Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Signal Processing 86 (2006) 698–709

SIGNAL PROCESSING

www.elsevier.com/locate/sigpro

Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models ☆

Israel Cohen*

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

Received 10 November 2004; received in revised form 8 June 2005; accepted 8 June 2005
Available online 20 July 2005

Abstract

In this paper, we develop and evaluate speech enhancement algorithms, which are based on supergaussian *generalized autoregressive conditional heteroscedasticity* (GARCH) models in the short-time Fourier transform (STFT) domain. We consider three different statistical models, two fidelity criteria, and two approaches for the estimation of the variances of the STFT coefficients. The statistical model is either Gaussian, Gamma or Laplacian; the fidelity criteria include minimum mean-squared error (MMSE) of the STFT coefficients and MMSE of the log-spectral amplitude (LSA); the spectral variance is estimated based on either the proposed GARCH models or the decision-directed method of Ephraim and Malah. We show that estimating the variance by the GARCH modeling method yields lower log-spectral distortion and higher perceptual evaluation of speech quality scores (PESQ, ITU-T P.862) than by using the decision-directed method, whether the presumed statistical model is Gaussian, Gamma or Laplacian, and whether the fidelity criterion is MMSE of the STFT coefficients or MMSE of the LSA. furthermore while a gaussian model is inferior to the supergaussian models when USING the decision-directed method, the Gaussian model is superior when using the garch modeling method.

Keywords: Speech enhancement; Speech modeling; GARCH; Time-frequency analysis

1. Introduction

Statistical modeling of speech signals in the short-time Fourier transform (STFT) domain has recently received much attention, but is still a puzzling problem [1]. Ephraim and Malah [2] proposed to model the individual STFT expansion coefficients of the speech signal as zero-mean statistically independent Gaussian random variables. It enables to derive useful minimum mean-squared error (MMSE) estimators for the short-term spectral amplitude (STSA), as well as the log-spectral amplitude (LSA) [2,3], and it underlies

☆This work was supported by E. and J. Bishop Research Fund.
*Tel.: +972 4 8294731; fax: +972 4 8295757.
E-mail address: icohen@ee.technion.ac.il.

0165-1684/$ - see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.sigpro.2005.06.005

the design of many speech enhancement algorithms, e.g. [4–8]. Martin [9] considered a Gamma speech model, under which the real and imaginary parts of the STFT coefficients are modeled as independent and identically distributed (iid) Gamma random variables. He assumed that distinct expansion coefficients are statistically independent, and derived their MMSE estimators. He showed that the Gamma model yields higher improvement in the segmental signal-to-noise ratio (SNR) than the Gaussian model.

Lotter et al. [7] proposed a parametric probability density function (pdf) for the magnitude of the expansion coefficients, which approximates, with a proper choice of the parameters, the Gamma and Laplacian densities. They derived a maximum a-posteriori (MAP) estimator for the speech spectral amplitude, and showed that under Laplacian speech modeling the MAP estimator demonstrates improved noise reduction compared with the STSA estimator of Ephraim–Malah. Martin and Breithaupt [10] showed that MMSE estimators for the STFT coefficients under Laplacian speech modeling have similar properties to those estimators derived under Gamma modeling, but are easier to compute and implement. Statistical models based on hidden Markov Processes (HMPs) try to circumvent the assumption of specific distributions [11,12]. The probability distributions of the speech and noise processes are estimated from long training sequences of the speech and noise samples, and then used jointly with a given fidelity criterion to derive an estimator for the speech signal. Unfortunately, the HMP-based speech enhancement relies on the types of training data [13]. It works best with the trained type of noise, but often worse with other type of noise. Furthermore, improved performance generally entails more complex models and higher computational requirements.

Recently, we introduced a novel approach for statistically modeling speech signals in the STFT domain [14]. This approach is based on generalized autoregressive conditional heteroscedasticity (GARCH) modeling, which is widely used for modeling the volatility of financial time-series such as exchange rates and stock returns [15,16]. Similar to financial time-series, speech signals in the STFT domain are characterized by heavy tailed distributions and volatility clustering. Specifically, when observing a time series of successive expansion coefficients in a fixed frequency bin, successive magnitudes of the expansion coefficients are highly correlated, whereas successive phases can be assumed uncorrelated. Hence, the expansion coefficients are clustered in the sense that large magnitudes tend to follow large magnitudes and small magnitudes tend to follow small magnitudes, while the phase is unpredictable.

In this paper, we develop and evaluate speech enhancement algorithms which are based on supergaussian GARCH models. We consider three different statistical models, two fidelity criteria, and two approaches for the estimation of the variances of the STFT coefficients. The statistical model is either Gaussian, Gamma or Laplacian; the fidelity criteria include MMSE of the STFT coefficients and MMSE of the LSA; the spectral variance is estimated based on either the proposed GARCH models or the decision-directed method of Ephraim and Malah [2]. We show that estimating the variance by the GARCH modeling method yields lower log-spectral distortion (LSD) and higher perceptual evaluation of speech quality (PESQ) scores (ITU-T P.862) than by using the decision-directed method, whether the presumed statistical model is Gaussian, Gamma or Laplacian, and whether the fidelity criterion is MMSE of the STFT coefficients or MMSE of the LSA. Furthermore, a Gaussian model is inferior to Gamma and Laplacian models if the speech variance is estimated by the decision-directed method. However, a Gaussian model is superior in the case speech variance is estimated by using the GARCH modeling method. Additionally, MMSE–LSA estimators yield lower LSD and higher PESQ scores than MMSE estimators of the STFT coefficients, whether the variance is estimated by using the GARCH modeling method or the decision-directed method. A subjective study of speech spectrograms and informal listening tests confirm that the quality of the enhanced speech obtained by using the GARCH modeling method is significantly better than that obtainable by using the decision-directed method.

The paper is organized as follows. In Section 2, we present supergaussian GARCH models for speech signals in the STFT domain. In Section 3, we discuss the problems addressed in this work. In Section 4, we derive recursive estimators for the STFT expansion coefficients of the speech signal. Finally, in Section 5, we evaluate the performances of MMSE estimators for the STFT coefficients and LSA under Gaussian, Gamma and Laplacian models, and compare the GARCH modeling method to the decision-directed method.

2. Statistical models

In this section, we introduce supergaussian GARCH models for speech signals in the STFT domain using the GARCH modeling approach proposed in [14].

