Real Time Implementation of Voice Activity Detection based on False Acceptance Regulation

Charaf Eddine Chelloug and Atef Farrouki

Siscom Laboratory, Department of Electronic
University of Constantine 1, 25000, Constantine, Algeria.
charafeddinechelloug@gmail.com, farroukiatef@yahoo.fr

Abstract: In speech compression systems, Voice Activity Detection (VAD) is frequently used to distinguish active voice from other noisy sounds. In this paper, a robust approach of VAD is presented to deal with non-stationary noisy environments. The proposed algorithm exploits adaptive thresholding technique to keep a desired False Acceptance (FA) rate. Iterative hypothesis tests, using signal energy, are implemented to discard or to accept the successive audio frames as active voice. According to the stationary property of the speech, we provide a smoothing method to obtain final VAD decisions. The main contribution of the proposed algorithm concerns its ability to automatically adjust the energy threshold according to the local noise estimator. We analyzed the proposed approach by presenting a comparison with the G.729-B via the NOIZEUS database. The VAD architecture is implemented on a Microcontroller-based system (MCU). Several tests have been conducted by performing real time acquisition via the Input/Output ports of the MCU-system.

Keywords: Voice Activity Detection, False Acceptance, Adaptive Thresholding, Non-Stationary Noise, Microcontroller Unit.

1. Introduction

The function of the Voice Activity Detection is to separate active voice from silence in utterances [1], it is an important component in many communication systems like mobile communications [2], speech coding [3], compression [4], digital hearing aids devices [5] and automatic speech recognition systems [6]. The VAD decisions of the ETSI-AMR are based on pitch and tone detection as well as correlated complex signal analysis [8]. The ETSI Advanced Front-End (AFE) coder has been adopted in 2002 for distributed speech recognition [9]. The VAD algorithm implemented in the ETSI-AFE uses Wiener filter and non-speech frame dropping to discard unvoiced intervals [9].

To enhance the robustness of VAD systems against strong noisy environments, various conventional statistical-based techniques were employed. Also, the widely used Gaussian plus Laplacian distribution is assumed to design noise-robust decision rules [10-14]. In [11], the Likelihood Ratio Test is used by assuming complex Laplacian and Gamma probability density functions (pdf) to model the DFT coefficients of the noisy speech. These methods act robustly when they are applied to relatively clean signals, while they exhibit poor performances in low Signal-to-Noise Ratio situations, or in presence of power noise transitions. To deal with these non-homogeneous situations, several speech enhancement techniques have been proposed, such as the Likelihood Ratio Sign Test (LRST), the Neyman–Pearson criterion [15] and the multiple observations Likelihood Ratio Test [16].

In this work, we propose a generalization of the VAD technique, presented in [14], for non-stationary noisy environments. Also, we introduce a VAD algorithm that maintains a constant False Acceptance probability within non speech periods. The iterative binary tests allow the rejection or the acceptance of the tested frame as active voice. The proposed speech detection is

Received: June 9th, 2018. Accepted: October 8th, 2020
DOI: 10.15676/ijeei.2020.12.3.13
based on the full-band energy as a decision variable. Ended, a moving window, containing
energies of free-speech frames, is updated only in the case of null-hypothesis decision (absence
of voice). By doing so, the above mentioned window, also called noise buffer, would likely contain
the estimation of the background noise, even when the algorithm processes voiced periods. To
initialize the noise buffer, we assume that the first 100 ms-frames of any call do not contain voiced
signals [12].

To remove discontinuities that may appear in speech intervals, we also develop a smoothing
procedure by taking into account the partial decisions related to the last contiguous frames.

The architecture of the proposed VAD is targeted on the STM32F746G-DISCO board [17] to
build a real time system exploiting the integrated components such as Codec, Serial Audio
Interface (SAI) and DMA channels. The obtained architecture has been evaluated via the
acquisition of real speech signals and analyzed by the software tools available on the MCU-
system.

2. Proposed VAD algorithm

A. Energy-based VAD

The diversity of both speech signals and background noise make the VAD problem quite
complicated in practical situations. The VAD generates a voice activity decision every 10ms
according to the specifications of the G.729 coder [7]. In general, four measures are required for
VAD decisions; namely, full-band energy, low-band energy, zero-crossing and line spectral
frequencies.

