# *Silence Detection and Removal Method Based on the Continuous Average Energy of Speech Signal*


Abderrahmane Adjila
*Department of Mathematics and Computer Science – Laboratory of Mathematics and Applied Sciences (LMSA)-Université de Ghardaia*
BP: 455, Noumerate, Ghardaia, Algeria
a.adjila@yahoo.fr

Maamar Ahfir
*Department of Electronics*
*Université Amar Telidji-Laghouat*
BP: 37G, Laghouat, Algeria
m.ahfir@mail.lagh-univ.dz

Djelloul Ziadi
*Groupe de Recherche Rouennais en Informatique Fondamentale (GR2IF)*
*Université de Rouen Normandie, France*
476800 Saint-Etienne-du-Rouvray
djelloul.ziadi@univ-rouen.fr



2021 International Conference on Information Systems and Advanced Technologies (ICISAT) | 978-1-6654-7824-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICISAT54145.2021.9678476



***Abstract*—The speech signal processing is a very important domain in digital signal processing. This is because a variety of noise signals could degrade the original speech signal and make it unclear to user. This paper contributes to the literature by suggesting a method to detect and remove silence from the original speech signal based on the continuous average energy of the signal. Deleting the silence and voiceless segments from the speech signal are very beneficial to growth the overall performance and accuracy of the system in many domains of applications such as speech recognition and automatic speech segmentation. The results for a database which contains English, Arabic and French speech signals shows a better performance and robustness in noisy environment. The proposed method also has a less complexity compared to the recent method based on multi-scale product and its spectral centroid. In this research work the performance is evaluated using MATLAB tool.**



***Keywords—Digital Signal Processing, Speech segmentation, Silence detection , Silence removal, Continuous average energy (CAE)***


## I. INTRODUCTION

Silence detection and removal from speech signals is an essential pre-processing block in many applications systems such as: automatic speech recognition, speaker identification, multimedia communication, .... It does not only reduce the amount of processing and increase the accuracy of the speech processing systems, but also helps to reduce the end-to-end delay seen by the users in real time audio communication. Silence segment is a frame with all samples equal to zero. This means that the energy level is equal to zero. However noise from the microphone or the environment during recording may create some small variation. No data is being transferred in silence segments of the speech signal, so it is very necessary to detect and remove these silence segments from the speech signal. Once it deleted then it'll get omitted from the further processing. Segmentation is a voice activity detection of the recorded speech signal separating it into voiced, non-voiced and silence segments. Despite its thresholds setting limitation and less robustness in low SNRs (Signal to Noise Ratio) conditions, short time energy has been used for a long time as a very simple and fast method for voice activity detection [1, 2, 3, 4, 8, 10].

Many alternative energy-based methods have been developed to overcome the above limitations, but they are more complex for real time implementation, such as those described in [5, 6, 7, 9].

In this paper, a simple and fast method for silence detection and removal from speech signals is proposed and evaluated in the context of a restricted database which contains English, Arabic and French speech signals.

## II. PROPOSED METHOD

Knowing that the square of the samples of a given signal (Equation (1)), is used to evaluate its energy in the temporal field, and comparing it to the Shannon energy (Equation (2)), as illustrated in Fig. 1 showing that the squared signal is more representative in high amplitude samples.

$$E = s^2(t) \qquad (1)$$

$$E = -s^2(t).\log s^2(t) \qquad (2)$$

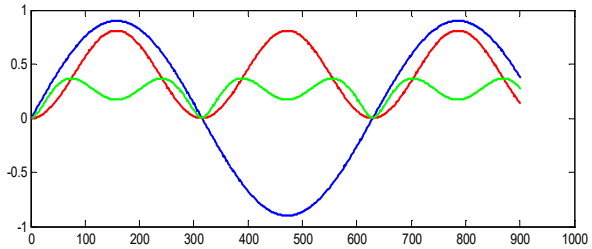
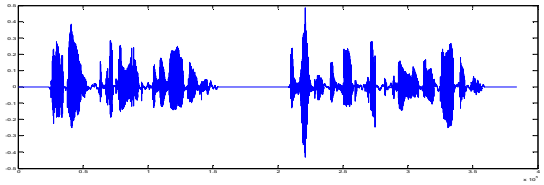


