Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı, S. 157-163, Ekim 2019

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue, pp. 157-163, October 2019

Research Article

Examination of Energy Based Voice Activity Detection Algorithms for Noisy Speech Signals

Selma Özaydın[1*]

[1] [1]Çankaya Üniversitesi, Mühendislik Fakültesi, Elektronik ve Haberleşme Mühendisliği Bölümü, Ankara, Türkiye, (ORCID: 0000-0002-4613-9441)

(Bu yayın HORA 2019 kongresinde sözlü olarak sunulmuştur.)

(First received 1 August 2019 and in final form 24 October 2019)

(**DOI:** 10.31590/ejosat.637741)

ATIF/REFERENCE: Özaydın, S. (2019). Examination of Energy Based Voice Activity Detection Algorithms for Noisy Speech Signals. *European Journal of Science and Technology*, (Special Issue), 157-163.

Abstract

This paper examines the behavior of two different energy-based voice activity detector (VAD) algorithms for noisy input signals. The examined detectors use time-domain methods to find speech boundaries. Time-domain short time energy features and/or zero-crossing rate of speech signals are used to evaluate the performance of the methods. In the first stage of both algorithms, time-domain short-time energy (STE) features are calculated for each speech segment. Then energy ratios and threshold values are used to detect any voicing activity of speech signals. The decision threshold value is calculated by evaluating the average STE of an initial silence period. The effectiveness of the selected methods is tested for clean and noisy speech samples. The methods are tested using the noisy speech signals under different SNR levels. The results indicated that both methods achieve a reasonable accuracy as low as an SNR value nearly 0dB with a slowly decreasing performance. But, under 0dB SNR, both methods lose their effectiveness against noisy conditions.

Keywords: Voice activity detection, Speech analysis, Speech/silence classification, Endpoint detection, Noise measurement

Enerji Tabanlı Konuşma Aktivitesi Belirleme Algoritmalarının Gürültülü Konuşma Sinyalleri için İncelenmesi

Öz

Bu çalışmada, iki farklı enerji tabanlı konuşma bölgesi aktivasyonu detektör (KAD) algoritmasının gürültülü giriş sinyallerine karşı davranışları incelenmektedir. İncelenen KAD detektörleri, konuşma sınırlarını etkin bir şekilde belirlemek için zaman düzlemindeki metotları kullanmaktadır. Zaman düzlemi kısa zaman aralığında enerji hesabı ve/veya sıfır geçiş oranı, metotların performansını değerlendirmede kullanılmaktadır. Her iki algoritmanın ilk aşamasında, zaman düzleminde her bir konuşma alt kesitinde enerji değerleri hesaplanmaktadır. Enerji oranları ve eşik değerler, konuşma sinyalinin aktif bölgelerini belirlemede kullanılmaktadır. Karar eşik değeri, konuşma sinyalinin başında sessiz bir bölge aralığında hesaplanmaktadır. Seçilen metotların etkinliği temiz ve gürültülü konuşma sinyal örnekleri için test edilmiştir. Metotlar, değişik SNR seviyelerinde gürültülü konuşma sinyalleri kullanarak test edilmiştir. Sonuçlar göstermiştir ki, 0dB SNR seviyesine kadar yavaşça azalan performansla her iki metot etkinliklerini koruyabilmekte, ancak 0dB SNR seviyesi altında her iki metot etkinliğini kaybetmektedir.

Anahtar Kelimeler: Konuşma Aktivite belirleme, Konuşma analizi, Konuşma/sessiz bölge sınıflandırma, Sınır değer belirleme, Gürültü hesaplama

*Corresponding Author : Çankaya Üniversitesi, Mühendislik Fakültesi, Elektronik ve Haberleşme Mühendisliği Bölümü, Ankara, Türkiye, ORCID: 0000-0002-4613 9441, selmaozaydin@cankaya.edu.tr

