A Robust Statistical Approach for Detecting of Speech and Non-speech Intervals

Sadegh Rezaei[ii] ; Rasool Tahmasbi[iii]

ABSTRACT

This paper presents a robust algorithm for Voice Activity Detection (VAD) based on Variance Gamma Distribution (VGD) and adaptive threshold function. 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, the parameters of the distributions are estimated, recursively. Soft and hard detection results are obtained by comparing the Multiple Observation Likelihood Ratio Test (MOLRT) results with an adaptive threshold function. The simulation results show that the proposed VAD is able to operate down to -5 dB and in nonstationary environments.

KEYWORDS

Adaptive Threshold Function, Bayes Factor, Estimation Theory, Variance Gamma Distribution.

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 decades, numerous researchers have studied different strategies for detecting speech in noise and the influence of the VAD on the performance of speech processing systems; For example, Sohn [1] proposed a robust VAD algorithm based on a statistical Likelihood Ratio Test (LRT) involving a single observation vector. Later, [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. Ramirez [9] proposed an optimum likelihood ratio test (LRT) involving multiple and independent observations.

In [10], the distribution of speech is assumed to be Laplacian but in the present method, it is assumed to follow VGD; because it is a generalization of Laplacian Distribution (LD) and in specific case the Variance Gamma Distribution (VGD) becomes LD.

In [11] the signal is first decorrelated using an orthogonal transformation and then a Hidden Markov Model (HMM) is employed. In the proposed method, the results are achieved without any decorrelating method. This is a useful aspect from computational point of view.

Tahmasbi and Rezaei [12] used a Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model with an empirical adaptive threshold function. They showed that the empirical adaptive threshold function has better results than those of [11]. They also showed that using a Binary Markov Chain Model with an adaptive threshold function caused an improvement in the results of VAD [13]. Davis et. al. [16] described a statistical method for voice activity detection using a signal-to-noise ratio measure. Their method employed a low-variance spectrum estimate and determines an optimal threshold based on the estimated noise statistics. Ramírez et. al. [17] presented a method based on a contextual likelihood ratio test defined over a multiple observation window. Their approach defined a maximum *a posteriori* statistical test in which all the global hypothesis on the multiple observation window containing up to one speech to non-speech or non-speech to speech transitions were considered. Using change point detection in GARCH model, Tahmasbi and Rezaei [18] showed that the voice activity detection is equivalent to the change point problem in GARCH models.

This paper is organized as follows. In section 2, the MOLRT is discussed. Section 3 presents statistical

[ii] S. Rezaei is with the Faculty of Mathematics and Computer Science, Shahid Chamran University, Ahvaz , Iran (e-mail: srezaei@aut.ac.ir).

[iii] R. Tahmasbi is Ph.D. student of the Faculty of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran (e-mail: rtahmasbi@ieee.org).

properties of speech, noise and adaptive threshold function. Section 4 illustrates the experimental results of the proposed method.

2. Multiple Observation Likelihood Ratio Test

For distinguishing speech and silence intervals, we do as follows:

Assume that X_t is an m-dimensional vector of noise and noisy speech at time t, i.e,

$$X_t = [x(t), x(t-1), \ldots, x(t-m+1)]^T, \qquad (1)$$

where, $(.)^T$ denotes the transpose operation. For detecting speech and non-speech intervals, we use the below hypothesis via MOLRT:

$$\begin{cases} H_0 : X_t & is\ silence \\ H_1 : X_t & is\ speech \end{cases}. \qquad (2)$$

In a two-hypothesis test, the optimal decision rule that minimizes the error probability is the Bayes classifier. In the LRT, it is assumed that the number of observations is fixed and represented by a vector X_t. The performance of the decision procedure can be improved by incorporating more observations to the statistical test. In a two-class classification problem, a multiple observation likelihood ratio test (MOLRT) can be defined by

$$L_m(X_t) = \frac{P_{H_1}((x(t), x(t-1), \ldots, x(t-m+1) \mid H_1)}{P_{H_0}((x(t), x(t-1), \ldots, x(t-m+1) \mid H_0)}, \qquad (3)$$

where $P_{H_1}(. \mid H_1)$ and $P_{H_0}(. \mid H_0)$ are probability distribution function (pdf) of speech and noise, respectively. Since we choose between two models H_0 (Gaussian model) and H_1 (VGD model), so this ratio is also equivalent to the Bayes factor [15]. If the observations are independent then,

$$L_m(X_t) = \prod_{i=0}^{m-1} \frac{P_{H_1}(x(t-i) \mid H_1)}{P_{H_0}(x(t-i) \mid H_0)}. \qquad (4)$$