The full-band energy by frame, $E_{FB}$, corresponding to the logarithm of the normalized first
autocorrelation coefficient is given by [7]

$$E_{FB} = 10 \log \left( \frac{1}{N} R_x(0) \right)$$

(1)

where $E_{FB}$ is the full-band energy by frame, $R_x(0)$ is the autocorrelation coefficient and $x_i$ are the sampled speech plus noise
amplitudes, thus

$$R_x(0) = \sum_{j=1}^{N} x_j^2$$

(2)

We assume that speech and additive noise are described by Laplacian and Gaussian models
respectively [6]. Also, under silence hypothesis ($H_0$), the AWGN samples are assumed
Independent and Identically Distributed (IID) random variables. Consequently, the unvoiced
samples are distributed according to the centered Gaussian probability density function (pdf) with
the variance $\sigma^2$, also referred to as $G(0, \sigma/2)$.

In the rest of this study, we use $R_x(0)$ as a decision random variable instead of $E_{FB}$. In fact
Eq.(1) refers to a monotonically increasing function, consequently $E_{FB}$ and $R_x(0)$ are equivalents
in term of binary decision. Thus, we define the decision variable for the $k^{th}$ frame of length $2M$
samples as

$$E_k = \sum_{i=1}^{2M} x_i^2$$

(3)

The IID processes $x_i \{i=1,...,2M\}$ represent the AWGN samples within a silence frame. By
defining $y_i$ as the sum of two squared normal random variables $y_i = x_{2i-1}^2 + x_{2i}^2$, we observe
that $y_i$ is governed by an exponential pdf of standard deviation $\sigma$, then

$$f_{y_i}(y) = \frac{1}{\sigma} \exp \left( -\frac{y}{\sigma} \right) \quad \{ i = 1, ..., M \}$$

(4)

By substituting $y_i$ in (3), $E_k$ becomes the sum of $M$ IID exponentially-distributed variables.
Therefore, under $H_0$-hypothesis, it is shown that $E_k$ follows a Gamma distribution [18] of
parameters $(M, \sigma)$

$$f_{E_k}(e) = \frac{e^{M-1} \exp(-\frac{e}{\sigma})}{\Gamma(M)\sigma^M}$$

(5)

At iteration $k$, the $k^{th}$ hypothesis test provides a binary decision to declare whether the frame
under investigation is unvoiced ($H_0$) or contains noisy speech ($H_1$). To realize such a detection
scheme, we define the False Acceptance rate $F_A$, as the probability to decide $H_1$, whereas the frame
under processing contains noise only ($H_0$), that is

$$F_A = \text{Prob}(E_k > Th/H_0)$$

(6)
Next, we compare the energy $E_k$ of the $k$th frame to the adaptive threshold $Th$ and a partial decision $d_k$ is made in favor of $H_1$ (voiced) or $H_0$ (unvoiced) in accordance to the following statistical test

$$E_k \gtrless_{H_0} Th$$

(7)

The adaptive threshold $Th$ is closely related to the noise level, thus, the background level should be estimated by considering only the last frames having been declared unvoiced. Therefore, the adaptive threshold can be defined as a local mean level

$$Th = T \cdot Z$$

(8)

where $Z = \sum_{i=1}^{N_0} E_i$ is the estimated mean level of the background and $E_i$’s are energies of the $H_0$-frames. The scaling factor $T$ is set to keep a desired $FA$ probability during silence periods ($H_0$)

### B. Constant false acceptance rate

To calculate $T$ that guarantees a constant probability of False Acceptance, we should derive mathematical expression for the probability $FA$. By substituting $Th$ in (6) and using the pdf $f_z$ and $f_{E_k}$ of $E_k$ and $Z$ respectively, we obtain the probability $FA$ as

$$FA = \int_{-\infty}^{\infty} f_z(z) \int_{-\infty}^{T_h} f_{E_k}(e) \, de \, dz$$

(9)

Resolving the contour integral yields to [18]

$$FA = -\sum \left[ \right]$$

(10)

where $\Gamma(M) = (M-1)!$ and we use the fact that

$$\frac{d^L}{d\omega^L} (U, V) = \sum_{i=0}^{L} \binom{L}{L-i} U^{(L-i)} \cdot V^{(i)}$$

(15)

By using (15) into (14), we obtain after some algebraic manipulations the false acceptance rate to be
\[ F_A = \sum_{i=0}^{M-1} \left( \frac{MN_0^{-1+i}}{MN_0^{-1}} \right) \frac{T^i}{(1+T)^{MN_0+i}} \]  

We note that the factor T should be beforehand determined by solving numerically equation (16).

