AR-GARCH in Presence of Noise: Parameter Estimation and its Application to Voice Activity Detection

Saman Mousazadeh and Israel Cohen, *Senior Member, IEEE*

***Abstract*—This paper presents a new method for voice activity detection (VAD) based on the autoregressive-generalized autoregressive conditional heteroscedasticity (AR-GARCH) model. The speech signal is modeled as an AR-GARCH process in the time domain, and the likelihood ratio is computed and compared to a threshold. The time-varying variance of the speech signal needed for computing the likelihood function under speech presence hypothesis, is estimated using the AR-GARCH model. The model parameters are estimated using a novel technique based on the recursive maximum likelihood (RML) estimation. The variance of the additive noise, a critical issue in designing a VAD, is estimated using the improved minima controlled recursive averaging (IMCRA) method, which is properly modified to be applicable to noise variance estimation in the time domain. The performances of the VAD and the parameter estimation method are examined under several conditions. Experimental results indicate the robustness of the AR-GARCH based VAD both to noise variations and low signal to noise ratio (SNR) conditions.**

***Index Terms*—AR-GARCH, voice activity detector, parameter estimation, noisy data, nonstationary noise**

I. Introduction

VOICE ACTIVITY DETECTION (VAD) is an integral part of all speech communication systems. Examples of such systems are audio conferencing, hands-free telephony [1] and discontinuous speech transmission [2]. During the last few decades, many researchers have dealt with this issue. The modern VADs are derived from statistical model-based algorithms in which the likelihood ratio test (LRT) is applied to a set of hypotheses. These VADs are developed in [3]-[6]. In all of these methods, the speech signal is transformed to the short-time Fourier transform (STFT) domain and different models are applied to the speech data in the STFT domain. For instance, Sohn et al. [3] and Ramirez et al. [4] assumed that the spectral coefficients of the noise and speech signal are complex gaussian random variables. Chang et al. [5] utilized the complex Laplacian and Gamma probability density functions (pdfs) to model the distributions of the speech and noise spectra. Shin et al. [6] applied the generalized gamma distribution to model the distribution of the clean speech spectrum. Davis et al. [7] proposed the statistical VAD method that makes no assumption about the distributions of the speech spectra.

Recently, Cohen [8] modeled the speech signal in the STFT domain as a complex GARCH process and used this model for speech enhancement. He showed that the time varying variance of the speech signal can be estimated using a complex GARCH model with Gaussian innovations. The well-known decision-directed method of Ephraim and Malah [9] can be derived as a special case using GARCH modeling. Solvang et al. [10] used the AR-GARCH model for VAD. In their work, they represent the AR part of the AR-GARCH model with a state-space to obtain the appropriate linear prediction error series. By applying the GARCH model to the residuals, they estimate the conditional variance sequences corresponding to the voice activity parts. To detect voice activity, they establish an appropriate threshold for the conditional variance sequences. However, their method is not computationally efficient and cannot be used in real time speech signal processing. Furthermore, they assume that the model is identified from frames with speech being present, but in the application at hand these frames are unknown to the user in advance.

In this paper, we present a new VAD using AR-GARCH modeling of the speech signal in the time domain. This model relies on the fact that speech signals in the time domain can be modeled as AR processes [11] and also demonstrate both the variability clustering and the heavy-tail behavior, that are the main properties of GARCH processes. Our VAD is based on the LRT, so we need to compute the likelihood function of the observations under the speech presence and absence hypotheses. The time varying variance of the speech signal needed for computing the likelihood function under speech presence hypothesis is estimated using the AR-GARCH model. To be able to use this model, one should know the parameters of the model or estimate them from the available data. Since the available data is often corrupted by additive noise, we present a novel technique for parameter estimation of the AR-GARCH model in presence of additive noise. To compute the likelihood function of the observations under speech presence and absence hypotheses, we need to know the variance of the additive noise, which might be time varying. The improved minima controlled recursive averaging (IMCRA) [12] method has been modified and utilized for estimating the variance of the corrupting noise in the time domain.

The paper is organized as follows. In Section 2, we introduce the AR-GARCH model and show its applicability to speech signals. In Section 3, we introduce our parameter

This work was supported by the Israel Science Foundation under Grant 1085/05. The authors are with the Department of Electrical Engineering, Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel. E-mail addresses: samanm@tx.technion.ac.il (S. Mousazadeh), icohen@ee.technion.ac.il (I. Cohen).

estimation algorithm. In Section 4, we introduce our VAD, which is based on the AR-GARCH modeling of the speech signal in the time domain. Finally in Section 4, we examine the performance of our method using several simulations.

II. AR VERSUS AR-GARCH MODELING OF SPEECH SIGNALS

The GARCH model was first introduced by Bollerslev [13] as an extension of the ARCH model developed by Engle [14] in economic data modeling. Since then, many researchers have tried to expand and use these models in several applications. AR-GARCH is one of these extensions [15]. The AR(p)-GARCH$(1,1)$ process is a filtered version of a GARCH$(1,1)$ process with an all-pole filter. This model is defined by the following three equations,

$$x_t = \sum_{i=1}^{p} \alpha_i x_{t-i} + \epsilon_t \tag{1}$$

$$\epsilon_t = \sigma_{t|t-1} v_t \tag{2}$$

$$\sigma^2_{t|t-1} = \beta_0 + \beta_1 \epsilon^2_{t-1} + \beta_2 \sigma^2_{t-1|t-2} \tag{3}$$

where t is the time index, α_i's and p in (1) are the parameters and the order of the AR part respectively, v_t's in (2) are zero-mean independent identically distributed random variables with unity variance, β_i's in (3) are the parameters of the GARCH$(1,1)$ model, and $\sigma^2_{t|t-1}$ is the one-sample-ahead conditional variance of the clean signal. We assume that v_t's are Gaussian random variables because this distribution models the speech data better than the Laplace or Gamma distributions [8].

We use the AR-GARCH to model the speech signal in the time domain, because it can model two main characteristics of the speech signals. Practical evidence shows that the spectrum of speech signals is very similar to that of AR processes. Since the spectrum of AR-GARCH processes depends only on the coefficients of the AR part, we can use the AR-GARCH model to model the spectral characteristics of speech signals. Another common property of speech signals and AR-GARCH processes is that they both have heavy tail pdfs. Fig. 1 shows the spectrum of a typical clean speech signal consisting of both male and female speakers from the TIMIT database [16], together with the spectrum of a synthetic AR(5)-GARCH$(1,1)$ process. The spectrums are estimated using the Welch method [17]. The synthetic AR-GARCH process is simulated using the parameters estimated from the clean speech signal using the RML method (discussed in the next section). From this figure, it is obvious that the spectrum of the speech signal is very similar to that of the synthetic AR-GARCH process. Fig. 2 displays the quantile-quantile plot (QQ-plot) of the quantiles of the speech signal versus the quantiles of the synthetic AR-GARCH samples. The purpose of the QQ plot is to determine whether the samples of two processes come from the same distribution. If the samples do come from the same distribution (same shape), even if one distribution is shifted and re-scaled from the other (different location and scale parameters), the plot will be linear. A reference line passing through the first and third quartiles is helpful for judging whether the points are linear. We provide in Fig. 3 the spectrum of a synthetic AR(5) process with Gaussian residual samples together with the spectrum of a clean speech signal. In Fig. 4, we also provide the QQ-plot of the quantiles of the speech signal versus the quantiles of the synthetic AR(5) process with Gaussian residual samples. From Figs. 1 to 4, we see that although both AR and AR-GARCH models can model the spectral properties of speech signals very well, the QQ-plot of the quantiles of the speech signal versus the quantiles of the AR-GARCH process is linear in a wider range when comparing the quantiles of the speech signal versus the quantiles of the AR process. This means that an AR-GARCH model can model the speech signal in the time domain much better than the widely-used AR model.

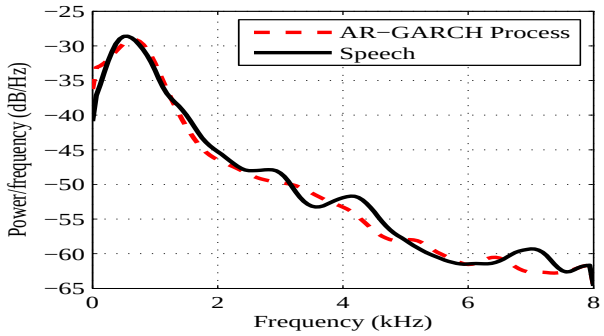

Fig. 1. Spectrum of a typical clean speech signal together with the spectrum of a synthetic AR-GARCH process.

