Design of a Voice Activity Detection Algorithm based on Logarithmic Signal Energy

Selma Ozaydin
Cankaya University, 06790,Ankara,Turkey
ORCID: 0000-0002-4613-9441

2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA) | 978-1-6654-5600-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICECTA57148.2022.9990492

Abstract—This article presents a new method for calculating the signal energies of speech segments in voice activity detection algorithms. In the study, the μ-law signal compression method is adapted to calculate short-term signal energies. A simple voice activity detection (VAD) algorithm is designed to demonstrate the effectiveness of the proposed method. The same VAD algorithm was also run with two different conventional energy calculation formulas and the performance of each VAD was evaluated using time-domain short-time energy features. The G729 standard VAD algorithm was also used for performance comparison. During the test of the analyzed detectors, many kinds of input speech signals with various types of background environmental noise, such as restaurants, vehicles, and streets, were tested. Using the new energy calculation method, the VAD detector has improved detection accuracy compared to VAD detectors based on the other two energy methods and was able to effectively identify voice-active regions even in noisy conditions at low SNR levels. The results revealed that the VAD detector designed with the proposed new energy calculation formula outperforms traditional energy-based voice activity detection methods and provides noticeable increases in detection rate even under adverse conditions.

Keywords— *voice activity detection, speech analysis, endpoint detection, feature analysis, signal energy calculation*

I. Introduction

Speech processing algorithms need a voice activity detection (VAD) decision to avoid wrongly detection of non-speech segments as speech for several types of acoustic background noise. The usage of a VAD algorithm and classification of a speech signal into voice-active and silence regions minimizes the processing time and reduces the computational cost in a digital speech processing application by separating voice activity regions from background noise or silence parts. A VAD algorithm increases the accuracy of a speech processing model and reduces the bit rate since fewer bits are required for silence representation. With the usage of a VAD, only speech segments are processed in an input signal. This allows the saving of bandwidth in a transmission line [1].

The voice activity detection (VAD) algorithms are also known as endpoint detection algorithms. Two main parts of an endpoint detection algorithm are the extraction of discriminative acoustic features and then speech/non-speech decision based on a set of classification rules. Discrimination between noisy speech frames and noise only frames is the most important problem of a VAD algorithm in adverse conditions. Endpoint detection algorithms in speech processing. Over the years, different approaches have been proposed for voice activity detection which can be operated in the time domain or frequency domain and evaluated according to their accuracies and computation costs. There are many kinds of algorithms in the literature using the feature vectors such as zero-crossing rate (ZCR) and/or energy [2, 3], fundamental frequency [4], autocorrelation [5, 6], spectrum [7], cepstrum [8], wavelets [9], principal component analysis [10] or statistical model-based algorithms [10, 11]. Most of these methods are sensitive to SNR variations in the noisy speech signal and assume that noise is stationary over certain periods of time. Lately, unsupervised models have been proposed [12-14] to overcome the requirement of large amounts of labeled training data for supervised models. For the VAD application in telephony and multimedia communications, The ITU-T VAD standard Rec. G.729 Annex-B was developed [15]. The other two VADs (options 1 and 2) for mobile communication systems are the adaptive multi-rate (AMR) speech codec standardized by ETSI [3, 17]

Signal energy based VAD algorithms estimates the frame energy within each analyzed frame by assuming that the energy of voiced regions is relatively higher than that of background noise regions. Time-domain energy calculation in VAD algorithms take advantage of the simplicity of computation and thus keep time delay at a minimum. For a fixed threshold VAD algorithm, an energy threshold value is defined at the beginning of the algorithm and used in the definition of start and endpoints of the speech. Therefore, the amplitude of a speech in a frame is an important parameter to classify the frames as voice-active or inactive (silence). While energy-based VAD algorithms take advantage of low complexity, simplicity, they are very sensitive to the background noise. Therefore, noise-robust algorithms are required to define endpoints correctly. The application of a logarithmic scale adjusted to a μ value increases low-amplitude signals while suppressing peak amplitudes depending on the selected μ value [18].