1. Introduction

Digital speech processing applications try to separate voice-active speech periods from inactive (silence) ones to minimize the process time. Therefore, VAD algorithms have been used widely in many speech processing applications such as speech coding, speech recognition, audio conferencing, echo cancelation or Voice over Internet Protocol (VoIP) algorithms. If an algorithm succeeds to locate the endpoints of an utterance accurately, it can label an interval of a signal as silence or voice-activated. Especially for a speech recognition system, false detection of voice-active regions will have a degradation effect on the recognition result. The proper detection of the beginning and end regions of a speech utterance by separating it from any background noise is an important problem in a speech recognition system. In a VoIP application, a VAD algorithm increases the bandwidth requirement of a voice session. During an application, a VAD decision is given according to a pre-defined threshold value. VAD algorithms can be operated in time domain or frequency domain. Time domain algorithms usually use energy and zero crossing rate (ZCR) parameters while frequency domain algorithms use spectrum information. When compared to the frequency domain algorithms, time domain algorithms are computationally simple. The simplicity of an algorithm gives chance to kept time delay at a minimum. The energy of a speech frame gives a piece of information for the activity of a frame and energy threshold value is used in decision making. For time domain VAD algorithms, the amplitude of a speech in a frame is an important parameter to classify the frames as voice-active or inactive. In voice-active parts, speech sounds can be divided into three classes of phonemes as voiced, unvoiced, and plosives according to their modes of excitation, where voiced phonemes (vowels) have quasi-periodic pulses whereas unvoiced phonemes (consonants) are considered as random pulses. Plosive sounds have an excitation similar to unvoiced sounds. The peak amplitude of voiced phonemes is much higher than the magnitude of unvoiced and plosive phonemes. Although unvoiced and plosive sounds have lower amplitude than voiced sounds, they contain important information for speech, especially when detecting the beginning and endpoints of a speech. For an unvoiced speech at the beginning or endpoints of an utterance while energy is close to silence energy, there is a sharp increase in the ZCR. On the other hand, for a voiced speech the energy is radically higher than the silence energy. For the utterances beginning or ending with weak fricatives such as (/f, th, h/), voiced fricatives becoming devoiced, weak plosive bursts such as (/p, t, k/), ending with nasal sounds such as (/n, m/), or some voiced sounds the final /i/ becoming unvoiced in the word such as "three" (/th-r-i/) or "binary" (/b-al-n-e-r-i/), it is difficult problem to locate to the accurate points for a VAD algorithm. [1-3]

A speech processing system must provide adequate quality for small amplitude signals consisting of unvoiced phonemes. Besides, speech signals are generally composed of relatively fewer voiced phonemes than unvoiced phonemes. When STE is used to evaluate the VAD of a speech signal, uniform calculation of STE in each frame provides inadequate quality for small amplitude unvoiced signals due to their close values to a decision threshold especially in case of background noise. The correct definition of unvoiced sounds is very important to extract the correct information in a speech. As a result, some VAD detection algorithms may need to apply a backward/forward search algorithm to exactly detect the beginning/endpoints of each utterance by of frames [1-5]. An endpoint detection in speech processing is to detect the presence of speech especially in a background of noise. With accurate detection of talkspurt boundaries, the required amount of speech processing can be reduced. Time-domain VAD methods usually based on energy and ZCR calculation of the signals. These methods are simple to compute. If SNR of the input speech signal is high, the energy of any lowest-level speech sound exceeds the background noise energy and a simple energy measure can detect the sound. However, such ideal recording conditions are not practical for real-world applications. The goal of an endpoint algorithm is simplicity, robustness to background noise and reliable location of acoustic events. Therefore, the VAD algorithm performance is evaluated in terms of delay, sensitivity, and accuracy. For a VAD algorithm, it is essential to find the exact word boundaries even in noisy environments. The time-domain analysis of a speech signal introduces low complexity to locate the endpoints of an utterance in a speech signal. Therefore, energy-based VAD algorithms in the literature take advantage of simplicity. Besides, they do not require to assume noise characteristics. On the other hand, they are sensitive to the noise and it is necessary first to define the background silence of input waveform. Although studies are comparing different energy-based algorithms in the literature, there is no single study measuring and comparing their performances against different kinds of noises. In this paper, two different energy-based VAD algorithms are tested with different levels of noisy input signals and their noise performance is measured with SNR calculation. The first method uses both energy calculation and ZCR for evaluation while the second one uses only energy calculation while making a VAD decision. During the evaluation of the first method (VAD1), the speech signal is divided into some segments and the zero-crossing rate and energy values are calculated to separate the voiced and unvoiced parts of speech. The second method (VAD2) only uses energy value to make a VAD decision and has an advantage of adaptive thresholding against changing noise conditions in an input signal [1-5,8]