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

$$Y_{tk} = X_{tk} + D_{tk}, \tag{1}$$

where t is the time frame index ($t = 0, 1, \ldots$) and k is the frequency-bin index ($k = 0, 1, \ldots, K-1$). Let H_0^{tk} and H_1^{tk} denote, respectively, hypotheses of signal absence and presence in the noisy spectral coefficient Y_{tk}, and let s_{tk} denote a binary state variable which indicates signal presence or absence, i.e., $s_{tk} = 0$ under H_0^{tk}, and $s_{tk} = 1$ under H_1^{tk}. Let $\sigma_{tk}^2 \triangleq E\{|D_{tk}|^2\}$ denote the variance of a noise spectral coefficient D_{tk}, and let $\lambda_{tk} \triangleq E\{|X_{tk}|^2|H_1^{tk}\}$ denote the variance of a speech spectral coefficient X_{tk} under H_1^{tk}. The variances of the speech coefficients are hidden from direct observation, in the sense that even under perfect conditions of zero noise ($D_{tk} = 0$ for all tk), the values of $\{\lambda_{tk}\}$ are not directly observable. Therefore, the approach proposed in [14] is to assume that $\{\lambda_{tk}\}$ themselves are random variables, and to introduce *conditional* variances which are estimated from the available information (*e.g.*, the clean spectral coefficients through frame $t-1$, or the noisy spectral coefficients through frame t). Let $\mathcal{X}_0^\tau = \{X_{tk} | t = 0, \ldots, \tau,\ k = 0, \ldots, K-1\}$ represent the set of clean speech spectral coefficients up to frame τ, and let $\lambda_{tk|\tau} \triangleq E\{|X_{tk}|^2|H_1^{tk}, \mathcal{X}_0^\tau\}$ denote the *conditional* variance of X_{tk} under H_1^{tk} given the clean spectral coefficients up to frame τ. Then, our statistical models in the STFT domain rely on the following set of assumptions:

(1) Given $\{\lambda_{tk}\}$ and $\{s_{tk}\}$, the speech spectral coefficients $\{X_{tk}\}$ are generated by

$$X_{tk} = \sqrt{\lambda_{tk}} V_{tk}, \tag{2}$$

where $\{V_{tk}|H_0^{tk}\}$ are identically zero, and $\{V_{tk}|H_1^{tk}\}$ are statistically independent complex random variables with zero mean, unit variance, and iid real and imaginary parts:

$$H_1^{tk}:\ E\{V_{tk}\} = 0, \quad E\{|V_{tk}|^2\} = 1,$$
$$H_0^{tk}:\ V_{tk} = 0. \tag{3}$$

(2) The pdf of V_{tk} under H_1^{tk} is determined by the specific statistical model. Let $V_{Rtk} = \Re\{V_{tk}\}$ and $V_{Itk} = \Im\{V_{tk}\}$ denote, respectively, the real and imaginary parts of V_{tk}. Let $p(V_{\rho tk}|H_1^{tk})$ denote the pdf of $V_{\rho tk}$ ($\rho \in \{R, I\}$) under H_1^{tk}. Then, for a Gaussian model [9]

$$p(V_{\rho tk}|H_1^{tk}) = \frac{1}{\sqrt{\pi}} \exp\left(-V_{\rho tk}^2\right), \tag{4}$$

for a Gamma model

$$p(V_{\rho tk}|H_1^{tk}) = \frac{1}{2\sqrt{\pi}} \left(\frac{3}{2}\right)^{1/4} |V_{\rho tk}|^{-1/2} \times \exp\left(-\sqrt{\frac{3}{2}}|V_{\rho tk}|\right) \tag{5}$$

and for a Laplacian model

$$p\left(V_{\rho tk}|H_1^{tk}\right) = \exp\left(-2|V_{\rho tk}|\right). \tag{6}$$

(3) The conditional variance $\lambda_{tk|t-1}$, referred to as the *one-frame-ahead conditional variance*, is a random process which evolves as a GARCH(1, 1) process:

$$\lambda_{tk|t-1} = \lambda_{\min} + \mu|X_{t-1,k}|^2 + \delta(\lambda_{t-1,k|t-2} - \lambda_{\min}), \tag{7}$$

where

$$\lambda_{\min} > 0, \quad \mu \geqslant 0, \quad \delta \geqslant 0, \quad \mu + \delta < 1 \tag{8}$$

are the standard constraints imposed on the parameters of the GARCH model [16]. The parameters μ and δ are, respectively, the moving average and autoregressive parameters of the GARCH(1,1) model, and $\lambda_{\min}$ is a lower bound on the variance of X_{tk} under H_1^{tk}.

(4) The noise spectral coefficients $\{D_{tk}\}$ are zero-mean statistically independent Gaussian random variables. The real and imaginary parts of D_{tk} are iid random variables $\sim \mathcal{N}\left(0, \frac{\sigma_{tk}^2}{2}\right)$.

The first assumption implies that the speech spectral coefficients $\{X_{tk}|H_1^{tk}\}$ are conditionally zero-mean statistically independent random variables given their variances $\{\lambda_{tk}\}$. The real and imaginary parts of X_t under H_1^t are conditionally iid random variables given λ_{tk}, satisfying

$$p\left(X_{\rho tk}|\lambda_{tk}, H_1^{tk}\right) = \frac{1}{\sqrt{\lambda_{tk}}} p\left(V_{\rho tk} = \frac{X_{\rho tk}}{\sqrt{\lambda_{tk}}}\middle| H_1^{tk}\right), \quad \rho \in \{R, I\}. \tag{9}$$

3. Problem formulation

The problem of spectral enhancement is generally formulated as deriving an estimator $\hat{X}_{tk}$ for the speech spectral coefficients, such that the expected value of a certain distortion measure is minimized. Let $d(X_{tk}, \hat{X}_{tk})$ denote a distortion measure between X_{tk} and its estimate $\hat{X}_{tk}$, and let ψ_t represents the information set that can be employed for the estimation at frame t (e.g., the noisy data observed through time t). Let $\hat{p}_{tk} = P(H_1^{tk}|\psi_t)$ denote an estimate for the signal presence probability, let $\hat{\lambda}_{tk} = E\{|X_{tk}|^2|H_1^{tk}, \psi_t\}$ denote an estimate for the variance of a speech spectral coefficient X_{tk} under H_1^{tk}. Then, we consider an estimator for X_{tk} which minimizes the expected distortion given $\hat{p}_{tk}$, $\hat{\lambda}_{tk}$ and the noisy spectral coefficient Y_{tk}:

$$\min_{\hat{X}_{tk}} E\left\{d(X_{tk}, \hat{X}_{tk})|\hat{p}_{tk}, \hat{\lambda}_{tk}, Y_{tk}\right\}. \tag{10}$$