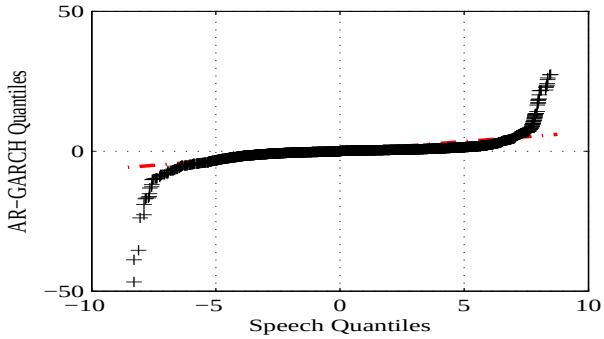

Fig. 2. QQ-plot of the quantiles of speech signal versus the quantiles of a synthetic AR-GARCH process.

III. AR-GARCH PARAMETER ESTIMATION

In this section we introduce a computationally efficient method for estimating the parameters of the AR-GARCH model from noisy observations. This method is then used in the next section for obtaining the proposed VAD.

Let y_t be a corrupted version of an AR-GARCH process with additive white Gaussian noise, i.e.,

$$y_t = x_t + n_t \qquad t = 1, 2, 3, ..., N \tag{4}$$

where N is the number of available data and n_t is the sequence of additive white Gaussian noise independent of x_t with unknown variance, i.e.,

$$n_t \sim \mathcal{N}\left(0, \sigma_t^2\right). \tag{5}$$

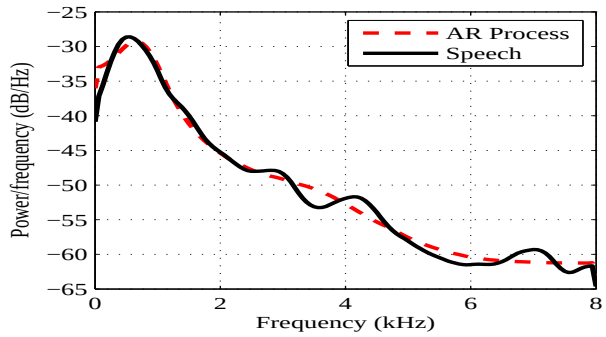

Fig. 3. Spectrum of a typical clean a speech signal together with the spectrum of a synthetic AR process.

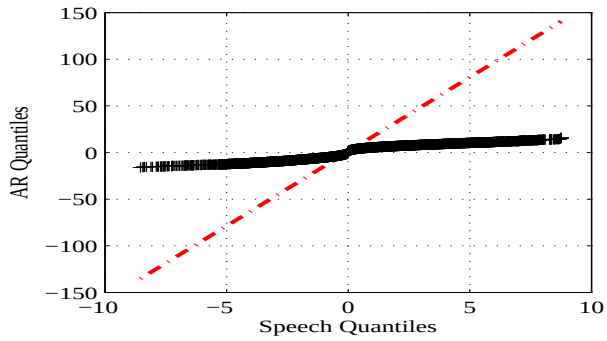

Fig. 4. QQ-plot of the quantiles of a speech signal versus the quantiles of a synthetic AR process with Gaussian residual.

Our purpose is to estimate the hidden states (x_t's) and the parameters of the model $\boldsymbol{\theta} = [\beta_0, \beta_1, \beta_2, \alpha_1, \alpha_2, ..., \alpha_p]^T$ using the available noisy data $y_t; t = 1, 2, 3, ..., N$.

The first step in obtaining the states and the parameter estimators is to estimate the variance of the corrupting noise (i.e., σ_t^2). Although most of the authors assume that the variance of the additive noise is known, in the current application (i.e., VAD) this assumption is not reasonable. The statistical properties of the noise are estimated from the speech-free frames and in this application these frames are unknown to the user in advance. Furthermore, the variance of the additive noise can be time varying. Hence, we need a method for estimating it adaptively. The IMCRA method is used for estimating the statistical properties of the additive noise (i.e., its variance). Note that the IMCRA method estimates the variance of the noise in the frequency domain (the STFT domain) while we need the variance of the noise in the time domain. In order to estimate the variance of the noise in the time domain using the estimates of the variance of the noise in the STFT domain, we use the well-known Parseval's relation. Suppose that the m-th frame consists of K samples of noisy speech signal and assume that the noise is stationary in this frame. Using Parseval's relation, we have,

$$\begin{aligned} \sum_{k=1}^{K} |N_m(k)|^2 &= K \sum_{t=(m-1)K+1}^{mK} n^2(t) \\ \sum_{k=1}^{K} E\left\{|N_m(k)|^2\right\} &= K \sum_{t=(m-1)K+1}^{mK} E\left\{n^2(t)\right\} \\ \sum_{k=1}^{K} \Gamma_m(k) &= K^2 \sigma_m^2 \end{aligned} \quad (6)$$

where $N_m(k)$ is the STFT coefficient of the noise in time frame number m and in the k-th frequency bin. $\Gamma_m(k)$ is the variance of the noise in the STFT domain in the m-th time frame and the k-th frequency bin. Thus, the estimate of the noise for a K sample time frame is as follows

$$\hat{\sigma}_t^2 = \frac{1}{K^2} \sum_{k=1}^{K} \hat{\Gamma}_m(k) \qquad t = (m-1)K + 1, 2, ..., mK \quad (7)$$

where $\hat{\Gamma}_m(k)$ is the estimate of the variance of the noise in the STFT domain in the m-th time frame and the k-th frequency bin obtained by the IMCRA method.

Using this estimate of the noise variance, we now introduce our state smoothing and parameter estimation method. The optimal estimator in the MMSE sense is the conditional mean, i.e.,

$$\hat{x}_t = E\left\{x_t \left| y_1^t \right.\right\} \quad (8)$$

where $y_1^t = \{y_\tau | \tau = 1, 2, ..., t\}$ is the set of observations up to time t. Let $x_1^t = \{x_\tau | \tau = 1, 2, ..., t\}$ and $\hat{x}_1^t = \{\hat{x}_\tau | \tau = 1, 2, ..., t\}$ be the set of clean samples (hidden states) and their estimates up to time t, respectively. Furthermore, as in [8], we assume that past estimates of the one-sample-ahead conditional variance and past estimates of the clean signal are sufficient statistics for one-sample-ahead conditional variance and clean signal estimation, i.e.,

$$\begin{aligned} \hat{x}_t &= E\left\{x_t | y_1^t\right\} \\ &= E\left\{x_t | \hat{x}_{t-p}^{t-1}, \hat{\sigma}_{t|t-1}^2, y_t\right\} \end{aligned} \quad (9)$$

$$\begin{aligned} \hat{\sigma}_{t|t-1}^2 &= E\left\{\sigma_{t|t-1}^2 | y_1^{t-1}\right\} \\ &= E\left\{\sigma_{t|t-1}^2 | \hat{x}_{t-p-1}^{t-2}, \hat{\sigma}_{t-1|t-2}^2, y_{t-1}\right\}. \end{aligned} \quad (10)$$

To estimate the clean signal, note that

$$\begin{aligned} \left(x_t | y_1^{t-1}\right) &\sim \left(x_t | \hat{x}_{t-p}^{t-1}, \hat{\sigma}_{t|t-1}^2\right) \\ &\sim \mathcal{N}\left(\sum_{i=1}^{p} \alpha_i \hat{x}_{t-i}, \hat{\sigma}_{t|t-1}^2\right) \end{aligned} \quad (11)$$

and

$$\left(n_t | y_1^{t-1}\right) \sim \mathcal{N}\left(0, \sigma_t^2\right). \quad (12)$$

Using (4), (9), (11), (12) and the well-known estimate of Gaussian signals in Gaussian noise [18] we get

$$\begin{aligned} \hat{x}_t &= E\left\{x_t | \hat{x}_{t-p}^{t-1}, \hat{\sigma}_{t|t-1}^2, y_t\right\} \\ &= \sum_{i=1}^{p} \alpha_i \hat{x}_{t-i} + \frac{\hat{\sigma}_{t|t-1}^2}{\hat{\sigma}_{t|t-1}^2 + \sigma_t^2} \left(y_t - \sum_{i=1}^{p} \alpha_i \hat{x}_{t-i}\right). \end{aligned} \quad (13)$$

Substituting (3) into (10) we get

$$\begin{aligned}\hat{\sigma}^2_{t|t-1}=&E\left\{\sigma^2_{t|t-1}|\hat{x}^{t-2}_{t-p-1},\hat{\sigma}^2_{t-1|t-2},y_{t-1}\right\}\\ =&E\left\{\beta_0+\beta_1\epsilon^2_{t-1}+\beta_2\sigma^2_{t-1|t-2}|\hat{x}^{t-2}_{t-p-1},\hat{\sigma}^2_{t-1|t-2},y_{t-1}\right\}\\ =&\beta_0+\beta_1\underbrace{E\left\{\epsilon^2_{t-1}|\hat{x}^{t-2}_{t-p-1},\hat{\sigma}^2_{t-1|t-2},y_{t-1}\right\}}_{u_{t-1}}+\beta_2\hat{\sigma}^2_{t-1|t-2}.\end{aligned} \tag{14}$$