C. Proposed VAD Approach

The pre-computed factor T allows to keep a constant FA probability against unvoiced regions. In turn, the partial decisions \( d_k \) are processed by the smoothing block in order to provide the final VAD binary decision. The smoothing principle consists of accepting an interval as speech if \( P \) successive frames, preceding the current one, have been declared \( H_1 \) i.e.\( (d_{k-1}=1, d_{k-2}=1, \ldots, d_{k-P}=1) \).

Similarly, when a voiced period is processed, the algorithm decides \( H_0 \) only if the \( P \) previous frames are detected as noise. In fact, the above mentioned conditions have been added according to the stationary property of speech signals and then avoiding eventual discontinuities. Below, we present the necessary explanation about the working principle of our VAD system, as described by Figure 1.

The analog signal is sampled at 8 kHz, next the framing process (10 ms) leads to \( 2M=80 \) samples by frame. The first \( N_0 \) incoming \( H_0 \)-frames (free speech) are loaded into a First-In First-Out (FIFO) noise buffer to set the initial background level \( (Th= TZ) \). Thereafter, at the \( k^{th} \) step \( (k > N_0) \), a partial decision \( d_k \) is provided by applying the test defined in (7) to decide if the \( k^{th} \) frame corresponds to speech plus noise \( (H_1) \) or noise only \( (H_0) \).

Next, the generated \( (P+1) \) partial decisions \( (d_k) \) are stored into a shift register. Indeed, the \( P \) decisions preceding the current one \( (d_k) \), are continually checked to perform the smoothing if the above mentioned particular conditions are met.

- **Smoothing**
  The role of this task is to remove undesirable discontinuities that may appear in both speech and non-speech regions, by generating smoothed final VAD decisions. However, we note that the smoothing task will introduce a slight latency of few tens milliseconds in the VAD output (Delay of \( P \) frames). Because of the low dynamic of the audio signal, the above delay cannot have any impact on the real time mode. Smoothing works by reversing few partial binary decisions already declared \( H_0 \) inside a voiced period (among \( P \) successive frames). Indeed, during a positive transition from \( H_0 \) to \( H_1 \), at least \( P \) successive \( H_1 \)-frames should be detected to be classified as active voice \( (VAD=1) \). In a similar way, during a negative transition from \( H_1 \) to \( H_0 \), the smoothed VAD is set to zero in the case where \( P \) successive \( H_0 \)-frames have been detected. The smoothing of \( d_k \)'s becomes effective only if the above mentioned conditions are met.

- **Noise buffer updating**
  This procedure consists of inserting the most representative unvoiced energies. As shown in Figure 1, whenever \( d_{k-P}=0 \), the energy of the detected noise frame \( (E_{k-P}) \) is inserted into the FIFO Noise-Buffer. By doing so, the updated content of the buffer becomes available for the next hypothesis test.

- **Scale factor adjustment**
  In the presence of uniform noise environment, where the power level is constant, the dynamic threshold \( Th = TZ \) guaranties a constant FA via the pre-computed value of T (see Eq.16). However, in non-homogeneous situations caused by abrupt power noise transitions during the same session call, the proposed VAD exhibits intolerable false acceptance rate leading to poor detection performances. To circumvent this problem, the factor T is automatically corrected whenever a transition region is under processing. A long-term set of energies is stored in a FIFO Buffer over the \( fu \) frames preceding the one under processing (Figure 1.b). We assume that \( E_{max} \) and \( E_{min} \) represent extreme values of the FIFO register of size \( fu \). For every set of \( fu \) frames, the moving energy difference \( (E_{max} - E_{min}) \) is compared to a predefined energy deviation threshold \( \Delta E \), to decide if an abrupt power transition has been occurred. Typically, a full-band energy deviation
above 15 dB indicates a significant change in the background noise environment [7]. Consequently, the scaling factor adjustment takes place according to the following condition

\[ E_{\text{max}} - E_{\text{min}} \geq \Delta_E \]  

\[ (17) \]

