Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®
Signal Processing 84 (2004) 2453–2459
SIGNAL PROCESSING
www.elsevier.com/locate/sigpro

Fast communication

Modeling speech signals in the time–frequency domain using GARCH

Israel Cohen*

Department of Electrical Engineering, Technion Israel Institute of Technology, Technion City, Haifai 32000, Israel

Received 8 July 2004

Abstract

In this paper, we introduce a novel modeling approach for speech signals in the short-time Fourier transform (STFT) domain. We define the conditional variance of the STFT expansion coefficients, and model the one-frame-ahead conditional variance as a generalized autoregressive conditional heteroscedasticity (GARCH) process. The proposed approach offers a reasonable model on which to base the estimation of the variances of the STFT expansion coefficients, while taking into consideration their heavy-tailed distribution.

Keywords: Speech modeling; Time–frequency analysis; GARCH

1. Introduction

Modeling speech signals in the short-time Fourier transform (STFT) domain is a fundamental problem in designing speech processing systems. The Gaussian model, proposed by Ephraim and Malah [7], describes the individual STFT expansion coefficients of the speech signal as zero-mean statistically independent Gaussian random variables. It facilitates a mathematically tractable design of useful speech enhancement algorithms in the STFT domain. However, the Gaussian approximation can be very inaccurate in the tail regions of the probability density function [9,11,12]. Martin [11] proposed to model the real and imaginary parts of the expansion coefficients as either Gamma or Laplacian random variables. Lotter and Vary [10] 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. Statistical models based on HMP try to circumvent the assumption of specific distributions [8]. The probability distribution is estimated from long training sequences of the speech samples.

In this paper, we introduce a novel modeling approach for speech signals in the STFT domain.

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

0165-1684/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2004.09.001

Autoregressive conditional heteroscedasticity (ARCH) models, introduced by Engle [6] and generalized by Bollerslev [1], are widely used in various financial applications such as risk management, option pricing, foreign exchange, and the term structure of interest rates [2]. They 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., heavy tail behavior) and volatility clustering, two important characteristics of financial time-series. Speech signals in the STFT domain demonstrate both "variability clustering" and heavy tail behavior. 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 variability of the expansion coefficients is 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. Modeling the time trajectories of the expansion coefficients as generalized ARCH (GARCH) processes offers a reasonable model on which to base the variance estimation, while taking into consideration the heavy-tailed distribution.

This paper is organized as follows. In Section 2, we review the GARCH model. In Section 3, we define *conditional variance* in the STFT domain, and model the one-frame-ahead conditional variance as a GARCH process. In Section 4, we address the problem of estimating the model parameters. Finally, in Section 5, we demonstrate the application of the proposed model to speech enhancement.

2. GARCH model

Let $\{y_t\}$ denote a real-valued discrete-time stochastic process, and let ψ_t denote the 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 minimum mean-squared error (MMSE) sense is obtained by subtracting from y_t its conditional expectation given the information through time $t-1$,

$$\varepsilon_t = y_t - E\{y_t|\psi_{t-1}\}. \tag{1}$$

The conditional variance (volatility) of y_t given the information through time $t-1$ is by definition the conditional expectation of ε_t^2,

$$\sigma_t^2 = \mathrm{var}\{y_t|\psi_{t-1}\} = E\{\varepsilon_t^2|\psi_{t-1}\}. \tag{2}$$

The ARCH and GARCH models provide a rich class of possible parametrization of conditional heteroscedasticity (i.e., time-varying volatility). They explicitly recognize the difference between the unconditional variance $E\{[y_t - E\{y_t\}]^2\}$ 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:

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

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

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 is a special case of the 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

$$\kappa > 0,\ \alpha_i \geqslant 0,\ \beta_j \geqslant 0, \quad i = 1, \ldots, q,\ j = 1, \ldots, p,$$

$$\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1.$$

The first three constraints are sufficient to ensure that the conditional variances $\{\sigma_t^2\}$ are strictly positive. The fourth constraint is a covariance stationarity constraint, which is necessary and sufficient for the existence of a finite unconditional variance of the innovations process [1].