The last step to complete the formulation of the MMSE estimation of the clean signal is to compute u_{t-1} as follows

$$\begin{aligned}u_{t-1}&=E\left\{\epsilon^2_{t-1}|\hat{x}^{t-2}_{t-p-1},\hat{\sigma}^2_{t-2|t-2},y_{t-2}\right\}\\ &=E\left\{\left(x_{t-1}-\sum_{i=1}^{p}\alpha_i x_{t-1-i}\right)^2\left|\hat{x}^{t-2}_{t-1-p},\hat{\sigma}^2_{t-1|t-2},y_{t-2}\right.\right\}\\ &=E\left\{\left(x_{t-1}-\sum_{i=1}^{p}\alpha_i \hat{x}_{t-1-i}\right)^2\left|\hat{\sigma}^2_{t-1|t-2},y_{t-1}\right.\right\}\end{aligned} \tag{15}$$

The last expectation in (15) can be easily computed using posterior pdf of the Bayesian general linear model [18] [equations (10.28)-(10.29)] as follows

$$\begin{aligned}u_{t-1}&=E\left\{\left(x_{t-1}-\sum_{i=1}^{p}\alpha_i \hat{x}_{t-1-i}\right)^2\left|\hat{\sigma}^2_{t-1|t-2},y_{t-1}\right.\right\}\\ &=\text{var}\left\{x_{t-1}\left|\hat{\sigma}^2_{t-1|t-2},y_{t-1}\right.\right\}\\ &\quad+\left(E\left\{x_{t-1}\left|\hat{\sigma}^2_{t-1|t-2},y_{t-1}\right.\right\}-\sum_{i=1}^{p}\alpha_i \hat{x}_{t-1-i}\right)^2\\ &=\left(\hat{\sigma}^2_{t-1|t-2}-\frac{\hat{\sigma}^4_{t-1|t-2}}{\hat{\sigma}^2_{t-1|t-2}+\sigma^2_{t-1}}\right)\\ &\quad+\left(\frac{\hat{\sigma}^4_{t-1|t-2}}{\hat{\sigma}^2_{t-1|t-2}+\sigma^2_{t-1}}\left(y_{t-1}-\sum_{i=1}^{p}\alpha_i \hat{x}_{t-1-i}\right)\right)^2\\ &=\frac{\hat{\sigma}^2_{t-1|t-2}\sigma^2_{t-1}}{\hat{\sigma}^2_{t-1|t-2}+\sigma^2_{t-1}}\\ &\quad+\left(\frac{\hat{\sigma}^2_{t-1|t-2}}{\hat{\sigma}^2_{t-1|t-2}+\sigma^2_{t-1}}\right)^2\left(y_{t-1}-\sum_{i=1}^{p}\alpha_i \hat{x}_{t-1-i}\right)^2.\end{aligned} \tag{16}$$

So far we have assumed that the parameters are known and we have found the MMSE estimate of the clean signal and the one-sample-ahead conditional variance. In real applications, this assumption is often not true, and we have to estimate the parameters of the model from the noisy observations. The ML estimate of the parameters is obtained by solving the following nonlinear optimization problem

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{ML}&=\max_{\boldsymbol{\theta}}\log\left(f\left(y_1^N;\boldsymbol{\theta}\right)\right)\\ \log\left(f\left(y_1^N;\boldsymbol{\theta}\right)\right)&=\sum_{t=2}^{N}\log\left(f\left(y_t|y_1^{t-1};\boldsymbol{\theta}\right)\right)\\ &\quad+\log\left(f(y_1;\boldsymbol{\theta})\right).\end{aligned} \tag{17}$$

where

$$\begin{aligned}\log\left(f\left(y_1^N;\boldsymbol{\theta}\right)\right)&=\sum_{t=2}^{N}\log\left(f\left(y_t|y_1^{t-1};\boldsymbol{\theta}\right)\right)\\ &\quad+\log\left(f(y_1;\boldsymbol{\theta})\right)\end{aligned} \tag{18}$$

and $f\left(y_1^N;\boldsymbol{\theta}\right)$ and $\hat{\boldsymbol{\theta}}_{ML}$ are, respectively, the likelihood function given the parameters and the maximum likelihood estimate of the parameters. The maximization problem described in (17) has no closed form solution, so it must be solved by numerical methods such as the steepest decent [19]. The main drawback of the steepest decent method is its high computational load making it inapplicable in real-time applications such as speech enhancement. In order to overcome this problem, a procedure like the one presented in [20] can be used to estimate the parameters together with the clean signal adaptively. The idea behind this method is to update the gradient of the likelihood function each time we receive new data. Next, the algorithm uses this estimate of the gradient vector to update the estimate of the parameter vector in the same way as the steepest decent method.

Under suitable regularity conditions described in [21] it can be shown that the average log-likelihood converges to the following limit

$$l\left(\boldsymbol{\theta}\right)=\lim_{k\to\infty}\frac{1}{k+1}\sum_{t=1}^{k}\log\left[f\left(\mathbf{y}^t|\mathbf{y}_0^{t-1};\boldsymbol{\theta}\right)\right]. \tag{19}$$

It can also be shown that $l\left(\boldsymbol{\theta}\right)$ admits $\boldsymbol{\theta}^*$ as a global maximum where $\boldsymbol{\theta}^*$ is the global maximum of the log-likelihood function in (17) [22]. To maximize $l\left(\boldsymbol{\theta}\right)$, one can use a Stochastic Approximation (SA) algorithm to update the parameter estimate at time t using the following recursion formula

$$\hat{\boldsymbol{\theta}}_t=\hat{\boldsymbol{\theta}}_{t-1}+\mu_t\nabla\log\left[f\left(\mathbf{y}_t|\mathbf{y}_0^{t-1};\hat{\boldsymbol{\theta}}\right)\right] \tag{20}$$

where $\hat{\boldsymbol{\theta}}_{t-1}$ is the parameter estimate at time $t-1$, and $\mathbf{g}_t=\nabla\log\left[f\left(\mathbf{y}_t|\mathbf{y}_0^{t-1};\hat{\boldsymbol{\theta}}\right)\right]$ is computed in the Appendix. This method is called recursive maximum likelihood (RML), and the algorithm for this method is summarized in Table I. In this table, the step size, μ_t is a positive non-increasing sequence, such that

$$\lim_{T\to\infty}\sum_{t=1}^{T}\mu_t=\infty \tag{21}$$

$$\sum_{t=1}^{\infty}\mu_t^2<\infty. \tag{22}$$

It can be shown that $\boldsymbol{\theta}_t$ will converge to the set of (global or local) maxima of $\log f\left(y_1^N;\boldsymbol{\theta}\right)$ [22]. It is worth mentioning that the RML method can also be used for adaptive parameter estimation (varying parameters). In this case, the choice of μ_t will be a trade-off between tracking capability (large μ_t) and low estimation noise around the parameter (small μ_t).

A. Considering the Stationarity Conditions

Often in signal processing applications, it is needed that the estimated model be stationary. Hence, we must maximize the

TABLE I
RECURSIVE ML ALGORITHM FOR ESTIMATING THE PARAMETERS OF A NOISY AR-GARCH PROCESS

Initialization:
Let $\hat{\boldsymbol{\theta}}_0 = \mathbf{0}$.
Let $u_t = \hat{\sigma}^2_{t-1|t-2} = \hat{x}_t = 0$ for $t = 1, 2, ..., p$.
for $t = p+1$ to N
Compute $\hat{\sigma}^2_{t|t-1} = \beta_0 + \beta_1 u_{t-1} + \beta_2 \hat{\sigma}^2_{t-1|t-2}$.
Compute $\mathbf{g}_t$ using (41).
$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t-1} + \mu_t \mathbf{g}_t$.
Compute
$u_t = \frac{\hat{\sigma}^2_{t|t-1}\sigma^2_t}{\hat{\sigma}^2_{t|t-1}+\sigma^2_t} + \left(\frac{\hat{\sigma}^2_{t|t-1}}{\hat{\sigma}^2_{t|t-1}+\sigma^2_t}\right)^2 \left(y_t - \sum_{i=1}^p \alpha_i \hat{x}_{t-i}\right)^2$.
$\hat{x}_t = \sum_{i=1}^p \alpha_i \hat{x}_{t-i} + \frac{\hat{\sigma}^2_{t|t-1}}{\hat{\sigma}^2_{t|t-1}+\sigma^2_t}\left(y_t - \sum_{i=1}^p \alpha_i \hat{x}_{t-i}\right)$.
end (for)

