A Fast and Soft Voice Activity Detector Using GARCH Model and Markov Model Chain

Sadegh Rezaei and Rasool Tahmasbi

ABSTRACT

This paper presents a robust algorithm for Voice Activity Detector (VAD) based on GARCH (Generalized Autoregressive Conditional Heteroscedasticity) Model, Variance Gamma Distribution (VGD) and binary Markov model.

GARCH models are new statistical methods that are used especially in economic time series. There is a consensus that speech signals exhibit variances that change through time. GARCH models are a popular choice to model these changing variances.

Speech signal is assumed to have VGD because the VGD has heavier tails than the Gaussian Distribution (GD) and the distribution of noise signal is assumed to be Gaussian.

In the proposed method, heteroscedasticity will be modeled by GARCH and then the parameters of the distributions will be estimated recursively. Finally, using a binary Markov model and comparing it with an adaptive threshold, leads to the derivation of soft and hard detection.

The simulation results show that the proposed VAD is able to operate down to -5 dB and in nonstationary environments.

KEYWORDS

Estimation Theory, GARCH Model, Heteroscedasticity, Markov Model Chain, Probability Distribution, Voice Activity Detection.

1. INTRODUCTION

Voice activity detection (VAD) refers to the ability of distinguishing voice from noise and is an integral part of a variety of speech communication system, such as speech coding [1], speech recognition, audio conferencing, hands-free telephony [2], speech enhancement [3], wireless communication [4], [5] and echo cancellation.

During the last years, numerous researchers have studied different strategies for detecting speech in noise and the influence of the VAD on the performance of speech processing systems. Sohn [1] proposed a robust VAD algorithm based on a statistical Likelihood Ratio Test (LRT) involving a single observation vector. Later, Cho [6] suggested an improvement based on a smoothed LRT.

It has been shown recently [7]–[8] that incorporating long-term speech information to the decision rule reports benefits for speech/pause discrimination in high noise environments. For example, Ramírez [9] proposed an LRT involving multiple and independent observations.

In [10] the method is a little different. But they incorporate long-term information, too: the signal is first decorrelated using an orthogonal transformation and then a Hidden Markov Model (HMM) is employed. Also, they assumed that the distribution of speech is Laplacian, because, in [11], it is shown that speech signal has Laplacian Distribution (LD). In this paper, we assume that the speech signal has a Variance Gamma Distribution (VGD), since it is a generalization of LD and in specific case ($\lambda = 1$) VGD becomes LD.

Tahmasbi and Rezaei [12] used a GARCH model with an empirical adaptive threshold function. They showed that the empirical adaptive threshold function has better results than of [10].

Note that approaches of [1] and [9] are performed in frequency domain, but [10] is performed in time domain. However it takes time to decorrelate signal via orthogonalization. Our proposed method is performed in time domain; like [10], it is necessary that the signal be uncorrelated. To decorrelate signals, we used Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model, which could model both noise and speech heteroscedasticity. However, estimating of GARCH parameters are time-consuming. So, a predefined estimation of the parameters is presented to solve this

S. Rezaei is with the Department of Statistics, Ahvaz University, Ahvaz, Iran (e-mail: srezaei@aut.ac.ir).
R. Tahmasbi, is student of Mathematical Statistics, Amirkabir University of Technology, Tehran, Iran (e-mail: rtahmaseby@yahoo.com).

problem.

In section 2, the GARCH model is introduced and is shown that it can model heteroscedasticity and also it is shown that every GARCH series are uncorrelated. In section 3, we review elements of Markov Model. Section 4 presents statistical aspects of speech and noise and adaptive threshold function. In section 5, algorithm of the proposed method is presented. And section 6 illustrates the experimental results of proposed method.

2. GARCH Model

GARCH models are new statistical methods that are used especially in economic time series. GARCH stands for Generalized Autoregressive Conditional Heteroscedasticity. Loosely speaking, you can think of heteroscedasticity as time-varying variance (i.e., volatility). Conditional implies a dependence on the observations of the immediate past, and autoregressive describes a feedback mechanism that incorporates past observations into the present. GARCH then is a mechanism that includes past variances in the explanation of future variances. More specifically, GARCH is a time series modeling technique that uses past variances and past variance forecasts to forecast future variances.

Definition: Let (Z_t) be a sequence of i.i.d. random variables such that Z_t have standard Gaussian distribution. (Y_t) is called GARCH(q,p) process if

$$Y_t = \sigma_t Z_t \qquad t \in \mathbb{Z} \tag{1}$$

where (σ_t) is a nonnegative process such that

$$\sigma_t^2 = \alpha_0 + \alpha_1 Y_{t-1}^2 + \ldots + \alpha_q Y_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + \ldots + \beta_p \sigma_{t-p}^2 \tag{2}$$

and

$$\alpha_0 > 0;\ \alpha_i \geq 0,\ i = 1,\ldots,q;\ \ \beta_i \geq 0,\ i = 1,\ldots,p. \tag{3}$$

For $p = 0$ the process reduced to ARCH(q) (Autoregressive Conditional Heteroscedasticity of order q).

