CCIT Report #478 April 2004

Speech Spectral Modeling and Enhancement Based on Autoregressive Conditional Heteroscedasticity Model

Israel Cohen

Abstract

In this paper, we introduce a novel approach for statistically modeling speech signals in the short-time Fourier transform (STFT) domain. The proposed model is based on autoregressive conditional heteroscedasticity (ARCH) modeling, which is widely-used for modeling the volatility of financial time-series such as exchange rates and stock returns. Generalized ARCH models account for excess kurtosis (*i.e.*, heavy-tailed distribution) and volatility clustering, two important characteristics of financial time-series. Speech signals in the STFT domain exhibit both "volatility clustering" and heavy tail behavior, and thus are well suited for such modeling. We define the conditional "volatility" of the STFT expansion coefficients, and propose to model the one-frame-ahead conditional variance of the expansion coefficients as a generalized ARCH process. Taking into account speech presence uncertainty, we derive recursive estimators for the variances and magnitudes of the STFT expansion coefficients. Experimental results show that the proposed model and speech enhancement algorithm yield a higher segmental signal-to-noise ratio, lower log-spectral distortion, and better Perceptual Evaluation of Speech Quality scores (PESQ, ITU-T P.862) than those obtained by using the Gaussian statistical model and the decision-directed estimation approach of Ephraim and Malah.

I. 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. Ephraim and Malah [1] proposed to model the individual STFT expansion coefficients of the speech signal as zero-mean

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

statistically independent Gaussian random variables. This model is motivated by the central limit theorem, as each expansion coefficient is a weighted sum of random variables resulting from the random sequence of speech samples. It facilitates a mathematically tractable design of useful speech enhancement algorithms in the STFT domain, *e.g.* [1]–[7]. However, the necessary conditions for the central limit theorem, *e.g.* that a particular few of the member random variables does not dominate the sum [8], are not satisfied for speech signals. Furthermore, the span of correlation within speech signals is often larger than the typical sizes of short-term frames used in speech enhancement applications [9]. Consequently, the Gaussian approximation can be very inaccurate in the tail regions of the probability density function [9]–[12].

Martin [9] proposed a Gamma speech model, in which the real and imaginary parts of the STFT expansion coefficients are modeled as independent and identically distributed (IID) Gamma random variables. He assumed that distinct expansion coefficients are statistically independent, and derived minimum mean-squared error (MMSE) estimators for the speech expansion coefficients under either Gaussian or 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. Alternatively, the real and imaginary parts of the speech STFT expansion coefficients are modeled as IID Laplacian random variables, and distinct expansion coefficients are likewise assumed statistically independent [9], [13]. Martin and Breithaupt [14] showed that MMSE estimators for the speech expansion coefficients derived under Laplacian modeling have similar properties to those estimators derived under Gamma modeling, but are easier to compute and implement.

The above-mentioned statistical models consider the variances of the speech STFT expansion coefficients as the model parameters, which have to be estimated from the noisy observed signal. Ephraim and Malah [1], [15] proposed three different methods for the estimation of the speech spectral variances. The first method is maximum-likelihood (ML) estimation, assuming that the variances are slowly time-varying parameters. This method results in musical residual noise, which is annoying and disturbing to the perception of the enhanced signal. The second method is "decision-directed" estimation, which is particularly useful when combined with the MMSE spectral, or log-spectral, magnitude estimators [1], [2], [16]. It results in colorless residual noise, but is heuristically motivated and its theoretical performance is unknown due to its

highly nonlinear nature. The third method [15] is maximum a-posteriori (MAP) estimation, assuming a specific heuristic first-order Markov model for generating a sequence of speech spectral variances. It involves a set of nonlinear equations, which are solved recursively by using the Viterbi algorithm. The computational complexity of the MAP estimator is relatively high, while it does not provide a significant improvement in the enhanced speech quality over the decision-directed estimator [15]. Therefore, the decision-directed approach has become the most acceptable estimation method for the variances of the speech STFT expansion coefficients.

Unfortunately, the decision-directed estimation approach heavily relies on the strong time-correlation between successive speech STFT expansion coefficients, whereas the underlying assumption in the above-mentioned models is that distinct expansion coefficients are statistically independent. Ephraim and Malah concluded their seminal paper [1] by stating that the full potential of their approach is not yet exploited, and better results may be obtained if the estimation of the speech spectral variances could be improved. They recognized the limit of their model, and conjectured that removing the statistical independence assumption may improve the speech enhancement results. Twenty years later, there still has not been found a statistical model for speech signals in the STFT domain, which allows reliable and efficient estimation of the variances and magnitudes of the expansion coefficients in noisy environments.

Recently [17] we proposed to relax the statistical model of Ephraim and Malah by considering *conditional* independence of the STFT expansion coefficients given their variances, where the sequence of variances at a given frequency is described as a random sequence. In this paper, pursuing this approach, we propose a novel statistical model for speech signals in the STFT domain, which is based on autoregressive conditional heteroscedasticity (ARCH) modeling. ARCH models, introduced by Engle [18] and generalized by Bollerslev [19], are widely-used for volatility modeling of financial time-series such as exchange rates and stock returns. They are successfully utilized in various financial applications such as risk management, option pricing, foreign exchange, and the term structure of interest rates [20]. The changes in volatility are important for understanding financial markets, since higher volatility is associated with a greater risk and investors require higher expected returns as compensation for holding riskier assets. Generalized autoregressive conditional heteroscedasticity (GARCH) models [19] explicitly parameterize the time-varying volatility in terms of past conditional variances and past squared innovations (prediction errors), while taking into account excess kurtosis (*i.e.*, heavier tailed

distribution than Gaussian) and volatility clustering, two important characteristics of financial time-series.

Speech signals, when transformed into the time-frequency domain by using the STFT, demonstrate both "volatility clustering" and heavy tail behavior. Consider a time series of successive expansion coefficients in a fixed frequency bin, then successive magnitudes of the expansion coefficients are highly correlated, whereas successive phases can be assumed uncorrelated [17]. Hence, large magnitudes tend to follow large magnitudes and small magnitudes tend to follow small magnitudes, while the spectral phase is practically unpredictable. Furthermore, expansion coefficients of speech signals do not have a Gaussian distribution, but rather Gamma-like distribution with significant heavy tail behavior [9]–[11]. Therefore, GARCH modeling can be tailored to speech signals in the STFT domain.

Here, we take into account the speech presence uncertainty, and explicitly define the *conditional variance* of the expansion coefficients. We show that the one-frame-ahead conditional variance is a MMSE estimator of the variance given past spectral components. We propose to model the one-frame-ahead conditional variance as a GARCH process, and derive recursive estimators for the variances and magnitudes of the STFT expansion coefficients. The performance of the proposed speech enhancement algorithm is evaluated, and compared to that obtained by using the conventional Gaussian statistical model and the decision-directed estimation approach. Experimental results show that the proposed method yields a higher segmental SNR, lower log-spectral distortion, and better Perceptual Evaluation of Speech Quality scores (PESQ, ITU-T P.862). A subjective study of speech spectrograms and informal listening tests confirm that by using the GARCH modeling method, weak speech components and unvoiced sounds are significantly more emphasized, and the enhanced speech is of higher quality.