In particular, restricting ourselves to a squared error distortion measure of the form

$$d\left(X_{tk}, \hat{X}_{tk}\right) = \left|g(\hat{X}_{tk}) - \tilde{g}(X_{tk})\right|^2, \tag{11}$$

where $g(X)$ and $\tilde{g}(X)$ are specific functions of X (e.g., $X|X|\log|X|$, $e^{j\angle X}$), the estimator $\hat{X}_{tk}$ is calculated from

$$\begin{aligned} g\left(\hat{X}_{tk}\right) &= E\left\{\tilde{g}(X_{tk})|\hat{p}_{tk}, \hat{\lambda}_{tk}, Y_{tk}\right\} \\ &= \hat{p}_{tk} E\left\{\tilde{g}(X_{tk})|H_1^{tk}, \hat{\lambda}_{tk}, Y_{tk}\right\} \\ &\quad + (1 - \hat{p}_{tk}) E\left\{\tilde{g}(X_{tk})|H_0^{tk}, Y_{tk}\right\}. \end{aligned} \tag{12}$$

The design of a particular estimator for X_{tk} requires the following specifications:

- Functions $g(X)$ and $\tilde{g}(X)$, which determine the fidelity criterion of the estimator.
- A conditional pdf $p(X_{tk}|\lambda_{tk}, H_1^{tk})$ for X_{tk} under H_1^{tk} given its variance λ_{tk}, which determines the statistical model.
- Estimators $\hat{\lambda}_{tk}$ and $\widehat{\sigma_{tk}^2}$ for the speech and noise spectral variances, respectively.
- An estimator $\hat{q}_{tk} = P(H_1^{tk}|\psi_t')$ for the a priori signal presence probability, where $\psi_t' = \psi_t \backslash Y_{tk}$ represents the information set known prior to having the measurement Y_{tk}.

In this work we consider MMSE estimators for the spectral coefficients and the LSA under Gaussian, Gamma and Laplacian models, while the speech spectral variance is estimated based on either the proposed GARCH models or the decision-directed method of Ephraim and Malah [2].

Generally, given an estimate for the a priori signal presence probability, the (a posteriori) signal presence probability can be obtained from Bayes' rule:

$$\hat{p}_{tk} = \frac{\hat{q}_{tk} P(Y_{tk}|H_1^{tk}, \psi_t')}{\hat{q}_{tk} P(Y_{tk}|H_1^{tk}, \psi_t') + (1 - \hat{q}_{tk}) P(Y_{tk}|H_0^{tk}, \psi_t')}. \tag{13}$$

However, to simplify the comparisons between the speech enhancement algorithms, we focus on implementations that assume speech presence (i.e., $\hat{p}_{tk} = 1$) whenever $20\log_{10}|X_{tk}| > \varepsilon$, where $\varepsilon = \max_{tk}\{20\log_{10}|X_{tk}|\} - 50$ confines the dynamic

range of the log-spectrum to 50 dB. In the other time-frequency bins, $\hat{p}_{tk}$ is set to zero. Furthermore, we assume knowledge of the noise variance σ_{tk}^2, which in practice can be estimated by using the *Minima Controlled Recursive Averaging* approach [6,17]. Our objectives in this work are as follows:

- Develop speech enhancement algorithms which are based on supergaussian GARCH models.
- Evaluate estimators for the speech variance which are based on GARCH models, with a comparison to variance estimation by the decision-directed method.
- Compare the performances of MMSE spectral and LSA estimators under Gaussian, Gamma and Laplacian models, while estimating the speech spectral variance by using the GARCH modeling or the decision-directed method.

4. Signal reconstruction

In this section, we assume that the model parameters μ, δ and $\lambda_{\min}$ are known, and derive recursive estimators for the speech variance under Gaussian, Gamma and Laplacian models. The speech is subsequently reconstructed by using MMSE spectral or LSA estimators.

4.1. Variance estimation

The speech variance estimation approach is closely related to the variance estimation approach introduced in [18,19]. We start with an estimate $\hat{\lambda}_{tk|t-1}$ that relies on the noisy observations up to frame $t-1$, and "update" the variance by using the additional information Y_{tk}. Then, the variance is "propagated" ahead in time, following the rational of Kalman filtering, to obtain a conditional variance estimate at frame $t+1$ from the information available at frame t. However, rather than using a heuristic propagation step, we propose propagation steps that are consistent with the supergaussian GARCH models.

Assuming an estimate $\hat{\lambda}_{tk|t-1}$ for the one-frame-ahead conditional variance of X_{tk} is available, an estimate for $\lambda_{tk|t}$ can be obtained by calculating its conditional mean under H_1^{tk} given Y_{tk} and $\hat{\lambda}_{tk|t-1}$. By definition, $\lambda_{tk|t} = |X_{tk}|^2$. Hence,

$$\begin{aligned}\hat{\lambda}_{tk|t} &= E\left\{|X_{tk}|^2 | H_1^{tk}, \hat{\lambda}_{tk|t-1}, Y_{tk}\right\} \\ &= E\left\{X_{Rtk}^2 | H_1^{tk}, \hat{\lambda}_{tk|t-1}, Y_{Rtk}\right\} \\ &\quad + E\left\{X_{Itk}^2 | H_1^{tk}, \hat{\lambda}_{tk|t-1}, Y_{Itk}\right\},\end{aligned} \tag{14}$$

where we have used that $X_{\rho tk}$ and $Y_{\rho' tk}$ are independent for $\rho \neq \rho'$ $(\rho, \rho' \in \{R, I\})$. Defining the a priori and a posteriori SNRs, respectively, by

$$\xi_{tk|t-1} \triangleq \frac{\lambda_{tk|t-1}}{\sigma_{tk}^2}, \quad \gamma_{\rho tk} \triangleq \frac{Y_{\rho tk}^2}{\sigma_{tk}^2}, \tag{15}$$

we can write for $Y_{\rho tk} \neq 0$ $(\rho \in \{R, I\})$

$$E\{X_{\rho tk}^2 | H_1^{tk}, \hat{\lambda}_{tk|t-1}, Y_{\rho tk}\} = G_{\mathrm{SP}}(\hat{\xi}_{tk|t-1}, \gamma_{\rho tk}) Y_{\rho tk}^2, \tag{16}$$

where $G_{\mathrm{SP}}(\xi, \gamma_\rho)$ represents the MMSE gain function in the spectral power domain [19]. The specific expression for $G_{\mathrm{SP}}(\xi, \gamma_\rho)$ depends on the particular statistical model. For a Gaussian model, the spectral power gain function is given by

$$G_{\mathrm{SP}}(\xi, \gamma_\rho) = \frac{\xi}{1+\xi}\left(\frac{1}{2\gamma_\rho} + \frac{\xi}{1+\xi}\right). \tag{17}$$