In ARCH(q) processes, the conditional variance is specified as a linear function of past sample variances only, whereas the GARCH(p,q) process allows lagged conditional variances to enter as well. This corresponds to some sort of learning mechanism [13].

To review elementary aspects of GARCH model, denote $E_{t-1}[.]$ as conditional expectation while condition is on the past information up to time $t-1$ (which is denoted by ϕ_{t-1}. I.e. $\phi_{t-1} = \sigma\{Z_{t-1}, Z_{t-2}, \ldots\}$, where $\sigma\{.\}$ is the sigma field generated by $\{.\}$). So

$$E_{t-1}[.] = E[. \mid \phi_{t-1}] \tag{4}$$

and conditional variance is

TABLE 1
MEAN OF GARCH PARAMETERS

	SNR(db)				
	15	10	5	0	-5
$\hat{\alpha}_0$	0.0001	0.0003	0.0006	0.0010	0.0015
$\hat{\alpha}_1$	0.3325	0.2490	0.1449	0.0672	0.0249
$\hat{\beta}_1$	0.6654	0.7316	0.8209	0.8958	0.9462

TABLE 2
VARIANCE OF GARCH PARAMETERS

	SNR(db)				
	15	10	5	0	-5
$\hat{\alpha}_0$	0.0001	0.0001	0.0001	0.0001	0.0001
$\hat{\alpha}_1$	0.0216	0.0208	0.0098	0.0047	0.0008
$\hat{\beta}_1$	0.0232	0.0211	0.0147	0.0096	0.0034

TABLE 3
PROBABILITY OF TRUE DETECTION (%)

SNR	15	10	5	0	-2	-5
P	98.52	97.61	97.04	94.28	90.74	75.91

TABLE 4
PROBABILITY OF TRUE DETECTION (%)

		Sohn	Gazor	Ramirez	Proposed
SNR (db)	5	91.05	94.21	95.35	97.01
	0	84.62	79.36	87.85	94.34
	-2	73.28	69.38	80.26	90.69

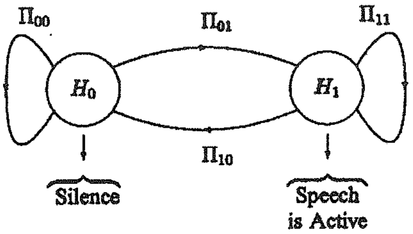

Figure 1: Binary Markov Model for VAD.

$$V_{t-1}[.] = V[. \mid \phi_{t-1}] \tag{5}$$

Now suppose (y_t) is a GARCH process and

$$\sum_{i=1}^{q} \alpha_i - \sum_{j=1}^{p} \beta_i < 1$$

So,

$$E(y_t) = E(\sigma_t Z_t) = E(\sigma_t)E(Z_t) = 0 \tag{6}$$

$$V(y_t) = E(y_t^2) = E(\sigma_t^2 Z_t^2) = E(\sigma_t^2)$$
$$= \alpha_0 + \sum_{i=1}^{q} \alpha_i E(y_t^2) + \sum_{j=1}^{p} \beta_i E(\sigma_t^2)$$

