Approach for Energy-Based Voice Detector with Adaptive Scaling Factor

Kirill Sakhnov, *Member, IAENG,* Ekaterina Verteletskaya, and Boris Simak

Abstract— **This paper presents an alternative energy-based algorithm to provide speech/silence classification. The algorithm is capable to track non-stationary signals and dynamically calculate instantaneous value for threshold using adaptive scaling parameter. It is based on the observation of a noise power estimation used for computation of the threshold can be obtained using minimum and maximum values of a short-term energy estimate. The paper presents this novel voice activity detection algorithm, its performance, its limitations, and some other techniques which deal with energy estimation as well.**

Index Terms—**Speech analysis, speech/silence classification, voice activity detection.**

I. Introduction

An important problem in speech processing applications is the determination of active speech periods within a given audio signal. Speech can be characterized by a discontinuous signal since information is carried only when someone is talking. The regions where voice information exists are referred to as ‘voice-active’ segments and the pauses between talking are called ‘voice-inactive’ or ‘silence’ segments. The decision of determining to what class an audio segment belongs is based on an observation vector. It is commonly referred to as a ‘feature’ vector. One or many different features may serve as the input to a decision rule that assigns the audio segment to one of the two given classes. Performance trade-offs are made by maximizing the detection rate of active speech while minimizing the false detection rate of inactive segments. However, generating an accurate indication of the presence of speech, or its absence, is generally difficult especially when the speech signal is corrupted by background noise or unwanted interference (impulse noise, atd.).

In the art, an algorithm employed to detect the presence or absence of speech is referred to as a voice activity detector (VAD). Many speech-based applications require VAD capability in order to operate properly. For example in speech coding, the purpose is to encode input audio signal such that the overall transferred data rate is reduced. Since information is only carried when someone is talking, clearly knowing when this occurs can greatly aid in data reduction. Another example is speech recognition. In this case, a clear indication of active speech periods is critical. False detection of active speech periods will have a direct degradation effect on the recognition algorithm. VAD is an integral part to many speech processing systems. Other examples include audio conferencing, echo cancellation, VoIP (voice over IP), cellular radio systems (GSM and CDMA based) and hands-free telephony [1-5].

Many different techniques have been applied to the art of VAD. In the early VAD algorithms, short-time energy, zero-crossing rate, and linear prediction coefficients were among the common feature used in the detection process [6]. Cepstral coefficients [7], spectral entropy [8], a least-square periodicity measure [9], wavelet transform coefficients [10] are examples of recently proposed VAD features. But in general, none will ever be a perfect solution to all applications because of the variety and varying nature of natural human speech and background noise.

Nevertheless, signal energy remains the basic component to the feature vector. Most of the standardized algorithms use energy besides other metrics to make a decision. Therefore, we decided to focus on energy-based techniques. It will be introduced an alternative way how to provide features extraction and threshold computation here. The present paper is organized as follows. The second section gives a general description of embodiment. The third section presents a review of earlier works. The fourth section will introduce the new algorithm. The fifth section reports the results of testing performed to evaluate the quality of the speech/silence classification, and the rest of the paper concludes the article.

II. Voice Activity Detection – The Principle

The basic principle of a VAD device is that it extracts measured features or quantities from the input signal and then compares these values with thresholds usually extracted from noise-only periods (see Fig. 1). Voice activity (VAD=1) is declared if the measured values exceed the thresholds. Otherwise, no speech activity or noise, silence (VAD=0) is present. VAD design involves selecting the features, and the way the thresholds are updated. Most VAD algorithms output a binary decision on a frame-by-frame basis where a “frame” of the input signal is a short unit of time such 5-40 ms. The accuracy of a VAD algorithm depends heavily on the decision thresholds. Adaptation of thresholds value helps to track time-varying changes in the acoustic environments, and hence gives a more reliable voice detection results.

Manuscript received January 9, 2009.

K. Sakhnov is with the Czech Technical University, Department of Telecommunication Engineering, Prague, 16627 Czech Republic (phone: (+420) 224-352-100; fax: (+420) 223-339-810; e-mail: sakhnk1@ fel.cvut.cz).