For a Gamma model [19,20],

$$\begin{aligned}G_{\mathrm{SP}}(\xi, \gamma_\rho) &= \frac{3}{(C_{\rho+} - C_{\rho-})^2} \\ &\times \frac{\exp(C_{\rho-}^2/4) D_{-2.5}(C_{\rho-}) + \exp(C_{\rho+}^2/4) D_{-2.5}(C_{\rho+})}{\exp(C_{\rho-}^2/4) D_{-0.5}(C_{\rho-}) + \exp(C_{\rho+}^2/4) D_{-0.5}(C_{\rho+})},\end{aligned} \tag{18}$$

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

$$C_{\rho\pm} \triangleq \frac{\sqrt{3}}{2\sqrt{\xi}} \pm \sqrt{2\gamma_\rho} \tag{19}$$

and $D_p(z)$ denotes the parabolic cylinder function [21, Eq. 9.240]. For a Laplacian model [19],

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

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

$$L_{\rho\pm} \triangleq \frac{1}{\sqrt{\xi}} \pm \sqrt{\gamma_\rho}, \tag{21}$$

and erfcx(x) is the scaled complementary error function, defined by

$$\mathrm{erfcx}(x) \triangleq \mathrm{e}^{x^2} \frac{2}{\sqrt{\pi}} \int_x^\infty \mathrm{e}^{-t^2}\, \mathrm{d}t. \tag{22}$$

Eq. (16) does not hold in the case $Y_{\rho tk} \to 0$, since $G_{\mathrm{SP}}(\xi, \gamma_\rho) \to \infty$ as $\gamma_\rho \to 0$, and the conditional variance of $X_{\rho tk}$ is generally not zero. For $Y_{\rho tk} = 0$ (or practically for $Y_{\rho tk}$ smaller in magnitude than a predetermined threshold) we use the following expressions [19]: For a Gaussian model

$$E\left\{ X_{\rho tk}^2 | H_1^{tk}, \hat{\lambda}_{tk|t-1}, Y_{\rho tk} = 0 \right\} = \frac{\hat{\xi}_{tk|t-1}}{1 + \hat{\xi}_{tk|t-1}} \sigma_{tk}^2, \tag{23}$$

for a Gamma model,

$$E\left\{ X_{\rho tk}^2 | H_1^{tk}, \hat{\lambda}_{tk|t-1}, Y_{\rho tk} = 0 \right\} = \frac{3D_{-2.5}\left(\sqrt{3}/2\sqrt{\hat{\xi}_{tk|t-1}}\right)}{8D_{-0.5}\left(\sqrt{3}/2\sqrt{\hat{\xi}_{tk|t-1}}\right)} \sigma_{tk}^2 \tag{24}$$

and for a Laplacian model,

$$E\left\{ X_{\rho tk}^2 | H_1^{tk}, \hat{\lambda}_{tk|t-1}, Y_{\rho tk} = 0 \right\} = \sqrt{\frac{2}{\pi}} \frac{\exp(1/2\hat{\xi}_{tk|t-1}) D_{-3}\left(\sqrt{2/\hat{\xi}_{tk|t-1}}\right)}{\mathrm{erfcx}\left(1/\sqrt{\hat{\xi}_{tk|t-1}}\right)} \sigma_{tk}^2. \tag{25}$$

From (16) to (25), we can define a function $f(\lambda, \sigma^2, Y_\rho^2)$ such that

$$E\left\{ X_{\rho tk}^2 | H_1^{tk}, \hat{\lambda}_{tk|t-1}, Y_{\rho tk} \right\} = f\left(\hat{\lambda}_{tk|t-1}, \sigma_{tk}^2, Y_{\rho tk}^2\right) \tag{26}$$

for all $Y_{\rho tk}$. Substituting (26) into (14), we obtain an estimate for $\lambda_{tk|t}$ given by

$$\hat{\lambda}_{tk|t} = f\left(\hat{\lambda}_{tk|t-1}, \sigma_{tk}^2, Y_{Rtk}^2\right) + f\left(\hat{\lambda}_{tk|t-1}, \sigma_{tk}^2, Y_{Itk}^2\right). \tag{27}$$

Eq. (27) is the update step of the recursive estimation, since we start with an estimate $\hat{\lambda}_{tk|t-1}$ that relies on the noisy observations up to frame $t-1$, and then update the estimate by using the additional information Y_{tk}.

To formulate the propagation step, we assume that we are given at frame $t-1$ an estimate $\hat{\lambda}_{t-1,k|t-2}$ for the conditional variance of $X_{t-1,k}$, which has been obtained from the noisy measurements up to frame $t-2$. Then a recursive MMSE estimate for $\lambda_{tk|t-1}$ can be obtained by calculating its conditional mean under $H_1^{t-1,k}$ given $\hat{\lambda}_{t-1,k|t-2}$ and $Y_{t-1,k}$:

$$\hat{\lambda}_{tk|t-1} = E\left\{ \lambda_{tk|t-1} | H_1^{t-1,k}, \hat{\lambda}_{t-1,k|t-2}, Y_{t-1,k} \right\}. \tag{28}$$

Substituting (7) into (28), we have

$$\hat{\lambda}_{tk|t-1} = \lambda_{\min} + \mu E\left\{ |X_{t-1,k}|^2 | H_1^{t-1,k}, \hat{\lambda}_{t-1,k|t-2}, Y_{t-1,k} \right\} + \delta(\hat{\lambda}_{t-1,k|t-2} - \lambda_{\min}). \tag{29}$$