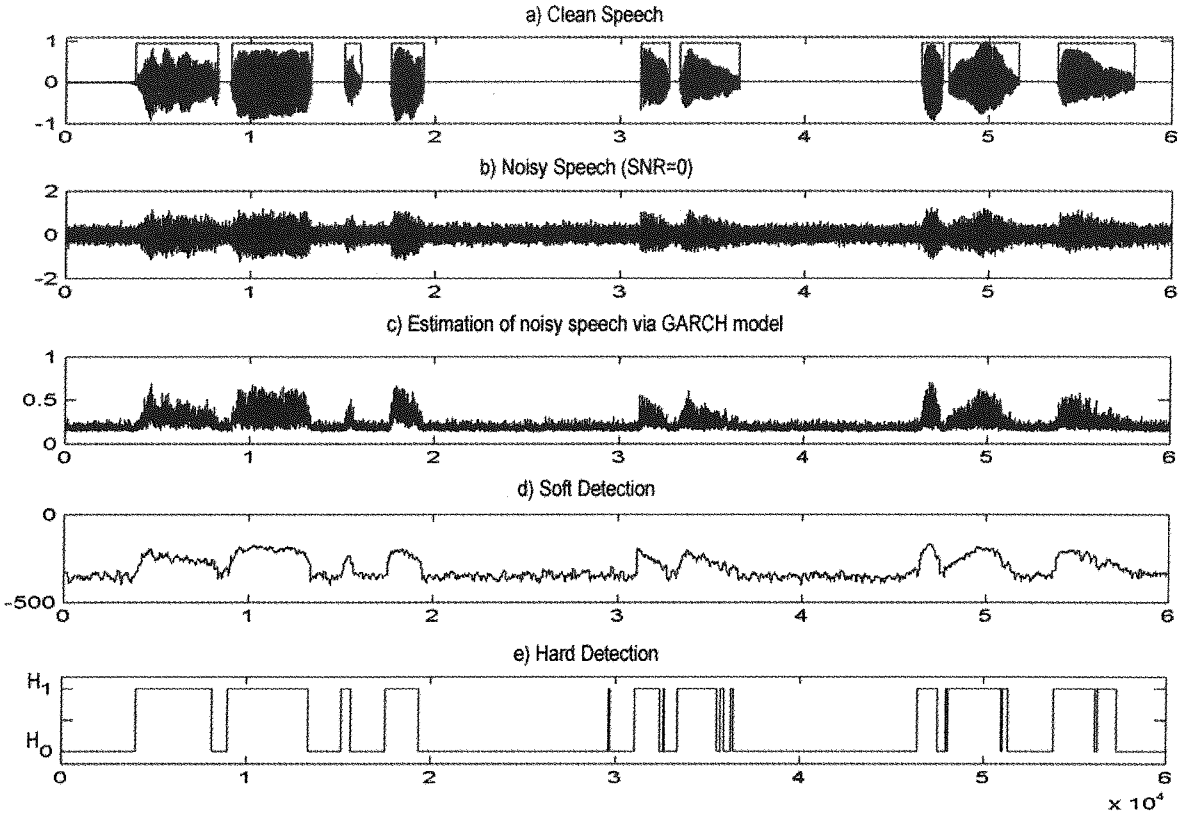

Figure 2: Results of the proposed VAD with 0 db SNR. (a) Clean speech. (b) Noisy speech with zero SNR. (c) Estimation of noisy speech via GARCH Model (d) Soft detection. (e) Hard detection.

$$\Rightarrow V(y_t)=\frac{\alpha_0}{1-\sum_{i=1}^{q}\alpha_i-\sum_{j=1}^{p}\beta_i} \qquad (7)$$

and its conditional expectation and variance is:

$$E[y_t \mid \phi_{t-1}]=E[\sigma_t Z_t \mid y_{t-1},y_{t-2},\ldots]=\sigma_t E[Z_t]=0 \quad (8)$$

$$\begin{aligned} V[y_t \mid \phi_{t-1}] &= E[y_t^2 \mid y_{t-1},y_{t-2},\ldots] \\ &= E[\alpha_0+\sum_{i=1}^{q}\alpha_i y_{t-i}^2+\sum_{j=1}^{p}\beta_i\sigma_{t-j}^2 \mid y_{t-1},y_{t-2},\ldots] \\ &= \alpha_0+\sum_{i=1}^{q}\alpha_i y_{t-i}^2+\sum_{j=1}^{p}\beta_i\sigma_{t-j}^2=\sigma_t^2. \qquad (9) \end{aligned}$$

As you can see, the mean and variance of a GARCH process are constant but its conditional variance is changing over time. So, in comparison with Autoregressive Moving Average (ARMA) process, GARCH could model the time-variation of variance (i.e., volatility).

Other interesting properties of GARCH process are:

i) $Cov(Y_t, Y_{t-k})=0, \quad \text{for } k>0;$

ii) (Y_t) has heavier tails than the Gaussian Distribution (GD).

Note that property (i) together with normality assumption of noises ensure that the (Y_t) are independent.

As [14] discussed when the mean level of a series stayed close to zero over the entire period and changes in variance (volatility) occurred, then this series could model through GARCH. It is clear that the speech signals have these properties (Fig 2.b) and several examples showed that it could be modeled through GARCH(1,1) and for example [15] used a GARCH(1,1) for modeling speech. So, we used a GARCH(1,1) to model the speech signal. Therefore,

$$\hat{\sigma}_t^2=\hat{\alpha}_0+\hat{\alpha}_1 Y_{t-1}^2+\hat{\beta}_1\hat{\sigma}_{t-1}^2 \qquad (10)$$

where $\hat{\alpha}_0$, $\hat{\alpha}_1$ and $\hat{\beta}_1$ should be estimated. (We call *garchfit* in Matlab to calibrate GARCH model).

The mean and variance of these three parameters for 100 different speeches with different SNR are shown in Table 1 and 2. As you see, the estimation values are related to the amount of SNR, i.e., less SNR more β_1 and less α_1 and visa versa.