Noting that the condition (17) should be tested, every fu frames, in order to reset the adaptive threshold TZ within the interval \([E_{\text{min}} , E_{\text{max}}]\) whenever a significant environment change is detected; in this case, the factor T is updated as follow

\[ T = \frac{E_{\text{min}} + \alpha \Delta_E}{Z} \]  

\[ (18) \]

where \(\alpha\) corresponds to the adjustment level parameter which has been empirically fixed through intensive simulation, using Noizeus Database [20].
3. Real time implementation

The proposed VAD method is implemented on the high performance MCU ARM Cortex®-M7 STM32F746NG board [17], which provides a single floating point unit supporting all ARM® single precision instructions and data types.

Figure.2 illustrates the function of the real time system. It consists of three fundamental parts: an audio source (from the microphone), a MCU board and the speaker. The analog speech signal from audio source is passed through the Analog to Digital (A/D) converter of the codec WM8994ECS/R to be sampled at 8 kHz with a 16 bits linear PCM (Pulse Code Modulation) data. The obtained digital signal will be frame-by-frame processed via the proposed real time VAD application. Finally, the output digital data is converted to an analog signal through the D/A converter and passed through the speaker, which allows a subjective evaluation of the studied VAD method.

A. Design of hardware implementation

The hardware implementation architecture is described in Figure.2, audio streaming data is digitized by A/D of the integrated codec WM8994ECS/R, the two blocks A and B are chosen to transfer obtained samples from Serial Audio Interface (SAI) to Direct Memory Access (DMA). The Code is developed using C language, compiled and loaded to the MCU board via USB port using embedded IDE of STM software.

![Figure 2. Real time system architecture on MCU board](image-url)
B. Data and memory mapping

The efficiency of a real time microcontroller-based application, in terms of area and power consumption depends, among others criterions, on the memory requirements of the implemented architecture. In the following, we first describe the memory mapping required to store and to process the incoming audio frames. We also present an overview of the main variables and constants used by the proposed VAD algorithm. Table 1 summarizes the signification, the size, the type and the address locations of the used memory resources.

The used sampling frequency is 8 KHz with Stereo mode, therefore each 10 ms we get one recorded data frame. Each frame is composed by two half blocks fulfilled and structured as detailed in Figure 3. Thus, 320 bytes should be online stored for the two audio input channels (right and left), however, we processed only 160 bytes corresponding to a single channel, and ignore the second data channel.

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Designation</th>
<th>Type</th>
<th>Number of bytes</th>
<th>Starting Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>Scaling factor</td>
<td>Float</td>
<td>4</td>
<td>0x20000114</td>
</tr>
<tr>
<td>P</td>
<td>At least P accepted voiced frames</td>
<td>unsigned short int</td>
<td>2</td>
<td>0x20000076</td>
</tr>
<tr>
<td>ΔE</td>
<td>Energy Variation threshold</td>
<td>Float</td>
<td>4</td>
<td>0x20000118</td>
</tr>
<tr>
<td>A</td>
<td>Adjustment level</td>
<td>Float</td>
<td>4</td>
<td>0x2000011c</td>
</tr>
<tr>
<td>Cf</td>
<td>Counter Frame</td>
<td>unsigned long int</td>
<td>4</td>
<td>0x20000120</td>
</tr>
<tr>
<td>frm_prsnt</td>
<td>Buffer containing samples of the current frame</td>
<td>Float</td>
<td>4*80</td>
<td>0x20000124</td>
</tr>
<tr>
<td>Rx</td>
<td>Autocorrelation coefficient</td>
<td>Float</td>
<td>4</td>
<td>0x20000264</td>
</tr>
<tr>
<td>Ef</td>
<td>Energy Full Band</td>
<td>Float</td>
<td>4</td>
<td>0x20000268</td>
</tr>
<tr>
<td>Z</td>
<td>mean level estimation of the background energy</td>
<td>Float</td>
<td>4*10</td>
<td>0x2000007c</td>
</tr>
<tr>
<td>VAD_buff</td>
<td>Partial Decisions Register</td>
<td>unsigned short int</td>
<td>2*(P+1)</td>
<td>0x200000cc</td>
</tr>
<tr>
<td>Ef_buff</td>
<td>FIFO Energy-Buffer of the last (P+1) frames</td>
<td>Float</td>
<td>4*4</td>
<td>0x200000dc</td>
</tr>
<tr>
<td>Emm</td>
<td>FIFO Energy-Buffer of the last fuframes</td>
<td>Float</td>
<td>4*15</td>
<td>0x200000a4</td>
</tr>
<tr>
<td>E_min</td>
<td>Energy Min</td>
<td>Float</td>
<td>4</td>
<td>0x20000104</td>
</tr>
<tr>
<td>E_max</td>
<td>Energy Max</td>
<td>Float</td>
<td>4</td>
<td>0x20000100</td>
</tr>
<tr>
<td>Th</td>
<td>Adaptive Threshold</td>
<td>Float</td>
<td>4</td>
<td>0x20000270</td>
</tr>
<tr>
<td>Vf</td>
<td>Final VAD decision</td>
<td>unsigned short int</td>
<td>2</td>
<td>0x20000074</td>
</tr>
</tbody>
</table>
4. Experimentation and results