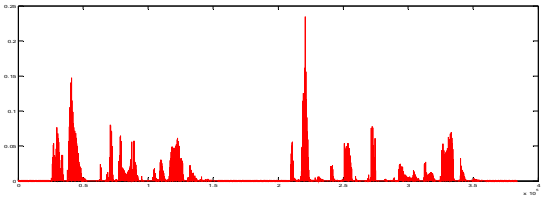

Fig.1. Temporal energy representation of the signal s (t) =sin (t): Original signal (blue) - Squared signal (red) - Shannon energy of the signal (green)

Energy representations of the speech signal (Fig. 2) highlight the relevance of the choice of our proposed method based originally on the square of the samples. Accordingly, we can see that there is not a big deference in silence zone detection while the squared signal is very simple in processing time perspective.

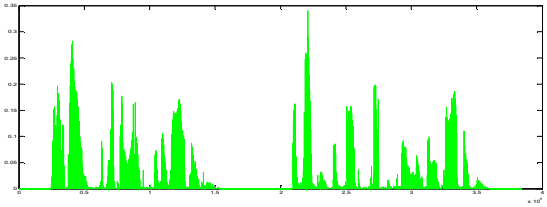


(a)


978-1-6654-7824-3/21/$31.00 ©2021 IEEE



(b)



(c)

Fig. 2 Energy representations of the speech signal. (a) speech signal (b) Squared speech signal (c) Shannon energy of the speech signal

The proposed method can be decomposed into three steps as shown in Fig. 3. The first step consists of calculating the continuous average energy for every 50 ms throughout the original signal (Equation (3)).

$$E_k = \frac{1}{N}\sum_{n=k}^{N+k-1} x^2(n). \qquad (3)$$

Where, $x\,(n)$ is the sampled sequence with the sampling frequency Fe of the original signal and $N$ represents the quantity of samples within 50 ms of the moved window for every sample $k$ of the normalized signal (i.e, a shift of 1 sample) which makes the signal overlapped by 98% of the moved window.

The second step consists of computing the normalized continuous average energy to obtain the signal's envelogram (Equation (4)).

$$E_m = \frac{E_k - avr(E_k)}{var(E_k)}. \qquad (4)$$

Where, $avr(\mathrm{E_k})$ is the mean value of $\mathrm{E_k}$ and $var(\mathrm{E_k})$ is its standard deviation.

Finally the third step consists to identify the maxima and indices of the lobes of $E_m$ and their boundaries, by the zero crossing method.

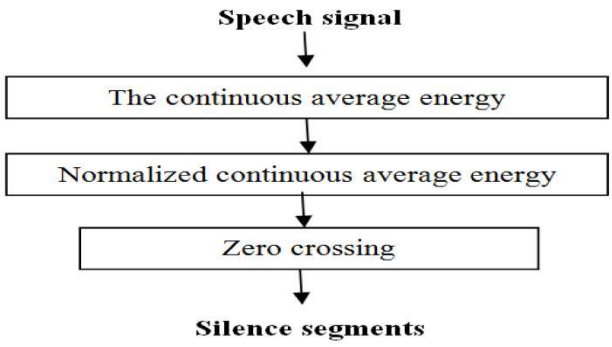



Fig. 3 Block diagram of the proposed method for silence detection

Fig. 4 and Fig. 5 show the CAE of a clean speech signal and the detected silence segments respectively.