Eq. (14) implies that $E\{|X_{t-1,k}|^2 | H_1^{t-1,k}, \hat{\lambda}_{t-1,k|t-2}, Y_{t-1,k}\} = \hat{\lambda}_{t-1,k|t-1}$. Substituting this into (29), we obtain

$$\hat{\lambda}_{tk|t-1} = \lambda_{\min} + \mu\hat{\lambda}_{t-1,k|t-1} + \delta\left(\hat{\lambda}_{t-1,k|t-2} - \lambda_{\min}\right). \tag{30}$$

Eq. (30) is called the propagation step, since the conditional variance estimates are propagated ahead in time to obtain a conditional variance estimate at frame t from the information available at frame $t-1$. The propagation and update steps are iterated as new data arrive, following the rational of Kalman filtering. The algorithm is initialized at the first frame, say $t = 0$, with $\hat{\lambda}_{0,k|-1} = \lambda_{\min}$ for all the frequency bins, $k = 0, \ldots, K-1$. Then, for $t = 0, 1, \ldots$, the estimate $\hat{\lambda}_{tk|t}$ is calculated by using the update (27), and $\hat{\lambda}_{t+1,k|t}$ is subsequently calculated by using the propagation (30).

4.2. MMSE spectral estimation

An MMSE estimator for X_{tk} is obtained by using the functions

$$g(\hat{X}_{tk}) = \hat{X}_{tk}, \quad \tilde{g}(X_{tk}) = \begin{cases} X_{tk} & \text{under } H_1^{tk}, \\ G_{\min} Y_{tk} & \text{under } H_0^{tk}, \end{cases} \tag{31}$$

where $G_{\min} \ll 1$ represents a constant attenuation factor. Substituting (31) into (12), we have

$$\hat{X}_{tk} = \hat{p}_{tk}\left[G_{\mathrm{MSE}}(\hat{\tilde{\xi}}_{tk|t}, \gamma_{Rtk}) Y_{Rtk} + jG_{\mathrm{MSE}}(\hat{\tilde{\xi}}_{tk|t}, \gamma_{Itk}) Y_{Itk}\right] + (1-\hat{p}_{tk}) G_{\min} Y_{tk}, \tag{32}$$

where $\hat{\tilde{\xi}}_{tk|t} = \hat{\lambda}_{tk|t}/\sigma_{tk}^2$ is an estimate for the a priori SNR, $\gamma_{\rho tk} = Y_{\rho tk}^2/\sigma_{tk}^2$ ($\rho \in \{R, I\}$) are the a posteriori SNRs, and $G_{\mathrm{MSE}}(\xi, \gamma)$ represents the MMSE spectral gain function under H_1. The specific expression for $G_{\mathrm{MSE}}(\xi, \gamma_\rho)$ depends on the particular statistical model. For a Gaussian model, the gain function is the Wiener filter given by [22]

$$G_{\mathrm{MSE}}(\xi) = \frac{\xi}{1+\xi}. \tag{33}$$

For a Gamma model, the gain function is given by [9]

$$G_{\mathrm{MSE}}(\xi, \gamma_\rho) = \frac{1}{C_{\rho+} - C_{\rho-}} \times \frac{\exp(C_{\rho-}^2/4) D_{-1.5}(C_{\rho-}) - \exp(C_{\rho+}^2/4) D_{-1.5}(C_{\rho+})}{\exp(C_{\rho-}^2/4) D_{-0.5}(C_{\rho-}) + \exp(C_{\rho+}^2/4) D_{-0.5}(C_{\rho+})}, \tag{34}$$

where $C_{\rho\pm}$ are defined by (19). For a Laplacian speech model, the gain function is given by [10]

$$G_{\mathrm{MSE}}(\xi, \gamma_\rho) = \frac{2}{L_{\rho+} - L_{\rho-}} \times \frac{L_{\rho+}\mathrm{erfcx}(L_{\rho+}) - L_{\rho-}\mathrm{erfcx}(L_{\rho-})}{\mathrm{erfcx}(L_{\rho+}) + \mathrm{erfcx}(L_{\rho-})}, \tag{35}$$

where $L_{\rho\pm}$ are defined by (21). Note that when the signal is surely absent (i.e., when $\hat{p}_{tk} = 0$), the resulting estimator $\hat{X}_{tk}$ reduces to a constant attenuation of Y_{tk} (i.e., $\hat{X}_{tk} = G_{\min} Y_{tk}$). This retains the noise naturalness, and is closely related to the "spectral floor" proposed by Berouti et al. [23].

4.3. MMSE log-spectral amplitude estimation

In speech enhancement applications, estimators which minimize the mean-squared error of the LSA have been found advantageous to MMSE spectral estimators [2,3,24]. An MMSE–LSA estimator is obtained by substituting into (12) the functions

$$g(\hat{X}_{tk}) = \log|\hat{X}_{tk}|, \quad \tilde{g}(X_{tk}) = \begin{cases} \log|X_{tk}| & \text{under } H_1^{tk}, \\ \log(G_{\min}|Y_{tk}|) & \text{under } H_0^{tk}. \end{cases} \tag{36}$$

Assuming a Gaussian model and combining the resulting amplitude estimate with the phase of the noisy spectral coefficient Y_{tk} yields

$$\hat{X}_{tk} = \left[G_{\mathrm{LSA}}(\hat{\tilde{\xi}}_{tk|t}, \gamma_{tk})\right]^{\hat{p}_{tk}} G_{\min}^{1-\hat{p}_{tk}} Y_{tk}, \tag{37}$$

where $\gamma_{tk} = \gamma_{Rtk} + \gamma_{Itk}$ denotes a posteriori SNR,

$$G_{\mathrm{LSA}}(\xi, \gamma) \triangleq \frac{\xi}{1+\xi} \exp\left(\frac{1}{2}\int_{\vartheta}^{\infty} \frac{\mathrm{e}^{-x}}{x}\,\mathrm{d}x\right) \tag{38}$$

represents the LSA gain function under H_1^{tk} which was derived by Ephraim and Malah [3], and ϑ is defined by $\vartheta \triangleq \xi\gamma/1+\xi$. Similar to the MMSE spectral estimator, the MMSE–LSA estimator reduces to a constant attenuation of Y_{tk} when the signal is surely absent (i.e., $\hat{p}_{tk} = 0$ implies $\hat{X}_{tk} = G_{\min} Y_{tk}$). However, for a fixed value of the a priori SNR, the LSA gain is a monotonically decreasing function of γ [3,25]. By contrast, the gain function $G_{\mathrm{MSE}}(\xi, \gamma_\rho)$ for a Gaussian model is independent of the a posteriori SNR, while for Gamma and Laplacian speech models the gain functions are *increasing* functions of the a posteriori SNR [19]. The behavior of $G_{\mathrm{LSA}}(\xi, \gamma)$ is related to the useful mechanism that counters the musical noise phenomenon [25]. 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. As a result, the MMSE–LSA estimator generally produces

lower levels of residual musical noise, when compared with MMSE spectral estimators.

