EVALUATION AND COMPARISON USING ACTIVITY SIGNALS OF SPEECH METHODS IN RIVER PLATE SPANISH USING BEPPA CORPUS

EVALUACIÓN Y COMPARACIÓN MEDIANTE EL USO DE SEÑALES DE ACTIVIDAD DE LOS MÉTODOS DE HABLA EN EL ESPAÑOL DEL RIO DE LA PLATA CON EL USO DE BEPPA CORPUS

Horderlin Vrangel Robles[1], Valentin Molina[2], Luis Martinez[2], Hermann Davila[2]

[1.] Dirección de Investigaciones, Grupo de Investigación - DEMA. Universidad del Sinú - Elías Bechara Zainúm. Montería, Colombia.

[2.] Ingeniería Biomédica, Grupo de Investigación en Ingeniería Clínica, GINIC-HUS. Escuela Colombiana de Carreras Industriales, ECCI. Bogotá, Colombia.

Recibido: 10 de Febrero de 2015
Aceptado: 20 de Febrero de 2016

*Correspondencia del autor: Horderlin Vrangel Robles. E-mail: horderlin_robles@unisinu.edu.co

ABSTRACT

The results obtained after comparing several algorithms which use basic methods of signal processing for speech activity detection of voice or VAD (Voice Activity Detection-VAD), were assessed in order to determine their effectiveness. The algorithms presented in this article are short-time or spectral energy based endpoint detection algorithm, the zero crossing rate method, and the higher order differential (High Order Difference, HOD) method. First, an introduction of the concept of VAD is presented and the need to apply such language algorithms in River Plate is Spanish. Then a summary of the state of the art techniques and algorithms for detecting voice activity is shown with evidence and experiments used to implement algorithms with BEPPA corpus (Evaluation Battery for Patients with Auditive Prostheses, BEPPA – in Spanish).

Keywords: Voice activity detection, BEPPA, short time energy, Zero Cross Rate, High Order Difference.

RESUMEN

Los resultados obtenidos luego de comparar varios algoritmos que utilizan métodos básicos de procesamiento de señales para la detección de actividad de voz (VAD, por el término en inglés) se evaluaron para determinar su efectividad. Los algoritmos presentados en este artículo son de corta duración o algoritmos de detección de punto final a base de energía espectral, el método de la tasa de cero cruce y el método de diferenciales de orden mayor (HOD, por el término en inglés). Primero, se presenta una introducción del concepto VAD y de la necesidad de aplicar dichos algoritmos de lenguaje al español del Rio de la Plata. Luego un resumen de las técnicas del estado del arte y algoritmos para detectar la actividad de voz se muestra con la evidencia y los experimentos utilizados para implementar algoritmos con Batería de Evaluación para Pacientes con Prótesis Auditiva (BEPPA) corpus.

Palabras claves: detección de actividad de voz, BEPPA, energía de corta duración, Tasa Cero Cruz, diferenciales de orden mayor.

INTRODUCTION

Currently, analysis, representation and modeling of speech signals is one of the most studied topics in biomedical engineering with applications ranging from automatic speech recognition to brain-computer interfaces (Brain Computer Interface - BCI). One of the key steps in these applications is the detection of activity (Voice Activity Detection VAD), which is necessary to reduce noise and/or classify the voice. To date each of the methods developed in current research in VAD has been evaluated using English databases with some exceptions in the Spanish and Chinese languages. Furthermore, the English language has phonetic and acoustic differences when compared to River Plate Spanish and therefore these algorithms may have increased rates of misclassification. Therefore, it is necessary to evaluate the effectiveness of these techniques when used for River Plate Spanish using the battery developed at the Faculty of Engineering of the National University of Entre Rios known as BEPPA (Evaluation Battery for Patients with Auditive Prostheses, BEPPA – in spanish). In communication, speech can be characterized as a discontinuous medium due to breaks, which are a unique feature compared to other multimedia signals such as video, audio and data. The regions where voice information is classified are active voice (voiced and unvoiced) and pauses which are called silent or inactive regions[1].