From now on, let us denote $x(t)$ as estimation of noisy speech via GARCH. I.e. $x(t)=\hat{\sigma}_t$.

3. Binary Markov Model

As the correlation of signal samples in consecutive frames is strong, the sequences of frame hypothesis states can be modeled as a first-order Markov process, which simply assumes that the current state only depends on the current data and the previous state. The use of Markov model can also prevent clipping effect of weak speech, because when decision is made, the previous state is also taken into consideration.

Here, we use a two-state discrete (binary) Markov model to estimate the probability of Voice Being Active (VBA) at each time frame. As shown in Fig.1, these states are H_0 (Silence) and H_1 (Speech, i.e., VBA). To obtain this, we consider that the sequence of hypothesis is a binary Markov process with the following probability transition matrix:

$$\Pi = \begin{bmatrix} \Pi_{00} & \Pi_{10} \\ \Pi_{01} & \Pi_{11} \end{bmatrix} \tag{11}$$

where Π_{00} denotes the probability of silence when there was silence in the previous time instance, Π_{10} denotes the probability of silence when there was speech in the previous time instance and so on. In HMM, this matrix is often determined by $\Pi_{ij} = T_{ij} / \sum_q T_{iq}$, where T_{pq} is a measured number of transitions from p th state to q th state [10].

4. Statistical Aspects

After estimating noisy speech through GARCH model, the result is a series of data that have heavier tails than the Gaussian and are independent - by normality assumption of noises. In such situation some distribution like VGD is used [17]-[18].

For distinguishing speech or silence we do as follows: Assume that X_t is an m-dimensional vector of data which is estimation of noisy speech via GARCH model at time t, i.e.,

$$X_t = [x(t), x(t-1), \ldots, \; x(t-m+1)]^T \tag{12}$$

where $(.)^T$ denotes the transposing operation.

Then, we use the below hypothesis via binary Markov model:

$$\begin{cases} H_0 : X_t \;\; \textit{is silence} \\ H_1 : X_t \;\; \textit{is speech} \end{cases} \tag{13}$$

Before starting the next section, we review the elementary features of the distribution of speech and noise, prior and posterior voice activity probabilities.

A. Distribution of speech

A random variable X is said to be Variance Gamma Distribution (VGD) with parameters λ, α, β and μ if its density is given by

$$f_X(x) = C.\left|x-\mu\right|^{\lambda-0.5} K_{\lambda-0.5}(\alpha\left|x-\mu\right|)e^{\beta\,(x-\mu)} \tag{14}$$

with

$$C = \frac{\gamma^{2\lambda}}{\sqrt{\pi}\Gamma(\lambda)(2\alpha)^{\lambda-0.5}}, \;\; \gamma^2 = \alpha^2 - \beta^2 \tag{15}$$

$K_\lambda(.)$ is the modified Bessel function of the third kind (see [19], Ch. 11) and $\Gamma(.)$ is gamma function. The parameter domain is restricted to $\lambda > 0$ and $\alpha > |\beta|$. If $\beta = 0$ then the distribution is symmetric and α, λ are shape parameters. The moment generating function of X is given by

$$L(z) = e^{\mu\, z}(\frac{\gamma}{\gamma_z})^{2\lambda}, \;\; z \in R \tag{16}$$

where $\gamma_z = \sqrt{\alpha^2 - (\beta + z)^2}$. Therefore, we obtain the centralized moments by $L^{(n)}(0) = E(X^n)$:

$$E(X) = \mu + 2\frac{\beta\lambda}{\gamma^2} \tag{17}$$

$$V(X) = \frac{2\lambda}{\gamma^2}(1 + 2(\frac{\beta}{\gamma})^2) \tag{18}$$

In [20] the estimation of the parameters of VGD via moment matching method is presented. Denote υ and κ as skewness and kurtosis, respectively [20]. So their estimation is:

$$\hat{\lambda} = {}^{3}\!/_{\kappa} - 3 \tag{19}$$

$$\hat{\beta} = \frac{\upsilon}{(\kappa - 3)\sqrt{V(X)}} \tag{20}$$

$$\hat{\alpha} = \sqrt{\frac{6}{(\kappa - 3)V(X)} + \frac{\upsilon^2}{(\kappa - 3)^2 V(X)}} \tag{21}$$