5. Experimental results

In this section, the performances of the MMSE spectral and LSA estimators are evaluated under Gaussian, Gamma and Laplacian models, while the speech variance is estimated by using either the GARCH modeling or the decision-directed method. The evaluation includes two objective quality measures, and informal listening tests. The first quality measure is LSD, in dB, which is defined by

$$\mathrm{LSD} = \left[\frac{1}{|\mathscr{H}_1|}\sum_{tk\in\mathscr{H}_1}(20\log_{10}|X_{tk}| - 20\log_{10}|\hat{X}_{tk}|)^2\right]^{1/2}, \tag{39}$$

where $\mathscr{H}_1 = \{tk|20\log_{10}|X_{tk}| > \varepsilon\}$ denotes the set of time-frequency bins which contain the speech signal, $|\mathscr{H}_1|$ denotes its cardinality, and $\varepsilon = \max_{tk}\{20\log_{10}|X_{tk}|\} - 50$ confines the dynamic range of the log-spectrum to 50 dB. The second quality measure is the PESQ score (ITU-T P.862).

The speech signals used in our evaluation are taken from the TIMIT database [26]. They include 20 different utterances from 20 different speakers, half male and half female. The speech signals are sampled at 16 kHz and degraded by white Gaussian noise with SNRs in the range [0,20] dB. The noisy signals are transformed into the STFT domain using half overlapping Hamming analysis windows of 32 ms length. The Gaussian, Gamma and Laplacian GARCH models (i.e., the parameters μ, δ and $\lambda_{\min}$) are estimated independently for each speaker from the clean signal of that speaker, as described in the Appendix. Eight different speech enhancement algorithms are then applied to each noisy speech signal, as summarized in Table 1. The presumed statistical model is either Gaussian, Gamma or Laplacian, the speech variance is estimated by using either the GARCH modeling method or the decision-directed method, and the fidelity criterion is either MMSE of the spectral coefficients or MMSE of the LSA. The

Table 1
List of the evaluated speech enhancement algorithms

Algorithm #	Statistical model	Variance estimation	Fidelity criterion
1	Gaussian	GARCH	MMSE
2	Gamma	GARCH	MMSE
3	Laplacian	GARCH	MMSE
4	Gaussian	Decision-directed	MMSE
5	Gamma	Decision-directed	MMSE
6	Laplacian	Decision-directed	MMSE
7	Gaussian	GARCH	MMSE–LSA
8	Gaussian	Decision-Directed	MMSE–LSA

decision-directed estimate of the speech variance is given by [2,25]

$$\hat{\lambda}_{tk}^{\mathrm{DD}} = \max\{\alpha|\hat{X}_{t-1,k}|^2 + (1-\alpha)(|Y_{tk}|^2 - \sigma_{tk}^2), \xi_{\min}\sigma_{tk}^2\}, \tag{40}$$

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 $\xi_{\min}$ is a lower bound on the a priori SNR. These parameters are set to the values $\xi_{\min} = -15$ dB and $\alpha = 0.98$ as specified in [2,3,25]. The noise spectral variance σ_{tk}^2 is estimated by averaging over time the spectral power values of the noise signal itself. Speech presence is determined (i.e., $\hat{p}_{tk} = 1$) whenever $20\log_{10}|X_{tk}| > \varepsilon$. The attenuation factor $G_{\min}$ during speech absence is -20 dB. In practice, the noise signal is unknown, and the noise spectral variance can be estimated by using the *Minima Controlled Recursive Averaging* approach [17]. Furthermore, the speech presence probability is estimated from the noisy spectral measurements [6].

Table 2 shows the results of the LSD obtained by using the different algorithms for various SNR levels.[1] The results of the PESQ scores are presented in Table 3. The results show that:

- The MMSE–LSA estimator yields lower LSD and higher PESQ scores than the MMSE

[1] Note that the LSD results in [14] are slightly different, since a different formulation of the log-spectral distortion is used.

Table 2
Log-spectral distortion obtained by using different variance estimation methods (GARCH modeling method vs. decision-directed method), statistical models (Gaussian vs. supergaussian) and fidelity criteria (MMSE vs. MMSE–LSA)

Variance estimation:		GARCH modeling method				Decision-directed method			
Statistical model:		Gaussian		Gamma	Laplacian	Gaussian		Gamma	Laplacian
Fidelity criterion:		MMSE	MMSE–LSA	MMSE	MMSE	MMSE	MMSE–LSA	MMSE	MMSE
	0	7.77	4.85	8.03	7.91	18.89	11.35	17.76	18.14
Input	5	5.78	4.04	6.93	6.45	17.29	11.03	15.73	16.26
SNR	10	4.14	3.27	5.35	4.85	13.87	9.13	11.83	12.48
[dB]:	15	2.50	2.25	3.23	2.92	9.19	6.05	6.95	7.59
	20	1.30	1.28	1.55	1.44	4.88	3.13	2.88	3.34

Table 3
PESQ scores obtained by using different variance estimation methods (GARCH modeling method vs. decision-directed method), statistical models (Gaussian vs. supergaussian) and fidelity criteria (MMSE vs. MMSE–LSA)

Variance estimation:		GARCH modeling method				Decision-directed method			
Statistical model:		Gaussian		Gamma	Laplacian	Gaussian		Gamma	Laplacian
Fidelity criterion:		MMSE	MMSE–LSA	MMSE	MMSE	MMSE	MMSE–LSA	MMSE	MMSE
	0	2.52	2.55	2.47	2.48	1.91	2.21	1.98	1.96
Input	5	2.97	2.98	2.90	2.91	2.30	2.61	2.38	2.36
SNR	10	3.37	3.38	3.28	3.31	2.70	2.99	2.77	2.75
[dB]:	15	3.67	3.69	3.59	3.62	3.09	3.31	3.17	3.15
	20	3.88	3.89	3.83	3.85	3.53	3.64	3.62	3.60

spectral estimators, whether the variance is estimated by using the GARCH modeling method or the decision-directed method.