An equivalent log-LRT can be defined by taking logarithms

$$l_m(X_t) = \sum_{i=0}^{m-1} \ln \frac{P_{H_1}(x(t-i) \mid H_1)}{P_{H_0}(x(t-i) \mid H_0)}, \qquad (5)$$

and

$$l_m(X_{t+1}) = \sum_{i=0}^{m-1} \ln \frac{P_{H_1}(x(t+1-i) \mid H_1)}{P_{H_0}(x(t+1-i) \mid H_0)}. \qquad (6)$$

Note that (5) is sometimes called as the weight of evidence

Subtracting (5) and (6), we have,

$$l_m(X_{t+1}) - l_m(X_t) = \ln \frac{P_{H_1}(x(t+1) \mid H_1)}{P_{H_0}(x(t+1) \mid H_0)} - \ln \frac{P_{H_1}(x(t-m+1) \mid H_1)}{P_{H_0}(x(t-m+1) \mid H_0)}. \qquad (7)$$

By defining

$$\Phi(k) = \ln \frac{P_{H_1}(x(k) \mid H_1)}{P_{H_0}(x(k) \mid H_0)}, \qquad (8)$$

the MOLRT can be recursively computed

$$l_m(X_{t+1}) = l_m(X_t) - \Phi(t-m+1) + \Phi(t+1) \qquad (9)$$

and the decision rule is defined by

$$X_t = \begin{cases} speech & l_m(X_t) \geq \eta \\ Silence & l_m(X_t) < \eta \end{cases}, \qquad (10)$$

where η, is a threshold value. This cutoff value determines the boundary between those samples resulting in a test statistic that leads to rejecting the null hypothesis and those lead to a decision not to reject the null hypothesis.

As Ramirez [9] discussed, the use of the MOLRT for voice activity detection is mainly motivated by two factors: 1) the optimal behavior of the so-defined decision rule and 2) a multiple observation vector for classification defines a reduced variance LRT reporting clear improvements in robustness against the acoustic noise present in the environment.

3. Distributions and Adaptive Threshold Function

As discussed before, speech signal is assumed to have VGD because the VGD has heavier tails than the Gaussian distribution and the distribution of noise signal is assumed to be Gaussian.

A. Distribution of speech

A random variable X is said to be Variance Gamma Distribution (VGD) with parameters $\lambda, \alpha, \beta, \mu$, if its density is given by

$$f_X(x) = C.\left|x-\mu\right|^{\lambda-0.5} K_{\lambda-0.5}(\alpha\left|x-\mu\right|)e^{\beta(x-\mu)} \qquad (11)$$

with

$$C = \frac{\gamma^{2\lambda}}{\sqrt{\pi}\Gamma(\lambda)(2\alpha)^{\lambda-0.5}}, \quad \gamma^2 = \alpha^2 - \beta^2. \qquad (12)$$

$K_\lambda(.)$ is the modified Bessel function of the third kind and its integral representation is

TABLE 1
PROBABILITY OF TRUE DETECTION (%)

SNR	15	10	5	0	-2	-3	-5
P	98.54	97.62	97.05	94.34	90.70	86.17	75.91

TABLE 2
PROBABILITY OF TRUE DETECTION (%)

		Sohn	Gazor	Ramírez	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

TABLE 3
SENSITIVE ANALYSIS RESULTS

τ	SNR						
	15	10	5	0	-2	-3	-5
-126	95.61	94.87	93.51	91.59	87.75	83.95	72.80
-128	97.80	96.65	96.95	92.98	89.23	85.34	74.62
-130	98.54	97.62	97.05	94.34	90.70	86.17	75.91
-132	97.22	95.62	96.91	92.69	88.97	84.19	74.55
-134	96.23	94.49	93.42	91.31	87.34	82.96	72.47