For evaluation purpose, we compared our approach to the VAD of the standard G.729 through the well-known NOIZEUS Database [20]. To simulate non-stationary noisy situations, we generated power-controlled synthetic AWGN to be added to the NOIZEUS clear utterances. In addition to that, several real time tests were applied via online audio acquisition, by exploiting the MCU sound devices.

A. NOIZEUS database

The NOIZEUS database [20] consists of thirty clear utterances, corrupted by different real world noises taken from AURORA database [21], namely: restaurant, car, street, airport, babble, exhibition hall and train station, at different SNR levels 0 dB, 5 dB, 10 dB, 15 dB, all files are saved in wav format with 8 KHz sampling and 16 bit PCM coding.

To evaluate our system comparatively to the G.729 VAD, we have taken into account the usual criteria [10];
- Correct: Correct decisions for both voiced and noised frames.
- TR (True Rejection rate): obtained by combining the front end clipping (FEC) and the mid speech clipping (MSC) indicators.
- FAex (False Acceptance rate): obtained by combining carry over (Over) and noise detected as speech (NDS) indicators.

Table 2 summarizes the results for the whole database, by using the following values; $T=0.13$, $P=3$, $N_0=8$ and a prescribed False Acceptance value of $0.06 \times 10^{-3}$.

Both the proposed VAD and the G.729 have been applied to the whole NOIZEUS database. The obtained results using above criterions are summarized in Table 2. We notice that the proposed VAD acts robustly, particularly in presence of high noise power (0 dB and 5 dB). Indeed, a high value of the Correct parameter shows a satisfactory detection performance, while a low FA indicates a good compression rate. However, in usual moderate noisy environments, we note that the suggested VAD and the G.729-B exhibit almost the same performances. Let us recall that the criterions Correct, TR and FA are just necessary to validate the mathematical derivation (section 2.A), and then, to show the behavior of the proposed system against several operational situations. To be completely accurate, the comparison with the G.729-B should take into account others...
important criterions such as the bandwidth consumption, where the G729 is known to be a reference in this area.