The paper is organized as follows. The second section gives a theoretical background about the STE, ZCR and the VAD algorithms under examination. The third section presents the test methods and the results of tests performed to evaluate the selected VAD methods. The last section concludes the article by evaluating the methods in the scope of test results.

2. Theoretical Background

The general block diagram of a VAD process can be briefed as in Figure 1. In this algorithm, after the initial processing of an input signal, feature extraction is performed. These features can be time-domain parameters such as STE, ZCR or linear prediction coefficients or frequency domain parameters such as cepstral or spectral coefficients. Then a VAD analysis is performed according to the features and a calculated threshold value. The voiced or unvoiced decision is made according to the decision algorithm.

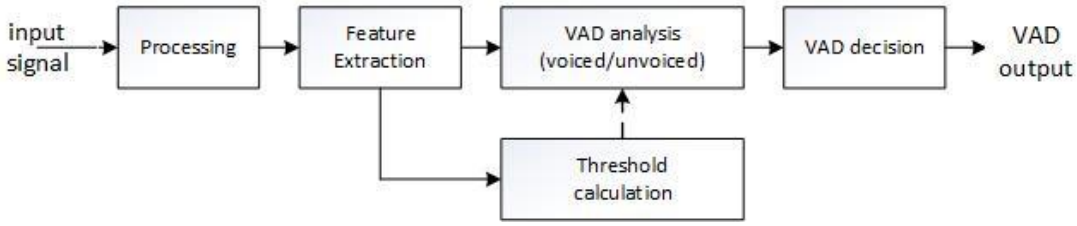

Figure 1. A block diagram of a basic VAD process

For the examined algorithms in this paper which are called as VAD1 and VAD2, two basic methods used to find features are ZCR and STE. These are briefed as follows.

2.1. Zero Crossing Rate

The magnitude of a speech signal in a frame passes through zero points of the axis in a number of times. While there is a small amount of sign changes for a voiced speech frame due to the excitation of the vocal tract by the periodic flow, the zero-crossing count increases rapidly for an unvoiced speech due to the noise like airflow (Figure 2). it is also considered as an indicator of frequency. Therefore, zero-crossing count is used for the segmentation of a speech signal as voiced or unvoiced. The ZCR is defined as the average value of sign change of a signal in a frame. For a 20ms analysis window, typical values of the ZCR for a voiced speech region can be taken as less than 0.1, while it is taken as greater than 0.3 for an unvoiced speech. An example of a voiced speech zero-crossing count or unvoiced speech count is seen in Figure 3. The ZCR of a windowed signal can be defined as the equation (1) and (2). [6,7]

A definition for zero-crossings rate is,

$$Z_n = \sum_{-\infty}^{\infty} 0.5|sgn[x(m)] - sgn[x(m-1)]| \,.\, w(n-m) \quad (1)$$

where

$$sgn[x(n)] = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases} \quad (2)$$

On the other hand, if there is some background noise in a speech signal, it is difficult to separate voiced, unvoiced or silence region in a speech easily just by evaluating the ZCR as can be seen in Figure 3. Therefore, ZCR cannot be an only VAD method especially in case of a background noise in a speech signal. In VAD algorithms, it is usually used as a supportive technique to improve decision rule.