2. STATE OF THE ART

The algorithms developed to classify the voiced speech signals (voice), the unvoiced speech (unvoiced) signals and the silent (silence) regions are techniques known as voice activity detection (Voice Activity Detection VAD). The first time when VAD was first defined or talked about was in the early 1960's with the speech interpolation system (Time Assignment Speech Interpolation, TASI) (1,2). A first approach to the problem of detecting voice activity was the location of the endpoint, so, in (3) an algorithm is proposed to solve this problem using two characteristics of speech signals, the energy and zero-crossing rate. Other algorithms used are those of linear prediction coefficients (Linear Prediction Coding, LPC) (4) and of frequency of the signal using the least squares estimator (Least-Square Estimator Periodicity, LPSE) (5). Tucker in (6) proposes using an LPSE VAD to detect speech. This technique only worked for a SNR above 0 dB, but showed better performance than the autocorrelation estimator SNR 5dB below. On the other hand in (7), the proposed VAD detects the intervals where the voice appears using the cepstral analysis. Then the researcher designed a VAD in global systems for mobile communications (Global System for Mobile Communications, GSM) cited in (8), which exhibits good performance under stationary noise environments. Soon after, a low complexity robust algorithm based on measuring the energy spectrum of the voice (9) was developed, which was adopted as part of Annex B to Recommendation G.729B of the International Telecommunication Union (International Telecommunication Union, ITU) and in 1996 for a standardized speech coding scheme (10). Moreover, in (11) a VAD algorithm based on wavelet transform (Wavelet Transform, WT) using its flexibility in temporal-frequency resolution to calculate the parameters of the VAD decision is proposed. A new technique capable of working under very low SNR (less than 10dB) is proposed in (12). One of its main advantages lies in the insensitivity to changes in noise levels. In (13) an algorithm is proposed that combines methods as geometrically adaptive energy thresholds (Threshold Energy Adaptive Geometrically, GAET), and the measuring of the periodicity LPSE and zero crossing rate of the signal is proposed. This method presents more robust information than previous VAD's, because it is working in conditions with SNR between 10dB to -10dB and under and is not sensitive to changes in noise levels. At present, algorithms are fused to achieve better performance and better classifications, such as described in (14), where the speech signal is decomposed into 4 sub-bands using the discrete wavelet transform (Discrete Wavelet Transform, DWT) and the Teager energy operator (Teager Energy Operation, TEO) is applied to each sub-band DWT coefficients. Results show that the algorithm is not affected by variations in noise levels, overcoming the method implemented with the transformed perceptual wavelet packet (Perceptual Wavelet Packet Transform PWPT) proposed in (15).

3. IMPLEMENTATION OF ALGORITHMS

According to its own properties and production models, the voice signal can assume different characteristics to discriminate speech and silent periods. Some of the most widely used assumptions in most algorithm voice activity detections are listed below (16):

[1] 978-1-4799-7666-9/14/$31.00 ©2014 IEEE

- The ambient noise is additive to the speech signal.
- The segment of the speech signal has a higher energy value than the ambient noise segment.
- Speech is stationary for short periods of time, for example T <40ms.
- The voice is not stationary over longer periods, eg T> 0; 5s.
- Ambient noise is stationary for much longer periods, eg T> 2s.
- The voice has more regular noise components than noise.

Using these hypotheses, among others, voice activity detection algorithms were proposed and implemented to discriminate speech and silent periods in the time domain.

A. Short-time Energy

As mentioned above, one of the considerations of the speech signal is that it's considered stationary for short periods of time, ie less than 40ms, being stationary, it is assumed that the same acoustic characteristics are present in the sample window. Now, a segment of the speech signal is defined as the product of a shifted window and the sequence of values of the speech signal, which are (17):

$$F_{s,m}[n] = s[n]w[m-n] \qquad (1)$$

Furthermore, short time energy (Short Time Energy, STE) can be defined as:

$$E_{s,w}(n) = \sum_{m=-3}^{3} (s[m]w[n-m])^2 \qquad (2)$$

To calculate the energy of real signals, adjust the formula to the intervals where the equation is no longer zero. The speech signal s [n] = $s_n \in R$ (not all zero) for n = 1, 2, ... , N and is 0 for n <1 ∧ n> N, we also have that w [m] = $w_m \in R$ (not all zero) for m = 1, 2, ... , M and is 0 for all other m; i.e. the energy of short time to real signals is:

$$E_{s,w}(n) = \sum_{m=n-m}^{n-1} (s[m]w[n-m])^2 \qquad (3)$$

where M is the length of the w window (18).