Many financial time-series such as exchange rates and stock returns exhibit a volatility clustering phenomenon, i.e., large changes tend to follow large changes of either sign and small changes tend to follow small changes. Eq. (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 $\{\beta_j\}$. Furthermore, the innovations of financial time-series are typically distributed with heavier tails than a Gaussian distribution. Bollerslev [1] showed that GARCH models are appropriate for heavy-tailed distributions.

3. Speech modeling

Let $x(n)$ denote a speech signal, and let its STFT expansion coefficients be given by

$$X_{tk} = \sum_{n=0}^{K-1} x(n+tM)h(n)\,\mathrm{e}^{-j(2\pi/K)nk}, \tag{6}$$

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). Let H_0^{tk} and H_1^{tk} denote, respectively, hypotheses of signal absence and presence, and let s_{tk} denote a binary state variable which indicates signal presence, i.e., $s_{tk} = 0$ when H_0^{tk}, and $s_{tk} = 1$ when H_1^{tk}. Let $\sigma_{tk}^2 \triangleq E\{|X_{tk}|^2|H_1^{tk}\}$ denote the variance of an expansion coefficient X_{tk} under H_1^{tk}.

We assume that given $\{\sigma_{tk}\}$ and $\{s_{tk}\}$, the expansion coefficients $\{X_{tk}\}$ are generated by

$$X_{tk} = \sigma_{tk} V_{tk}, \tag{7}$$

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

$$E\{V_{tk}|H_1^{tk}\} = 0, \quad E\{|V_{tk}|^2|H_1^{tk}\} = 1. \tag{8}$$

Accordingly, the expansion coefficients $\{X_{tk}|H_1^{tk}\}$ are *conditionally* zero-mean statistically independent Gaussian random variables given their standard deviations $\{\sigma_{tk}\}$. The real and imaginary parts of X_{tk} under H_1^{tk} are *conditionally* IID random variables given σ_{tk}.

Let $\mathscr{X}_0^\tau = \{X_{tk}|t = 0, \ldots, \tau, k = 0, \ldots, K-1\}$ represent the set of expansion coefficients up to frame τ, and let $\sigma_{tk|\tau}^2 \triangleq E\{|X_{tk}|^2|H_1^{tk}, \mathscr{X}_0^\tau\}$ denote the *conditional* variance of X_{tk} under H_1^{tk} given the expansion coefficients up to frame τ. Then, we assume that the one-frame-ahead conditional variance $\sigma_{tk|t-1}^2$ evolves according to a GARCH(1, 1) process:

$$\sigma_{tk|t-1}^2 = \sigma_{\min}^2 + \mu|X_{t-1,k}|^2 + \delta(\sigma_{t-1,k|t-2}^2 - \sigma_{\min}^2), \tag{9}$$

where

$$\sigma_{\min}^2 > 0, \quad \mu \geqslant 0, \quad \delta \geqslant 0, \quad \mu + \delta < 1 \tag{10}$$

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 $\sigma_{\min}^2$ is a lower bound on the variance of X_{tk} under H_1^{tk}.

The variances $\{\sigma_{tk}^2\}$ are hidden from direct observation, in the sense that even under perfect conditions of zero noise, their values are not directly observable, but have to be estimated from the available information. If the available information includes the set of expansion coefficients up to frame $t-1$, then an MMSE estimator for σ_{tk}^2 can be obtained by

$$\widehat{\sigma_{tk}^2} = E\{\sigma_{tk}^2|H_1^{tk}, \mathscr{X}_0^{t-1}\}. \tag{11}$$

From (7), (8) and the definition of the conditional variance $\sigma^2_{tk|\tau}$, we have

$$\begin{aligned}\sigma^2_{tk|t-1} &\triangleq E\{|X_{tk}|^2|H_1^{tk},\mathscr{X}_0^{t-1}\}\\ &= E\{\sigma^2_{tk}|V_{tk}|^2|H_1^{tk},\mathscr{X}_0^{t-1}\}\\ &= E\{\sigma^2_{tk}|H_1^{tk},\mathscr{X}_0^{t-1}\}E\{|V_{tk}|^2|H_1^{tk}\}\\ &= \widehat{\sigma^2_{tk}}. \end{aligned} \tag{12}$$

Therefore, given $\mathscr{X}_0^{t-1}$, the conditional variance $\sigma^2_{tk|t-1}$ is an MMSE estimator for σ^2_{tk}.