log-likelihood function under stationarity conditions. The AR-GARCH process is strictly stationary with finite second order moment if and only if the following conditions hold [15]:
i) $\beta_0 > 0$ and $\beta_i \geq 0 \ \forall \ 1 \leq i \leq 2$
ii) $\beta_1 + \beta_2 < 1$
iii) All the roots of $1 - \sum_{i=1}^p \alpha_i z^{-i}$ must be inside the unit circle.
It is obvious that the third condition is highly nonlinear and cannot be easily applied to nonlinear optimization [19]. To overcome this problem, we use an alternative condition for the stationarity of AR processes that the roots of $1 - \sum_{i=1}^p \alpha_i z^{-i}$ are inside the unit circle (so that the AR process is stationary) if, and only if, $|\gamma_i| \leq 1 \ \forall \ 1 \leq i \leq p$ where γ_i's are the reflection coefficients of the AR model [17]. The reflection coefficients of the AR model can be computed uniquely from the AR coefficients recursively using the Levinson-Durbin recursion algorithm as follows [17]

$$\begin{aligned} \alpha_{i,k} &= \alpha_{i,k-1} + \gamma_k \alpha_{k-i,k-1}, \quad 1 \leq i < k \\ \alpha_{k,k} &= \gamma_k, \quad k = 1, ..., p \\ \alpha_i &= -\alpha_{i,p}, \quad i = 1, ..., p. \end{aligned} \tag{23}$$

Therefore, it can be concluded that the AR-GARCH process is strictly stationary with finite second order moment if, and only if, i, ii, and the following conditions hold:
iv) $-1 < \gamma_i < 1$,$\forall \ 1 \leq i \leq p$.
Obviously, three inequality constraints (i.e., i, ii and iv) are linear, and the ML estimate of the parameters can be found by solving the following optimization problem

$$\begin{aligned} \hat{\boldsymbol{\phi}} &= \arg \max_{\boldsymbol{\phi}} \left\{\log\left(f\left(y_1^N; \boldsymbol{\phi}\right)\right)\right\} \\ &\quad \text{s.t.} \\ &\quad \beta_0 > 0 \ \text{ and } \ \beta_i \geq 0 \ \ \forall 1 \leq i \leq 2 \\ &\quad \beta_1 + \beta_2 < 1 \\ &\quad 1 < \gamma_i < -1 \ \ \forall 1 \leq i \leq p \end{aligned} \tag{24}$$

where $\boldsymbol{\phi} = [\beta_0, \beta_1, \beta_2, \gamma_1, \gamma_2, ..., \gamma_p]^T$ is the vector of parameters consisting of the parameters of the GARCH part and reflection coefficients of the AR part.

It is clear that the maximization problem described in (24) has no closed form solution and must be solved using numerical methods such as the gradient projection method [19]. As previously stated, this optimization problem has a high computational load; therefore, we use the RML method in order to decrease computational complexity. In order to use these numerical methods, we must compute the gradient of the log-likelihood function of the current observation conditioned on past observations with respect to the recently defined vector of parameters $(\boldsymbol{\phi})$. The first three elements of the gradient vector are computed in the Appendix (equations (37) through (41)), and the other elements can be computed using the following equations

$$\begin{aligned} g_{t,i+3} &= \frac{\partial}{\partial \gamma_i}\left(\log\left(f\left(y_t | y_0^{t-1}; \boldsymbol{\phi}\right)\right)\right) \\ &= \sum_{k=1}^p \frac{\partial}{\partial \alpha_k}\left(\log\left(f\left(y_t | y_0^{t-1}; \boldsymbol{\phi}\right)\right)\right) \frac{\partial \alpha_k}{\partial \gamma_i} \\ &\qquad i = 1, 2, ..., p \end{aligned} \tag{25}$$

where the partial derivative of the log-likelihood function with respect to the AR parameters (i.e., $\frac{\partial}{\partial \alpha_k}\left(\log\left(f\left(y_t | y_0^{t-1}; \boldsymbol{\phi}\right)\right)\right)$) is computed in the Appendix (equation (41)), and the partial derivative of α_k with respect to γ_i (i.e. $\frac{\partial \alpha_k}{\gamma_i}$) can be obtained recursively from the following equations [23]:

$$\frac{\partial \alpha_{k,i}}{\partial \gamma_i} = \begin{cases} \alpha_{i-k,i-1} & \text{if } 1 \leq k < i \\ 1 & \text{if } k = i \end{cases}$$

$$\frac{\partial \alpha_{k,j}}{\partial \gamma_i} = \begin{cases} \frac{\partial \alpha_{k,j-1}}{\partial \gamma_i} + \gamma_j \frac{\partial \alpha_{j-k,j-1}}{\partial \gamma_i} & \text{if } 1 \leq k < j \\ 0 & \text{if } k = i \text{ and } i < k \leqslant p. \end{cases}$$

$$\frac{\partial \alpha_k}{\partial \gamma_i} = -\frac{\partial \alpha_{k,p}}{\partial \gamma_i} \tag{26}$$

In the next section, we use the stationarity conditions along with the RML parameter estimation method in order to propose a new VAD.

IV. VAD BASED ON AR-GARCH MODELING OF SPEECH SIGNALS IN TIME DOMAIN

In this section, we introduce our VAD, which is based on an AR-GARCH model. We assume that the speech signal in the time domain can be modeled by an AR-GARCH process. Let y_t be the noisy samples of the speech signal in the time domain divided by the noise variance, σ_t^2. The noise variance, σ_t^2, is estimated by the IMCRA method and equations (6)-(7). This makes the proposed VAD a constant false alarm rate (CFAR) VAD, without affecting the overall performance of the algorithm. Suppose that we have a time frame and we want to decide whether it consists of speech or not. Let m be the frame index, K be the number of samples in each frame, and $0 \leq \delta \leq 1$ be the overlap factor between adjacent frames. The likelihood ratio (LR) for the t-th sample, given the observations up to $t-1$, can be written as follows

$$\lambda_t = \frac{f\left(y_t | y_1^{t-1}; \boldsymbol{\theta}, \boldsymbol{H}_1\right)}{f\left(y_t | y_1^{t-1}; \boldsymbol{\theta}, \boldsymbol{H}_0\right)} \tag{27}$$

where $\boldsymbol{H}_1$ and $\boldsymbol{H}_0$ are the speech presence and absence hypotheses, respectively. The decision rule for the m-th time

frame is established from the geometric mean of the likelihood ratios for the individual time sample [3], which is given by

$$\begin{aligned}\log\Lambda_m &= \frac{1}{K+2K\delta}\sum_{t=t_{1,m}}^{t_{2,m}}\log\lambda_t \\ &= \frac{1}{K+2K\delta}\sum_{t=t_{1,m}}^{t_{2,m}}\log\frac{f\left(y_t|y_1^{t-1};\boldsymbol{\theta},\boldsymbol{H}_1\right)}{f\left(y_t|y_1^{t-1};\boldsymbol{\theta},\boldsymbol{H}_0\right)}\underset{H_0}{\overset{H_1}{\gtrless}}T_h\end{aligned} \tag{28}$$

where $t_{1,m}=(m-1-\delta)K+1$ and $t_{2,m}=(m+\delta)K$. Suppose that the estimate of the model parameters ($\hat{\boldsymbol{\theta}}_t$), the one-sample-ahead conditional variance of the speech signal ($\hat{\sigma}^2_{t|t-1}$), and the estimate of the clean signal ($\hat{x}_t$) are obtained using the RML procedure. The numerators of the terms in the right side of (27) can be computed using (37). Assuming that the noise is white Gaussian with known variance, the denominators of the terms in the right side of (27) can be computed by

$$\begin{aligned}f\left(y_t|y_1^{t-1};\boldsymbol{\theta},\boldsymbol{H}_0\right) &= f\left(y_t;\boldsymbol{\theta},\boldsymbol{H}_0\right) \\ &= \frac{1}{\sqrt{2\pi\sigma_t^2}}\exp\left(\frac{-y_t^2}{2\sigma_t^2}\right).\end{aligned} \tag{29}$$

Another issue that must be taken into account is the correlation between consecutive samples of the speech signal. The sequence of speech activity states can be modeled as a first-order Markov process which may help to prevent clipping of weak speech [3]. Based on the total probability theorem and Bayes' rule, we can easily derive the soft VAD rule as follows [24]

$$P_{m|m}=\frac{\Lambda_m P_{m|m-1}}{\Lambda_m P_{m|m-1}+(1-P_{m|m-1})} \tag{30}$$

where m is the index of a time frame consisting of k samples, Λ_m is the likelihood ratio of the m-th frame, which can be computed by (27), and $P_{m|m}$ and $P_{m|m-1}$ are the soft decision rule with and without using the information provided by the m-th frame data, respectively. The relationship between $P_{m|m}$ and $P_{m|m-1}$ is similar to the relationship between the current state and the previous state in the Markov model and is given by [24]

$$P_{m+1|m}=h_{0,1}(1-P_{m|m})+h_{1,1}P_{m|m} \tag{31}$$

where $0<h_{0,1}<1$ and $0<h_{1,1}<1$ denote the probabilities of speech activity in the current frame when there is silence or speech in the previous frame, respectively. The issue of selecting the values of $h_{0,1}$ and $h_{1,1}$ is a trade-off between the probability of false alarm ($\mathrm{P_{fa}}$) and the probability of missed detection ($\mathrm{P_{miss}}$). This means that a smaller value of $h_{0,1}$ results in less noise detected as signal (smaller probability of false alarm) at the expense of more front-end speech being detected as noise (greater $\mathrm{P_{miss}}$) and that a greater value of $h_{1,1}$ cause less middle speech clipping at the expense of more front-end speech being detected as noise.