B. Zero Cross Rate.

The zero crossing rate (ZCR, Zero Cross Rate) is one of the basic acoustic features that can easily calculate and implement the VAD's. ZCR is defined as the weighted average of the number of times the speech signal sign changes within a segment of the speech signal (18). The representation of the operator in terms of a linear filter is:

$$Z_{s,w}(n) = \sum_{m=n-M+1}^{n-1} 0,5\,|sgn\{s[m]\} - sgn\{s[m-1]\}|w[n-m] \qquad (4)$$

where

$$sgn\{s[m]\} = \begin{cases} 1 & s[m] \geq 0, \\ -1 & s[m] < 0 \end{cases} \qquad (5)$$

In general, the ZCR of clunks (not voiced speech) and ambient noise is greater than for voiced sounds (voiced speech). The ZCR is often used in combination with the STE for endpoint detection (18). In particular, the ZCR is used to detect the start and end positions of unvoiced sounds.

C. *High order difference.*

The hardest part in the VAD is to distinguish unvoiced sounds of silence (18). One way to achieve this is to use higher order differential (HOD, High Order Difference) voice as a characteristic in the time domain. This method was discussed in (19) and no literature used before or after the higher order differential is used to find the endpoint of the voice activity, since there are no scientific articles found regarding this topic. The idea of the author is applied to a differential signal segment previously multiplied by the window function of the speech signal and adding their absolute values, ie:

$$HOD_{s,w}(n) = \sum_{m=n-M+1}^{n-1} \frac{d\,;\, s[m]w[n-m];}{dm} \qquad (6)$$

The HOD used to identify unvoiced sounds is easier than the ZCR(19).

4. TESTS AND EXPERIMENTS

The tests were implemented using the BEPPA corpus, which was designed to study River Plate Spanish adult audiometric testing and hearing aid selection and performance (20). Recordings of 3 male voices and 3 female speakers between 18 and 45 years, natives of Argentina, belonging to the region of Río de la Plata were performed. BEPPA has a sampling frequency of 48 kHz and 16 bit resolution (21).

It was also necessary to segment and perform a subsampling of the BEPPA corpus, it was then implemented with a MATLAB algorithm that converts data cho-

sen from 48 kHz to 16 kHz. In the BEPPA corpus the following experimental sentences were implemented and subsampled(in Spanish) in order to perform the signal analysis experiments:

- Dawn was clouded but it cleared up in the afternoon.
- The pain was very great but short lived.
- It is too heavy, no?
- The highway was empty but we still arrived late.
- The meat is raw and I do not like that way.
- The children fell asleep, did you realize this?
- What a terrible problem, don't you think?
- Do you remember the number or do I have to check?
- They are all on vacation.
- Did you take the remedy or did you forget?

The first implemented VAD uses the STE to locate the beginning and end of the speech signal. The method uses two steps to calculate the detection thresholds:

1. Uses a higher threshold τ_u to determine the start of the final index.
2. Expands the boundaries to reach the lower threshold τ_l.

Figure 1 illustrates the algorithm used.

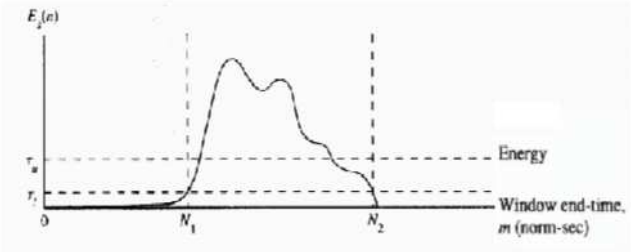

Figure.1 Location thresholds for VAD employing short-term energy (3).

This method was proposed by (3) one of the most used in the literature for its versatility and low computational cost.

After testing with the decision using only the power of short time, an algorithm that combines the zero crossing rate and energy of short-time signal was implemented. The VAD detection algorithm performs the following steps:

1. Using a high threshold τ_u to determine the start and end point energy.
2. Extend the limits to reach the lower threshold τ_l.
3. Extend the limits further to reach the threshold of ZCR τ_{zc}.

The method is illustrated in Figure 2 and was first proposed by (3).