THE VALUE IN EACH CELL IS THE PROBABILITY OF TRUE DETECTION (%) FOR DIFFERENT THRESHOLD VALUES AND DIFFERENT SNR'S.

$$K_\lambda(x) = \frac{1}{2}\int_0^\infty y^{\lambda-1}\exp\left(-\frac{x}{2}(y+y^{-1})\right)dy, \quad x>0, \quad (13)$$

(See [14], Chapter 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 Laplace transform of f is given by

$$L(z) = e^{\mu z}(\frac{\gamma}{\gamma_z})^{2\lambda}, \quad z \in R \qquad (14)$$

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

$$E(X) = \mu + 2\frac{\beta\lambda}{\gamma^2}, \qquad (15)$$

$$Var(X) = \frac{2\lambda}{\gamma^2}(1 + 2(\frac{\beta}{\gamma})^2). \qquad (16)$$

In [10], it is shown that the decorrelated speech components have zero-mean LD. A specific case of VGD ($\beta = \mu = 0$ and $\lambda = 1$) is zero-mean LD.

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

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

$$\hat{\mu}_t = \bar{x}_t = 1/m\sum_{i=0}^{m-1}x(t-i), \qquad (17)$$

$$\hat{\alpha}_t^2 = 2\lambda/\hat{\sigma}_t^2, \qquad (18)$$

$$\hat{\sigma}_t^2 = \widehat{Var}(X_t) = 1/(m-1)[\sum_{i=0}^{m-1}x^2(t-i) - m\hat{\mu}_t^2], \quad (19)$$

and can be recursively computed via

$$\hat{\mu}_{t+1} = \hat{\mu}_t + R/m, \qquad (20)$$

$$\hat{\sigma}_{t+1}^2 = \hat{\sigma}_t^2 + R/(m-1)\ [T - R/m - 2\hat{\mu}_t], \qquad (21)$$