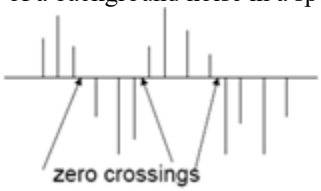

Figure 2. Definition of zero crossings

2.2. Short Time Energy

The STE of a speech signal reflects the amplitude variations in it. The voiced speech regions have higher energy than unvoiced speech parts. The peak magnitude of a speech signal changes with time. Because magnitudes in voiced regions much higher than unvoiced parts, time-domain analysis can provide information about voiced/unvoiced (VUV) decision of a signal. During the short time analysis of a speech signal, a windowing of length N is performed as a first step and its window size is taken between 20-50ms to reflect the variations in amplitude. Then, STE of the windowed signal is calculated for each window. The threshold value for a VAD decision is computed by selecting a silence period at the beginning of the speech signal. The STE (En) of an mth frame of length N is defined by equation (3) as the sum of the magnitude where n is time index and w represents the window. [6,7]

$$E_n = \sum_{-\infty}^{\infty} [\mathrm{x(m).\,w(n-m)}]^2 \quad , \qquad m = 1, \dots, M \quad (3)$$

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \quad (4)$$

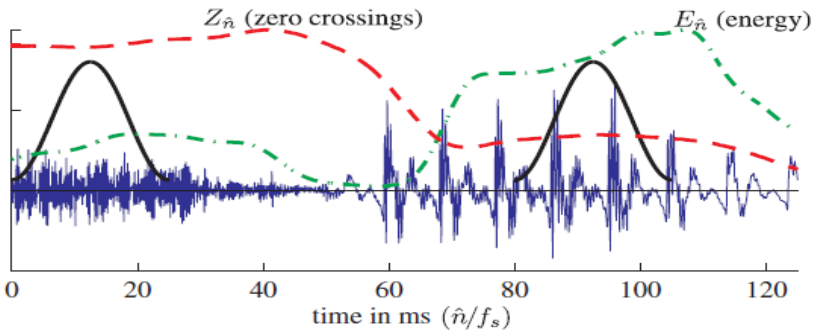

Figure 3. A part of speech waveform with short-time energy and zero-crossing rate ([7])

$$E_{rms}(\mathrm{m}) = \sqrt{(\tfrac{1}{\mathrm{N}}) \sum_{\mathrm{n=m.N}}^{\mathrm{m.N+N-1}} \mathrm{x(n)\text{^}2}} \quad , \qquad m = 1, \dots, M \tag{5}$$

As can be seen from Figure xxx above, STE and ZCR change slowly in a frame and an average value can be used for a frame. ZCR of an unvoiced frame is higher than that of the voiced frame. Conversely, the energy of an unvoiced frame is much lower than that of the voiced frame. Thus, the STE and ZCR are two important analysis parameters for the time-domain VAD analysis of a speech signal. The statistical distributions of the energy and zero-crossing rate are used to derive thresholds for voiced/unvoiced decision.

3. Algorithms under Evaluation

In this paper, two methods are considered to examine their noisy signal performance. The first algorithm examined in this paper (VAD1) is based on the ZCR and STE calculation [1]. The method combines the STE and ZCR for classification and divides the speech samples into some segments to evaluate voiced and unvoiced parts. The results suggest that zero crossing rates are low for the voiced part and high for the unvoiced part whereas the energy is high for the voiced part and low for the unvoiced part. Therefore, STE and ZCR are proved effective in the separation of voiced and unvoiced speech. After the endpoint detection process to separate speech and silence parts, a frame by frame processing is performed and speech is divided into short frames. Then, a speech threshold is determined from the silence energy region which stays constant through analysis frames. At the beginning of the algorithm, a small period of speech is selected to calculate an initial threshold value from the background noise. The threshold value is used to separate speech and non-speech segments in a speech signal. The threshold calculation is fixed and not adapted to changing the noisy environment. To calculate threshold value, the initial estimate of the silence energy (threshold) is calculated by taking the means of the energies in a small period interval as can be seen in equation (6), in which Eint represents initial energy threshold mean value and f represent the number of frames selected. From the sum of the mean zero-crossing rate during silence, a threshold value is defined for zero crossings. ZCR is used to recover some low energy phoneme information which are below the energy thresholds. Hamming window in equation (4) is used in the method. During the VUV evaluation, STE and ZCR are calculated in each frame. Then, a decision algorithm is executed and if ZCR is small and STE is high, the frame is accepted as voiced. If not, the frame is selected as unvoiced. If the decision is not clear, the algorithm decreases the window size. STE and ZCR are recalculated in these sub-divided frames. Because frame length is selected as 50ms in the original paper [1], we selected the same frame rate for an 8kHz sampled speech.