E. Verteletskaya is with the Czech Technical University, Department of Telecommunication Engineering, Prague, 16627 Czech Republic (e-mail: vertee1@ fel.cvut.cz).

B. Simak is with the Czech Technical University, Department of Telecommunication Engineering, Prague, 16627 Czech Republic (e-mail: simak@ fel.cvut.cz).

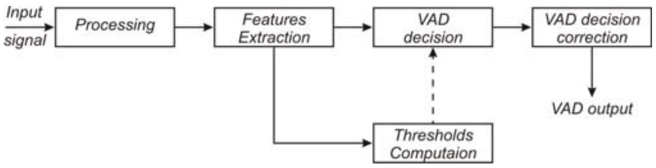


Figure 1. A block diagram of a basic VAD design.

It should be mentioned as well that a general guideline for a good VAD algorithm for all speech enhancement (i.e., noise reduction) systems is to keep the duration of clipped segments below 64 ms and no more than 0.2 % of the active speech clipped [G.116].

A. *Desirable Aspects of VAD Algorithms*

In speech enhancement systems, a reliable VAD is often a keystone component, for instance, for noise estimation and for adaptive echo cancellation. So the list of desirable aspects of good VAD algorithms for speech enhancement is the following:

- VAD algorithm must implement a good decision rule that exploits the properties of speech to consistently classify segments of speech into inactive and active.
- It should adapt to non-stationary background noise to enhance robustness.
- The computational complexity of VAD algorithm must be low to suit real-time applications.
- VAD must have minimum errors of misclassifying speech as noise.
- Toll quality voice should be achieved after applying VAD algorithm.

The assumptions on the VAD algorithm proposed here is based on the following characteristics

- Speech is quasi-stationary. Its spectral form changes over short periods, e.g. 20-30ms.
- Background noise is relatively stationary, changing very slowly with time.
- Energy of the speech signal is usually higher than background noise energy; else speech will be unintelligible.

B. *Choice of Frame Duration*

Speech samples that are transmitted should be stored in a signal-buffer first. The length of the buffer may vary depending on the application. For example in the AMR Option 2 VAD divides the 20-ms frames into two subframes of 10 ms [2]. A frame is judged to be active if at least one subframe is active there. Through this paper a 10 ms frame with 8 kHz sampling, linear quantization (8/16 bits linear PCM) and single channel (mono) recording will be used. The advantage of using linear PCM is that the voice data can be transformed to any other compressed code (G.711, G.723, and G.729). Frame duration of 10 ms corresponds to 80 samples in time domain representation.

Let x(i) be the i-th sample of speech. If the length of the frame was N samples, then the j-th frame can be represented as,

$$f_j = \{x(i)\}_{i=(j-1)\cdot N+1}^{j\cdot N} \tag{1}$$

C. *Energy of Frame*

The most common way to calculate the full-band energy of a speech signal is

$$E_j = \frac{1}{N} \cdot \sum_{i=(j-1)\cdot N+1}^{j\cdot N} x^2(i) \tag{3}$$

where, E_j – energy of the j-th frame and fj is the j-th frame is under consideration.

D. *Initial Value of Threshold*

The starting value for the threshold is important for its evolution, which tracks the background noise. Though an arbitrary initial choice of the threshold can be used, in some cases it may result in poor performance. Two methods were proposed for finding a starting threshold value [11].

Method 1: The VAD algorithm is trained for a small period using a prerecorded speech samples that contain only background noise. The initial threshold level for various parameters then can be computed from these speech samples. For example, the initial estimate of energy is obtained by taking the mean of the energies of each frame as in

$$E_r = \frac{1}{\upsilon} \cdot \sum_{m=0}^{\upsilon} E_m \tag{4}$$

where, E_r – initial threshold estimate, υ – number of frames in prerecorded sample.

This method can not be used for most real-time applications, because the background noise can vary with time. Thus it would be used the second method given below.

Method 2: Though similar to the previous method, here it is assumed that the initial 100 ms of any call does not contain any speech. This is a plausible assumption given that users need some reaction time before they start speaking. These initial 100 ms are considered inactive and their mean energy is calculated using Eq.4.