Our final voice activity detector is obtained by comparing $P_{m|m}$ to a predetermined threshold, i.e.,

$$H(m)=\begin{cases}H_1, & \text{if } P_{m|m}\geq T_h\\ H_0, & \text{otherwise}\end{cases} \tag{32}$$

where T_h determines the trade-off between the probability of false alarm ($\mathrm{P_{fa}}$) and the probability of detection ($\mathrm{P_d}$). The proposed VAD algorithm is summarized in Table II. In this table, ϕ^i is the i-th element of the vector $\boldsymbol{\phi}$. As stated in Table II, the VAD algorithm first utilizes the IMCRA method to estimate the noise variance. A new observation sequence (y_t) is then defined by dividing the the noisy speech sequence by the estimated noise variance obtained by the IMCRA method. Then, the RML method is used to estimate the parameters of the AR-GARCH model together with the hidden states and the one-sample-ahead variance of the speech signal. Finally the likelihood ratio is computed using the estimated one-sample-ahead variance of the speech signal and is compared to a threshold.

TABLE II
THE PROPOSED VAD ALGORITHM USING AR-GARCH MODELING

Initialization:
Use the IMCRA method presented in section II, to estimate the variance of the noise $\hat{\sigma}_t^2$.
Let y_t be the noisy speech signal divided by $\hat{\sigma}_t$.
Let $\sigma_t^2=1$.
Let $\boldsymbol{\phi}_0=\mathbf{0}$.
Let $u_t=\hat{\sigma}^2_{t-1|t-2}=\hat{x}_t=0$ for $t=1,2,...,p+1$.
for $t=p+2$ to N.
Compute the AR parameters from the reflection coefficients using (23).
Compute $\hat{\sigma}^2_{t|t-1}=\beta_0+\beta_1 u_{t-1}+\beta_2\hat{\sigma}^2_{t-1|t-2}$.
Compute $\mathbf{g}_t$ using (41), (25) and (26).
$\hat{\boldsymbol{\phi}}_t=\hat{\boldsymbol{\phi}}_{t-1}+\mu_t\mathbf{g}_t$.
If $\hat{\phi}_t^i\leq 0; 1\leq i\leq 3$ then $\hat{\phi}_t^i=0$.
If $\hat{\phi}_t^i\geq 1; 1\leq i\leq 3$ then $\hat{\phi}_t^i=1$.
If $\hat{\phi}_t^i\leq -1; 4\leq i\leq 3+p$ then $\hat{\phi}_t^i=-1$.
If $\hat{\phi}_t^i\geq 1; 4\leq i\leq 3+p$ then $\hat{\phi}_t^i=1$.
Compute
$u_t=\frac{\hat{\sigma}^2_{t|t-1}\sigma_t^2}{\hat{\sigma}^2_{t|t-1}+\sigma_t^2}+\left(\frac{\hat{\sigma}^2_{t|t-1}}{\hat{\sigma}^2_{t|t-1}+\sigma_t^2}\right)^2\left(y_t-\sum_{i=1}^p\alpha_i\hat{x}_{t-i}\right)^2$.
$\hat{x}_t=\sum_{i=1}^p\alpha_i\hat{x}_{t-i}+\frac{\hat{\sigma}^2_{t|t-1}}{\hat{\sigma}^2_{t|t-1}+\sigma_t^2}\left(y_t-\sum_{i=1}^p\alpha_i\hat{x}_{t-i}\right)$.
end (for)
Select K, $h_{0,1}$ and $h_{1,1}$ and let $P_{1|0}=\frac{1}{2}$.
for $m=1$ to $\left[\frac{N}{K}\right]-1$.
Compute
$\Lambda_m=\exp\left(\frac{1}{K+2K\delta}\sum_{t=t_{1,m}}^{t_{2,m}}\log\frac{f\left(y_t|y_1^{t-1};\boldsymbol{\theta},\boldsymbol{H}_1\right)}{f\left(y_t|y_1^{t-1};\boldsymbol{\theta},\boldsymbol{H}_0\right)}\right)$
$P_{m|m}=\frac{\Lambda_m P_{m|m-1}}{\Lambda_m P_{m|m-1}+(1-P_{m|m-1})}$.
Update $P_{m+1|m}=h_{0,1}(1-P_{m|m})+h_{1,1}P_{m|m}$.
Decide whether the m-th frame contains speech or not by comparing $P_{m|m}$ with a threshold.
end (for)

V. PERFORMANCE EVALUATION

In this section, we evaluate the performance of the proposed parameter estimation method and VAD under several conditions. In the first experiment, we compare the performance of the constrained ML parameter estimation method (i.e., optimization problem in (24) solved using gradient the projection method [19]) with those of two different ML methods. The first method, denoted by MLClean, employs clean data (unavailable in practical situations) for estimating the parameter. The second method, denoted by MLNoisy, utilizes the noisy data for

estimating the parameter but falsely handles the data as clean. In this experiment, we use ten different AR(2)-GARCH$(1,1)$, processes corrupted with zero-mean complex Gaussian white noise with two different SNR levels. The number of available data (N) is set to 1024. The process v_t is a zero-mean Gaussian white process with unity variance. The reflection coefficients of the AR part are chosen randomly and uniformly from the interval $(-1,1)$, and the parameters of the GARCH$(1,1)$ are chosen randomly and uniformly from the interval $(0,1)$ such that the processes will be stationary. In this experiment, the additive corrupting noise variance is assumed to be known. For evaluating the performance of the proposed method we use mean square errors (MSE) in estimation of the parameters (which were estimated using 1000 realizations) and MSE for the estimate of $\sigma^2_{t|t-1}$, which are given by

$$\text{MSE} = \frac{1}{10}\sum_{j=1}^{10}\frac{1}{10^3}\sum_{i=1}^{1000}\frac{\left(\hat{\theta}_{k,(i,j)}-\theta_{k,j}\right)^2}{\left(\theta_{k,j}\right)^2} \tag{33}$$

$$E_{\sigma_t^2} = \frac{1}{10^4}\sum_{j=1}^{10}\sum_{i=1}^{1000}\frac{1}{N-1}\sum_{t=2}^{N}\left(\sigma^2_{t|t-1,(i,j)}-\hat{\sigma}^2_{t|t-1,(i,j)}\right)^2 \tag{34}$$

where i is the iteration index, j is the index for different parameters set, $\theta_{k,j}$ is the k-th element of the vector of parameters of the j-th process, $\hat{\theta}_{k,(i,j)}$ is the estimate of $\theta_{k,j}$ in i-th iteration, and $\hat{\sigma}^2_{t|t-1,(i,j)}$ is computed as follows

$$\hat{\sigma}^2_{t|t-1,(i,j)} = \hat{\theta}_{0,(i,j)} + \hat{\theta}_{1,(i,j)}\epsilon^2_{t-1} + \hat{\theta}_{2,(i,j)}\hat{\sigma}^2_{t-1|t-2,(i,j)}. \tag{35}$$

The results are given in Table III. It is obvious that the proposed method yields better performance over the MLNoisy method often used in real world problems. An important point to be emphasized here is that all of these methods (i.e. constrained ML, MLNoisy, MlNoisy) have the similar computational loads. This makes these methods unsuitable for real-time applications. We give these results for the sake of comparison.