The paper is organized as follows. In Section II, we review the autoregressive conditional heteroscedasticity models. In Section III, we formulate a novel approach for statistically modeling speech signals in the STFT domain. In Section IV, we derive recursive estimators for the variances and magnitudes of the STFT expansion coefficients. In Section V, we address the problem of estimating the model parameters. Finally, in Section VI, we demonstrate the improved performance of proposed speech enhancement algorithm, compared to that obtained by using the decision-directed estimation approach.

II. Autoregressive Conditional Heteroscedasticity

Let $\{y_t\}$ denote a real-valued discrete-time stochastic process, and let ψ_t denote an information set available at time t (*e.g.*, $\{y_t\}$ may represent a sequence of observations, and ψ_t may include the observed data through time t). Then, the innovation (prediction error) ε_t at time t in the MMSE sense is obtained by subtracting from y_t its conditional expectation given the information ψ_{t-1},

$$\varepsilon_t = y_t - E\left\{y_t \mid \psi_{t-1}\right\} . \tag{1}$$

The conditional variance (volatility) of y_t given ψ_{t-1} is by definition the conditional expectation of ε_t^2,

$$\begin{aligned} \sigma_t^2 &= \operatorname{var}\left\{y_t \mid \psi_{t-1}\right\} \\ &= E\left\{\varepsilon_t^2 \mid \psi_{t-1}\right\} . \end{aligned} \tag{2}$$

Changes in the conditional variance are quite important for understanding financial markets, since higher volatility is associated with a greater risk and investors require higher expected returns as compensation for holding riskier assets. The ARCH model introduced by Engle [18], and the GARCH model proposed by Bollerslev [19] as a generalization of Engle's model, provide a rich class of possible parametrization of conditional heteroscedasticity (*i.e.*, time-varying volatility). The ARCH and GARCH models explicitly recognize the difference between the unconditional variance $E\left\{[y_t - E\{y_t\}]^2\right\}$ and the conditional variance σ_t^2, allowing the latter to change over time. The fundamental characteristic of these models is that magnitudes of recent innovations provide information about future volatility.

Let $\{z_t\}$ be a zero-mean unit-variance white noise process with some specified probability distribution. Then a GARCH model of order (p, q), denoted by $\varepsilon_t \sim \mathrm{GARCH}(p, q)$, has the following general form

$$\varepsilon_t = \sigma_t \, z_t \tag{3}$$

$$\sigma_t^2 = f\left(\sigma_{t-1}^2, \ldots, \sigma_{t-p}^2, \varepsilon_{t-1}^2 \ldots, \varepsilon_{t-q}^2\right) \tag{4}$$

where σ_t is the conditional standard deviation given by the square root of (4). That is, the conditional variance σ_t^2 is determined by the values of p past conditional variances and q past squared innovations, and the innovation ε_t is generated by scaling a white noise sample with

the conditional standard deviation. The ARCH(q) model, introduced by Engle [18], is a special case of the $\mathrm{GARCH}(p, q)$ model with $p = 0$.

The most widely-used GARCH model specifies a linear function f in (4) as follows,

$$\sigma_t^2 = \kappa + \sum_{i=1}^{q} \alpha_i \, \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \, \sigma_{t-j}^2 \,, \tag{5}$$

where the values of the parameters are constrained by

$$\begin{aligned} &\kappa > 0 \,, \quad \alpha_i \geq 0 \,, \quad \beta_j \geq 0 \,, \qquad i = 1, \ldots, q \,, \quad j = 1, \ldots, p \,, \\ &\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1 \,. \end{aligned}$$

The first three constraints are sufficient to ensure that the conditional variances $\{\sigma_t^2\}$ are strictly positive. The forth constraint is a covariance stationarity constraint, which is necessary and sufficient for the existence of a finite unconditional variance of the innovations process [19]. Mandlebrot [21] observed that many financial time-series such as exchange rates and stock returns exhibit volatility clustering phenomenon, *i.e.* large changes tend to follow large changes of either sign and small changes tend to follow small changes. Equation (5) captures the volatility clustering phenomenon, since large innovations of either sign increase the variance forecasts for several samples. This in return increases the likelihood of large innovations in the succeeding samples, which allows the large innovations to persist. The degree of persistence is determined by the lag lengths p and q, as well as the magnitudes of the coefficients $\{\alpha_i\}$ and $\{\beta_j\}$.

An important attribute of financial time series is excess kurtosis, *i.e.*, the probability distributions exhibit heavier tails than the Gaussian distribution. Bollerslev [19] showed that GARCH models account also for heavy tail behavior of the innovations process. Specifically, he showed that the standard $\mathrm{GARCH}(1, 1)$ process, which is defined by

$$\varepsilon_t \mid \psi_{t-1} \sim N(0, \sigma_t^2) \tag{6}$$

$$\sigma_t^2 = \kappa + \alpha_1 \, \varepsilon_{t-1}^2 + \beta_1 \, \sigma_{t-1}^2 \,, \tag{7}$$

generates data with excess kurtosis. Bollerslev found that a necessary and sufficient condition for the existence of the $2n$th moment $E\{\varepsilon_t^{2n}\}$ of the standard GARCH$(1, 1)$ process is

$$\beta_1^n + \sum_{k=1}^{n} \frac{n! \, (2k-1)!!}{k! \, (n-k)!} \, \alpha_1^k \, \beta_1^{n-k} < 1 \tag{8}$$

where $(2k-1)!! \stackrel{\triangle}{=} 1 \cdot 3 \ldots (2k-1)$. Accordingly, $3\,\alpha_1^2 + 2\,\alpha_1\,\beta_1 + \beta_1^2 < 1$ is necessary and sufficient for the existence of the forth-order moment. Under this constraint, the second and fourth order moments are given by

$$\begin{aligned} E\left\{\varepsilon_t^2\right\} &= \frac{\kappa}{1-\alpha_1-\beta_1}\,. \\ E\left\{\varepsilon_t^4\right\} &= \frac{3\,\kappa^2(1+\alpha_1+\beta_1)}{(1-\alpha_1-\beta_1)(1-\beta_1^2-2\,\alpha_1\,\beta_1-3\,\alpha_1^2)}\,. \end{aligned}$$

The kurtosis "excess" is therefore

$$\frac{E\left\{\varepsilon_t^4\right\}}{\left(E\{\varepsilon_t^2\}\right)^2} - 3 = \frac{6\,\alpha_1^2}{1-\beta_1^2-2\,\alpha_1\,\beta_1-3\,\alpha_1^2} \tag{9}$$

which is greater than zero by imposing the constraint on the existence of the forth-order moment.