III. E-VAD Algorithms – A Literature Review

Scenario: the energy of the signal is compared with the threshold depending on the noise level. Speech is detected when the energy estimation lies over the threshold. The main classification rule is,

$$\begin{aligned} &\mathit{if} \quad (E_j \succ k \cdot E_r), \quad \mathit{where} \quad k \succ 1 \\ &\qquad \mathit{current\ \ frame\ is\ ACTIVE} \\ &\mathit{else} \\ &\qquad \mathit{current\ \ frame\ is\ INACTIVE} \end{aligned} \tag{5}$$

In this equation, Er represents the energy of noise frames, while $k \cdot E_r$ is the 'Threshold' being used in the decision-making. Having a scaling factor, 'k' allows a safe band for the adaptation of Er, and therefore, the threshold.

A hang-over of several frames is also added to compensate for small energy gaps in the speech and to make sure the end of the utterance, often characterized by a decline of the energy (especially for unvoiced frames), is not clipped.

A. *LED: Linear Energy-Based Detector*

This is the simplest energy-based method that was first described in [12]. Since a fixed threshold would be 'deaf' to varying acoustic environments around the speaker, an adaptive threshold is more suitable. The rule to update the threshold value was specified as,

$$E_{rnew} = (1-p)\cdot E_{r\ old} + p\cdot E_{silence} \tag{6}$$

Here, $E_{r\ new}$ – is the updated value of the threshold, $E_{r\ old}$ – is the previous energy threshold, and $E_{silence}$ – is the energy of the most recent noise frame.

The reference E_r is updated as a convex combination of the old threshold and the current noise update. Parameter 'p' is chosen considering the impulse response of Eq.(6) as a first order filter (0<p<1) [12].

B. ALED: Adaptive Linear Energy-Based Detector

The drawback of LED is coefficient 'p' in Eq.(6) being insensitive to the noise statistics. The threshold value E_r can be computed alternatively based on the second order statistics of inactive frames [11]. A noise buffer of the most recent 'm' silence frames should be used then. Whenever a new noise frame is detected, it is added to the buffer and the oldest one is removed. The variance of the buffer, in terms of energy is given by

$$\sigma = VAR[E_{silence}] \tag{9}$$

A change in the background noise is detected by comparing the energy of the new inactive frame with a statistical measure of the energies of the past 'm' inactive frames.

To understand the mechanism, consider first the instant of addition of a new inactive frame to the noise buffer. The variance, just before the addition, is denoted by σ_{old}. After the addition of the new inactive frame, the variance is σ_{new}. A sudden change in the background noise would mean

$$\sigma_{new} \succ \sigma_{old} \tag{10}$$

Thus, a new rule to vary 'p' in Eq.(6) can be set in steps as per Table I (refer to algorithm LED to chose the range of 'p').

Table I. Value of 'p' depending on $\frac{\sigma_{new}}{\sigma_{old}}$.

$\frac{\sigma_{new}}{\sigma_{old}} \geq 1.25$	0.25
$1.25 \geq \frac{\sigma_{new}}{\sigma_{old}} \geq 1.10$	0.20
$1.10 \geq \frac{\sigma_{new}}{\sigma_{old}} \geq 1.00$	0.15
$1.00 \geq \frac{\sigma_{new}}{\sigma_{old}}$	0.10

The coefficient 'p' in Eq.(6) now depends on variance of $E_{silence}$. It would make the threshold to respond faster to changes in the background environment. The classification rule for the signal frames continues to be the same as in Eq.(5).

C. LED II: Linear Energy-Based Detector with double threshold

Another VAD design is in application of two different thresholds for speech and silence periods separately. It avoids switching when the energy level is near to the single threshold. This algorithm works as it is described below. First the noise level is estimated using sliding window and defined as [13],

$$E_{r\ new} = \lambda_1 \cdot E_{r\ old} + (1-\lambda_1)\cdot E_j \tag{11}$$

for active segments and

$$E_{r\ new} = \lambda_2 \cdot E_{r\ old} + (1-\lambda_2)\cdot E_j \tag{12}$$

for inactive segments, respectively.