TABLE III
MSE IN PARAMETER ESTIMATION AND MSE IN ESTIMATION OF $\sigma^2_{t|t-1}$ FOR DIFFERENT SNR LEVELS

Method:	MLClean	MLNoisy		Proposed	
SNR	——	5dB	10dB	5dB	10dB
β_0	0.1791	26.0113	3.4975	3.8168	1.2206
β_1	0.1023	0.7089	0.4879	0.4192	0.2400
β_2	0.1521	0.6499	0.4412	0.7246	0.5376
α_1	0.0048	0.1102	0.0203	0.0082	0.0065
α_2	0.0002	0.0596	0.0086	0.0003	0.0002
E^2_σ	0.0086	2.2257	0.3973	0.2009	0.0554

In the second experiment, we compare the performance of the proposed VAD with those of four commonly used VADs (i.e., Sohn et al. [3], Ramirez et al. [4], Chang et al. [5] and Shin et al. [6]). The speech signal used in our evaluation is taken from the TIMIT database [16]. We use a speech signal of 12s long consisting of different English sentences from both male and female speakers sampled at 16 kHz. The speech signal is corrupted by a computer generated white Gaussian noise, and the overall SNR is set to 5 dB. The parameters of the proposed VAD are given in Table IV. These parameters are chosen experimentally to achieve the highest performance.

TABLE IV
PARAMETER VALUES USED FOR THE PROPOSED VAD

K	μ_t	$h_{0,1}$	$h_{1,1}$	p	δ
256	$\frac{0.1}{\|\mathbf{g}_t\|}$	0.80	0.90	1	.90

The results of the simulations are depicted in Fig. 5 and Fig. 6. The clean speech signal together with a hand-marked VAD is depicted in Fig. 1. The hand-marked VAD is obtained as follows. In the m-th time frame we assume that the speech is present whenever $10\log_{10} S_m > \max_m\{10\log_{10} S_m\} - 22$ and the speech is absent whenever $10\log_{10} S_m < \max_m\{10\log_{10} S_m\} - 23$ where

$$S_m = \frac{1}{k}\sum_{t=(m-1)k+1}^{mk} x_t^2.$$

No decision is made in other frames. We chose these hand-marked VAD thresholds experimentally by two qualitative criteria. The speech presence threshold (-22 in this case) is chosen such that the listeners do not sense any essential difference between the original clean signal and the speech signal passed through the hand-marked VAD. The speech absence threshold (-23 in this case) is chosen such that the listeners do not sense any speech signal component in that part of speech which is labeled as silence.

In this experiment, the unknown variance of the additive noise is estimated using IMCRA for all of the reference methods. Fig. 6 shows the receiver operating characteristic (ROC) of the four mentioned methods together with that of the proposed method. The ROCs show the probability of detection (P_d) versus the probability of false alarms (P_d) where P_d is the probability that a signal frame is classified as H_1 and P_fa is the probability that a noise frame is classified as H_1. These ROCs are obtained by 100 repetitions of the experiment with the same speech signal and different white Gaussian noise. From these curves, it is obvious that our method outperforms the VADs proposed in [3] and [4] for all values of probability of false alarm. Fig. 6 also shows that although the performance of the method proposed in [5] and [6] is a little bit better than that of the proposed method for low probability of false alarm, the performance of the proposed method is much better than that of the method proposed in [5] and [6] for high probability of false alarm. In Table V, we also provide the probability of detection for different values of probability of false alarm and different SNR levels for the above mentioned methods. In Table V, the best performance in each column (i.e., different SNRs) is indicated by bold numbers. As can be seen from the data in this table, the performance of the proposed method is higher than that of other methods except for low probability of false alarm in high SNRs, as mentioned before. As can be seen from the first column of Table V, the proposed VAD

also outperforms the reference VADs even in low probability of false alarm in low SNRs.

In the third experiment, the effect of the order of the autoregressive part on the performance of the VAD is investigated. The ROC curves for different values of p are depicted in Fig. 9. These curves are obtained by the same setup explained in the second experiment (i.e., white Gaussian noise and SNR=5 dB and the same set of parameters for the proposed VAD) but a different speech signal is used. The speech signals were again chosen from the TIMIT database, half from male and half from female speakers with total length of 20s. As can be seen from this figure, the performances of the different AR-GARCH models are approximately the same for $1 \leq p \leq 4$ and get worse for a higher order. This is because increasing the order (increasing the number of parameters) makes the estimation of the parameters more difficult, and larger error in estimation of the parameters impairs the performance of the proposed VAD. Thus, we chose the first order in other simulations because this model is computationally simpler than higher order models.

In the fourth experiment, we investigate the effect of other kinds of noise on the performance of the proposed VAD. In this simulation the parameters of the proposed VAD are chosen from Table II. Two noise types (factory and babble) are employed in this experiment. These two noise types are generally difficult to deal with. For all the reference methods, the IMCRA method is used for estimating the variance of the noise in the STFT domain. Figs. 7 and 8 show the receiver operating characteristic (ROC) of the four mentioned methods together with that of the proposed method for factory and babble noise, respectively. We also provide in Tables VI and VII detection probability for different values of probability of false alarm and for different SNR levels. In these tables, the best performance in each column (i.e., different SNRs) is indicated by bold numbers. The figures and tables are obtained using the same method explained in the second experiment. The advantage of the proposed method over the reference methods is obvious from these results.

In the final experiment, we evaluate the performance of the proposed VAD in presence of an additive sinusoidal component. A 20s long speech signal is chosen from the TIMIT database and corrupted by a sinusoidal signal of 60 Hz frequency. The SNR is set to 5 dB. The ROC curves for the reference methods are depicted in Fig. 10. For all reference methods, the IMCRA method is used for estimating the variance of the additive noise. We also provide in Table VIII detection probability for different values of probability of false alarm and for different SNR levels. In this table, the best performance in each column (i.e., different SNRs) is indicated by bold numbers. It is apparent that the performance of the proposed VAD is substantially higher than that of the reference VADs.

VI. Conclusions

We have presented a new VAD based on AR(p)-GARCH $(1,1)$ modeling of the speech signal. We also introduced a novel procedure based on the ML estimation method for parameter estimation of the AR$(p)-$GARCH $(1,1)$ model in

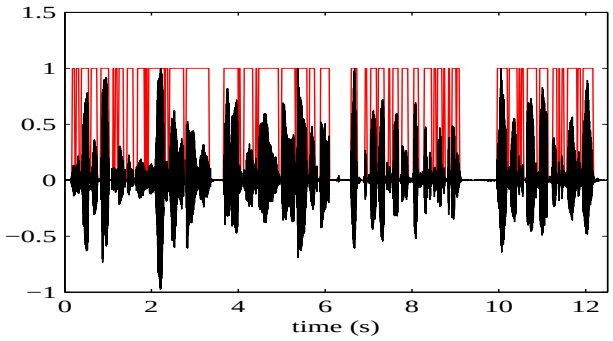

Fig. 5. Clean speech signal together with hand-marked VAD

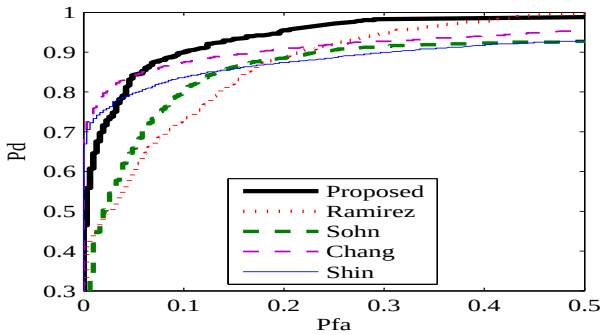

Fig. 6. Comparison of ROC curves of the proposed method with those of reference methods for Gaussian noise and SNR=5dB

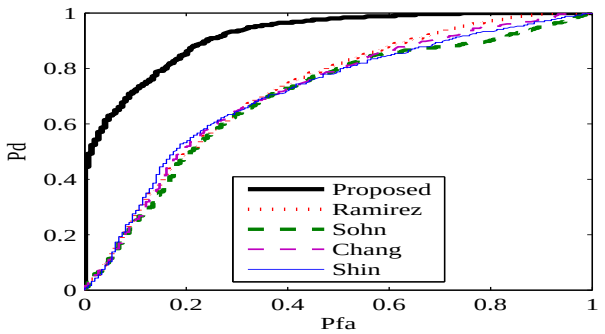

Fig. 7. Comparison of ROC curves of the proposed method with those of reference methods for factory noise and SNR=5dB

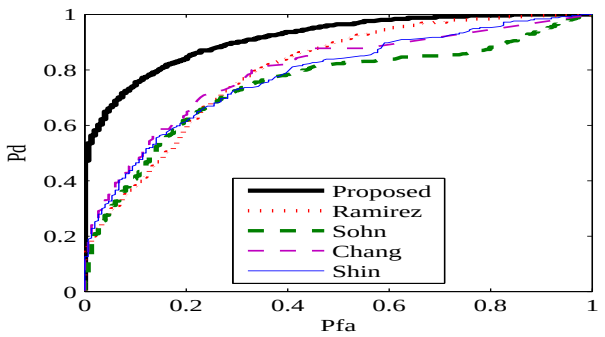

Fig. 8. Comparison of ROC curves of the proposed method with those of reference methods for babble noise and SNR=5dB