Speech signals in the STFT domain demonstrate both heavy-tailed distribution [9], [10], [12] and "volatility clustering". Magnitudes of successive expansion coefficients in the same frequency bin are highly correlated, whereas the corresponding phases can be assumed uncorrelated [17]. Hence, large magnitudes tend to follow large magnitudes and small magnitudes tend to follow small magnitudes, while the spectral phase ("sign" of the innovation) is unpredictable. Therefore, GARCH modeling is well-suited for speech signals in the STFT domain.

III. Spectral Modeling

In this section, we formulate a novel modeling approach for speech signals in the STFT domain, which utilizes the GARCH model. We take into account the speech presence uncertainty, and define the *conditional variance* of spectral components under signal presence hypothesis. For simplicity, the *conditional* distribution of the real and imaginary parts of the spectral components is assumed Gaussian, and the sequence of the conditional variances is modeled as a $\mathrm{GARCH}(1,1)$ process.

Let $x(n)$ and $d(n)$ denote speech and uncorrelated additive noise signals, respectively, where n is a discrete-time index. The observed signal $y(n)$ is transformed into the time-frequency domain by applying the STFT. Specifically,

$$Y_{tk} = \sum_{n=0}^{K-1} y(n+tM)h(n)\; e^{-j\frac{2\pi}{K}\,n\,k} \tag{10}$$

where t is the time frame index ($t = 0, 1, \ldots$), k is the frequency-bin index ($k = 0, 1, \ldots, K-1$), $h(n)$ is an analysis window of size K (*e.g.,* Hamming window), and M is the framing step

(number of samples separating two successive frames). In the time-frequency domain we have $Y_{tk} = X_{tk} + D_{tk}$, where $\{X_{tk}\}$ are the signal components and $\{D_{tk}\}$ are the noise components.

In accordance with the Gaussian statistical model, proposed by Ephraim and Malah [1], we assume that the noise spectral components $\{D_{tk}\}$ are zero-mean statistically independent Gaussian random variables. However, we do not make a similar assumption with regard to the *speech* spectral components $\{X_{tk}\}$, since the latter are highly correlated. Recently [17] we proposed to relax the statistical model of Ephraim and Malah by considering *conditional* independence of the speech spectral components given their variances, where the sequence of variances at a given frequency k is described as a random sequence. Here, pursuing this approach, we propose to model the variance sequence as a random GARCH process.

Let H_0^{tk} and H_1^{tk} denote, respectively, hypotheses of signal absence and presence in the noisy spectral component 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 $\lambda_{tk} \stackrel{\triangle}{=} E\left\{|X_{tk}|^2 \,|\, H_1^{tk}\right\}$ denote the variance of a speech spectral component X_{tk} under H_1^{tk}. We assume that given $\{\lambda_{tk}\}$ and $\{s_{tk}\}$, the speech spectral components $\{X_{tk}\}$ are generated by

$$X_{tk} = \sqrt{\lambda_{tk}}\, V_{tk} \tag{11}$$

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

$$\begin{aligned} H_1^{tk}: &\quad E\left\{V_{tk}\right\} = 0\,,\; E\left\{|V_{tk}|^2\right\} = 1\,, \\ H_0^{tk}: &\quad V_{tk} = 0\,. \end{aligned} \tag{12}$$

Accordingly, the speech spectral components $\left\{X_{tk} \,|\, H_1^{tk}\right\}$ are *conditionally* zero-mean statistically independent Gaussian random variables given their variances $\{\lambda_{tk}\}$. The real and imaginary parts of X_{tk} under H_1^{tk} are *conditionally* IID random variables given λ_{tk}.

Let $\mathcal{X}_0^{\tau} = \{X_{tk} \,|\, t = 0, \ldots, \tau,\; k = 0, \ldots, K-1\}$ represent the set of clean speech spectral components up to frame τ, and let $\lambda_{tk|\tau} \stackrel{\triangle}{=} E\left\{|X_{tk}|^2 \,|\, H_1^{tk}\,,\, \mathcal{X}_0^{\tau}\right\}$ denote the *conditional* variance of X_{tk} under H_1^{tk} given the clean spectral components up to frame τ. Then, for $\tau \geq t$ we clearly have $\lambda_{tk|\tau} = |X_{tk}|^2$. For $\tau = t-1$, we assume that the one-frame-ahead conditional variance $\lambda_{tk|t-1}$ evolves according to a $\mathrm{GARCH}(1,1)$ process:

$$\lambda_{tk|t-1} = \lambda_{\min} + \mu\,|X_{t-1,k}|^2 + \delta\left(\lambda_{t-1,k|t-2} - \lambda_{\min}\right) \tag{13}$$

where

$$\begin{gathered}\lambda_{\min} > 0 \\ \mu \geq 0\,, \quad \delta \geq 0 \\ \mu + \delta < 1\end{gathered} \tag{14}$$

are the standard constraints imposed on the parameters of the GARCH model. 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}. Note that $\lambda_{\min}$ in (13) is related to κ in (7) by $\lambda_{\min} = \kappa/(1-\delta)$, which is strictly positive under the constraints $\kappa > 0$, $\mu \geq 0$, $\delta \geq 0$, $\mu + \delta < 1$. We use (13) rather than (7) for convenience to make the lower bound on the variance an explicit parameter of the model.

The variances of the speech spectral component are generally unknown, and have to be estimated from the available information. If the available information include the set of clean spectral components up to frame $t-1$, then a MMSE estimator for λ_{tk} can be obtained by

$$\hat{\lambda}_{tk} = E\left\{\lambda_{tk} \,\middle|\, H_1^{tk}\,,\, \mathcal{X}_0^{t-1}\right\}\,. \tag{15}$$

From (11), (12) and the definition of the conditional variance $\lambda_{tk|\tau}$ we have

$$\begin{aligned}\lambda_{tk|t-1} &\stackrel{\triangle}{=} E\left\{|X_{tk}|^2 \,\middle|\, H_1^{tk}\,,\, \mathcal{X}_0^{t-1}\right\} = E\left\{\lambda_{tk}\,|V_{tk}|^2 \,\middle|\, H_1^{tk}\,,\, \mathcal{X}_0^{t-1}\right\} \\ &= E\left\{\lambda_{tk} \,\middle|\, H_1^{tk}\,,\, \mathcal{X}_0^{t-1}\right\} E\left\{|V_{tk}|^2 \,\middle|\, H_1^{tk}\right\} = \hat{\lambda}_{tk}\,.\end{aligned} \tag{16}$$

Therefore, given $\mathcal{X}_0^{t-1}$, the conditional variance $\lambda_{tk|t-1}$ is a MMSE estimator for λ_{tk}. In practice, the available information is the set of *noisy* spectral components up to frame t, rather than the *clean* spectral components up to frame $t-1$. Hence, an estimate for λ_{tk}, and ultimately an estimate for X_{tk}, have to be derived from the available noisy data.