λ_1 [0.85,0.95] and λ_2 [0.98,0.999] are the adaptation factors. They define a low-pass filtering. The value of the decay defined by λ_1 is fixed according to following constraints: it should be small enough to track noise variation, but greater than the speech variation. It is made so to avoid the adaptation following the variation of the energy when speech is present. This leads to decays between 60 ms and 200 ms, when the sampling period for the energy is 10 ms. λ_2 is fixed with similar constraints: the decay must be big enough to avoid tracking the variation of the speech energy, but small enough to adapt to variations in the background noise, which leads to values between 500 ms to one second [13].

The noise and speech thresholds are defined as,

$$\begin{aligned} T_{silence\ new} &= E_{r\ new} + \delta_{silence} \\ T_{speech\ new} &= E_{r\ new} + \delta_{speech} \end{aligned} \tag{13}$$

where, $\delta_{silence}$ [0.1,0.4] and δ_{speech} [0.5,0.8] are additive constants used to determine the thresholds. When the energy is greater than the speech threshold, speech is detected and when the energy is lower than the noise threshold no-speech is detected. Thus, the use of double threshold reduces the problem of sudden variations in the VAD's output which may be obtained if a single threshold is used.

IV. Dynamical VAD - Description

It occurs that in classical energy-based algorithms, detector can not track the threshold value accurately, especially when speech signal is mostly voice-active and the noise level changes considerably before the next noise level re-calibration instant. The 'dynamical' VAD was proposed to provide its classification more accurately in comparisson with abovementioned techniques. The main idea behind this algorithm was that the threshold level is estimated without the need of voice-inactive segments by using minimums and maximums of the speech energy. In the rest of this section we will present the algorithm and discuss some of its statistical properties.

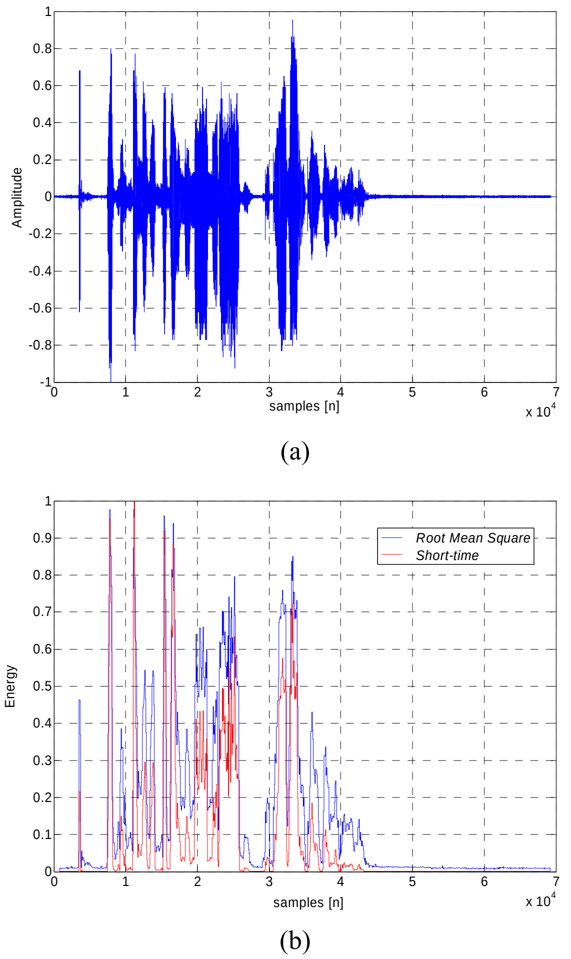
A. RMS Energy

Another common way to calculate the energy of a speech signal is the *root mean square energy (RMSE)*, which is the square root of the average sum of the squares of the amplitude of the signal samples. It is given as,

$$E_j = \left[\frac{1}{N}\cdot \sum_{i=(j-1)\cdot N+1}^{j\cdot N} x^2(i)\right]^{\frac{1}{2}} \tag{14}$$

all the abbreviations here are the same as in Eq.(3).

The 'dynamical' VAD is based on the observation that the power estimate of a speech signal exhibits distinct peaks and valleys (see Fig. 2).While the peaks correspond to speech activity the valleys can be used to obtain a noise power estimate. Therefore, the RMSE is more appropriate.