TABLE V
PROBABILITY OF DETECTION VERSUS PROBABILITY OF FALSE ALARM FOR DIFFERENT SNR LEVELS AND DIFFERENT METHODS (WHITE NOISE)

	$P_{fa}=5\%$			$P_{fa}=10\%$			$P_{fa}=20\%$			$P_{fa}=40\%$		
SNR	0db	5db	10db	0db	5db	10db	0db	5db	10db	0db	5db	10db
$P_d$for Proposed Method	**80.49**	**84.12**	85.93	**84.43**	**87.10**	89.13	**87.31**	**89.66**	90.72	**87.63**	**91.47**	**92.96**
$P_d$for Ramirez Method	51.60	57.78	64.18	61.09	64.61	70.04	69.51	73.13	76.97	73.99	77.83	81.56
$P_d$for Sohn Method	56.93	66.52	69.51	64.29	72.92	75.69	69.40	77.51	80.28	71.43	79.32	81.77
$P_d$for Chang Method	67.59	82.52	**88.70**	70.36	84.22	**90.19**	73.88	86.25	**91.26**	77.83	88.27	92.43
$P_d$for Shin Method	63.33	78.68	86.99	65.99	80.92	88.27	69.62	82.94	89.45	73.88	84.97	90.41

TABLE VI
PROBABILITY OF DETECTION VERSUS PROBABILITY OF FALSE ALARM FOR DIFFERENT SNR LEVELS AND DIFFERENT METHODS (FACTORY NOISE)

	$P_{fa}=5\%$			$P_{fa}=10\%$			$P_{fa}=20\%$			$P_{fa}=40\%$		
SNR	0db	5db	10db	0db	5db	10db	0db	5db	10db	0db	5db	10db
$P_d$for Proposed Method	**54.69**	**73.03**	**81.34**	**59.06**	**76.87**	**84.86**	**64.39**	**80.28**	**87.31**	**70.04**	**82.94**	**88.59**
$P_d$for Ramirez Method	6.82	14.29	23.24	12.05	21.32	30.92	20.90	31.98	40.09	35.93	45.31	51.81
$P_d$for Sohn Method	7.68	13.86	22.81	12.47	21.00	30.92	21.00	31.77	40.94	34.75	43.92	52.24
$P_d$for Chang Method	7.25	14.50	26.23	12.47	22.17	34.33	22.39	33.26	43.39	35.82	44.46	52.67
$P_d$for Shin Method	7.46	14.71	26.87	14.61	23.56	35.18	24.63	33.48	44.67	36.89	44.78	53.52

TABLE VII
PROBABILITY OF DETECTION VERSUS PROBABILITY OF FALSE ALARM FOR DIFFERENT SNR LEVELS AND DIFFERENT METHODS (BABBLE NOISE)

	$P_{fa}=5\%$			$P_{fa}=10\%$			$P_{fa}=20\%$			$P_{fa}=40\%$		
SNR	0db	5db	10db	0db	5db	10db	0db	5db	10db	0db	5db	10db
$P_d$for Proposed Method	**53.20**	**73.13**	**80.81**	**57.89**	**75.59**	**84.22**	**63.22**	**79.32**	**86.67**	**69.08**	**82.73**	**88.38**
$P_d$for Ramirez Method	27.29	35.61	40.72	33.90	41.90	46.16	43.07	50.11	53.30	53.30	60.98	65.14
$P_d$for Sohn Method	27.08	34.97	38.38	33.48	42.22	45.74	42.43	51.17	55.54	50.85	58.53	61.94
$P_d$for Chang Method	28.78	39.77	44.14	34.65	45.95	52.45	43.39	55.76	60.23	51.60	61.09	62.47
$P_d$for Shin Method	27.93	39.34	44.56	34.01	45.42	51.28	41.26	52.45	59.49	49.15	58.74	62.15

TABLE VIII
PROBABILITY OF DETECTION VERSUS PROBABILITY OF FALSE ALARM FOR DIFFERENT SNR LEVELS AND DIFFERENT METHODS (60 HZ ADDITIVE SINUSOIDAL COMPONENT)

	$P_{fa}=5\%$			$P_{fa}=10\%$			$P_{fa}=20\%$			$P_{fa}=40\%$		
SNR	0db	5db	10db	0db	5db	10db	0db	5db	10db	0db	5db	10db
$P_d$for Proposed Method	**88.06**	**88.91**	**90.09**	**90.51**	**91.04**	**92.32**	**91.26**	**93.18**	**94.78**	**93.92**	**94.56**	**95.10**
$P_d$for Ramirez Method	44.99	45.84	49.89	49.79	51.39	54.48	57.57	58.32	61.30	69.08	69.72	72.28
$P_d$for Sohn Method	12.79	18.55	29.96	21.54	29.10	39.98	36.67	42.22	52.67	52.03	55.01	60.45
$P_d$for Chang Method	25.59	36.46	41.15	36.03	44.24	49.36	47.55	56.08	61.62	61.09	65.67	66.74
$P_d$for Shin Method	34.33	39.02	46.06	44.14	48.51	54.48	55.33	60.13	63.22	64.61	66.31	69.19

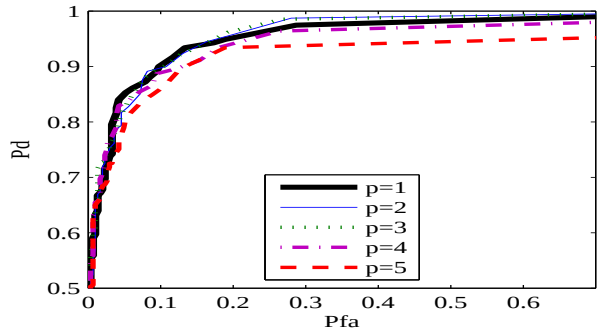

Fig. 9. ROC curves of the proposed VAD method for different AR orders

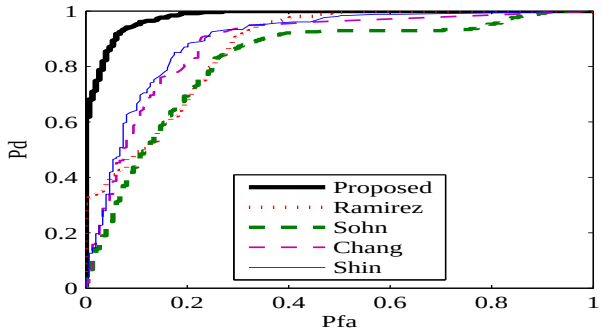

Fig. 10. Comparison of ROC curves of the proposed method with those of reference methods in presence of a 60 Hz Additive Sinusoidal Component and SNR=5dB

presence of additive noise. We presented an adaptive version of parameter estimation method, namely, the RML method. In this method, upon receiving a new sample, we updated the likelihood function together with its gradient vector and used the steepest descent method to numerically find the maximum of the log-likelihood function. The proposed VAD is obtained by comparing the likelihood ratio to a threshold. The likelihood ratio of the observations was computed using the estimated variance of the speech signal, which was obtained by the RML algorithm. We used the IMCRA method for estimating the variance of the additive noise, which can be used under nonstationary conditions. This enabled our VAD

to follow variations in the variance of the additive noise. Simulation results have demonstrated the high performance of the proposed method and particularly its advantage in nonstationary noise environments.

APPENDIX: COMPUTATION OF THE GRADIENT VECTOR

In this apendix, we compute the gradient of the pdf of the current observation conditioned on past observations, i.e., $\mathbf{g}_t = \nabla f\left(y_t|y_1^{t-1};\boldsymbol{\theta}\right)$. To find the pdf of the current observation conditioned on past observations, i.e., $f\left(y_t|y_1^{t-1};\boldsymbol{\theta}\right)$, note that the sequence of the clean signal is independent of the noise sequence; hence, using (11) and (12) we have

$$\left(y_t = x_t + n_t|y_1^{t-1};\boldsymbol{\theta}\right) \sim \mathcal{N}\left(\sum_{i=1}^{p}\alpha_i\hat{x}_{t-i}, \hat{\sigma}^2_{t|t-1} + \sigma^2\right) \quad (36)$$

or equivalently

$$\log f\left(\left(y_t|y_1^{t-1};\boldsymbol{\theta}\right)\right) = \log\frac{1}{\sqrt{2\pi}} - \frac{1}{2}\log\left(\hat{\sigma}^2_{t|t-1}+\sigma^2\right) - \frac{1}{2}\frac{\left(y_t - \sum_{i=1}^{p}\alpha_i\hat{x}_{t-i}\right)^2}{\hat{\sigma}^2_{t|t-1}+\sigma^2}. \quad (37)$$