IV. Spectral Enhancement

In this section, we assume that the model parameters μ, δ and $\lambda_{\min}$ are known, and derive recursive estimators for the speech spectral variance λ_{tk} and the spectral component X_{tk} given the *noisy* measurements up to frame t. We also assume knowledge of the noise spectrum, which in practice can be estimated by using the *Minima Controlled Recursive Averaging* approach [22].

Let $\mathcal{Y}_0^t = \{Y_{\tau k} \,|\, \tau = 0, \ldots, t,\, k = 0, \ldots, K-1\}$ represent the set of noisy spectral components up to frame t, and let ψ_t denote the information employed for the recursive estimation at frame t. To retain the computational complexity of the implementation manageable,

ψ_t does not include the complete set of spectral measurements up to frame t, but only a few estimated variables from the previous frame $(t-1)$ and the new spectral measurements $\{Y_{tk} \,|\, k = 0, \ldots, K-1\}$. Suppose that the available information at frame t is an estimate $\hat{\lambda}_{tk|t-1}$ for the one-frame-ahead conditional variance of X_{tk}, and the new noisy spectral components $\{Y_{tk} \,|\, k = 0, \ldots, K-1\}$. Then a MMSE estimate for $\lambda_{tk|t}$ can be obtained by calculating its conditional mean under H_1^{tk} given Y_{tk} and $\hat{\lambda}_{tk|t-1}$:

$$\hat{\lambda}_{tk|t} = E\left\{ \lambda_{tk|t} \;\middle|\; H_1^{tk}, \hat{\lambda}_{tk|t-1}, Y_{tk} \right\}. \tag{17}$$

By definition, $\lambda_{tk|t} = |X_{tk}|^2$. Hence

$$\begin{aligned} \hat{\lambda}_{tk|t} &= E\left\{ |X_{tk}|^2 \;\middle|\; H_1^{tk}, \hat{\lambda}_{tk|t-1}, Y_{tk} \right\} \\ &= \mathrm{var}\left\{ X_{tk} \,|\, H_1^{tk}, \hat{\lambda}_{tk|t-1}, Y_{tk} \right\} + \left| E\left\{ X_{tk} \;\middle|\; H_1^{tk}, \hat{\lambda}_{tk|t-1}, Y_{tk} \right\} \right|^2 . \end{aligned} \tag{18}$$

Let $\sigma_{tk}^2 \triangleq E\left\{ |D_{tk}|^2 \right\}$ denote the variance of a noise spectral component D_{tk}. Then, the assumption, that $\left\{ X_{tk} \,|\, H_1^{tk}, \lambda_{tk} \right\}$ and $\{ D_{tk} \,|\, \sigma_{tk}^2 \}$ are statistically independent Gaussian complex variables, implies that the conditional distribution of $X_{tk}|\lambda_{tk}$ under H_1^{tk} given Y_{tk} is Gaussian with mean and variance

$$E\left\{ X_{tk} \;\middle|\; H_1^{tk}, \lambda_{tk}, Y_{tk} \right\} = \frac{\lambda_{tk}}{\lambda_{tk} + \sigma_{tk}^2} Y_{tk} \tag{19}$$

$$\mathrm{var}\left\{ X_{tk} \,|\, H_1^{tk}, \lambda_{tk}, Y_{tk} \right\} = \frac{\lambda_{tk}}{\lambda_{tk} + \sigma_{tk}^2} \sigma_{tk}^2 . \tag{20}$$

Substituting (19) and (20) into (18), we have

$$\hat{\lambda}_{tk|t} = \frac{\hat{\lambda}_{tk|t-1}}{\hat{\lambda}_{tk|t-1} + \sigma_{tk}^2} \left(\sigma_{tk}^2 + \frac{\hat{\lambda}_{tk|t-1} \, |Y_{tk}|^2}{\hat{\lambda}_{tk|t-1} + \sigma_{tk}^2} \right) . \tag{21}$$

We call (21) the "update" step, 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}. This step can be expressed in terms of the *a priori* and *a posteriori* SNRs, which are defined by

$$\xi_{tk|\tau} \triangleq \frac{\lambda_{tk|\tau}}{\sigma_{tk}^2}, \quad \gamma_{tk} \triangleq \frac{|Y_{tk}|^2}{\sigma_{tk}^2} . \tag{22}$$

Dividing both sides of (21) by σ_{tk}^2, we have

$$\hat{\xi}_{tk|t} = \frac{\hat{\xi}_{tk|t-1}}{\hat{\xi}_{tk|t-1} + 1} \left(1 + \frac{\hat{\xi}_{tk|t-1} \, \gamma_{tk}}{\hat{\xi}_{tk|t-1} + 1} \right) . \tag{23}$$

Computation of the update step requires the estimate $\hat{\lambda}_{tk|t-1}$. Suppose 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 $Y_{t-1,k}$:

$$\hat{\lambda}_{tk|t-1} = E\left\{\lambda_{tk|t-1} \;\middle|\; H_1^{t-1,k}, \hat{\lambda}_{t-1,k|t-2}, Y_{t-1,k}\right\}. \tag{24}$$

Substituting (13) into (24), we have

$$\hat{\lambda}_{tk|t-1} = \lambda_{\min} + \mu\, E\left\{|X_{t-1,k}|^2 \;\middle|\; H_1^{t-1,k}, \hat{\lambda}_{t-1,k|t-2}, Y_{t-1,k}\right\} + \delta\left(\hat{\lambda}_{t-1,k|t-2} - \lambda_{\min}\right). \tag{25}$$

Equation (18) implies that $E\left\{|X_{t-1,k}|^2 \;\middle|\; H_1^{t-1,k}, \hat{\lambda}_{t-1,k|t-2}, Y_{t-1,k}\right\} = \hat{\lambda}_{t-1,k|t-1}$. Substituting this into (25), we obtain

$$\hat{\lambda}_{tk|t-1} = \lambda_{\min} + \mu\, \hat{\lambda}_{t-1,k|t-1} + \delta\left(\hat{\lambda}_{t-1,k|t-2} - \lambda_{\min}\right). \tag{26}$$

We call (26) 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, following the rational of Kalman filtering, to recursively predict and update the conditional variance estimates for the speech spectral components as new data arrive. 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 step (21), and $\hat{\lambda}_{t+1,k|t}$ is subsequently calculated by using the propagation step (26).