(a)

(b)

Figure 2. Short-time vs. Root Mean Square energy.

B. Threshold

Threshold estimation is based on energy levels, E_{min} and $E_{max,}$ obtained from the sequence of incoming frames. These values are stored in a memory and the threshold is calculated as,

$$Threshold = k_1 \cdot E_{\max} + k_2 \cdot E_{\min} \quad (15)$$

Where, k_1 and k_2 are factors, used to interpolate the threshold value to an optimal performance. If the current frame's energy is less than the threshold value the frame is marked as inactive. However this does not mean that the transmission immediately will be halted. There is also a hangover period that should consist of more than four inactive frames before the transmission is to be stopped. If the energy increases above the threshold the communication is resumed again.

Since low energy anomalies can occur there is a prevention needed for this. The parameter E_{min} is slightly increased for each frame and this is defined by,

$$E_{\min}(j) = E_{\min}(j-1) \cdot \Delta(j) \quad (17)$$

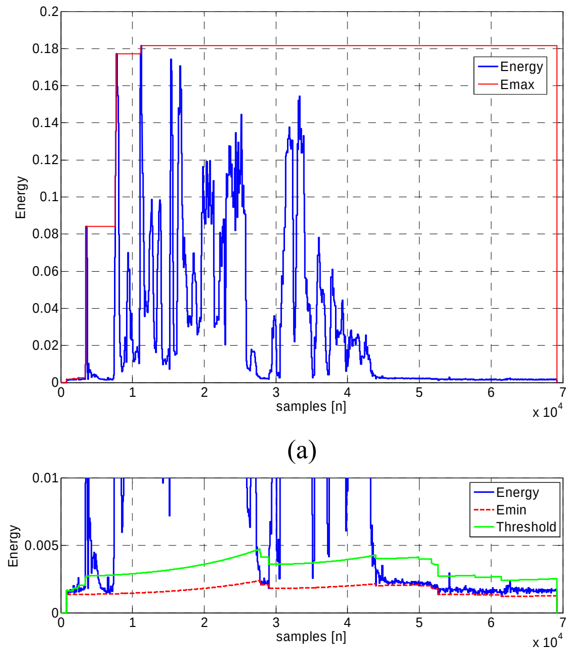
The parameter Δ for each frame is defined as,

$$\Delta(j) = \Delta(j-1) \cdot 1.0001 \quad (18)$$

C. Algorithm Enhancement - Scaling Factor

It is possible to introduce Eq.(15) as a convex combination of a single parameter λ (i.e., λ = k2):

$$Threshold = (1-\lambda) \cdot E_{\max} + \lambda \cdot E_{\min} \quad (19)$$



(a)

(b)

Figure 3. RMS energy, maximum energy, minimum energy and threshold curves.

Here, λ – a scaling factor controlling estimation process. Voice detector performs reliably when λ is in the range of [0.950,…,0.999]. However, the values for different types of signals could not be the same and a priori information has still been necessarily to set up λ properly. The equation below shows how to make the scaling factor to be independent and resistant to the variable background environment

$$\lambda = \frac{E_{\max} - E_{\min}}{E_{\max}} \quad (20)$$

Figure 3 depicts the curves estimated from the speech signal shown in Fig. 2 (a). It can be seen how the algorithm tracks energy levels and calculates corresponding threshold value. A flowchart of the whole embodiment is given in Fig. 4 respectively. The results of testing performed to evaluate the quality of the proposed algorithm together with described energy-based algorithms will be discussed through the next section.

V. Experimental Results - Discussion

A. Database

Described VAD algorithms were evaluated using speech data (short monologues and numbers) from Czech Speech database. The test templates used varied in loudness, speech continuity, background noise and accent. The data was recorded in a quiet environment, sampled at rate of 8 kHz, and quantized to 16 bit per sample. The utterances tested were drawn from eight speakers, four male and four female.

B. Computation

MATLAB environment was used to test the algorithms developed on various sample signals. The speech data is segmented into 20ms frames (160 samples per frame). For each frame RMS energy and threshold value are computed. The values of thresholds Emax a Emin are also renewed every frame, based on comparison of current frame energy to initial E_{max} a E_{min}. The algorithm is working as it is shown in Fig. 4.