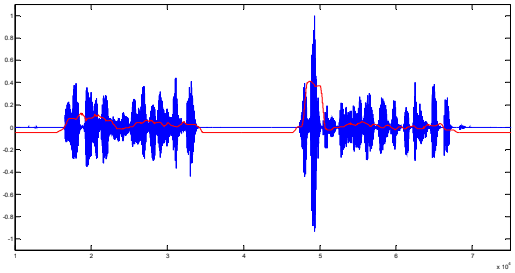


Fig. 4 Continuous average energy (red) – speech signal (blue)

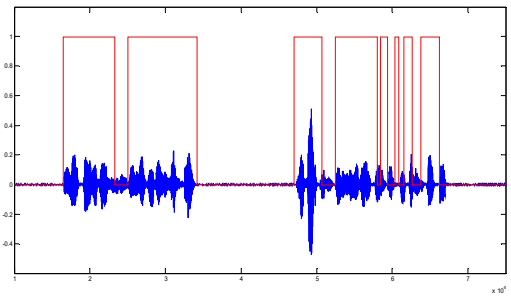


Fig. 5 Detected segments (red) – speech signal (blue)

## III. PERFORMANCE TESTING

The method was evaluated in the context of varied database which contains 13 speech signals for male and female speakers, with duration about 10 seconds and stored in wave format: 4 English, 2 Arabic and 7 French. All the files were digitized with a resolution of 32 bit and 48 kHz sampling rate and down sampling later to 8 kHz.

The testing was performed through four different criteria, reflecting the Silence Detection (SD) performance. Fig. 6 shows the different evaluation criteria.

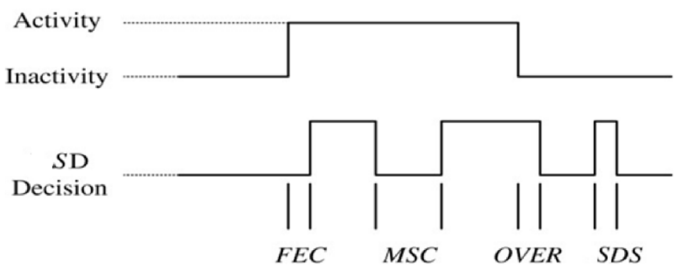



Fig. 6 Evaluation Criteria

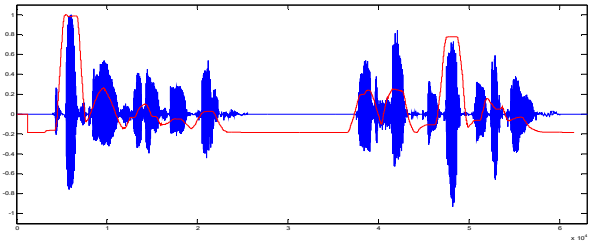
- CORRECT: decision made by the SD is correct.
- FEC (Front-End Clipping): Clipping introduced when passing from speech to silence.
- MSC (Mid-Speech Clipping): Clipping introduced when silence misclassified as speech in an utterance.

- OVER (Carry Over): speech detected as silence because the SD flag remaining active when passing from silence to speech.
- SDS (Speech detected as silence): speech detected as silence within a speech period.

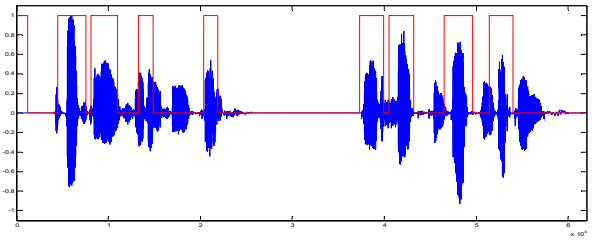
The parameter SDS is not used because of short silence, where only silence of duration superior or equal to 0.5 s is considered.

FEC and MSC are signs of proper rejection, while SDS and OVER are signs of fake acceptance. CORRECT parameter indicates the quantity of correct decisions made. Thus all the four parameters FEC, MSC, SDS and OVER must be minimized and the CORRECT parameter need to be maximized to gain the best overall system performance.