We are now interested in estimating the speech spectral component X_{tk} from the information ψ_t available at frame t, such that the expected value of a certain distortion measure is minimized:

$$\hat{X}_{tk} = \arg\min_{\hat{X}} E\left\{d\left(X_{tk}, \hat{X}\right) \mid \psi_t\right\}, \tag{27}$$

where $d\left(X_{tk}, \hat{X}_{tk}\right)$ is a given distortion measure between X_{tk} and $\hat{X}_{tk}$. Recall that given the variance λ_{tk} and the state variable s_{tk}, the speech spectral component X_{tk} is statistically independent of $\mathcal{Y}_0^{t-1}$, the information required to be extracted from past measurements for the recursive estimation is the estimates for λ_{tk} and the signal presence probability $P\left(H_1^{tk}\right)$. Let $\hat{p}_{tk} = P\left(H_1^{tk} \mid \mathcal{Y}_0^t\right)$ denote an estimate for the signal presence probability that is recursively calculated by using the noisy spectral measurement up to frame t, *e.g.*, [4], [5], [23]. Then given

$\hat{p}_{tk}$ and employing $\hat{\lambda}_{tk|t}$ as an estimate for the variance of X_{tk}, the estimator $\hat{X}_{tk}$ can be obtained from

$$\min_{\hat{X}_{tk}} E\left\{ d\left(X_{tk}, \hat{X}_{tk}\right) \,\middle|\, \hat{p}_{tk}\,, \hat{\lambda}_{tk|t}\,, Y_{tk} \right\}. \tag{28}$$

This estimation problem was already solved for several distortion measures, which are of interest in speech enhancement applications. In particular, assuming a squared error distortion measure of the form

$$d\left(X_{tk}, \hat{X}_{tk}\right) = \left| f(X_{tk}) - g(\hat{X}_{tk}) \right|^2 \tag{29}$$

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

$$\begin{aligned} g(\hat{X}_{tk}) &= E\left\{ f(X_{tk}) \,\middle|\, \hat{p}_{tk}\,, \hat{\lambda}_{tk|t}\,, Y_{tk} \right\} \\ &= \hat{p}_{tk}\, E\left\{ f(X_{tk}) \,\middle|\, H_1^{tk}\,, \hat{\lambda}_{tk|t}\,, Y_{tk} \right\} + (1-\hat{p}_{tk})\, E\left\{ f(X_{tk}) \,\middle|\, H_0^{tk}\,, Y_{tk} \right\}. \end{aligned} \tag{30}$$

A MMSE estimator for X_{tk} (Wiener filter) is obtained by using $f(X) = g(X) = X$:

$$\hat{X}_{tk} = \hat{p}_{tk}\, \frac{\hat{\xi}_{tk|t}}{1+\hat{\xi}_{tk|t}}\, Y_{tk}\,, \tag{31}$$

where $\hat{\xi}_{tk|t} = \hat{\lambda}_{tk|t} / \sigma_{tk}^2$ is an estimate for the *a priori* SNR. Using $f(X) = g(X) = |X|$ and combing the resulting spectral amplitude estimator with the phase of the noisy spectral component Y_{tk} yields [1]

$$\hat{X}_{tk} = \hat{p}_{tk}\, G_{\mathrm{SA}}\left(\hat{\vartheta}_{tk|t},\, \gamma_{tk}\right)\, Y_{tk}\,, \tag{32}$$

where $\hat{\vartheta}_{tk|t}$ is defined by $\hat{\vartheta}_{tk|t} = \frac{\hat{\xi}_{tk|t}}{1+\hat{\xi}_{tk|t}}\, \gamma_{tk}$, and

$$G_{\mathrm{SA}}\left(\vartheta,\, \gamma\right) = \frac{\sqrt{\pi\,\vartheta}}{2\,\gamma} \left[(1+\vartheta) I_0\left(\frac{\vartheta}{2}\right) + \vartheta\, I_1\left(\frac{\vartheta}{2}\right) \right] \exp\left(-\frac{\vartheta}{2}\right) \tag{33}$$

represents the spectral-amplitude gain function when the signal is surely present [1]. The functions $I_0(\cdot)$ and $I_1(\cdot)$ denote, respectively, the modified Bessel functions of zero and first order.

The optimally-modified log-spectral amplitude (OM-LSA) estimator [5] is obtained by using

$$g(\hat{X}_{tk}) = \log|\hat{X}_{tk}|\,, \quad f(X_{tk}) = \begin{cases} \log|X_{tk}|\,, & \text{under } H_1^{tk}\,, \\ \log\left(G_{\min}|Y_{tk}|\right)\,, & \text{under } H_0^{tk}\,, \end{cases} \tag{34}$$

where $G_{\min} \ll 1$ represents a constant attenuation factor. Substituting (34) into (30) and combing the resulting amplitude estimate with the phase of the noisy spectral component Y_{tk} yields

$$\hat{X}_{tk} = \Big[G_{\mathrm{LSA}}(\hat{\xi}_{tk|t}, \hat{\vartheta}_{tk|t})\Big]^{\hat{p}_{tk}} \; G_{\min}^{1-\hat{p}_{tk}} \, Y_{tk} \tag{35}$$

where

$$G_{\mathrm{LSA}}\,(\xi,\, \vartheta) = \frac{\xi}{1+\xi} \exp\left(\frac{1}{2}\int_{\vartheta}^{\infty} \frac{e^{-x}}{x} dx\right) \tag{36}$$

represents the log-spectral amplitude (LSA) gain function under H_1^{tk} which was derived by Ephraim and Malah [2]. Note that $\hat{X}_{tk}$ in (35) is not zero when the signal is surely absent, but it reduces to Y_{tk} attenuated by a constant factor (*i.e.*, $\hat{X}_{tk} = G_{\min}\, Y_{tk}$ when $\hat{p}_{tk} = 0$). The constant attenuation under H_0^{tk} retains the noise naturalness, and is closely related to the "spectral floor" modification of the spectral subtraction method, as proposed by Berouti, Schwartz and Makhoul [24].

V. Model Estimation

In this section we address the problem of estimating the model parameters μ, δ and $\lambda_{\min}$. The ML estimation approach is commonly used for estimating the parameters of a GARCH model [25]. We derive the ML function of the model parameters, by using the spectral components of the clean speech signal on some interval $t \in [0, T]$. For simplicity, we assume that the parameters are constant during the above interval and are independent of the frequency-bin index k. In practice, the speech signal can be divided into short time segments and split in frequency into narrow subbands, such that the parameters can be assumed to be constant in each time-frequency region. Furthermore, we generally do not have a direct access to the clean spectral components. However, the expectation-maximization (EM) algorithm [26], [27] can be utilized for solving this problem by iteratively estimating the clean spectral components and the model parameters from the noisy measurements.