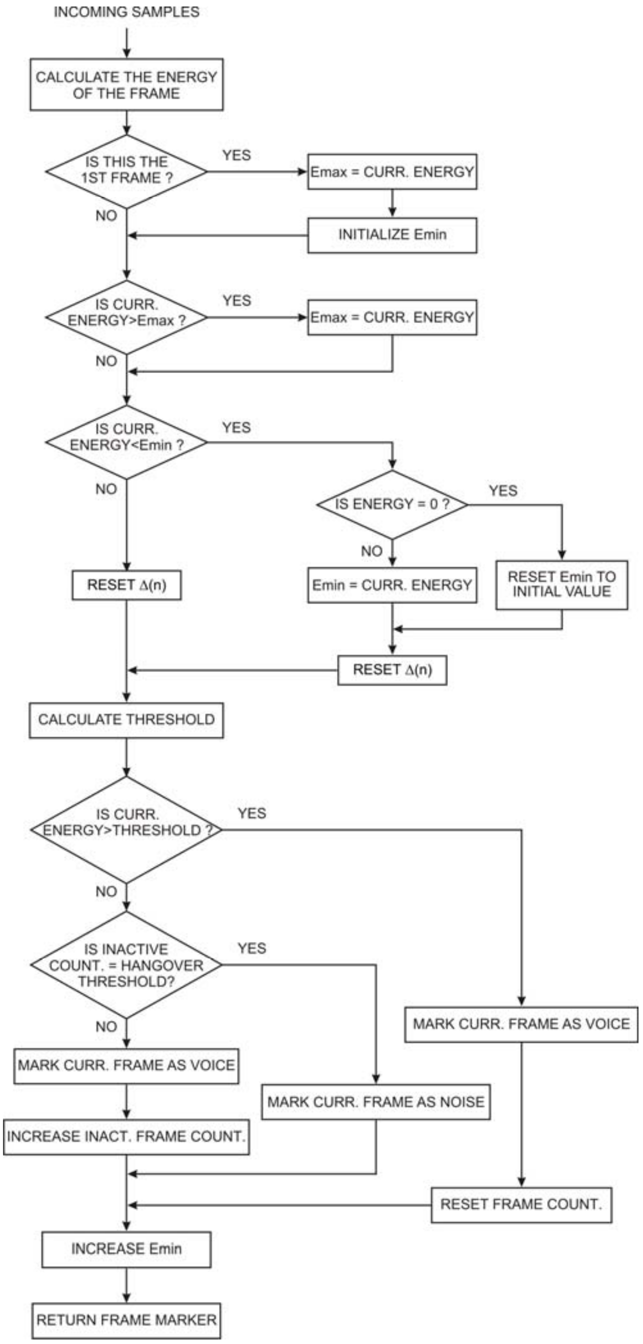


Figure 4. A flowchart of the proposed VAD.

C. *Experimental Results*

Performance of the algorithms was studied on the basis of the following parameters:

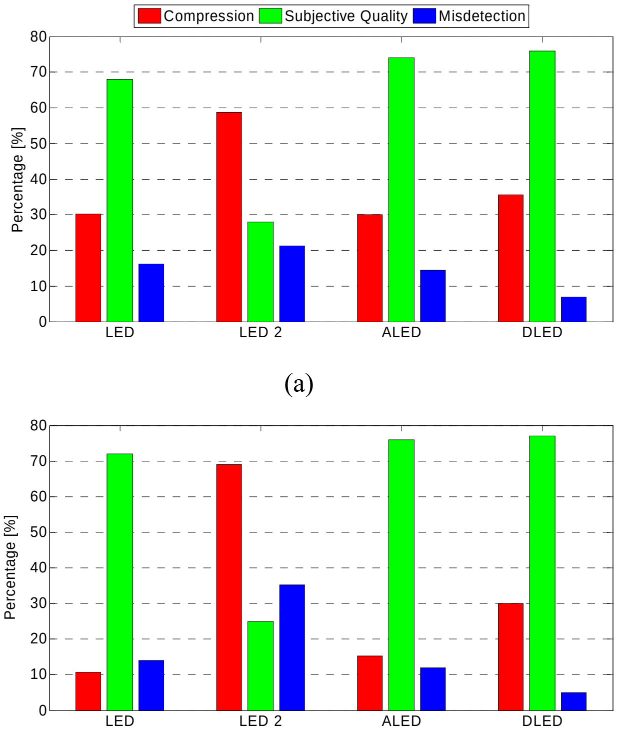
1. Percentage compression: The ratio of total 'inactive' frames detected to the total number of frames formed expressed as a percentage. A good VAD should have high percentage compression. It is necessary to note that the percentage compression also depends on the speech samples. If the speech signal was continuous, without any brakes, it would be unreasonable to expect high compression levels;