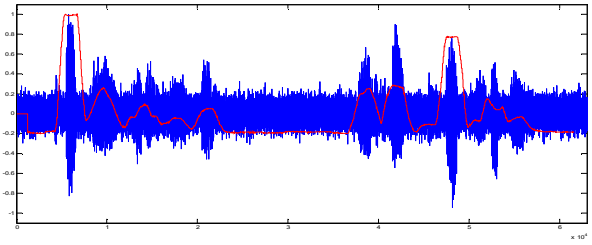
To evaluate the robustness in noisy environment of our algorithm, we measured the four performance parameters by adding a stationary noise in three different SNR levels (-10 dB, - 5 dB, 0 dB) to the clean speech signals. Fig. 7 shows an example of original speech signal for an English speaker and its continuous average energy followed by its detected silence segments in a different SNR levels.



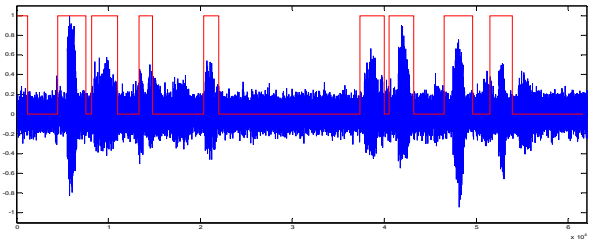
(a)



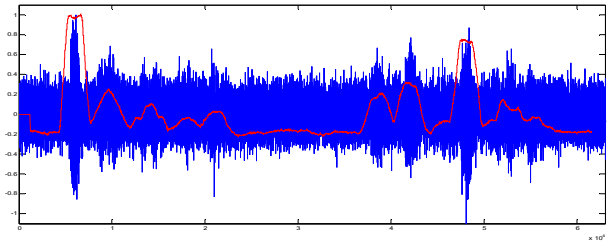
( b)



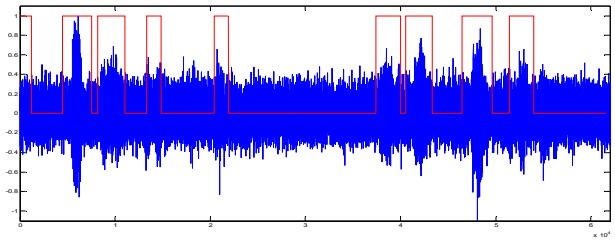
(c)



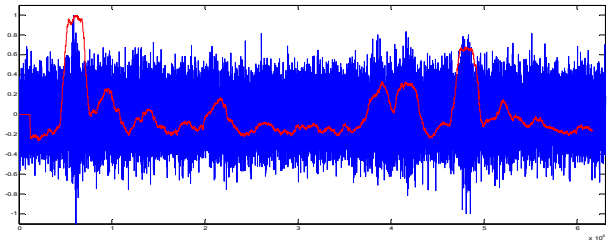
(d)



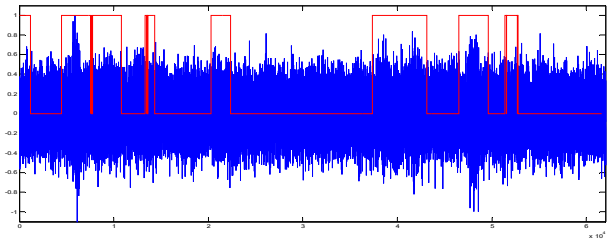
(e)



(f)



(g)



(h)

Fig. 7 Silence Detection for an English speech signal (a) clean speech signal and its CAE (b) Detected segments (c) white noise added at 0 dB SNR and its CAE (d) Detected segments (e) white noise added at -5 dB SNR and its CAE (f) Detected segments (g) white noise added at -10 dB SNR and its CAE (h) Detected segments.

When we compare our method to the method proposed in [9] which uses two features based on multi-scale product (MP) of the clean speech, namely the spectral centroid of MP and the zero crossings rate of MP, our method is very simple and less complex for implementation and have the best performance. The obtained results using the same database and the four testing parameters in different SNR levels are shown in TABLE.1 and TABLE.2.