Let $\mathcal{X}_0^T = \{X_{tk} \,|\, t = 0, \ldots, T,\; k = 0, \ldots, K-1\}$ denote the set of clean speech spectral components employed for the model estimation, let $\mathcal{H}_1 = \{tk \mid X_{tk} \neq 0\}$ denote the set of time-frequency bins in which the signal is present, and let $\boldsymbol{\phi} = \begin{bmatrix}\mu & \delta & \lambda_{\min}\end{bmatrix}$ denote the vector of unknown parameters. Then for $tk \in \mathcal{H}_1$, the conditional distribution of X_{tk} given its variance λ_{tk} is Gaussian:

$$p\,(X_{tk} \,|\, \lambda_{tk}\,) = \frac{1}{\pi\, \lambda_{tk}} \exp\left(-\frac{|X_{tk}|^2}{\lambda_{tk}}\right), \quad tk \in \mathcal{H}_1\,. \tag{37}$$

Furthermore, $\{X_{tk} \mid \lambda_{tk}\,,\, tk \in \mathcal{H}_1\}$ are statistically independent. We showed in (16) that the conditional MMSE estimate of λ_{tk} given the speech spectral components up to frame $t-1$ is $\lambda_{tk|t-1}$. The conditional variance $\lambda_{tk|t-1}$ can recursively be calculated from past spectral components $\mathcal{X}_0^{t-1}$ by using (13) and the parameter vector $\boldsymbol{\phi}$. Hence, the logarithm of the conditional density of X_{tk} given the clean spectral components up to frame $t-1$ can be expressed as

$$\log p\left(X_{tk} \;\middle|\; \mathcal{X}_0^{t-1}\,;\boldsymbol{\phi}\right) = -\frac{|X_{tk}|^2}{\lambda_{tk|t-1}} - \log \lambda_{tk|t-1} - \log \pi\,, \quad tk \in \mathcal{H}_1\,. \tag{38}$$

It is convenient to regard the speech spectral components in the first frame ($t=0$) as deterministic, with the values of $\lambda_{0,k|-1}$ in the first frame initialized to their minimal value $\lambda_{\min}$, and maximize the log-likelihood when conditioned on the first frame (for sufficiently large sample size, the spectral components of the first frame make a negligible contribution to the total likelihood). The log-likelihood conditional on the spectral components of the first frame is given by

$$\mathcal{L}(\boldsymbol{\phi}) = \sum_{tk\in\mathcal{H}_1 \cap t\in[1,T]} \log p\left(X_{tk} \;\middle|\; H_1^{tk}\,, \mathcal{X}_0^{t-1}\,;\boldsymbol{\phi}\right)\,. \tag{39}$$

Substituting (38) into (39) and imposing the constraints in (14) on the estimated parameters, the maximum-likelihood estimates of the model parameters can be obtained by solving the following constrained minimization problem

$$\begin{aligned} &\underset{\hat{\lambda}_{\min},\,\hat{\mu},\,\hat{\delta}}{\text{minimize}} \quad \sum_{tk\in\mathcal{H}_1 \cap t\in[1,T]} \left[\frac{|X_{tk}|^2}{\lambda_{tk|t-1}} + \log \lambda_{tk|t-1}\right] \\ &\text{subject to} \quad \hat{\lambda}_{\min} > 0\,,\; \hat{\mu} \geq 0\,,\; \hat{\delta} \geq 0\,,\; \hat{\mu} + \hat{\delta} < 1\,. \end{aligned} \tag{40}$$

Such problem is generally referred to as constrained nonlinear optimization or nonlinear programming. For given numerical values of the parameters, the sequences of conditional variances $\left\{\lambda_{tk|t-1}\right\}$ can be calculated from (13) and used to evaluate the series in (40). The result can then be minimized numerically by using the Berndt, Hall, Hall and Hausman [28] algorithm as in Bollerslev [19]. Alternatively, the function *fmincon* of the Optimization Toolbox in MATLAB® can be used to find the minimum of the constrained nonlinear function of the model parameters, similar to its use within the function *garchfit* of the GARCH Toolbox. The latter function provides ML estimates for the parameters of a univariate (scalar) one-state GARCH process. It cannot be used directly in the present work, since the spectral components are complex and generated from a two-state model (speech presence and absence states).

VI. Experimental Results

In this section, the performance of the proposed speech enhancement algorithm is evaluated, and compared to that obtained by using the decision-directed *a priori SNR* estimator. The evaluation includes three objective quality measures, and informal listening tests. The first quality measure is the segmental SNR, in dB, defined by [29]

$$SegSNR = \frac{10}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\log_{10}\frac{\sum_{n=0}^{N-1}x^2(n+tN/2)}{\sum_{n=0}^{N-1}\left[x(n+tN/2)-\hat{x}(n+tN/2)\right]^2} \tag{41}$$

where $\mathcal{T}$ represents the set of frames that contain speech, $|\mathcal{T}|$ its cardinality, and $N = 512$ is the number of samples per frame (corresponding to 32 ms half overlapping frames). The second quality measure is the log-spectral distortion (LSD), in dB, which is defined by

$$\mathrm{LSD} = \frac{1}{L}\sum_{t=0}^{L-1}\left\{\frac{1}{N/2+1}\sum_{k=0}^{N/2}\left[\mathcal{C}\left(20\log_{10}|X_{tk}|\right)-\mathcal{C}\left(20\log_{10}|\hat{X}_{tk}|\right)\right]^2\right\}^{\frac{1}{2}} \tag{42}$$

where L denotes the number of frames in the signal, and $\mathcal{C}$ confines the dynamic range of the log-spectrum to 50 dB (that is, $\mathcal{C}(x) = \max\{x, \epsilon\}$, where $\epsilon = \max_{tk}\{20\log_{10}|X_{tk}|\} - 50$). The third quality measure is the Perceptual Evaluation of Speech Quality (PESQ) score (ITU-T P.862).

The speech signals used in our evaluation are taken from the TIMIT database [30]. 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 milliseconds length. The GARCH model (*i.e.*, the parameters μ, δ and $\lambda_{\min}$) is estimated independently for each speaker from the clean signal of that speaker, as described in Section V. The proposed speech enhancement algorithm is then applied to each noisy speech signal using the corresponding model parameters and the OM-LSA estimator in (35) with $G_{\min} = -20$ dB. Alternatively, the *a priori* SNR ξ_{tk} is estimated by the decision-directed method [1]:

$$\hat{\xi}_{tk}^{\mathrm{DD}} = \max\left\{\alpha\frac{|\hat{X}_{t-1,k}|^2}{\sigma_{t-1,k}^2} + (1-\alpha)(\gamma_{tk}-1)\,,\,\xi_{\min}\right\}, \tag{43}$$

with the parameters $\xi_{\min} = -15$ dB and $\alpha = 0.98$ (these value were determined in [1], [2], [16] by simulations and informal listening tests). The noise spectral variance σ_{tk}^2 is estimated

by averaging over time the spectral power values of the noise signal itself. In practice, the noise signal is unknown, and the noise spectral variance can be estimated by using the *Minima Controlled Recursive Averaging* approach [22], which is particularly useful in nonstationary noise environments. Speech presence is determined (*i.e.*, $\hat{p}_{tk} = 1$) whenever $20 \log_{10} |X_{tk}| > \max_{tk} \{20 \log_{10} |X_{tk}|\} - 50$; In the other time-frequency bins, $\hat{p}_{tk}$ is set to zero and consequently the OM-LSA estimator reduces to $\hat{X}_{tk} = G_{\min} Y_{tk}$. In practice, the clean spectral components are obviously unknown, and the speech presence probability $p_{tk} = P(H_1^{tk})$ has to be estimated from the noisy spectral measurements [5].