$$\hat{\mu} = E(X) - \frac{\upsilon\sqrt{V(X)}}{\kappa - 3} \tag{22}$$

```
Assume that the vector Y is noisy speech and the vector X is its estimated GARCH model
and $\alpha_0,\alpha_1,\beta_1$ are brought out from Table 1 (or by garchfit command in Matlab).

$x(1)=\sqrt{|y(1)|}$;
For $t=2$ to $N$ (length of Y)
    $x(t)=sqrt(\alpha_0+\alpha_1 y^2(t-1)+\beta_1 x^2(t-1))$;          Eq. (10)
    $\hat{\mu}_t=$ Mean of $X_t$;                                      for $t>m$ use (30)
    $\hat{\sigma}_t^2=$ Variance of $X_t$;                             for $t>m$ use (31)
    $\hat{\alpha}_t=sqrt(\frac{2\lambda}{\hat{\sigma}_t^2})$;          Eq. (28)
    compute $L_m(X_t)$                                                 Eq. (43)
    compute $P_{t|t}=\frac{L_m(X_t)P_{t|t-1}}{L_m(X_t)P_{t|t-1}+(1-P_{t|t-1})}$     Eq. (42)
    $Soft\_Detection(t)=\log(P_{t|t}(X_t))-m\log(C)$;                  Eq. (49)
Next $t$
```

Figure 3:. Proposed VAD algorithm

In [20], it is shown that if υ(skewness) is close to zero, then these estimators could successfully obtain good approximation to λ,α,β,μ. On the other hand, in [11] it is shown that speech signal has a symmetric distribution ($\upsilon=0$). So

$$\hat{\beta}=0 \tag{23}$$

$$\hat{\mu}=E(X) \tag{24}$$

$$\hat{\lambda}={}^{3}\!/_{\kappa}-3 \tag{25}$$

$$\hat{\alpha}=\sqrt{\frac{2\hat{\lambda}}{V(X)}}. \tag{26}$$

As you can see, estimating of λ is related to κ(kurtosis). To simplify the computations, we assumed a fixed value for λ. For example, [10] assumed $\lambda=1$(LD is specific case of VGD), but one could consider all of the above equations.

Therefore, by assuming that speech distribution is symmetric, then for a fixed λ, the estimation of the parameters via the moment matching method is as follows:

If the vector X_t was estimated at time t, then the estimation of μ_t and α_t^2 is

$$\hat{\mu}_t=\overline{X}_t=1/m\sum_{i=0}^{m-1}x(t-i) \tag{27}$$

$$\hat{\alpha}_t=\sqrt{2\lambda/\hat{\sigma}_t^2} \tag{28}$$

$$\hat{\sigma}_t^2=V(X_t)=1/(m-1)[\sum_{i=0}^{m-1}x^2(t-i)-m\hat{\mu}_t^2] \tag{29}$$

and can be recursively computed via

$$\hat{\mu}_t=\hat{\mu}_{t-1}+R_t/m \tag{30}$$

$$\hat{\sigma}_t^2=\hat{\sigma}_{t-1}^2+R_t/(m-1)\ [T_t-R_t/m-2\hat{\mu}_{t-1}] \tag{31}$$

$$R_t=x(t)-x(t-m) \tag{32}$$

$$T_t=x(t)+x(t-m) \tag{33}$$

Note that R_t and T_t is defined to shorten above equations.

B. Distribution of Noise

We assume that the noise components are Gaussian. Therefore, its PDF is given by

$$f_X(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{34}$$

If the vector X_t was observed, then the estimation of μ_t and σ_t^2 are

$$\hat{\mu}_t=\overline{X}_t \tag{35}$$

$$\hat{\sigma}_t^2=V(X_t) \tag{36}$$

and can be recursively computed as same as above recursive estimations, i.e.,

$$\hat{\mu}_t = \hat{\mu}_{t-1} + R_t / m \tag{37}$$

$$\hat{\sigma}_t^2 = \hat{\sigma}_{t-1}^2 + R_t / (m-1) \, [T_t - R_t / m - 2\hat{\mu}_{t-1}] \tag{38}$$

C. *Prior and Posterior Voice Activity Probabilities*

The adaptive VAD algorithm starts with GARCH Model and then parameters of speech and noise are estimated. So, the prior probability, $P_{t|t-1}$ of VBA at time t is provided. i.e.,

$$\begin{aligned} P_{t|t-1} = P(&\text{speech is presented at time } t \\ &|\text{observation up to time } t\text{-}1). \end{aligned} \tag{39}$$

Note that a prior probability of VBA is initialized at $P_{1|0} = 1/2$ at time $t = 1$. Based on these assumptions the hypothesis (13) could be revised as follows:

$$\begin{cases} H_0 : X_t \text{ is silence with the prior} \\ \qquad \text{probability } (1 - P_{t|t-1}) \\ H_1 : X_t \text{ is speech with the prior} \\ \qquad \text{probability } P_{t|t-1} \end{cases} \tag{40}$$

Now the objective is to estimate recursively the posterior probability $P_{t|t}$

$$\begin{aligned} P_{t|t} \equiv P(&\text{speech is presented at time } t \\ &|\text{ observation up to time } t). \end{aligned} \tag{41}$$

This is soft decision rule that estimates the probability of VBA in each time instance, where $P_{t|t}$ and $P_{t|t-1}$ are estimated probability of VBA, with and without current information.