$$E_{int} = \frac{1}{\mathrm{f}} \sum_{\mathrm{i=1}}^{\mathrm{f}} E_i \tag{6}$$

The second algorithm (VAD2) is also an energy-based technique and presents an adaptive threshold valued VAD method [2]. This method does not use ZCR and only based on STE estimation. The algorithm calculates the energy threshold value dynamically by using an adaptive scaling parameter. Noise power estimation is done in each frame and used for adaptation of the threshold. In the paper, an adaptive thresholding issue is evaluated and some solutions are proposed. We used the linear energy-based detector with a double threshold. In this algorithm, energy is updated for both voiced and unvoiced frames according to updated threshold values. If energy is greater than the threshold, the algorithm assumes this as a voiced frame, otherwise, it assumes as an unvoiced frame. Because it uses different thresholding values for speech and silence detection, it tries to prevent the problem of sudden variations at the output. For the energy calculation of frames, the root mean square energy (Erms) in equation (5) is used. The Erms formula calculates the square root of the average sum of squares of the amplitude of the speech samples. Erms is more effective to extract peaks and valleys on power estimation of the speech signal. The threshold calculation is based on the minimum and maximum energy levels obtained from incoming frames. There is also a hangover period to check if the last four frames are inactive to stop transmission. If so, the algorithm waits until the energy is over the threshold value. To increase the performance of the algorithm against background noise, the scaling factor of the threshold is adapted according to a minimum and maximum energy values. The algorithm finds and labels the beginning points of the utterance as N1 when the lower threshold (ITL) is exceeded. Then, the endpoint of the utterance is labeled (N2) when the energy falls below ITL. When defining the exact N1 and N2 points, the algorithm requires a backward/forward search due to the possible uncertainty of definition during energies of small magnitude signals in an utterance.

4. Results and Discussion

MATLAB platform is used to test the algorithms. The methods have been tested on different SNR levels of background noises and for many input test data. The performance of the algorithms are analyzed based on the percentage of correct detection, robustness to background noise. Test samples are selected from TIMIT clean speech database and background noise is added artificially with 'awgn' command in Matlab. The SNR level of the noisy speech is calculated with 'snr' command in Matlab. Then, VAD performances of the algorithms are measured with SNR calculation by taking the VAD decision of clean speech test samples as reference. Test results show that while the VAD methods under evaluation perform VUV decision with a high correction rate up to SNR level is on the order of 0 dB, detection performance reduces fastly below this rate.

The first VAD method (VAD2) in [1] is examined for noisy signals. The results of VAD1 for clean speech can be seen in Figure 4.a and Figure 4.b. The sentence pronounced by a male is 'The pipe began to rust while new'. As can be seen in Figure 4, while the ZCR and STE separation capability and as a result VUV performance seems very good for clean speech data, ZCR performance reduces for noisy speech especially due to the noisy signal addition to the ZCR calculation as can be seen in Figure 5.a. Besides, the difference between the minimum and maximum values of STE reduces in a noisy signal. This reduced performance in STE and ZCR directly shows itself in the VUV separation capability of the VAD1 method. Even if it seems in Figure 5.b that VUV separation regions close to clean speech VUV regions, it is not easy to say that the method performs a reliable VUV decision due to the uncorrect VUV regions.