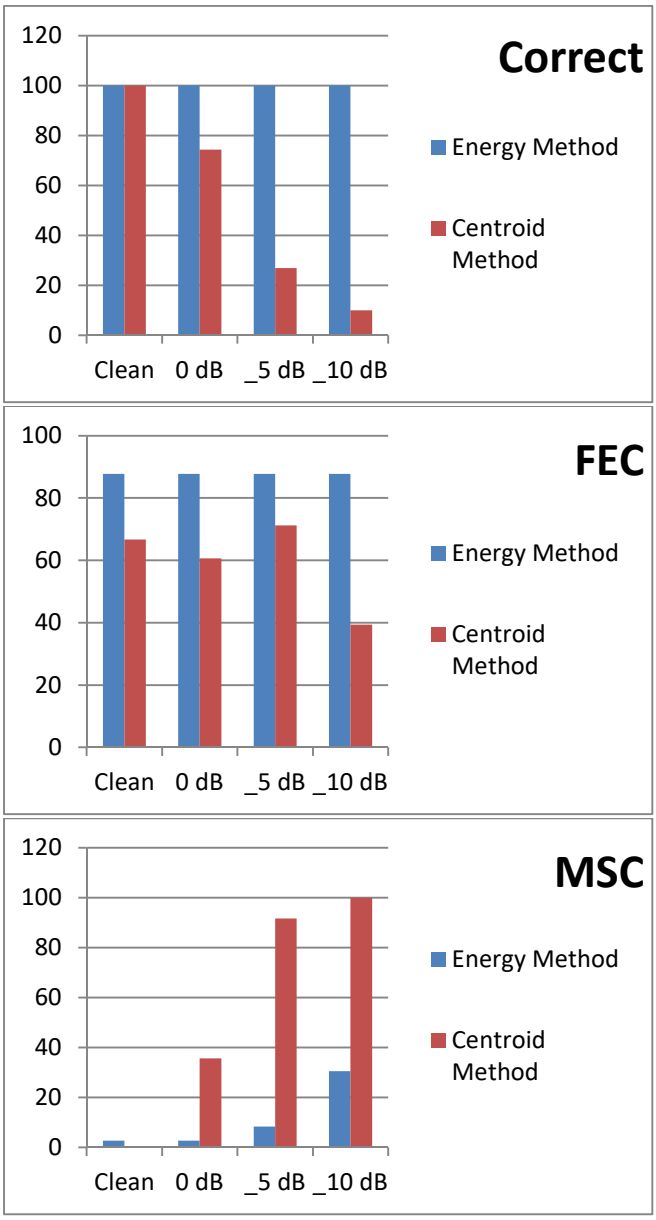
TABLE.1 Silence detection performance using the Continuous average energy method

| SNR / SD | **Clean** | **0 dB** | **-5 dB** | **-10 dB** |
|---|---|---|---|---|
| Correct % | **100** | **100** | **100** | **100** |
| FEC % | **87.17** | **87.17** | **87.17** | **87.17** |
| MSC % | **2.77** | **2.77** | **8.33** | **30.55** |
| OVER % | **27.88** | **27.88** | **25.32** | **25.32** |

TABLE.2 Silence detection performance using the spectral Centroid of MP

| SNR / SD | Clean | 0 dB | -5 dB | -10 dB |
|---|---|---|---|---|
| Correct % | **100** | **74.31** | **26.96** | **10** |
| FEC % | **66.66** | **60.60** | **71.21** | **39.39** |
| MSC % | **0** | **35.64** | **91.66** | **100** |
| OVER % | **52.77** | **80.55** | **100** | **100** |