For ease of notation, we define three intermediate variables as follows

$$\xi_{\alpha_i} = \frac{\partial u_{t-1}}{\partial\alpha_i} = -2\left(y_{t-1} - \sum_{i=1}^{p}\alpha_i\hat{x}_{t-1-i}\right)\left(\frac{\hat{\sigma}^2_{t-1|t-2}}{\hat{\sigma}^2_{t-1|t-2}+\sigma_t^2}\right)^2\hat{x}_{t-i-1} \quad (38)$$

$$\chi = \frac{\partial\log\left(f\left(y_t|y_0^{t-1};\boldsymbol{\theta}\right)\right)}{\partial\hat{\sigma}^2_{t|t-1}} = \frac{\partial\log\left(f\left(y_t|y_0^{t-1};\boldsymbol{\theta}\right)\right)}{\partial\sigma^2} = \frac{-1}{2}\left(\frac{\hat{\sigma}^2_{t|t-1}+\sigma_t^2-\left(y_t-\sum_{i=1}^{p}\alpha_i\hat{x}_{t-i}\right)^2}{\left(\hat{\sigma}^2_{t|t-1}+\sigma_t^2\right)^2}\right) \quad (39)$$

and

$$\chi_{\alpha_i} = \frac{\partial\chi}{\partial\alpha_i} = -\frac{y_t-\sum_{i=1}^{p}\alpha_i\hat{x}_{t-i}}{\left(\hat{\sigma}^2_{t|t-1}+\sigma_t^2\right)^2}x_{t-i} - \frac{1}{2}\beta_1\xi_{\alpha_i}\frac{2\left(y_t-\sum_{i=1}^{p}\alpha_i\hat{x}_{t-i}\right)-\left(\sigma^2_{t|t-1}+\sigma_t^2\right)}{\left(\hat{\sigma}^2_{t|t-1}+\sigma_t^2\right)^3}. \quad (40)$$

The gradient of the log-likelihood function of the current observation conditioned on past observations, i.e., $\mathbf{g}_t = \nabla\log\left(f\left(y_t|y_0^{t-1};\boldsymbol{\theta}\right)\right)$ can be computed as follows

$$\begin{aligned}\left[g_{t,1},g_{t,2},g_{t,3}\right]^T &= \chi\mathbf{a}\\ g_{t,i+3} &= \frac{\partial}{\partial\alpha_i}\left(\log\left(f\left(y_t|y_0^{t-1};\boldsymbol{\theta}\right)\right)\right)\\ &= \frac{y_t-\sum_{i=1}^{p}\alpha_i\hat{x}_{t-i}}{\hat{\sigma}^2_{t|t-1}+\sigma_t^2}\hat{x}_{t-i} + \chi\beta_1\xi_{\alpha_i}\\ &\quad ;i=1,2,...,p\end{aligned} \quad (41)$$

where $g_{t,i}$ is the i-th element of vector $\mathbf{g}_t$ and $\mathbf{a} = \left[1, u_{t-1}, \hat{\sigma}^2_{t-1|t-2}\right]^T$.

ACKNOWLEDGEMENT

The authors thank the anonymous reviewers for their constructive comments and helpful suggestions. The first author also thanks Rahmat and Marisa for their help in preparing this paper.

REFERENCES

[1] N. R. Garner, P. A. Barrett, D. M. Howard, and A. M. Tyrrell, "Robust noise detection for speech detection and enhancement," *Electron. Lett.*, vol. 33, pp. 270–271, Feb. 1997.

[2] A. Benyassine, E. Shlomot, H. Su, D. M. C. Lamblin, and J. Petit, "ITU-T recommendation G.729 annex B: a silence compression scheme for use with g.729 optimized forv.70 digital simultaneous voice and data applications," *IEEE Commun. Mag.*, vol. 35, pp. 64–73, Sep. 1997.

[3] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 16, pp. 1–3, Jan. 1999.

[4] J. Ramirez and J. C. Segura, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Lett.*, vol. 12, pp. 689–692, Oct. 2005.

[5] J. H. Chang and N. S. Kim, "Voice activity detection based on complex laplacian model," *Electron. Lett.*, vol. 39, no. 7, pp. 632–634, Apr. 2003.

[6] J. W. Shin, J. H. Chang, H. S. Yun, and N. S. Kim, "Voice activity detection based on generalized gamma distribution," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. I781–I784, 2005.

[7] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 412–424, Mar. 2006.

[8] I. Cohen, "Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models," *signal processing*, vol. 86, pp. 698–709, 2006.

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

[10] H. Solvang, K. Ishizuka, and M. Fujimoto, "Voice activity detection based on adjustable linear prediction and garch models," *Speech Communication*, vol. 50, pp. 476–486, 2008.

[11] S. M. J. Benesty and J. Chen, *Speech Enhancement*, S. M. J. Benesty and J. Chen, Eds. Springer, Berlin, Germany, 2005.

[12] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. on Speech and Audio proc.*, vol. 11, pp. 466–475, Oct. 2003.

[13] T. Bollerslev, "Generalized autoregressive conditional heteroscedasticity," *Journal of Econometrics*, vol. 31, pp. 307–327, 1986.

[14] R. Engle, "Autoregressive conditional heteroskedasticity with estimates of the variance of u.k. inflation," *Econometrica*, vol. 50, pp. 987–1008, 1982.

[15] W. Li, S. Ling, and M. McAleer, "Recent theoretical results for time series models with garch errors," *Journal of Economic Surveys*, vol. 16, pp. 245–269, 2002.

[16] 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) (prototype as of December 1988).

[17] P. Stoica and R. L. Moses, *Introduction to Spectral Analysis*, M. Horton, Ed. Prentice Hall, 1997.
[18] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation*, A. V. Oppenheim, Ed. Prentice Hall Signal Processing Series, 1993.
[19] D. G. Luenberger, *Linear and nonlinear programming*. Addison-Wesley Publishing Company, Inc, 1989.
[20] S. Mousazadeh and I. Cohen, "Simultaneous parameter estimation and state smoothing of complex garch process in the presence of additive noise," *Signal Processing*, vol. 90, pp. 2947–2953, Nov. 2010.
[21] V. Tadic and A. Doucet, "Exponential forgetting and geometric ergodicity for optimal filtering in general state-space models," *Stochastic Processes and Their Applications*, vol. 115, pp. 1408–1436, 2005.
[22] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*. MIT Press, Cambridge, 1983.
[23] P. D. Tuan, "Maximum likelihood estimation of the autoregressive model by relaxation on the reflection coefficients," *IEEE Transaction on Acoustics, Speech, and Signal Processing*, vol. 36, pp. 1363–1367, 1988.
[24] S. Gazor and W. Zhang, "A soft voice activity detector based on a laplacian gaussian model," *IEEE Trans. on Speech and Audio proc.*, vol. 11, pp. 498–505, Sep. 2003.

Saman Mousazadeh was born in Shiraz, Iran, in 1982. He received the B.S. and M.S. degrees, both in electrical engineering, from Shiraz University, Shiraz, Iran, in 2005 and 2008, respectively.

He is currently pursuing the M.Sc. degree in electrical engineering at the Technion – Israel Institute of Technology, Haifa, Israel. Since 2009, he has been a Project Supervisor with the Signal and Image Processing Lab (SIPL), Electrical Engineering Department, Technion. His research interests include speech, image and array signal processing. He received the student challenge award of the Acoustical Society of America (ASA) in 2006.

Israel Cohen (M'01-SM'03) received the B.Sc. (Summa Cum Laude), M.Sc. and Ph.D. degrees in electrical engineering from the Technion – Israel Institute of Technology, Haifa, Israel, in 1990, 1993 and 1998, respectively.

From 1990 to 1998, he was a Research Scientist with RAFAEL Research Laboratories, Haifa, Israel Ministry of Defense. From 1998 to 2001, he was a Postdoctoral Research Associate with the Computer Science Department, Yale University, New Haven, CT. In 2001 he joined the Electrical Engineering Department of the Technion, where he is currently an Associate Professor. His research interests are statistical signal processing, analysis and modeling of acoustic signals, speech enhancement, noise estimation, microphone arrays, source localization, blind source separation, system identification and adaptive filtering. He is a coeditor of the Multichannel Speech Processing section of the *Springer Handbook of Speech Processing* (Springer, 2007), a coauthor of *Noise Reduction in Speech Processing* (Springer, 2009), and a cochair of the 2010 International Workshop on Acoustic Echo and Noise Control.

Dr. Cohen received in 2009 the Muriel and David Jacknow award for Excellence in Teaching, and in 2010 the Alexander Goldberg Prize for Excellence in Research. He served as Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and IEEE SIGNAL PROCESSING LETTERS, and as Guest Editor of a special issue of the *EURASIP Journal on Advances in Signal Processing* on Advances in Multimicrophone Speech Processing and a special issue of the *EURASIP Speech Communication Journal* on Speech Enhancement.