$$R = x(t+1) - x(t-m+1), \quad T = x(t+1) + x(t-m+1). \qquad (22)$$

B. Distribution of noise

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

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}. \qquad (23)$$

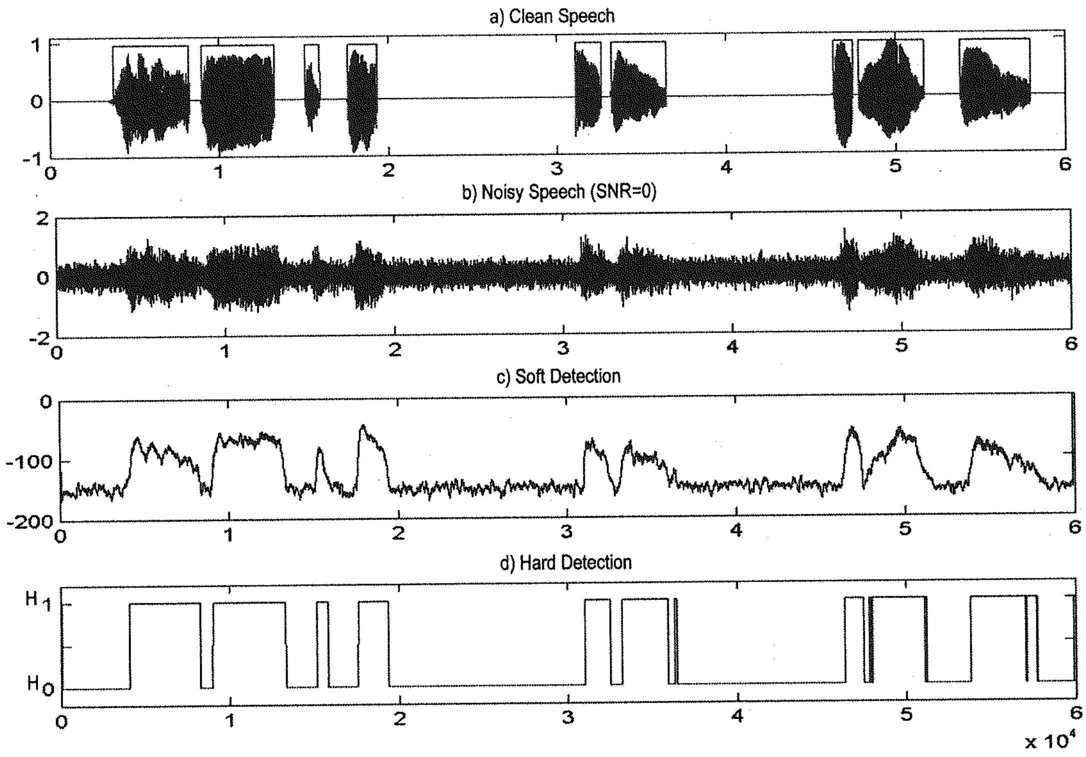

Figure 1: Results of the proposed VAD with 0 db SNR. (a) Clean speech. (b) Noisy speech with zero SNR. (c) Soft detection. (d) Hard detection.

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

$$\hat{\mu}_t = \overline{x}_t \qquad (24)$$

$$\hat{\sigma}_t^2 = \widehat{Var}(X_t)\,. \qquad (25)$$

These equations can be computed, recursively:

$$\hat{\mu}_{t+1} = \hat{\mu}_t + R/m\,, \qquad (26)$$

$$\hat{\sigma}_{t+1}^2 = \hat{\sigma}_t^2 + R/(m-1)\;[T - R/m - 2\hat{\mu}_t]\,. \quad (27)$$

C. Adaptive threshold

As discussed in Section 2, speech is active at time t if $l_m(X_t) \geq \eta$ (based on MOLRT). It is a better, however, to compare $l_m(X_t)$ with a threshold function, η_t, at each time t,

$$X_t = \begin{cases} \text{speech} & l_m(X_t) \geq \eta_t \\ \text{Silence} & l_m(X_t) < \eta_t \end{cases}, \qquad (28)$$

or speech is active if

$$\sum_{i=0}^{m-1} \ln f_{VG}(x(t-i)) - \sum_{i=0}^{m-1} \ln f_G(x(t-i)) \geq \eta_t \quad (29)$$

where $f_{VG}(.)$ and $f_G(.)$ are VGD and GD, respectively.

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

$$\eta_t = \tau + m\ln(C)\,, \qquad (30)$$

where τ is a constant threshold and C is defined in (12), and its parameters are estimated by the vector X_t at time t.

The constant threshold τ, is empirically chosen. The results of sensitive analysis for this constant threshold are presented in Table 3. In this table, the probability of true detection for different values of τ s and different values of SNR's are presented. As can be seen, there is no significant difference between the results.

If we define

$$g(x) = \left|x-\mu\right|^{\lambda-0.5} K_{\lambda-0.5}(\alpha\left|x-\mu\right|)e^{\beta\,(x-\mu)}, \quad (31)$$

then

$$f_{VG}(x) = C.g(x)\,. \qquad (32)$$

Therefore, using (30) and (32) in (29) will lead to derive a new approach as follows:

speech is active if

$$\sum_{i=0}^{m-1} \ln g(x(t-i)) - \sum_{i=0}^{m-1} \ln f_G(x(t-i)) \geq \tau\,. \qquad (33)$$

The left side of (33) can be viewed as a criterion for detecting silence or speech activity, and is a criterion for soft detection. The soft detection function will be compared with τ and hard detection will be derived.

As can be seen, reducing the number of computation in (33) is another advantage of the proposed method.

4. Experimental Results

In this section the results of the proposed method are presented. The speech signals are obtained from http://www.dailywav.com and sampled at 16 KHz. Fig. 1(a) shows the clean speech together with manually marked clean speech sample. Fig. 1(b) shows the noisy speech with SNR equal to 0. The results of the proposed method with different SNRs are presented in Table 1. The probability of true detection 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 left side of (6) is computed and shown in Fig. 1(c). This criterion is compared with $\tau=-130$ and hard decision is derived. The hard detection is shown in Fig. 1(d). To evaluate the performance of the proposed VAD, the speech and silence intervals marked manually. The hard decision of the VAD is compared with the manually marked intervals. The probability of true detection is equal to 0.94.