Based on the Baye's rule, we can easily derive soft VAD rule as follows:

$$P_{t|t} = \frac{L_m(X_t) P_{t|t-1}}{L_m(X_t) P_{t|t-1} + (1 - P_{t|t-1})} \tag{42}$$

where $L_m(X_t)$ is the Likelihood ratio at time t, i.e.,

$$L_m(X_t) = \prod_{i=0}^{m-1} \frac{P_{H_1}(x(t-i) | H_1)}{P_{H_0}(x(t-i) | H_0)} \tag{43}$$

where $P_{H_1}(. | H_1)$ and $P_{H_0}(. | H_0)$ are probability distribution function (PDF) of speech and noise, respectively. A hard decision rule is then given by

$$H(t) = \begin{cases} H_1 & \log(P_{t|t}(X_t)) \geq \eta \\ H_0 & \log(P_{t|t}(X_t)) < \eta \end{cases} \tag{44}$$

A prediction of the prior probability for next time instance, $P_{t+1|t}$ is required to perform the soft decision. This prediction is easily obtained based on the assumed Markov Model,

$$\begin{pmatrix} 1 - P_{t+1|t} \\ P_{t+1|t} \end{pmatrix} = \begin{pmatrix} \Pi_{00} & \Pi_{10} \\ \Pi_{01} & \Pi_{11} \end{pmatrix} \begin{pmatrix} 1 - P_{t|t} \\ P_{t|t} \end{pmatrix} \tag{45}$$

That is,

$$P_{t+1|t} = \Pi_{01}(1 - P_{t|t}) + \Pi_{11} P_{t|t} \tag{46}$$

D. *Adaptive Threshold*

As discussed before, speech is active at time t if $\log(P_{t|t}(X_t)) \geq \eta$ (based on HMM). It is better, however, to compare $\log(P_{t|t}(X_t))$ with a threshold function, η_t at each time t, i.e.,

$$X_t = \begin{cases} \text{speech} & \log(P_{t|t}(X_t)) \geq \eta_t \\ \text{Silence} & \log(P_{t|t}(X_t)) < \eta_t \end{cases} \tag{47}$$

It is natural to assume that the threshold function, η_t be a function of parameters of the speech distribution and m. An appropriate approach is to define

$$\eta_t = \tau + m \log(C) \tag{48}$$

where τ is a constant threshold and C is defined in (15) and its parameters had been estimated by X_t at time t. Note that (48) is obtained, empirically. So, speech is active if,

$$\log(P_{t|t}(X_t)) - m \log(C) \geq \tau \tag{49}$$

The left side of (49) can be viewed as a criterion for detecting silence or speech activity and is a criterion for soft detection. This criterion will be compared with τ and hard detection will be derived.

5. Proposed Voice Activity Detection Algorithm

For implementation of the proposed VAD, its algorithm is presented and shown in Fig 3.

This algorithm can divide into two parts: algorithms for $t \leq m$ and for $t > m$. Since for $t \leq m$, the length of vector X_t is less than m, so we can not use recursive Eq. (30-33) and (37-38) for estimating $\hat{\mu}_t, \hat{\sigma}_t^2$ and $\hat{\alpha}_t^2$. And should use (27-29) and (35-36) for $X = [x(t), x(t-1), \ldots, x(2), x(1)]$. Also, for computing soft detection, we use information up to time $t(< m)$. On the other hand, since log-likelihood is motivated by length of segment, so we multiply log-likelihood by m/t. But for $t > m$ since length of X_t is m so, use recursive (30-33) and (37-38).

6. Experimental Results

In this section, the results of the proposed method are presented. The speech signals are obtained from http://www.dailywave.com. Fig. 2(a) shows the clean speech together with manually marked clean speech sample. Fig. 2(b) shows the noisy speech with Gaussian noise and its SNR is equal to 0. GARCH estimation of noisy speech is shown in Fig. 2(c). Its estimated GARCH parameters are $\hat{\alpha}_0 = 0.002$, $\hat{\alpha}_1 = 0.12$ and $\hat{\beta}_1 = 0.83$. As discussed before, Table 1 and 2 show the mean and variance of GARCH parameters for different noisy speeches with Gaussian noise. Since estimation of GARCH parameters is related to SNR, so it is possible to use these approximated parameters instead of using an algorithm for finding exact estimation of the parameters.