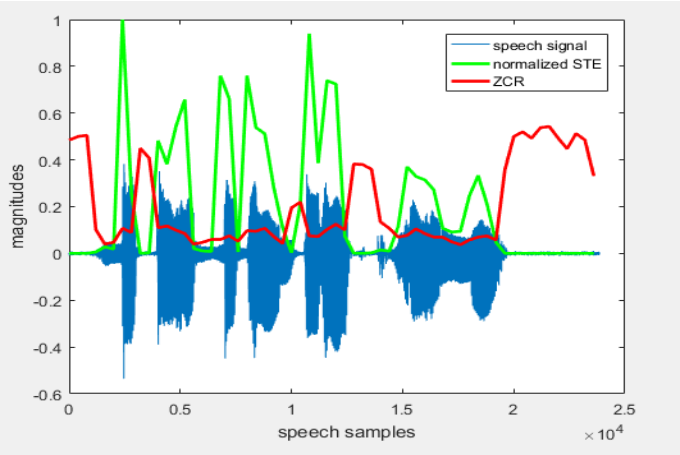

Figure 4.a. normalized STE and ZCR for clean speech signal for a male

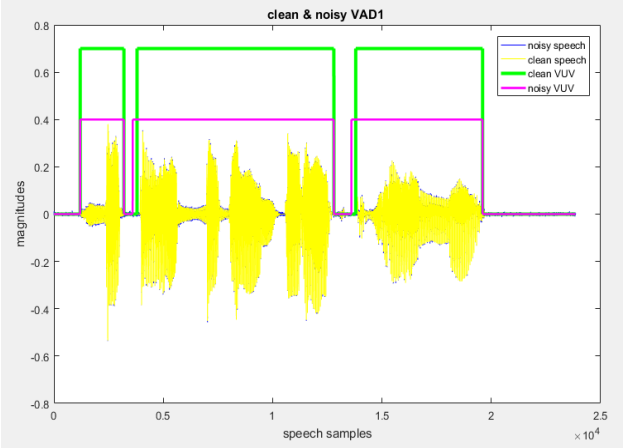

Figure 4.b VAD1 for clean speech signal

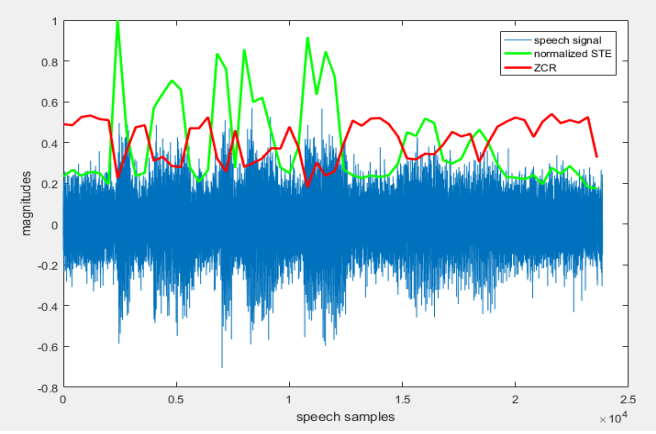

Figure 5.a VAD1 :normalized STE and ZCR for noisy speech signal for a male (SNR ≈ -1dB)

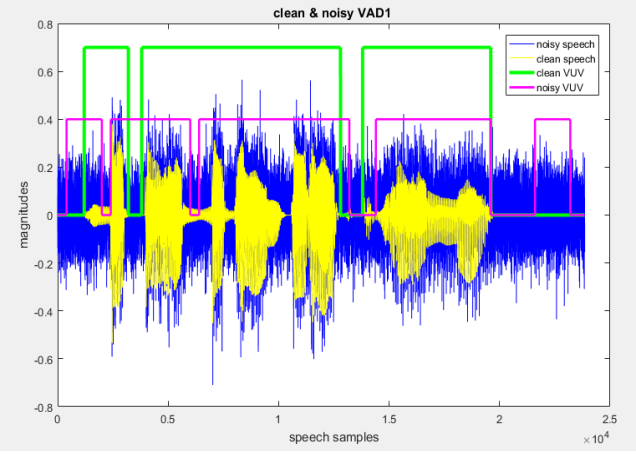

Figure 5.b VUV decisions in VAD1 for noisy speech signal for a male (SNR ≈ -1dB)

