho^n|\lambda_X,Y_\rho=0\}=\frac{1+(-1)^n}{2}n!\sqrt{\frac{2}{\pi}}\left(\frac{\lambda_D}{2}\right)^{\frac{n}{2}}\times\frac{\exp\left(\frac{1}{2\xi}\right)D_{-3}\left(\sqrt{\frac{2}{\xi}}\right)}{\mathrm{erfcx}\left(\frac{1}{\sqrt{\xi}}\right)}, \tag{43}$$

$$E\{X_\rho^n|\lambda_X,Y_\rho\}=n!\sqrt{\frac{2}{\pi}}\left(\frac{\lambda_D}{2}\right)^{\frac{n}{2}}\frac{\exp(F_{\rho-}^2/2)D_{-n-1}(\sqrt{2}F_{\rho-})+(-1)^n\exp(F_{\rho+}^2/2)D_{-n-1}(\sqrt{2}F_{\rho+})}{\mathrm{erfcx}(F_{\rho+})+\mathrm{erfcx}(F_{\rho-})}. \tag{40}$$

The relation between $L_{\rho\pm}$, which are defined by (13), and $F_{\rho\pm}$ is given by

$$F_{\rho\pm}=\begin{cases}L_{\rho\pm}, & \text{if } Y_\rho\geqslant 0,\\ L_{\rho\mp}, & \text{otherwise.}\end{cases} \tag{41}$$

Hence, we can rewrite (40) for $Y_\rho\neq 0$ as

$$E\{X_\rho^n|\lambda_X,Y_\rho\}=\frac{n!\sqrt{\frac{2^{n+1}}{\pi}}}{(L_{\rho+}-L_{\rho-})^n}\frac{\exp(L_{\rho-}^2/2)D_{-n-1}(\sqrt{2}L_{\rho-})+(-1)^n\exp(L_{\rho+}^2/2)D_{-n-1}(\sqrt{2}L_{\rho+})}{\mathrm{erfcx}(L_{\rho+})+\mathrm{erfcx}(L_{\rho-})}Y_\rho^n. \tag{42}$$

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 (42) by using (Gradshteyn and Ryzhik, 1980, Eq. 9.254.2) , and is given by (12). 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 (42) by using (Gradshteyn and Ryzhik, 1980, Eqs. 9.247.1, 9.254.1,2), and is given by (20). Note that for $Y_\rho=0$, (40) reduces to

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

References

Accardi, A.J., Cox, R.V., 1999. 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, AZ, 15–19 March 1999. pp. 201–204.

Breithaupt, C., Martin, R., 2003. 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.

Cappé, O., 1994. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans. Acoust. Speech Signal Process. 2 (2), 345–349.

Cohen, I., 2003. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. 11 (5), 466–475.

Cohen, I., 2004a. Speech enhancement using a noncausal a priori SNR estimator. IEEE Signal Process. Lett. 11 (9), 725–728.

Cohen, I., 2004b. On the decision-directed estimation approach of Ephraim and Malah. In: Proc. 29th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-2004, Montreal, Canada, 17–21 May 2004. pp. I-293–I-296.

Cohen, I., in press. Relaxed statistical model for speech enhancement and a priori SNR estimation. IEEE Trans. Speech Audio Process.

Cohen, I., Berdugo, B., 2001. Speech enhancement for non-stationary noise environments. Signal Process. 81 (11), 2403–2418.

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

Deller, J.R., Hansen, J.H.L., Proakis, J.G., 2000. Discrete-time Processing of Speech Signals, second ed. IEEE Press, New York.

Ephraim, Y., 1992a. A bayesian estimation approach for speech enhancement using hidden Markov models. IEEE Trans. Signal Process. 40 (4), 725–735.

Ephraim, Y., 1992b. Statistical-model-based speech enhancement systems. Proc. IEEE 80 (10), 1526–1555.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. ASSP-32 (6), 1109–1121.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 33 (2), 443–445.

Garofolo, J.S., 1988. 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).

Gradshteyn, I.S., Ryzhik, I.M., 1980. Table of Integrals, Series, and Products, fourth ed. Academic Press.

Lim, J.S., Oppenheim, A.V., 1979. Enhancement and bandwidth compression of noisy speech. Proc. IEEE 67 (12), 1586–1604.

Lotter, T., Vary, P., 2003. 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.

Lotter, T., Benien, C., Vary, P., 2003. 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.

Malah, D., Cox, R.V., Accardi, A.J., 1999. 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, AZ, 15–19 March 1999. pp. 789–792.

Martin, R., 2002. 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, FL, 13–17 May 2002. pp. I-253–I-256.

Martin, R., Breithaupt, C., 2003. 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.

McAulay, R.J., Malpass, M.L., 1980. Speech enhancement using a soft-decision noise suppression filter. IEEE Trans. Acoust. Speech Signal Process. ASSP-28 (2), 137–145.

Papamichalis, P.E., 1987. Practical Approaches to Speech Coding. Prentice-Hall Inc., Englewood Cliffs, NJ.

Porter, J., Boll, S., 1984. Optimal estimators for spectral restoration of noisy speech. In: Proc. IEEE Internat. Conf. Acoust. Speech, Signal Process. (ICASSP), San Diego, CA, 19–21 March 1984. pp. 18A.2.1–18A.2.4.

Quackenbush, S.R., Barnwell, T.P., Clements, M.A., 1988. Objective Measures of Speech Quality. Prentice-Hall Inc., Englewood Cliffs, NJ.

Scalart, P., Vieira-Filho, J., 1996. Speech enhancement based on a priori signal to noise estimation. In: Proc. 21th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-96, Atlanta, GA, 7–10 May 1996. pp. 629–632.

Sohn, J., Kim, N.S., Sung, W., 1999. A statistical model-based voice activity detector. IEEE Signal Process. Lett. 6 (1), 1–3.

Varga, A., Steeneken, H.J.M., 1993. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. 12 (3), 247–251.

Wolfe, P.J., Godsill, S.J., 2003. Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement. In: Digital audio for multimedia communications. EURASIP JASP 2003 (10), 1043–1051 (special issue).