The left side of (49) is computed and shown in Fig. 2(d). This criterion is compared with $\tau = -300$ and hard decision is derived. The hard detection is shown in Fig. 2(e).

To evaluate the performance of the proposed VAD, the speech and silence intervals marked manually then the hard decision of the VAD is compared with the manually marked intervals. For Fig 2, the Probability of True Detection (PTD) is equal to 0.94 and the results of the proposed method with different SNR are presented in Table 3. The PTD is more than 96 percent when SNR is greater than 2, and in very noisy cases (SNR= -5) is near 75 percent which are good results. So, the presented algorithm can be viewed as a robust algorithm.

The PTD, as evaluation criteria for the performance of the previous works and presented method are given in Table 4 with different SNR. As you see, the performance of the proposed method is more than Ramírez [9], Gazor [10] and Sohn [1]. Also, to prove the superiority of our algorithm, we examine our algorithm with non-Gaussian noises.

As same as Fig. 2, the results of proposed method with noises of t distribution (with 4 degree of freedom) and Beta Distribution (with parameters $\alpha = 2.5$ and $\beta = 1.5$) are shown in Fig. 4 and 5, respectively. Since Beta random numbers are not zero-mean, so we subtract their mean and then add these noises to clean speech signal. Also Fig. 6 shows the results of the proposed method with colored noise. To generate colored noise we passed white noise through a finite impulse response.

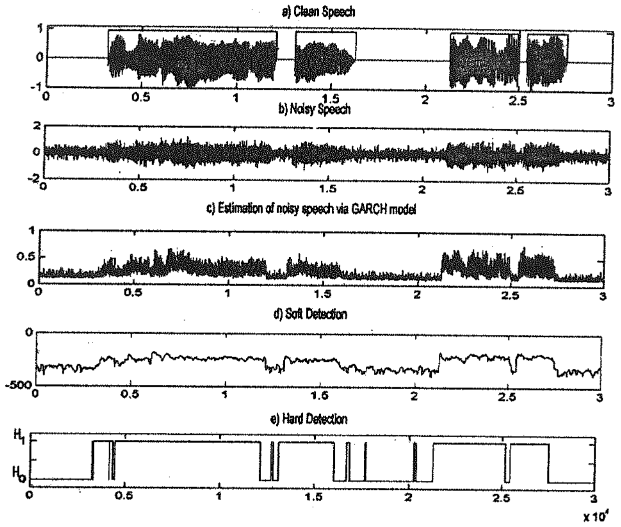

Figure 4: Results of the proposed VAD for t Distribution noise. (a) Clean speech. (b) Noisy speech. (c) Estimation of noisy speech via GARCH Model. (d) Soft detection. (e) Hard detection.

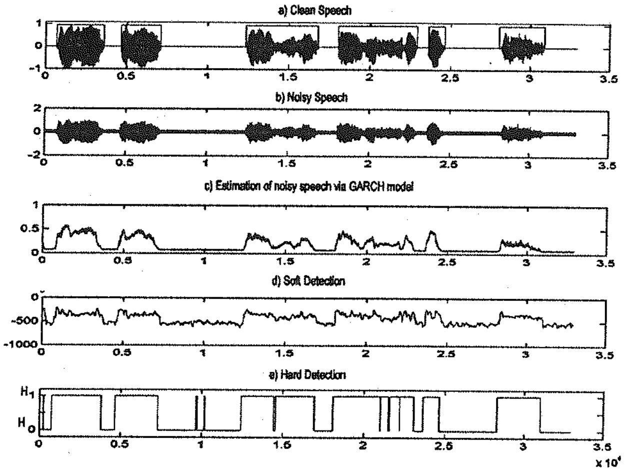

Figure 5: Results of the proposed VAD with Beta Distribution noise. (a) Clean speech. (b) Noisy speech. (c) Estimation of noisy speech via GARCH Model. (d) Soft detection. (e) Hard detection.

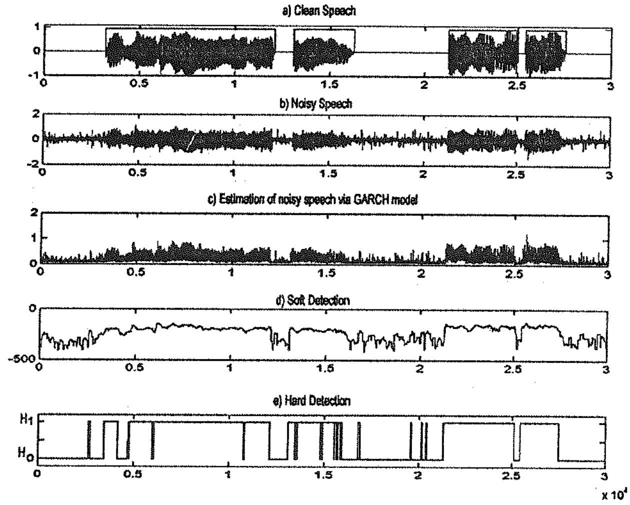

Figure 6: Results of the proposed VAD with colored noise. (a) Clean speech. (b) Noisy speech. (c) Estimation of noisy speech via GARCH Model. (d) Soft detection. (e) Hard detection.