For more clarity of the compared results, we use the graph presented in Fig. 8



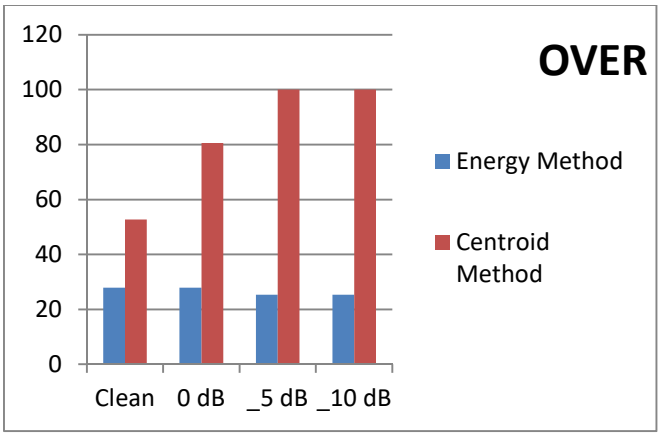




Fig. 8 The percentage of the four performance parameters for the two methods in three SNR levels

As shown in TABLE.1, TABLE.2 and Fig. 8 for clean speech, the two methods have almost the same performance score. In the noisy environment, Centroid method doesn't overpass the proposed method in any parameter except the *FEC* which is less important than the *CORRECT* and *MSC* parameters. In term of noise robustness, we note that the performance parameters score (*CORRECT*, *MSC, OVER)* in the proposed method for -10 dB SNR is better than the score obtained in Centroid method for 0 dB SNR.

## IV. CONCLUSION

In this research a novel proposition for silence detection and removal is suggested. Silence detection and removal is an important phase in systems like speaker identification and speech recognition. It increases the overall performance and minimize the processing time. It is concluded from the results of performance testing that the proposed method detect 100% unvoiced segments from the speech signal even in a very noisy environment (-10 dB) without corrupting the voiced segments. In the future work the proposed method should be compared to other methods and tested in more sized database to confirm the obtained results.

## REFERENCES


[1] B. Atal, L. Rabiner, , "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition", IEEE Transactions on ASSP, Vol-24, Issue: 3, pp 201 - 212 , Jun 1976.

[2] D. G. Childers, M. Hand, J. M. Larar, "Silent and Voiced/Unvoied/ Mixed Excitation(Four-Way), Classification of Speech", IEEE Transaction on ASSP, Vol-37, No-11, pp. 1771-74, Nov 1989.

[3] D. Enqing, L. Guizhong, Z. Yatong and C. Yu, "Voice Activity Detection Based on Short Terme Energy and Noise Spectrum Adaptation", 6th International Conference on Signal Processing, Vol 1, pp 464-467, 2002.

[4] S. Zhang, "An energy based adaptative voice detection approach", 8th International Conference on Signal Processing, Vol 1, 2006.

[5] T. R. Sahoo and S. Patra, "Silence Removal and Endpoint Detection of Speech Signal for Text Independent Speaker Identification", IJIGSP Vol. 6, No. 6, pp 27-35, May 2014.

[6] P.K. Ghosh, A. T. Tsiartas and S. Narayanan, "Robust Voice Activity Detection Using Long-Term Signal Variability", IEEE Transactions on ASLP, Vol 19, No 3, pp 600-613, March 2011.

[7] V.K. Prasad, T. Nagarajan and H. A. Murthy, "Automatic Segmentation of Continuous Speech Using Minimum-phase Group Delay", Speech Communication 42, Elsevier, pp 429-446, 2004.

[8] M. H. Moattar and M. M. Homayounpour, “A Simple but Efficien Real Time Voice Activity Detection Algorithm”, 17th European Signal Processing Conference (EUSIPCO 2009), pp 2549-2553, Glasgow, Scotland, August 24-28, 2009.

[9] M. A. Ben Messaoud, A. Bouzid, and N. Ellouze, “Automatic Segmentation of the Clean Speech Signal”, International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering Vol:9, No:1, pp 114-117, 2015 .

[10] M.Z. Belmecheri, M. Ahfir and I. Kale “Automatic Heart Sounds Segmentation based on the Correlation Coefficients Matrix for Similar Cardiac Cycles Identification”, Elsevier, BSPC, Vol 43, pp 300-310, 2018.