Table 2. Comparison with G.729-B VAD using NOIZEUS database

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>Approach</th>
<th>Correct TR</th>
<th>FA</th>
<th>FEC</th>
<th>MSC</th>
<th>OVER</th>
<th>NDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>G.729</td>
<td>80.440</td>
<td>5.701</td>
<td>13.861</td>
<td>1.020</td>
<td>4.691</td>
<td>1.221</td>
</tr>
<tr>
<td></td>
<td>Suggested</td>
<td>85.062</td>
<td>2.730</td>
<td>12.202</td>
<td>0.601</td>
<td>2.130</td>
<td>1.600</td>
</tr>
<tr>
<td>10</td>
<td>G.729</td>
<td>79.540</td>
<td>4.460</td>
<td>16.003</td>
<td>0.760</td>
<td>3.702</td>
<td>1.491</td>
</tr>
<tr>
<td></td>
<td>Suggested</td>
<td>83.704</td>
<td>3.811</td>
<td>12.491</td>
<td>0.730</td>
<td>3.081</td>
<td>1.603</td>
</tr>
<tr>
<td>5</td>
<td>G.729</td>
<td>78.931</td>
<td>3.692</td>
<td>17.380</td>
<td>0.512</td>
<td>3.180</td>
<td>1.670</td>
</tr>
<tr>
<td></td>
<td>Suggested</td>
<td>82.801</td>
<td>4.731</td>
<td>12.460</td>
<td>0.780</td>
<td>3.950</td>
<td>1.582</td>
</tr>
<tr>
<td>0</td>
<td>G.729</td>
<td>78.060</td>
<td>4.005</td>
<td>17.942</td>
<td>0.442</td>
<td>3.571</td>
<td>1.721</td>
</tr>
<tr>
<td></td>
<td>Suggested</td>
<td>80.922</td>
<td>6.110</td>
<td>12.961</td>
<td>0.810</td>
<td>5.302</td>
<td>1.600</td>
</tr>
<tr>
<td>Mean</td>
<td>G.729</td>
<td>79.240</td>
<td>4.461</td>
<td>16.300</td>
<td>0.680</td>
<td>3.780</td>
<td>1.530</td>
</tr>
<tr>
<td></td>
<td>Suggested</td>
<td>83.121</td>
<td>4.350</td>
<td>12.530</td>
<td>0.731</td>
<td>3.621</td>
<td>1.600</td>
</tr>
</tbody>
</table>

B. Non-stationary noisy situations

The proposed VAD-system is evaluated in non-stationary background by concatenating four clear audio signals from NOIZEUS to form a clear wav file of 10 seconds, this file is corrupted by synthetic white noise with different SNR (-5dB, 10dB, 25dB, 10dB, -5dB), each SNR level is applied during 2 seconds. In these experimental tests, it is essential to consider several situations ranging from weak to strong noisy environments. Low SNR intervals (-5 dB, 10 dB) have been generated with intending to focalize on the behavior of the proposed VAD system against highly noisy backgrounds. Such extreme situations may be encountered when a phone call occurs in usually noisy places (airports, train stations, street ...). Others factors which are intrinsic to the communication system can also lead to an intolerable noise level along with a poor call quality. For instance, IP-packets can suffer from RF disruptions and attenuations during the instant they are transmitted via wireless LAN networks, leading to strong noisy situations in VoIP communications.

Figure 4a and 4b depict the VAD decisions for G.729 and our suggested algorithm respectively. The obtained results are presented in Table 3.

Table 3. Proposed system Vs G.729-B in non-stationary situation

<table>
<thead>
<tr>
<th>VAD Method</th>
<th>Correct</th>
<th>TR</th>
<th>FA</th>
<th>FEC</th>
<th>MSC</th>
<th>Over</th>
<th>NDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>G.729-B</td>
<td>75.32</td>
<td>13.10</td>
<td>11.63</td>
<td>1.31</td>
<td>11.82</td>
<td>2.70</td>
<td>8.91</td>
</tr>
<tr>
<td>Proposed system</td>
<td>84.61</td>
<td>5.72</td>
<td>9.70</td>
<td>1.72</td>
<td>4.02</td>
<td>2.51</td>
<td>7.20</td>
</tr>
</tbody>
</table>
As summarized in table 3, the G.729-B method exhibits an intolerable increase of the critical parameter True Rejection TR (13.1%) and a relatively high value of the False Acceptance (11.6%), which causes a decrease of the Correct parameter (75.3%). In the other hand, we note a consequent improvement of the Correct parameter for our VAD algorithm (84.6%), which clearly lead to low values for both parameters TR (5.7%) and FA(9.7%). Through the obtained results, it clearly appears that the proposed detection scheme outperforms the G.729-B in terms of Correct decisions rate, true rejection and false acceptance rate, in particular in non-stationary situations.