It should be noted that in speech enhancement applications, the available information is generally the set of *noisy* spectral components up to frame t, rather than the *clean* spectral components up to frame $t-1$. In that case, an estimate for σ_{tk} is derived from the available noisy data [4].

4. Model estimation

In this section we address the problem of estimating the model parameters $\sigma^2_{\min}$, μ and δ. The maximum-likelihood (ML) estimation approach is commonly used for estimating the parameters of a GARCH model. We derive the ML function of the model parameters by using the expansion coefficients in 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.

Let $\mathscr{H}_1=\{tk|s_{tk}=1\}$ denote the set of time–frequency bins where the signal is present, and let $\boldsymbol{\phi}=[\sigma^2_{\min},\mu,\delta]$ denote the vector of unknown parameters. Then for $tk\in\mathscr{H}_1$, the conditional distribution of X_{tk} given its standard deviation σ_{tk} is Gaussian:

$$p(X_{tk}|\sigma_{tk})=\frac{1}{\pi\sigma^2_{tk}}\exp\left(-\frac{|X_{tk}|^2}{\sigma^2_{tk}}\right),\quad tk\in\mathscr{H}_1. \tag{13}$$

Furthermore, $\{X_{tk}|\sigma_{tk},tk\in\mathscr{H}_1\}$ are statistically independent. We showed in Section 3 that the MMSE estimate of σ^2_{tk} given the expansion coefficients up to frame $t-1$ is $\sigma^2_{tk|t-1}$. The conditional variance $\sigma^2_{tk|t-1}$ can recursively be calculated from past expansion coefficients $\mathscr{X}_0^{t-1}$ by using (9) and the parameter vector $\boldsymbol{\phi}$. Hence, the logarithm of the conditional density of X_{tk} given the expansion coefficients up to frame $t-1$ can be expressed for $tk\in\mathscr{H}_1$ as

$$\log p(X_{tk}|\mathscr{X}_0^{t-1};\boldsymbol{\phi})=-\frac{|X_{tk}|^2}{\sigma^2_{tk|t-1}}-\log\sigma^2_{tk|t-1}-\log\pi. \tag{14}$$

It is convenient to regard the expansion coefficients in the first frame $\{X_{0,k}\}$ as deterministic, and initialize the conditional variances in the first frame $\{\sigma^2_{0,k|-1}\}$ to their minimal values $\sigma^2_{\min}$. Then, the log-likelihood can be maximized when conditioned on the first frame (for sufficiently large sample size, the expansion coefficients of the first frame make a negligible contribution to the total likelihood). The log-likelihood conditional on the expansion coefficients of the first frame is given by

$$\mathscr{L}(\boldsymbol{\phi})=\sum_{tk\in\mathscr{H}_1\cap t\in[1,T]}\log p(X_{tk}|H_1^{tk},\mathscr{X}_0^{t-1};\boldsymbol{\phi}). \tag{15}$$

Substituting (14) into (15) and imposing the constraints in (10) on the estimated parameters, the ML estimates of the model parameters can be obtained by solving the following constrained minimization problem:

$$\begin{aligned}&\underset{\hat\sigma^2_{\min},\hat\mu,\hat\delta}{\text{minimize}} \quad \sum_{tk\in\mathscr{H}_1\cap t\in[1,T]}\left[\frac{|X_{tk}|^2}{\sigma^2_{tk|t-1}}+\log\sigma^2_{tk|t-1}\right]\\ &\text{subject to}\quad \hat\sigma^2_{\min}>0,\ \hat\mu\geqslant 0,\ \hat\delta\geqslant 0,\ \hat\mu+\hat\delta<1.\end{aligned} \tag{16}$$

Such a problem is generally referred to as constrained nonlinear optimization or nonlinear programming. For given numerical values of the parameters, the sequences of conditional variances $\{\sigma^2_{tk|t-1}\}$ can be calculated from (9) and used to evaluate the series in (16). The result can then be minimized numerically as in Bollerslev [1]. 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 expansion coefficients are complex and generated from a two-state model (speech presence and absence states).