- An MMSE spectral estimator derived under a Gamma statistical model performs better than MMSE spectral estimators derived under Gaussian or Laplacian models, but only if the speech variance is estimated by the decision-directed method. However, if the speech variance is estimated by using the GARCH modeling method, a Gaussian model is preferable to Gamma and Laplacian models.
- Speech variance estimation based on GARCH modeling yields lower LSD and higher PESQ scores than those obtained by using the decision-directed method, whether the statistical model is assumed Gaussian, Gamma or Laplacian, and whether the fidelity criterion is MMSE of the spectral coefficients or MMSE–LSA.
- The best performance in terms of minimum LSD and maximum PESQ scores is obtained when using the GARCH modeling method, a Gaussian model and an MMSE–LSA estimator. The worst performance is obtained when using the decision-directed method, a Gaussian model and an MMSE spectral estimator.

It is worthwhile noting that it is difficult, or even impossible, to derive analytical expressions for MMSE-LSA estimators under Gamma or Laplacian models. The GARCH modeling method facilitates MMSE-LSA estimation, while taking into consideration the heavy-tailed distribution.

6. Conclusion

We have introduced speech enhancement algorithms which are based on supergaussian GARCH

models in the STFT domain. We assumed that the conditional variances of the STFT expansion coefficients are random variables, and that the *one-frame-ahead* conditional variance evolves as a GARCH(1, 1) process. The variance of an expansion coefficient is recursively estimated by iterating propagation and update steps following the rational of Kalman filtering. We compared our variance estimation approach to the decision-directed method of Ephraim and Malah by evaluating the performances of MMSE spectral estimators under Gaussian, Gamma and Laplacian models, and MMSE–LSA estimator under a Gaussian model. We showed that the MMSE–LSA estimator yields lower log-spectral distortion and higher PESQ scores than the MMSE spectral estimators, whether the variance is estimated by using the GARCH modeling method or the decision-directed method. Furthermore, a Gamma model is preferable when using the decision-directed method, but a Gaussian model is preferable when using the GARCH modeling method. This is particularly important since it is difficult or even impossible to find analytical expressions for MMSE–LSA estimators under Gamma or Laplacian models. While the decision-directed method necessitates the derivation of the MMSE–LSA estimator under a Gamma model, the GARCH modeling method enables to retain the MMSE–LSA estimator derived under a Gaussian model.

It should be noted that the experimental results in this work are obtained under the assumption that signal presence is perfectly detected, that the noise spectral variance is known, and that the clean speech is available for the estimation of the model parameters. In practice, under signal presence uncertainty, the quality of the enhanced speech may be lower due to miss-detection of speech components ($\hat{p}_{tk}<1$ under H_1^{tk}), and some residual musical noise may be generated due to false-detection of speech components ($\hat{p}_{tk}>0$ under H_0^{tk}). In addition, estimating the model parameters from the noisy signal would degrade the performance due to model mismatch. Nevertheless, the experimental results show the potential of the proposed approach, and motivate a further research on the estimation of the signal presence probability and the model itself.

Acknowledgements

The author thanks Prof. David Malah and Prof. Yariv Ephraim for valuable discussions and helpful suggestions. He also thanks the anonymous reviewers for their helpful comments.

Appendix

For completeness, we briefly repeat the model estimation method employed in [14], with further consideration of the supergaussian GARCH models. Let $\mathscr{X}_0^{\mathrm{T}}$ denote the set of clean speech spectral coefficients employed for the model estimation, let $\mathscr{H}_1=\{tk|s_{tk}=1\}$ denote the set of time-frequency bins where the signal is present, and let $\boldsymbol{\phi}=[\mu,\delta,\lambda_{\min}]$ denote the vector of unknown parameters. Then, for a Gaussian model, the logarithm of the conditional density of X_{tk} given the clean spectral coefficients up to frame $(t-1)$ can be expressed as [14]

$$\log p\left(X_{tk}|\mathscr{X}_0^{t-1};\boldsymbol{\phi}\right) = -\frac{|X_{tk}|^2}{\lambda_{tk|t-1}}-\log\lambda_{tk|t-1}-\log\pi,\quad tk\in\mathscr{H}_1. \tag{41}$$

For a Gamma model we have

$$\begin{aligned}\log p(X_{tk}|\mathscr{X}_0^{t-1};\boldsymbol{\phi}) &= -\sqrt{\frac{3}{2\lambda_{tk|t-1}}}(|X_{Rtk}|+|X_{Itk}|)\\ &\quad-\frac{1}{2}\log|X_{Rtk}\,X_{Itk}|-\frac{1}{2}\log\lambda_{tk|t-1}\\ &\quad-\frac{1}{2}\log\frac{3}{32\pi^2}\end{aligned} \tag{42}$$

and for a Laplacian model we obtain

$$\log p(X_{tk}|\mathscr{X}_0^{t-1};\boldsymbol{\phi}) = -\frac{2}{\sqrt{\lambda_{tk|t-1}}}(|X_{Rtk}|+|X_{Itk}|)-\log\lambda_{tk|t-1}. \tag{43}$$

For sufficiently large sample size, the spectral coefficients of the first frame make a negligible contribution to the total likelihood. Therefore, the values of $\lambda_{0,k|-1}$ in the first frame are initialized to their minimal value $\lambda_{\min}$, and the log-likelihood is

maximized when conditioned on the first frame. The log-likelihood conditional on the spectral coefficients of the first frame is given by

$$\mathscr{L}(\boldsymbol{\phi}) = \sum_{tk\in\mathscr{H}_1\cap t\in[1,T]} \log p\left(X_{tk}|H_1^{tk}, \mathscr{X}_0^{t-1}; \boldsymbol{\phi}\right). \quad (44)$$