The VAD algorithm proposed in this article gives chance to exactly define the low amplitude voicing activity points by increasing signal visibility even in a noisy background in time plane analysis. For any logarithmically scaled magnitudes in the algorithm, the first calculation of non-linear energy value and the rate of this nonlinear energy calculation to the linear energy is used in the decision algorithm of the proposed method. The quality of the proposed VAD was measured by

evaluating its detection rate, robustness against background noise and low computational complexity.

This paper is organized as follows; The Section 2 presents a general description of a VAD algorithm and the proposed VAD algorithm is introduced. The third section presents the test methods to evaluate the selected VAD methods and the results of the tests performed. Sect.4 concludes the article by evaluating the methods within the scope of the test results.

II. METHODOLOGY

The basic principle of a VAD algorithm is, as shown in Fig. 1, after a preprocessing process on the input signal, the extraction of feature vectors to be used for VAD decision in each analysis window, then from the speech-free silent (or noise-containing) regions of speech sections according to the specified threshold value. VAD algorithms perform their operations on speech signal segments allocated to analysis windows during pre-processing. In each analysis window, if a situation occurs above the threshold value that is compared at the decision stage, they produce a two-state result as 'unvoiced' (VAD = 0) if it does not appear 'voiced' (VAD = 1). Sections without speech are also called noise in various VAD algorithms. Analysis window length varies in each algorithm and varies in the range of [5-40ms]. The accuracy and reliability of the VAD algorithms depend on the chosen threshold as well as the applied methods. In some VAD applications, while the threshold value is fixed, some other VAD algorithms update the threshold value according to the base noise [16].

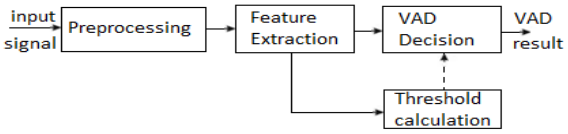

Fig. 1. Block diagram of a general VAD algorithm

In energy based VAD algorithms, the general approach for threshold value calculation is assuming that the average energy (E_r) of the signals in the selected analysis windows is calculated, assuming that there is no speech within the signal in a certain period of time (v is the number of analysis window) as in (1). Then, the signal energy (E_j) in an j^{th} analysis window is compared with E_r and the decision is made according to the scenario in (2). (where the k factor is a fixed value used to determine a reliable threshold value). In addition, there are algorithms that continually adapt the threshold value based on the background noise conditions.

$$E_r = \frac{1}{v} \cdot \sum_{m=0}^{v} E_m \tag{1}$$

$$VAD_{decision}: \begin{cases} 1, & if\ (E_j > k.E_r) \\ 0, & otherwise \end{cases} \quad (k > 1) \tag{2}$$

In this paper, the proposed VAD algorithm (called as the MuVAD method in this article), VADe2, VADrms and G.729 VAD algorithms which were evaluated for comparison purposes all produce VAD decisions in 10ms analysis windows.

A. G.729 VAD method

G.729-B VAD [3, 15] is a standard VAD encoder by ITU-T for fixed telephony and multimedia communications. In G.729 VAD algorithm, VAD decision is made by looking at 4 main parameters such as differential power calculation in the range of [0-1kHz], all band differential power calculation, line spectrum coefficients and zero crossing rate (*ZCR*). However, since the *ZCR* and energy calculation method used gives poor performance for low SNR input signals, G-729-B has low performance for noisy signals.

B. The proposed MuVAD method

The μ -law compression method [18] has been used for many years for the purpose of compressing directly quantized audio signals in communication systems and it is defined by the (3).

$$f(x) = sgn(x).ln(1 + \mu\,|x|) \,/\, ln\,(1 + \mu) \tag{3}$$

where μ is the compression parameter (μ : 0-255), and x ($-1 \leq x \leq 1$) is the normalized speech signal to be compressed. The method is applied to our VAD algorithm to boost up low amplitude signals (x) by first converting them to *f(x)* function as in (3) and evaluating their energy values.