C. Real time tests

In order to evaluate the online behavior of the proposed method using the MCU board, experimental audio signals embedded in real-world noisy environments (babble crow of students) have been recorded. The curves presented in Figure 4 have been displayed in real time while recording the signal via the STM-STUDIO tool [22]. The first curve shows the full-band energy versus the adaptive threshold. We observe that the VAD is set to 1 whenever the full band energy exceeds the adaptive threshold, which corresponds to a voiced frame. Noting that the VAD outputs are illustrated in red on the third curve, while the second curve shows the recorded signal. As mentioned before, we only keep frames declared as speech and set unvoiced frames to 0, in order to get a subjective evaluation and to guarantee that the signal has not been degraded after the VAD processing.
Figure 5. Real-time monitoring for a recorded signal.

D. Real time execution

The suggested VAD algorithm is implemented and tested on the MCU board in order to validate its real time operation mode. To this end, several tests have been executed using a sampling frequency of 8 KHz. After acquisition of the last sample from each frame (S80, S160, S240 ...), as shown in Figure 6, the full data frame are passed to our VAD algorithm to provide the final decision. By placing break points in our implemented code program, the average cycle counter has been estimated to be 200 clock cycles. With a CPU clock of 49.142 MHz, the processing time has been determined to be 4μs, which is sufficient regarding to the sampling frequency used in speech processing (125 μs at 8kHz or 62,5μs at 16 kHz).

Figure 6. Real-time VAD-Timing
5. Conclusion

Through this article, we propose a low complexity VAD approach, using adaptive thresholding to detect voiced periods in presence of non-stationary noise. The attractive property of the considered Voice Activity Detection scheme is its aptitude to keep a prescribed false acceptance rate against silence periods. An exact expression of the probability of false acceptance ($F_A$) has been derived in order to set the scaling factor value, and then to control the $F_A$ rate. However, the presence of abrupt noise transitions in non-stationary noisy situation leads to an intolerable false acceptance rate ($F_A$) of the proposed system. To deal with this problem, the scale factor should be adjusted whenever some particular conditions are met. Also, to avoid undesirable discontinuities in both speech and silence intervals, a smoothing procedure has been carried out by analyzing contiguous binary decisions preceding the frame being processed. Using NOIZEUS data base, synthetic audio signals have been generated to show the VAD performances under different noise conditions in non-homogeneous situations.

The results show that the proposed VAD acts robustly against non-homogeneous noisy environments, where different SNR situations have been simulated within the same call period. By considering usual criterions, namely; TR, Correct, and $F_A$ measures. The proposed VAD and the G.729-B perform similarly in moderate environments. However, we note that the proposed system slightly outperforms the normalized G.729-B algorithm in highly noisy backgrounds.

Finally, several real time tests have been carried out by implementing the VAD-Algorithm. The architecture is targeted on the MCU-board STM32F746NG. The behavior analysis has been done through online audio acquisition via the I/O ports available on the target board. Online operations show a reduced computational complexity, along with an overall latency of 4 μs. Computational requirements seem to be widely sufficient up to 16 kHz sampling.

6. References

[8]. ETSI, 'Voice activity detector (VAD) for adaptive multirate (AMR) speech traffic channels', ETSI EN 301 708 v.7.1.1, December 1999.
[9]. ETSI, 'Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms', ETSI, Sophia Antipolis, France, ETSIES 202 050 Rec., 2002.


Charaf Eddine Chelloug was born in Constantine, Algeria, on April 24, 1988. He received the Master degree in Communication systems from university of Constantine, Algeria, on 2013. He is currently working towards the Ph.D. degree at the same university. Since 2014, he is a research assistant in SISCOM (Signaux et Systèmes de Communication) laboratory at Constantine university. Prior to that he worked as a development engineer at a mobile network company "Huawei". His current research includes speech processing and mobile communication systems.

Atef Farrouki received the engineering degree in electrical engineering from ENITA, Algiers, Algeria, in 1992. He received the Ph.D. degree in signal processing from the University of Constantine, Algeria, 2004. Prior to that, he has been working as an engineer at the national company of aeronautic. Currently, he is a member the SISCOM laboratory and Professor with the University of Constantine, where he has been teaching computer hardware and communication systems and involved in research in signal processing since 1998. His current research interest lies in the field of radar detection and speech processing systems. Prof. A. Farrouki is a reviewer of IET and has served in the technical program committees of several international conferences.