2. Subjective Speech Quality: The quality of the samples was rated on a scale of 1 (poorest) to 5 (the best) where 4 represents toll grade quality. The input signal was taken to have speech quality 5. The speech samples after compression were played to independent jurors randomly for an unbiased decision;

3. Objective Assessment of Misdetection: The number of frames which have speech content, but were classified as 'inactive' and number of frames without speech content but classified as 'active' are counted. The ratio of this count to the total number of frames in the sample is taken as the 'misdetection' percentage. This gives a quantitative measure of VAD performance.

From figures it can be observed the following percentage of compression, subjective quality and misdetection for different speech templates. Each figure shows the response of all the above algorithms for a particular type of input signal:

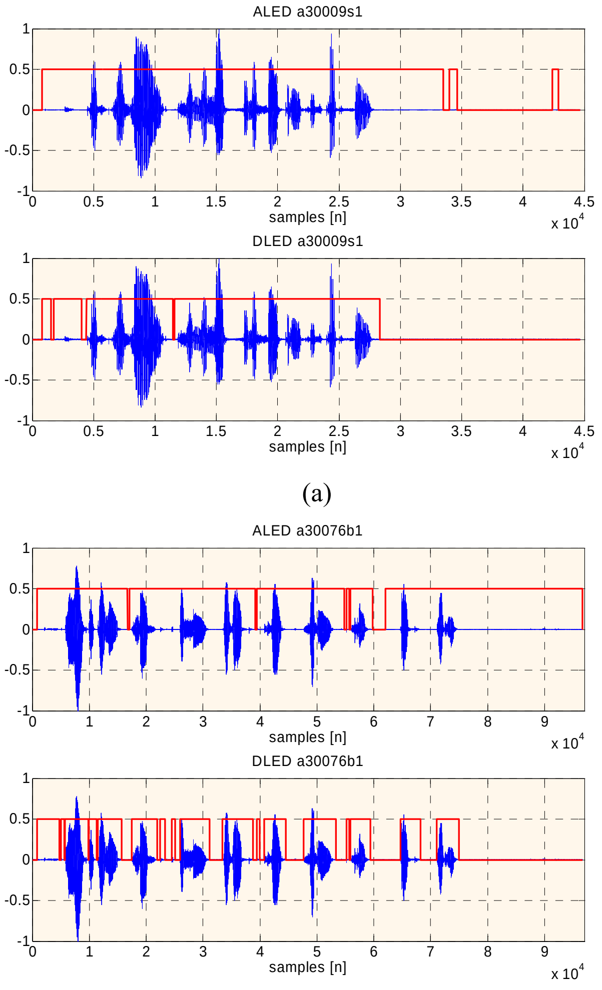
- Compression: the LED 2 has the highest percentage of compression for both different templates compared to other algorithms (see Fig. 5, for comparison). The proposed 'dynamical' linear energy-based detector (DLED) takes the second place, leaving behind LED and ALED. However, inspite of its high compression rate, the LED 2 has an inadmissible percentage of the active speech segments clipped. For this reason, the quality of the output signal becomes unacceptable.
- Subjective Quality: for all algorithms, except the LED 2, the speech quality was nearly the same. Because the most common misdetection mistake in case of the LED and ALED was marking 'inactive' frames as 'active'. It was reflected on the percentage of compression and did not lead to the poor quality of speech.
- Misdetection: with respect to the rate of misdetection, the DLED outperformed LED and ALED algorithms. The LED 2 has the worse results. In Fig. 6, it can be observed the way how two algorithms work. The proposed VAD compared to another one performs more accurately classifying speech frames.