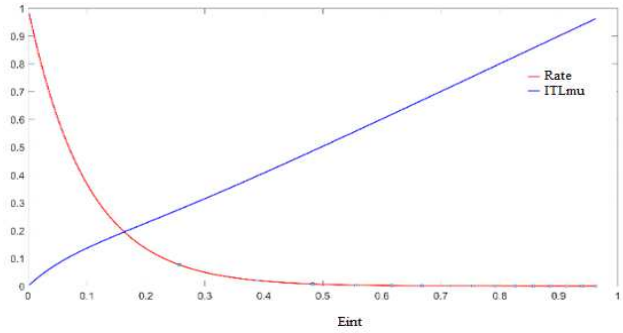

Fig. 2. ITL_{mu} and *Rate* against E_{int}

f(x) samples have the same sign with the samples of the original input signal *x*. In the study carried out in this article, Equation 3 and *f(x)* values are obtained from the input signal *x* and amplitude squared energy values of *f(x)* values are calculated by using (7) and the VAD decision process is operated according to the determined threshold value.

The performance parameters for an endpoint algorithm can be simplicity, robustness to background noise and exact definition of word boundaries even in noisy environments. To achieve these requirements, an energy-based measurement method is used in this paper to locate the voiced sections of the speech. With the MuVAD method, μ-law signal compression method in (3), and the energy calculation method in (7) that operates in the time plane, derived by using the amplitude square energy calculation method in (6), the input signal in each analysis window based on the calculation of their energy.

To find the parts containing an utterance in a speech signal, it is first necessary to identify the background noise of the waveform. For this reason, at the beginning of the algorithm, it is assumed that for the initial threshold value calculation

(E_{int}), the input speech signal ($x[n]$) does not contain an utterance through the selected f analysis window as in (4).

In this silence period, time-domain short time energy values of the background noise were calculated according to (7) and there is an average threshold value (E_{int}) according to (4). Then, for speech signals in other analysis windows, energy values calculated according to the threshold. *FE* in (7) and E_{int} in (4) were both used to determine the speech activity regions of speech signals.

$$E_{int} = \frac{1}{f}\sum_{i=1}^{f} FE_i \tag{4}$$

$$Erms(m) = \left[\frac{1}{N}\cdot\sum_{n=m.N}^{m.N+N-1} x^2(n)\right]^{\frac{1}{2}} \tag{5}$$

$$E(m) = \frac{1}{N}\sum_{n=m.N}^{m.N+N-1}|x[n]|^2 \tag{6}$$

$$FE(m) = \frac{1}{N}\sum_{n=m.N}^{m.N+N-1}|f(x[n])|^2 \tag{7}$$

In the MuVAD method, based on the energy calculation in (6), the energy formula ($FE(m)$) in (7) was used after converting the $x[n]$ input signal samples into $f(x)$ by (3). A fixed thresholding (ITL_{mu}) as in (9) was used to represent the effectiveness of the MuVAD method.

$$Rate = e^{-10.E_{int}} \tag{8}$$

$$ITL_{mu} = (1 + Rate) \times E_{int} \tag{9}$$

Since there is a nonlinear relationship between the input signal $x[n]$ and the MuVAD energy value $f(x[n])$, a nonlinear factor value is required to set a linear proportionally increasing threshold value relative to the E_{int} with the increasing noise. In this regard, the non-linear *'Rate'* value in (8) has been calculated experimentally as the most appropriate multiplier value for ITL_{mu} threshold calculation in (9). The graph showing the ITL_{mu} threshold change and *Rate* change according to the increasing E_{int} value is given in Fig. 2.

The proposed MuVAD method is based on a simple computation of energy values in adjacent frames. The algorithm finds the beginning points of the utterance when the lower threshold (*ITL*) is exceeded. Then, the endpoint of the utterance is found when the energy falls below *ITL*. This new energy-based algorithm does not require any backward/forward search and *ZCR* analysis due to the capability of boost up of energies of small magnitude signals in an utterance. The energy calculation method proposed here produces a fairly high energy signal for high amplitude input signals, while the amount of energy for low energy input signals remains relatively low and therefore is not easily triggered by a background noise of the noisy input signals.