Substituting either (41), (42) or (43) into (44) and imposing the constraints in (8) on the estimated parameters, the maximum-likelihood estimates of the model parameters are obtained by solving the following constrained minimization problems: For a Gaussian model [14]

$$\underset{\hat{\lambda}_{\min}\hat{\mu},\hat{\delta}}{\text{minimize}} \sum_{tk\in\mathscr{H}_1\cap t\in[1,T]} \left[\frac{|X_{tk}|^2}{\lambda_{tk|t-1}} + \log\,\lambda_{tk|t-1}\right], \quad (45)$$

for a Gamma model

$$\underset{\hat{\lambda}_{\min},\,\hat{\mu},\hat{\delta}}{\text{minimize}} \sum_{tk\in\mathscr{H}_1\cap t\in[1,T]} \times\left[\sqrt{\frac{3}{2\lambda_{tk|t-1}}}(|X_{Rtk}| + |X_{Itk}|) + \frac{1}{2}\log|X_{Rtk}X_{Itk}| + \frac{1}{2}\log\lambda_{tk|t-1}\right], \quad (46)$$

and for a Laplacian model

$$\underset{\hat{\lambda}_{\min},\hat{\mu},\hat{\delta}}{\text{minimize}} \sum_{tk\in\mathscr{H}_1\cap t\in[1,T]} \times\left[\frac{2}{\sqrt{\lambda_{tk|t-1}}}(|X_{Rtk}| + |X_{Itk}|) + \log\,\lambda_{tk|t-1}\right], \quad (47)$$

where the above minimizations are subject to the constraints

$$\hat{\lambda}_{\min} > 0, \quad \hat{\mu} \geqslant 0, \quad \hat{\delta} \geqslant 0, \quad \hat{\mu} + \hat{\delta} < 1. \quad (48)$$

For given numerical values of the parameters, the sequences of conditional variances $\{\lambda_{tk|t-1}\}$ are calculated from (7) and used to evaluate the series in (45), (46) or (47). The result is then minimized numerically by using the Berndt et al. [27] algorithm as in Bollerslev [28].

References

[1] Y. Ephraim, I. Cohen, Recent advancements in speech enhancement, in: The Electrical Engineering Handbook, third ed., CRC Press, to be published.

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

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

[4] A.J. Accardi, R.V. Cox, A modular approach to speech enhancement with an application to speech coding, in: Proceedings of the 24th IEEE International Conference on Acoustics Speech Signal Processing, ICASSP-99, Phoenix, Arizona, 15–19 March 1999, pp. 201–204.

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

[6] I. Cohen, B. Berdugo, Speech enhancement for non-stationary noise environments, Signal Processing 81 (11) (November 2001) 2403–2418.

[7] T. Lotter, C. Benien, P. Vary, Multichannel speech enhancement using bayesian spectral amplitude estimation, in: Proceedings of the 28th IEEE International Conference on Acoustics Speech Signal Processing, ICASSP-03, Hong Kong, 6–10 April 2003, pp. I832–I835.

[8] P.J. Wolfe, S.J. Godsill, Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement, (special issue) EURASIP JASP on Digital Audio for Multimedia Communications 2003 (10) (September 2003) 1043–1051.

[9] R. Martin, Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors, in: Proceedings of the 27th IEEE International Conference Acoustics Speech Signal Processing, ICASSP-02, Orlando, Florida, 13–17 May 2002, pp. I-253–I-256.

[10] R. Martin, C. Breithaupt, Speech enhancement in the DFT domain using Laplacian speech priors, in: Proceedings of the 8th International Workshop on Acoustic Echo and Noise Control (IWAENC), Kyoto, Japan, 8–11 September 2003, pp. 87–90.

[11] B.H. Juang, L.R. Rabiner, Mixture autoregressive hidden Markov models for speech signals, IEEE Trans. Acoust., Speech Signal Process. ASSP-33 (6) (December 1985) 1404–1413.

[12] Y. Ephraim, N. Merhav, Hidden Markov processes, IEEE Trans. Inf. Theory 48 (6) (June 2002) 1518–1568.

[13] H. Sameti, H. Sheikhzadeh, L. Deng, R.L. Brennan, HMM-based strategies for enhancement of speech signals embedded in nonstationary noise, IEEE Trans. Speech Audio Process. 6 (5) (September 1998) 445–455.

[14] I. Cohen, Modeling speech signals in the time-frequency domain using GARCH, Signal Processing 84 (12) (December 2004) 2453–2459.

[15] R.F. Engle, (Ed.), ARCH Selected Readings, Oxford University Press Inc., New York, 1995.

[16] T. Bollerslev, R.Y. ChouKenneth, F. Kroner, ARCH modeling in finance: a review of the theory and empirical evidence, J. Econometrics 52 (1–2) (April–May 1992) 5–59.

[17] I. Cohen, Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging, IEEE Trans. Speech Audio Process. 11 (5) (September 2003) 466–475.
[18] I. Cohen, Relaxed statistical model for speech enhancement and a priori SNR estimation, IEEE Trans. Speech Audio Process., to appear.
[19] I. Cohen, Speech enhancement using supergaussian speech models and noncausal a priori SNR estimation, Speech Communication, to appear.
[20] C. Breithaupt, R. Martin, MMSE estimation of magnitude-squared DFT coefficients with supergaussian priors, in: Proceedings of the 28th IEEE International Conference on Acoustics Speech Signal Processing, ICASSP-03, Hong Kong, 6–10 April 2003, pp. I_896–I_899.
[21] I.S. Gradshteyn, I.M. Ryzhik, Table of Integrals, Series, and Products, fourth ed., Academic Press, 1980.
[22] J.S. Lim, A.V. Oppenheim, Enhancement and bandwidth compression of noisy speech, Proc. IEEE 67 (12) (December 1979) 1586–1604.
[23] M. Berouti, R. Schwartz, J. Makhoul, Enhancement of speech corrupted by acoustic noise, in: Proceedings of the fourth IEEE International Conference on Acoustics Speech Signal Processing, ICASSP-79, Washington, DC, 9–11 April 1979, pp. 208–211.
[24] J. Porter, S. Boll, Optimal estimators for spectral restoration of noisy speech, in: Proceedings of the IEEE International Conference Acoustics Speech, Signal Processing (ICASSP), San Diego, California, 19–21 March 1984, pp. 18A.2.1–18A.2.4.
[25] O. Cappé, Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor, IEEE Trans. Acoust., Speech Signal Process. 2 (2) (April 1994) 345–349.
[26] J.S. Garofolo, Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database, Technical report, National Institute of Standards and Technology (NIST), Gaithersburg, Maryland (prototype as of December 1988).
[27] E.K. Berndt, B.H. Hall, R.E. Hall, J.A. Hausman, Estimation and inference in nonlinear structural models, Ann. Economic Social Measurement 4 (1974) 653–665.
[28] T. Bollerslev, Generalized autoregressive conditional heteroskedasticity, J. Econometrics 31 (3) (April 1986) 307–327.