(a)

(b)

Figure 5. Discontinuous telephone speech (a) monologue (b) numbers.



(a)

(b)

Figure 6. Example telephone speech (a) numbers (b) monologue.

VI. Conclusion

This article is a forecast on voice activity detection algorithms employed to detect the presence/absence of speech components in audio signal. A new alternative energy-based VAD to provide speech/silence classification was presented. The aim of this work was to show the principle of the proposed algorithm, compare it to other known energy VADs, discuss its advantages and possible drawbacks.

The algorithm has several features, which characterizes its behaviour: the root-mean square energy is used to calculate the power of a speech segment; estimation of threshold is based on the observation that the short-time energy exhibits distinct peaks and valleys corresponding to speech activity or silence periods; an adaptive scaling factor, λ, makes the threshold to be independent on signal characteristics and resistant to the variable environment as well.

It is easy to realize that the expounded algorithm is very independent and easily can be integrated into most VADs used by speech coders and other speech enhancement systems.

References

[1] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd, "The voice activity detector for the pan-European digital cellular mobile telephone service, " *in IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, (Glasgow, Scotland), pp. 369-372, May 1989.

[2] A. Benyassine, E. Shlomot, and H.-Y. Su, "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, " *IEEE Commun. Mag.*, vol. 35, pp. 64-73, Sept. 1997.

[3] E. Ekudden, R. Hagen, I. Johansson, and J. Svedberg, "The adaptive multi-rate speech coder, " *in Proc. IEEE Workshop on Speech Coding for Telecommunications*, (Porvoo, Finalnd), pp. 117-119, June 1999.

[4] ETSI TS 126 094 V3.0.0 (2000-01), 3G TS 26.094 version 3.0.0 Release 1999, Universal Mobile Telecommunications System (UMTS); Mandatory Speech Codec speech processing functions AMR speech codec; Voice Activity Detector (VAD), 2000.

[5] TIA/EIA/IS-127, Enhanced Variable Rate Codec, Speech Service Option 3 for Wide-band Spread Spectrum Digital Systems, Jan. 1996.

[6] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced- silence classi_cation with applications to speech recognition, " *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 24, pp. 201-212, June 1976.

[7] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstral fea-tures, " *in Proc. of IEEE Region 10 Annual Conf. Speech and Image Technologies for Computing and Telecommunications*, (Beijing), pp. 321-324, Oct. 1993.

[8] S. A. McClellan and J. D. Gibson, "Spectral entropy: An alternative indicator for rate allocation, " *in IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, (Adelaide, Australia), pp. 201-204, Apr. 1994.

[9] R. Tucker, "Voice activity detection using a periodicity measure, " *IEE Proc.-I*, vol. 139, pp. 377-380, Aug. 1992.

[10] J. Stegmann and G. Schroder, "Robust voice-activity detection based on the wavelet transform, " *in Proc. IEEE Workshop on Speech Coding for Telecommunications*, (Pocono Manor, PN), pp. 99-100, Sept. 1997.

[11] Venkatesha Prasad, R. Sangwan, A. Jamadagni, H.S. Chiranth, M.C. Sah, R. Gaurav, V., "Comparison of voice activity detection algorithms for VoIP", *proc. of the Seventh International Symposium on Computers and Communications – ISCC 2002*, (Taormina, Italy), pp. 530-532, 2002.

[12] P. Pollak, P. Sovka and J. Uhlir, "Noise System for a Car", *proc. of the Third European Conference on Speech, Communication and Technology – EUROSPEECH'93*, (Berlin, Germany), pp. 1073-1076, Sept. 1993.

[13] P. Renevey, A. Drygajlo, "Entropy Based Voice Activity Detection In Very Noisy Conditions", *proc. of the Seventh European Conference on Speech Communication and technology – EUROSPEECH 2001*, (Aalborg, Denmark), pp.1883-1886, 2001.