III. EXPERIMENTAL STUDY

The VAD algorithm was designed to test the effectiveness of the MuVAD and other energy-based VAD methods under the same conditions. For this, in the energy calculation step in the VAD algorithm, Equation 7 for the MuVAD energy calculation method, Equation 6 for the VAD algorithm based on the amplitude-square energy calculation method called VADrms, and Equation 5 for the VAD based on mean square energy calculation method of the amplitudes called VADrms was used. Subsequently, the performances of each VAD method were measured using sounds with different background acoustic noises and their performance were measured. Fixed threshold method was used for all the VAD algorithms. Here, threshold calculation of VADe2 and VADrms methods were performed by using the Equation 1 and the factorial value k in (2) is selected as 2 for a reliable decision. Threshold calculation of MuVAD method was performed by using Equation 9 due to its logarithmic signal calculation formula in (3). For low noise conditions, ITLmu value is about 2 as seen in (9).

During the experimental study, MATLAB platform was used to design and test the algorithms. The test data-set was chosen to have different acoustic noises. The VAD methods used in the experimental study were evaluated with clean and noisy speech signals (NOISEUS database) [19], containing 30 different phonetically balanced sentences pronounced by three male and three female speakers. Clean speech files in the database are used as reference and 8 different noisy speech files (car, restaurant, babble, airport, street, station exhibition and train sounds) in 4 different SNR levels (0dB, 5dB, 10dB, 15dB) are used for testing. The sounds in the database are sampled at 8kHz. As a result, nearly 50minutes of sound test data with different acoustic noises for each SNR level were applied for each VAD algorithm during the tests. The performance of the algorithms was analyzed on the basis of robustness against background noise. (μ: 255) was selected to evaluate the maximum effectiveness of the MuVAD method.

TABLE I. (a) HR0 (b) HR1 and (c) P_d results of tested VAD methods for different SNR noise levels

%HR1 results				
SNR(dB)	***VADe2***	***VADrms***	***MuVAD***	***G.729***
15	71	78	86	98
10	70	66	78	95
5	78	69	84	88
0	48	26	60	76

(a)

%HR0 results				
SNR(dB)	***VADe2***	***VADrms***	***MuVAD***	***G.729***
15	100	99	95	53
10	97	98	91	48
5	76	83	61	49
0	85	96	76	51

(b)

%Pd results				
SNR(dB)	***VADe2***	***VADrms***	***MuVAD***	***G.729***
15	82	86	90	80
10	80	79	83	76
5	77	75	75	72
0	63	53	66	66

(c)

Because the noises were artificially added to the clean speech signal in NOIZEUS corpus, the clean speech sample of a noisy speech was used as reference for the evaluation of the performance of VADs during all of the applied test scenarios. The reference VAD was compared with the VAD outputs of each algorithm for each environmental noise with

4 different SNR values of input noisy speeches as can be seen in Table I. The Table presents the results of the VAD frames with the proposed MuVAD method and some other energy based VAD methods in the literature. The performance of each method in this article was evaluated by measuring the detection parameters such as its total accuracy in detecting speech as speech (speech hit rate, HR1), in detecting noise as noise (non-speech hit rate,HR0) and the total probability of detection (P_d) value. The aim was to minimize the restructured error variance and maximize the accuracy. From the test results in Table I, it can be seen that very good improvements was achieved with the proposed method. To summarize, we can say that there is a noticeable performance increase with the MuVAD technique when we compared with the other tested VAD algorithms. Besides, it was observed from the VAD results that it could perform an effective VAD operation even at 0dB SNR input level.

HR0 rates of VADe2 and VADrms remain higher than HR1 rates for all SNR levels. It can also easily be seen in Table I that, HR1, HR0 and detection rates (P_d) of MuVAD appears to be higher than VADe2 and VADrms for almost all input noise levels. Besides, there is no large reductions in the P_d despite the increased noise level in the input speech signal. it seems remain noise-robust to changing background noise. MuVAD method succeeds higher P_d for all kinds of input noisy signals and remains robust to changing environmental conditions in background noise.