The second VAD method (VAD2) in [2] is examined for noisy signals. The same speech signals are used for performance comparison. According to the test results, it is seen that there is a good quality of VUV decision and more precise boundaries than VAD1 method for the clean speech signal. On the other hand, for a noisy signal with an SNR nearly -1.35dB, the VAD2 method loses its VAD decision capability. The results can be seen in Figure 6 and Figure 7 for clean speech and noisy speech samples, respectively. From the results, it can be said that the second algorithm (VAD2) more robust the background noise increase and even if its performance reduces in noisy speech input, we have not met a false detection in non-speech regions. If we compare VAD1 and VAD2 according to tested

noisy speech data, we can say that VAD1 is more sensitive to background noise due to the false detection in non-speech regions especially for the noise levels SNR value less than 0dB.

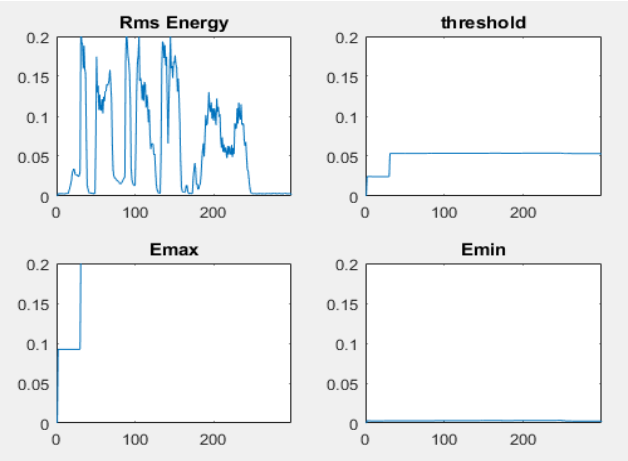

Figure 6.a. rms energy, threshold and energy limit values for a clean speech

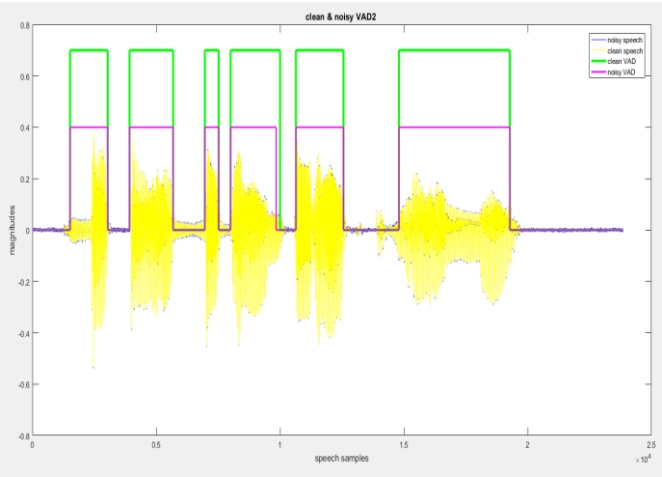

Figure 6.b. VAD1 for clean speech signal

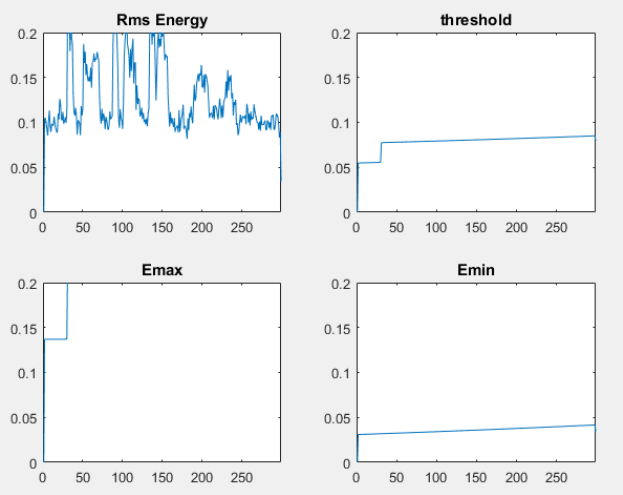

Figure 7.a. rms energy, threshold and energy limit values for a noisy speech (SNR ≈ -1.35dB)