5. Experimental results

In this section, we demonstrate the application of the proposed model to speech enhancement. The evaluation includes three objective quality measures, namely segmental SNR, log-spectral distortion (LSD) and perceptual evaluation of speech quality (PESQ) score (ITU-T P.862). The speech signals, taken from the TIMIT database, include 20 different utterances from 20 different speakers, half male and half female. The signals are sampled at 16 kHz and degraded by white Gaussian noise with SNR 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 parameters of the GARCH model are estimated independently for each speaker from the clean signal of that speaker, as described in Section 4. A speech enhancement algorithm, which is based on the proposed model and log-spectral amplitude (LSA) estimation [5], is applied to each noisy speech signal (the details of the algorithm can be found in [4], and are not presented here due to space limitations). Alternatively, the variances of the speech STFT expansion coefficients are estimated by using the decision-directed method [7] with parameters as suggested in [3] (specifically, the weighting factor is $\alpha = 0.98$ and the lower bound on the a priori signal-to-noise ratio (SNR) is $\xi_{\min} = -15$ dB).

Table 1 summarizes the results of the segmental SNR, the LSD and the PESQ scores achieved by both algorithms. It shows that speech enhancement based on the proposed GARCH modeling method yields a higher segmental SNR, lower LSD, and higher PESQ scores than that based on 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 proposed modeling method is better than that obtained by using the decision-directed method. Fig. 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 more emphasized in the signal enhanced by the proposed method than in the signal enhanced by using the decision-directed method.

6. Conclusion

We have proposed a novel approach for statistically modeling speech signals in the STFT domain. It provides an explicit model for the conditional variance and conditional distribution of the expansion coefficients. It offers a reasonable model on which to base the estimation of the variances of the STFT expansion coefficients, while taking into consideration their heavy-tailed distribution. To capture a more significant heavy tail behavior, the Gaussian distribution of $\{V_{tk}|H_1^{tk}\}$ may be replaced with a heavy tail

Table 1
Segmental SNR, log-spectral distortion, and PESQ scores obtained by using the GARCH modeling and the decision-directed methods

Input SNR (dB)	GARCH modeling method			Decision-directed method		
	SegSNR	LSD	PESQ	SegSNR	LSD	PESQ
0	7.29	4.47	2.55	6.73	4.74	2.21
5	10.78	3.15	2.98	9.62	4.07	2.61
10	14.69	2.26	3.39	12.80	3.50	2.98
15	18.89	1.61	3.69	16.32	2.82	3.31
20	23.03	1.14	3.89	20.03	2.13	3.64