CONCLUSIONS

A new logarithmic based energy calculation method is proposed to be used in VAD detectors. The effectiveness of the proposed method was tested under different background noise conditions and compared with other conventional time-domain energy calculation methods. The results showed that the new VAD detector provided noticeable increases in detection rate even under adverse conditions. Future works include comparing the new detector with some other energy-based VAD detectors in the literature.

REFERENCES

[1] Ramirez, J., Gorriz, J. M. and Segura, J. C. "Voice Activity Detection. Fundamentals and Speech Recognition System Robustness", In book:Robust Speech Recognition and Understanding,.I-Tech Education and Publishing, ISBN: 978-3-902613-08-0., June 2007.

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

[3] Beritelli, F., et al. "Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors." in IEEE Signal Processing Letters, vol. 9, no. 3, pp. 85-88, doi: 10.1109/97.995824, March 2002.

[4] Wu, B. F. and Wang, K. C. "Robust endpoint detection algorithm based on the adaptive band partitioning spectral entropy in adverse environments." , IEEE Transactions Speech Audio Processing, Vol. 13, pp. 762–775., 2005.

[5] Sadjadi, S. O. and Hansen, J. H.L. "Unsupervised Speech Activity Detection Using Voicing Measures and Perceptual Spectral Flux", IEEE Signal Processing Letters, Vol. 20, pp. 197-200., March 2013.

[6] Marzinzik, M. and Kollmeier, B. "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics", IEEE Trans. Speech Audio Process, Vol. 10, pp. 109-118., 2002.

[7] Kristjansson, T., Deligne, S. and Olsen, P. Voicing features for robust speech detection . INTERSPEECH. ,, pp. 369–372., 2005.

[8] Chung, K. and Oh, S. Y. "Voice Activity Detection Using an Improved Unvoiced Feature Normalization Process in Noisy Environments.", Wireless Personal Communications, Vol. 89, pp. 1-13.,3, 2015.

[9] Chen, S. H., et al. "Robust voice activity detection using perceptual wavelet-packet transform and Teager energy operator.", Pattern Recognition Letters, Vol. 28, pp. 1327-1332., 11, 2007.

[10] Tahmasbi, R. and Rezaei, "Change point detection in GARCH models for voice activity detection". IEEE Trans. Audio, Speech, Lang. Process., Vol. 16, pp. 1038–1046., Jul. 2008.

[11] Davis, A., Nordholm, S. and Togneri, R. "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold", IEEE Trans. Audio, Speech, Lang. Process., Vol. 14, pp. 412–424., Mar. 2006.

[12] Ferroni, G., et al. "A Deep Neural Network approach for Voice Activity Detection in multi-room domestic scenarios.", Killarney, Ireland. International Joint Conference on Neural Networks (IJCNN), 2015.

[13] Ali, Z. and Talha, M. "Innovative Method for Unsupervised Voice Activ¬ity Detection and Classification of Audio Segments",. IEEE Access, Vol. 6, pp. 15494-15504., 2018.

[14] Bie, F., et al. "DNN-based Voice Activity Detection for Speaker Recognition." CLST Technical Report. pp. 1-11., 2015.

[15] Enyassine, A., Shlomot, E. and Su, H. Y. "ITU-T recommendation G.729 annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data application". BSept. IEEE Commun. Mag., Vol. 35, pp. 64–73., 1997.

[16] Ozaydin, S. “Examination of Energy Based Voice Activity Detection Algorithms for Noisy Speech Signals”, European Journal of Science and Technology (EJOSAT), Special Issue, pp. 157-163, DOI: 10.31590/ejosat.637741, October 2019,

[17] Standard, "European. GSM 06.94. Digital cellular telecommunication system (Phase 2+); voice activity detector VAD for adaptive multi rate (AMR) speech traffic channels; general description. ETSI",, Tech. Rep.. V.7.0.0. Feb. 1999.

[18] Brokish, C. W. and Lewis, M. "A-Law and μ-Law Companding Implementations Using the TMS320C54x", Texas ınstruments, 1997.

[19] Loizou," P.NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms", Speech Communication, Vol. 49, pp. 588-601., 2017.