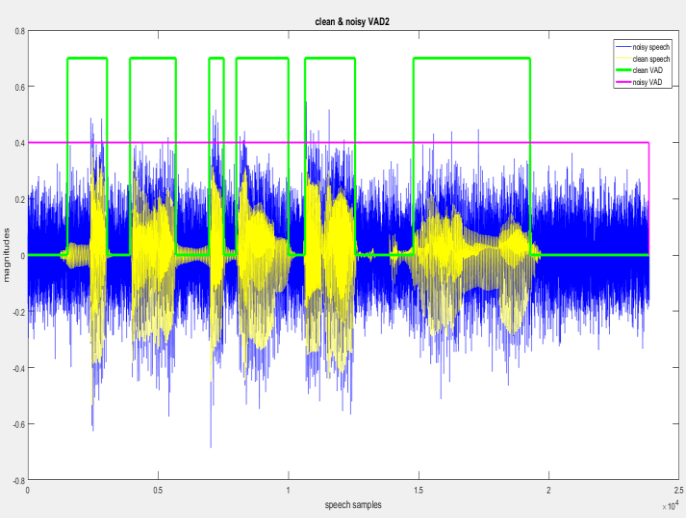

Figure 7.b. VAD1 for clean speech signal

4. Conclusion and Recommendations

Energy-based VAD algorithms have an advantage of simplicity when compared to frequency-domain algorithms. Time-domain algorithms perform a good detection rate when clean speech data is used for detection. But it is essential to see their performance in the noise. This study examined the performances of two different time domain VAD algorithms with objective measurement methods. We analyzed these two different energy-based voice activity detection algorithms with speech data having different SNR levels of noise. This work aimed to analyze the effect of background noise on time-domain energy-based algorithms. The test results showed that both methods perform good quality of detection for the clean speech signal. On the other hand, when the noise amount is increased, the second method which is based on adaptive thresholding performed better performance. On the other hand, it should be noted that their noise performance is examined with limited test data and evaluations are made on this basis. To make a comprehensive noise performance evaluation of the methods, test data should be increased and different environmental test conditions should be evaluated as a further study.

References

[1] R. G. Bachu, S. Kopparthi, B. Adapa and B. D. Barkana (2010), Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy, January, 2010, Advanced Techniques in Computing Sciences and Software Engineering, pp 279-282, 2010; DOI 10.1007/978-90-481-3660-5_47

[2] K.Sakhnov, E.Verteletskaya and B. Simak (2009), Dynamical Energy-Based Speech/Silence Detector for Speech Enhancement Applications, Proceedings of the World Congress on Engineering 2009 Vol I, WCE 2009, July 1 - 3, London, U.K., ISBN: 978-988-17012-5-1

[3] L. R. Rabiner ; M. R. Sambur (1975), An algorithm for determining the endpoints of isolated utterances, The Bell System Technical Journal (Volume: 54 , Issue: 2 , Feb. 1975), (ISSN: 0005-8580), DOI: 10.1002/j.1538-7305.1975.tb02840.x, pp. 297 – 315,

[4] Prasad, V. (2002), Comparison of voice activity detection algorithms for VoIP, Proceedings - International Symposium on Computers and Communications, ·DOI: 10.1109/ISCC.2002.1021726, pp.62-65,

[5] Pollak, P., Sovka, P., Uhlir, J. (1993), Noise Suppression System for a Car, proc. of the Third European Conference on Speech, Communication and Technology – EUROSPEECH'93, (Berlin, Germany), p. 1 073–1 076, vol.5, Sept..

[6] A. M. Kondoz (1999), Digital Speech. New York: John Wiley and Sons,

[7] L. R. Rabiner and R. W. Schafer (2007), Introduction to Digital Speech Processing, Foundations and Trends in Signal Processing. Boston: Now Publishers Inc.,

[8] P.Renevey, A.Drygajlo, (2001), Entropy based voice activity detection in very noisy conditions, in Proc. Eurospeech 2001, pp.1887-1890