Table I shows the results of the segmental SNR achieved by the proposed and the decision-directed *a priori* SNR estimators. The results of the LSD and the PESQ mean opinion score are presented, respectively, in Tables II and III. The results show that the proposed estimator yields a higher segmental SNR, lower LSD, and higher PESQ scores than the decision-directed estimator under all tested environmental conditions. 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 much better than that obtained by using the decision-directed method. In particular, weak speech components and unvoiced sounds are better preserved. Figure 1 demonstrates the spectrograms and waveforms of the clean signal, noisy signal (SNR $= 5$ dB) and the enhanced speech signals obtained by using the two methods. It shows that weak speech components and unvoiced sounds are significantly more emphasized in the signal enhanced by the proposed method than in the signal enhanced by using the decision-directed estimator.

VII. Conclusion

We have proposed a novel approach for statistically modeling speech signals in the STFT domain, and enhancing speech degraded by uncorrelated additive noise. Our approach builds on advances in stochastic financial models of volatility and conditional heteroscedasticity. It provides an explicit model for the conditional variance and conditional distribution of the expansion coefficients. It takes into account the correlation between successive expansion coefficients, heavy tails of the probability distributions, and persistence in variability. The correlation between successive expansion coefficients is considered by parameterizing the conditional variances in terms of past conditional variances and past power values of the expansion coefficients. Excess

kurtosis and persistence in variability are natural outcomes of modeling the one-frame-ahead conditional variance as a GARCH process. These aspects conform to the observations that the STFT expansion coefficients of speech signals have probability distributions with heavier tails than a Gaussian distribution [9]–[11], and that variability of expansion coefficients persists in the sense that large magnitudes tend to follow large magnitudes and small magnitudes tend to follow small magnitudes while the phase is unpredictable.

We assumed that the one-frame-ahead conditional variance evolves as a standard $\mathrm{GARCH}(1,1)$ process, with Gaussian conditional distribution. To capture a more significant heavy tail behavior of the *unconditional* probability distribution of the expansion coefficients, the Gaussian distribution may be replaced with a heavy-tailed distribution, such as Gamma, Laplacian or student-t. Furthermore, GARCH models of higher orders may be utilized. However, the choice of the particular distribution and order of the GARCH model is a matter of trial and error.

We derived recursive estimators for the variances and magnitudes of the STFT expansion coefficients. The variance of an expansion coefficient is recursively estimated by iterating propagation and update steps following the rational of Kalman filtering. Maximum-likelihood estimates of the model parameters are obtained by solving a constrained nonlinear minimization problem, similar to the estimation problem of standard GARCH models. The performance of the proposed speech enhancement algorithm was compared to that obtained by using the conventional Gaussian model and the decision-directed estimation approach. Using the proposed method, weak speech components and unvoiced sounds are significantly more emphasized and the enhanced speech is of higher quality.

It should be noted that the experimental results in this work are obtained under the assumption that signal presence is perfectly detected. That is, for each time-frequency bin tk we know in advance whether a desired speech component X_{tk} is present or absent in the noisy component Y_{tk}. Therefore, whenever speech is present we apply the log-spectral gain function (see (36)) to the noisy spectral component, and whenever speech is absent we simply attenuate the noisy spectral component by a constant factor. In practice, under signal presence uncertainty the signal presence probability $p_{tk} = P\left(H_1^{tk}\right)$ is estimated, and the quality of the enhanced speech may be lower due to miss-detection of speech components ($\hat{p}_{tk} < 1$ under H_1^{tk}). Furthermore, some residual musical noise may be generated due to false-detection of speech components ($\hat{p}_{tk} > 0$ under H_0^{tk}). In addition, we assumed that the clean signal is available for the estimation of the

model parameters. In practice, the performance of the proposed algorithm will be lower, since the model has to be estimated from the noisy signal. Nevertheless, the experimental results show the potential of the proposed model, and motivate a further research on the estimation of the signal presence probability and the model itself.

ACKNOWLEDGEMENT

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

REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[2] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-33, no. 2, pp. 443–445, April 1985.

[3] A. J. Accardi and R. V. Cox, "A modular approach to speech enhancement with an application to speech coding," in *Proc. 24th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-99*, Phoenix, Arizona, 15–19 March 1999, pp. 201–204.

[4] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detector," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, January 1999.

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

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

[7] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *special issue of EURASIP JASP on Digital Audio for Multimedia Communications*, vol. 2003, no. 10, pp. 1043–1051, September 2003.

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

[9] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proc. 27th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-02*, Orlando, Florida, 13–17 May 2002, pp. I–253–I–256.

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

[11] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204–207, July 2003.

[12] ——, "A soft voice activity detector based on a laplacian-gaussian model," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 5, pp. 498–505, September 2003.

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

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

[15] Y. Ephraim and D. Malah, "Signal to noise ratio estimation for enhancing speech using the Viterbi algorithm," Technion - Israel Institute of Technology, Haifa, Israel, Technical Report, EE PUB 489, March 1984.

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

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

[18] R. F. Engle, "Autoregressive conditional heteroskedasticity with estimates of the variance of united kingdom inflation," *Econometrica*, vol. 50, no. 4, pp. 987–1007, July 1982.

[19] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, vol. 31, no. 3, pp. 307–327, April 1986.

[20] T. Bollerslev, R. Y. ChouKenneth, and F. Kroner, "ARCH modeling in finance: A review of the theory and empirical evidence," *Journal of Econometrics*, vol. 52, no. 1-2, pp. 5–59, April-May 1992.

[21] B. Mandelbrot, "The variation of certain speculative prices," *Journal of Business of the University of Chicago*, vol. 36, pp. 394–419, October 1963.

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

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

[24] R. S. M. Berouti and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. 4th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-79*, Washington, DC, 9–11 April 1979, pp. 208–211.

[25] R. F. Engle, Ed., *ARCH Selected Readings*. New York: Oxford University Press Inc., 1995.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society (B)*, vol. 39, no. 1, pp. 1–38, 1977.

[27] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. New York: Wiley, 1997.