The probabilities of true detection as evaluation criteria for the performance of the previous works and presented method are given in Table 2 with different SNRs. As can be sen , the performance of the proposed method is more than Ramírez [9], Gazor [10] and Sohn [1].

5. Conclusion

The objective of this paper is to exploit the properties of heavy tailed distribution and adaptive threshold function in order to attain a robust and fast 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. A sensitive analysis on the threshold values shows that this parameter is robust against its value defined on a suitable interval and against environmental noise.

The complexity of the proposed algorithm is very low due the fact that all the algorithms are computed recursively.

6. References

[1] Sohn, J., Kim, N. S., and Sung, W., *"A statistical model-based voice activity detection,"* IEEE Signal Processing Letters, Vol. 6, No. 1, 1–3, 1999.

[2] Garner, N. R., Barrett, P. A., Howard, D. M., and Tyrrell, A. M., *"Robust noise detection for speech detection and enhancement,"* Electronics Letters, Vol. 33, No. 4, 270–271, 1997.

[3] Rezayee, A. and Gazor, S., *"an Adaptive KLT Approach for Speech Enhancement,"* IEEE Transactions on Speech and Audio Processing, Vol. 9, 87–95, 2001.

[4] Beritelli, F., Casale, S., and Cavallaero, A., 1998, *"A robust voice activity detector for wireless communications using soft computing,"* IEEE Journal on Selected Areas in Communications, Vol. 16, 1818–1829.

[5] Freeman, D. K., Cosier, G., Southcott, C. B., and Boyd, I., *"The voice activity detector for the pan European digital cellular mobile telephone service,"* IEEE International Conference on Acoustics, Speech and Signal Processing, 369-372, 1989.

[6] Cho, Y. D., Al-Naimi, K., and Kondoz, A., *"Improved voice activity detection based on a smoothed statistical likelihood ratio,"* IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2, 737–740, 2001.

[7] Ramírez, J., Segura, J. C., Benítez, M. C., de la Torre, A., and Rubio, A , *"Efficient voice activity detection algorithms using long-term speech information, Speech Commun.,"* Vol. 42, No. 3–4, 271–287, 2004.

[8] Ramírez, J., Segura, J. C., Benítez, M. C., de la Torre, A., and Rubio, A.,Feb. *"A new Kullback-Leibler VAD for speech recognition in noise,"* IEEE Signal Processing Letters, Vol. 11, No. 2, 666–669, 2004.

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

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

[11] Gazor, S. and Zhang, W., *"A Soft Voice Activity Detector Based on a Laplacian–Gaussian Model,"* IEEE Transactions on Speech and Audio Processing, Vol. 11, No. 5, 498– 505, 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, Vol. 15, No. 4, 1129-1134, 2007.

[13] S. Rezaei, and R. Tahmasbi, "A Fast and Soft Voice Activity Detector Using GARCH Model and Markov Model Chain," Amirkabir Journal of Science and Technology, No. 18 (66-D), pp. 65-72, 2007.

[14] Abramowitz, M., and Stegun, I., Handbook of Mathematical Functions, New York: Dover Publ., (Ch apter 11), 1968.

[15] H. Jeffreys, *The Theory of Probability* (3e), Oxford (1961); p. 432

[16] A. Davis, S. Nordholm, and R. Togneri, *"Statistical Voice Activity Detection Using Low- Variance Spectrum Estimation and an Adaptive Threshold,"* Vol. 14, Num 2, 2006.

[17] J. Ramírez, J. C. Segura, J. M. Górriz, and L. García, *"Improved Voice Activity Detection Using Contextual Multiple Hypothesis Testing for Robust Speech Recognition,"* Vol. 15, Num. 8, 2007.

[18] R. Tahmasbi, S. Rezaei, *"Change Point Detection in GARCH Models for Voice Activity Detection,"* in press.