Note that for Fig. 3 and 5, SNR is 5db and for Fig 4 it is 10 db.

The PTD for these noisy speeches are 0.96, 0.94 and 0.93, respectively.

7. Conclusion

The objective of this paper is to exploit the properties of new statistical tools such as GARCH model and heavy tailed distribution for finding a robust algorithm for VAD in the presence of high level of noise. The results show that the performance of the presented VAD has been improved when we take advantage of the adaptive threshold function.

The complexity of the proposed algorithm is very low due the fact that it is performed in time domain and estimations can be computed recursively.

8. References

[1] J. Sohn, N. S. Kim, and W. Sung, "A Statistical Model-Based Voice Activity Detection," *IEEE Signal Processing Lett.*, Vol. 6, No. 1, pp. 1–3, Jan. 1999.

[2] 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, No. 4, pp. 270–271, Feb. 1997.

[3] A. Rezayee and S. Gazor, "An Adaptive KLT Approach for Speech Enhancement," *IEEE Trans. Speech Audio Processing*, Vol. 9, pp. 87–95, Feb. 2001.

[4] F. Beritelli, S. Casale, and A. Cavallaero, "A Robust Voice Activity Detector for Wireless Communications Using Soft Computing," *IEEE J. Select. Areas Commun.*, Vol. 16, pp. 1818–1829, Dec. 1998.

[5] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd, "The Voice Activity Detector for the Pan European Digital Cellular Mobile Telephone Service," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, May 1989, pp. 369-372.

[6] Y. D. Cho, K. Al-Naimi, and A. Kondoz, "Improved Voice Activity Detection Based on a Smoothed Statistical Likelihood Ratio," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Vol. 2, 2001, pp. 737–740.

[7] J. Ramírez, J. C. Segura, M. C. Benítez, A. de la Torre, and A. Rubio, "A New Kullback-Leibler VAD for Speech Recognition in Noise," *IEEE Signal Process. Lett.*, Vol. 11, No. 2, pp. 666–669, Feb. 2004.

[8] J. Ramírez, J. C. Segura, M. C. Benítez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, Vol. 42, No. 3–4, pp. 271–287, databases and a wide range of noises and SNR conditions. On 2004.

[9] J. Ramírez and J. C. Segura, "Statistical Voice Activity Detection Using a Multiple Observation Likelihood Ratio Test" *IEEE Signal Process. Lett.*, Vol. 12, No. 10, pp. 689- 692, Oct 2005.

[10] S. Gazor and W. Zhang, "A Soft Voice Activity Detector Based on a Laplacian–Gaussian Model," *IEEE Trans. Speech Audio Processing*, Vol. 11, No. 5, pp. 498– 505, Sep 2003.

[11] S. Gazor and W. Zhang, "Speech Probability Distribution," *IEEE Signal Processing Lett.*, Vol. 10, pp. 204–207, July 2003.

[12] R. Tahmasbi and S. Rezaei, "A Soft Voice Activity Detection Using GARCH Model and Variance Gamma Distribution," *IEEE Trans. Audi, Speech and Language Processing*, in Press, 2007.

[13] T. Bollerslev, "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, Vol 31, pp. 307-327, 1986.

[14] D. Pena, G. C. Tiao and R. S. Tsay, *A Course in Time Series Analysis*, New York: JOHN WILY & SONS, 2001, Ch. 1 and 9.

[15] I. Cohen, "Modeling Speech Signals in Time-Frequency Domain Using GARCH", Signal Processing, Vol 84, 2453-2459, 2004.

[16] A. McNeil, R. Frey, and p. Embrechts, *Quantitative Risk Management: Concepts, Techniques and Tools*, Princeton University Press, 2005, Ch. 4.

[17] D. B. Madan, P. P. Carr, E. C. Chang, "The Variance Gamma Process and Option Pricing," *European Finance Review 2*: 79–105, 1998.

[18] E. Daal and D. Madan, "An Empirical Examination of the Variance-Gamma Model for Foreign Currency Option", *Journal of Business*, Vol. 78, pp. 134-176, 2005.

[19] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, New York: Dover Publ., 1968, Ch. 11.

[20] E. Seneta, "Fitting the Variance-Gamma to Financial Data", *Journal of Applied Probability*, special Vol. 41A, 177-187, 2004.