[28] E. K. Berndt, B. H. Hall, R. E. Hall, and J. A. Hausman, "Estimation and inference in nonlinear structural models," *Annals of Economic and Social Measurement*, vol. 4, pp. 653–665, 1974.

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

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

LIST OF TABLES

LIST OF FIGURES

TABLE I

SEGMENTAL SNR OBTAINED BY USING THE GARCH MODELING AND THE DECISION-DIRECTED METHODS.

Input SNR	GARCH modeling method				Decision-Directed method			
[dB]	Mean	Best	Worst	Median	Mean	Best	Worst	Median
0	7.29	8.67	5.68	7.34	6.73	8.09	5.39	6.65
5	10.78	12.25	8.97	10.81	9.62	11.47	8.07	9.59
10	14.69	16.12	12.76	14.77	12.80	14.85	10.95	12.85
15	18.89	20.29	17.03	18.95	16.32	18.32	14.41	16.23
20	23.03	24.33	21.48	23.03	20.03	21.65	18.43	19.97

TABLE II

LOG-SPECTRAL DISTORTION OBTAINED BY USING THE GARCH MODELING AND THE DECISION-DIRECTED METHODS.

Input SNR	GARCH modeling method				Decision-Directed method			
[dB]	Mean	Best	Worst	Median	Mean	Best	Worst	Median
0	4.47	2.70	6.15	4.47	4.74	3.27	6.27	4.90
5	3.15	1.92	4.46	3.10	4.07	2.81	5.75	4.25
10	2.26	1.36	3.35	2.22	3.50	2.32	5.09	3.56
15	1.61	0.97	2.50	1.56	2.82	1.75	4.20	2.79
20	1.14	0.67	1.82	1.09	2.13	1.27	3.29	2.06

TABLE III

PESQ SCORES OBTAINED BY USING THE GARCH MODELING AND THE DECISION-DIRECTED METHODS.

Input SNR	GARCH modeling method				Decision-Directed method			
[dB]	Mean	Best	Worst	Median	Mean	Best	Worst	Median
0	2.55	2.90	2.36	2.52	2.21	2.39	2.09	2.19
5	2.98	3.46	2.80	2.95	2.61	2.80	2.49	2.58
10	3.39	3.81	3.05	3.35	2.98	3.15	2.76	3.01
15	3.69	4.05	3.20	3.69	3.31	3.47	2.98	3.33
20	3.89	4.22	3.48	3.92	3.64	3.91	3.28	3.67

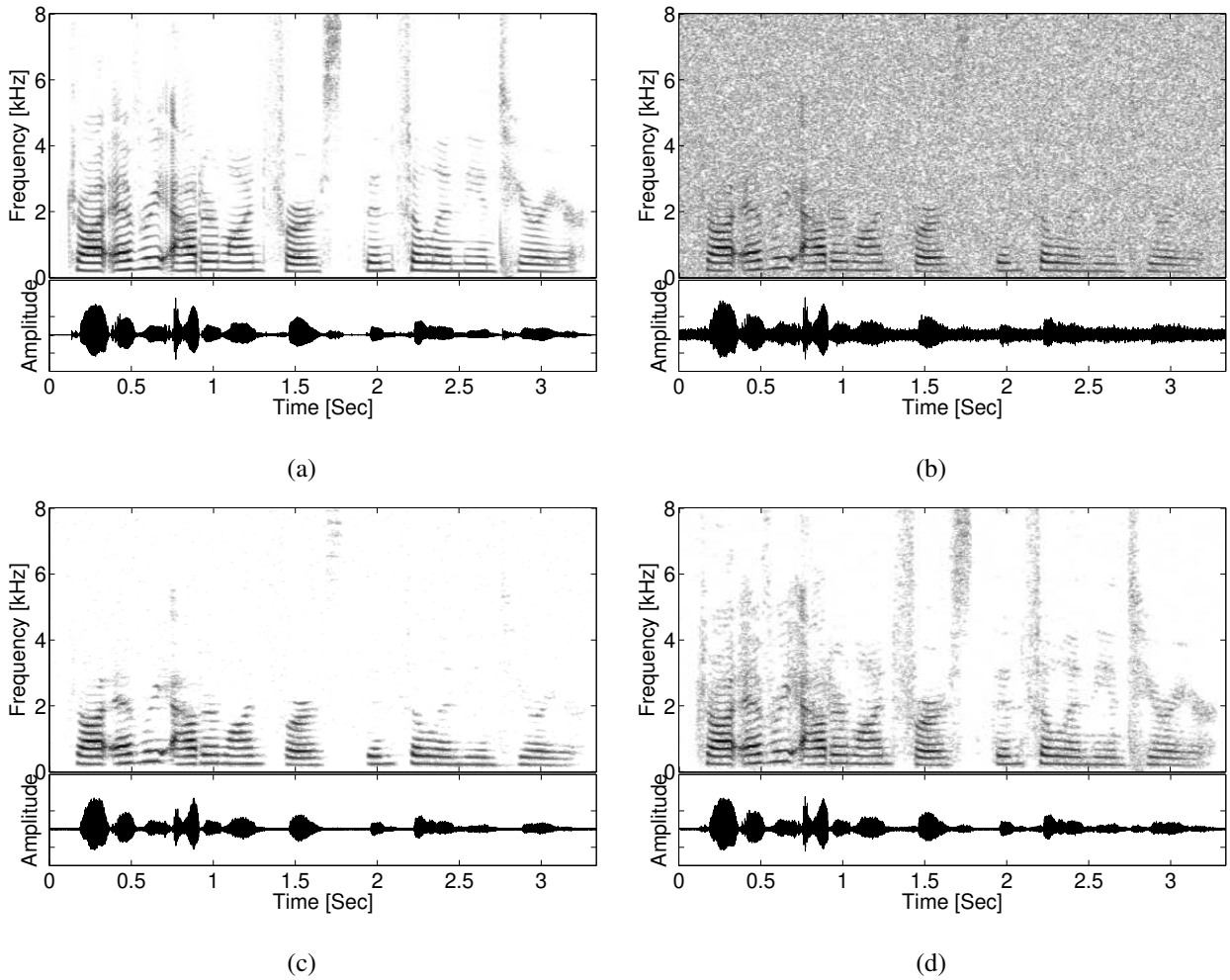

Fig. 1. Speech spectrograms and waveforms. (a) Original clean speech signal: "Draw every outer line first, then fill in the interior."; (b) noisy signal (SNR $= 5$ dB, SegSNR $= 3.75$ dB, LSD $= 12.17$ dB, PESQ $= 1.80$); (c) speech enhanced using the decision-directed method (SegSNR $= 11.04$ dB, LSD $= 3.28$ dB, PESQ $= 2.69$); (d) speech enhanced using the GARCH modeling method (SegSNR $= 11.78$ dB, LSD $= 2.56$ dB, PESQ $= 2.88$).