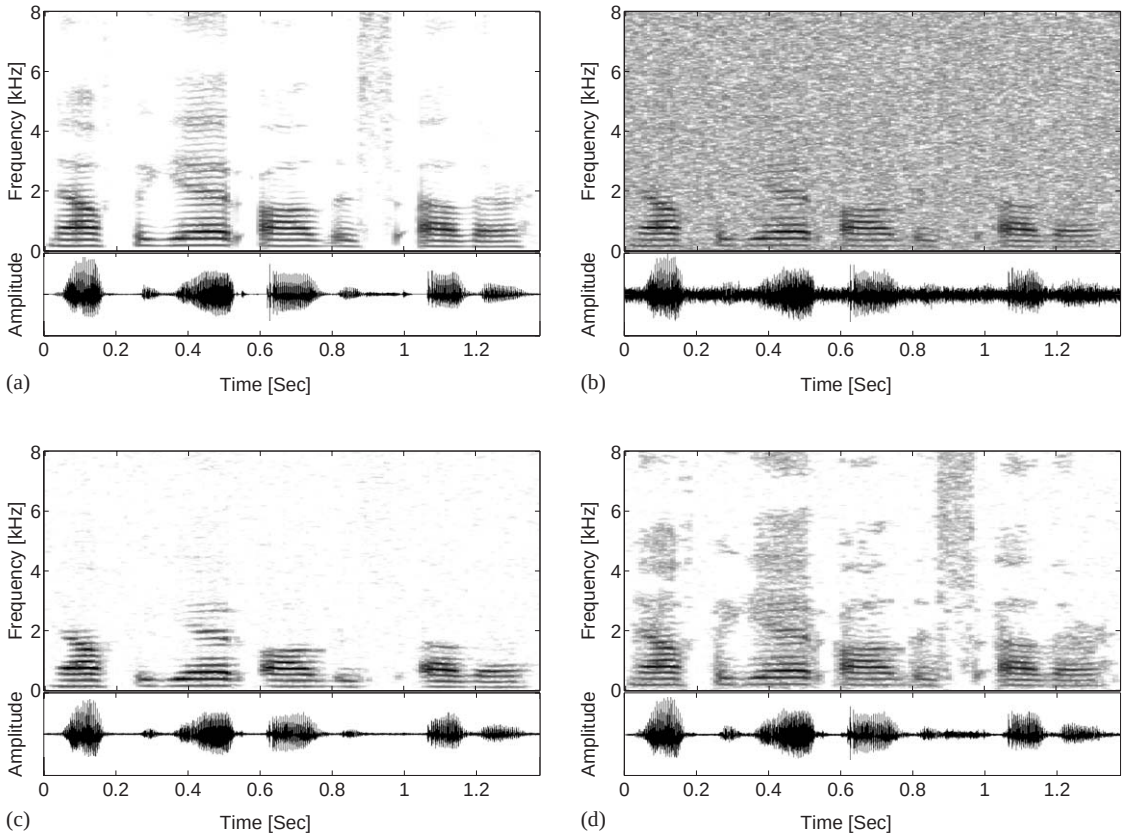

Fig. 1. Speech spectrograms and waveforms. (a) Original clean speech signal: "Now forget all this other."; (b) noisy signal (SNR = 5 dB, SegSNR = 4.35 dB, LSD = 13.75 dB, PESQ = 1.76); (c) speech enhanced using the decision-directed method (SegSNR = 11.47 dB, LSD = 3.28 dB, PESQ = 2.56); (d) speech enhanced using the GARCH modeling method (SegSNR = 12.25 dB, LSD = 3.00 dB, PESQ = 2.88).

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.

Acknowledgements

The author thanks Prof. Murat Kunt and the anonymous reviewers for their helpful comments.

References

[1] T. Bollerslev, Generalized autoregressive conditional heteroskedasticity, J. Econometrics 31 (3) (April 1986) 307–327.

[2] 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.

[3] 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.

[4] I. Cohen, Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity model, also Technical Report, EE PUB 1425, Technion, Israel Institute of Technology, Haifa, Israel, April 2004, submitted for publication.

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

[6] R.F. Engle, Autoregressive conditional heteroskedasticity with estimates of the variance of united kingdom inflation, Econometrica 50 (4) (July 1982) 987–1007.

[7] Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, IEEE Trans. Acoust. Speech Signal Process. ASSP-32 (6) (December 1984) 1109–1121.
[8] Y. Ephraim, N. Merhav, Hidden Markov processes, IEEE Trans. Inf. Theory 48 (6) (June 2002) 1518–1568.
[9] S. Gazor, W. Zhang, Speech probability distribution, IEEE Signal Process. Lett. 10 (7) (July 2003) 204–207.
[10] 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. I_832–I_835.
[11] R. Martin, Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors, in: Proceedings of the 27th IEEE International Conference on Acoustics Speech Signal Processing, ICASSP-02, Orlando, Florida, 13–17 May 2002, pp. I-253–I-256.
[12] J. Porter, S. Boll, Optimal estimators for spectral restoration of noisy speech, in: Proceedings of the IEEE International Conference on Acoustics Speech, Signal Processing, San Diego, California, 19–21 March 1984, pp. 18A.2.1–18A.2.4.