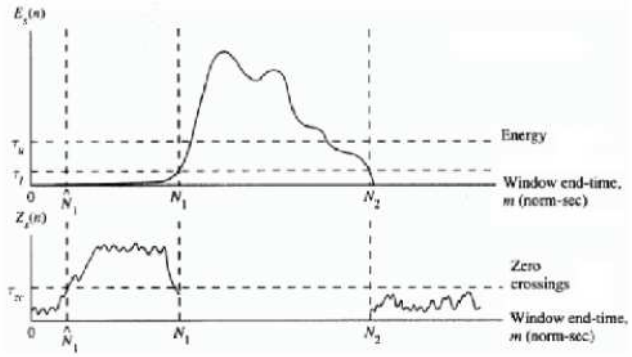

Figure. 2 VAD Algorithm with STE and ZCR (3).

It can be noticed that the ZCR improves detection of voice activity. Finally, HOD was used for the detection of the segments that are silent or of unvoiced speech. In order to be implemented, the algorithm must perform the following steps:

1. Calculate the energy and the sum of the absolute values of the differentials of order n.
2. Select a weighting factor (weight) w from [0,1] to calculate the new curve:

$$VH = w \times energy + (1\text{-}w) \times HOD \qquad (7)$$

3. Finding a relation ρ to calculate the threshold τ *VH* and determine the final index of the sample. threshold equals:

$$VH_{min} + (VH_{max} - VH_{min}) \times \rho \qquad (8)$$

As can be seen, this method depends on three parameters to determine n, w, ρ. Typical values of these parameters are: n = 4, w = 5 and ρ = 0; 125, according to (19), but nevertheless vary according to the dataset (voice).

5. RESULTS

The results obtained using the short-time power are shown in Figure 3.

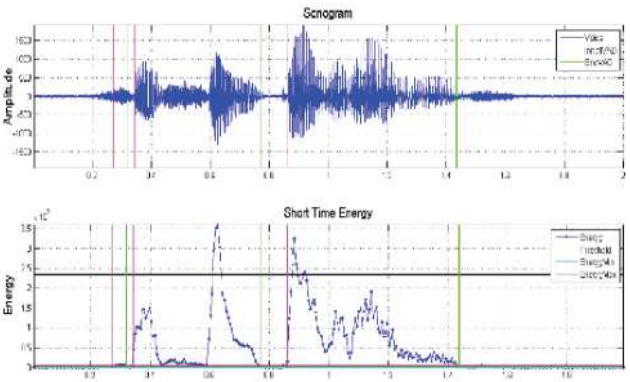

Figure. 3 VAD using the STE

Clearly, it can be observed that the threshold values are critical when locating the start and endpoints of the speech signal. When the thresholds of minimum and maximum energy were changed, erroneous detections in the algorithm were obtained. Many problems also presented themselves when trying to detect unvoiced or silent speech. The most suitable value for the threshold found in the analysis, was 0.03 for the BEPPA battery, the index was multiplied to the maximum value of the energy of the speech signal.

When implementing the algorithm that combines the STE and ZCR, the improvement was evident in the detection of unvoiced speech segments, in Figure 4 the results are shown.

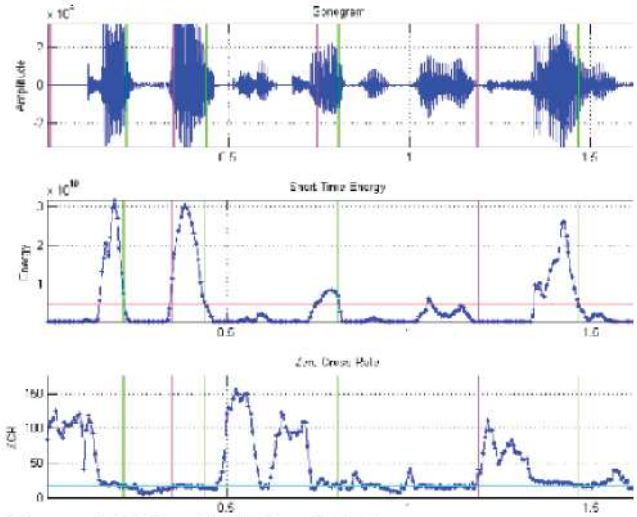

Figure. 4 VAD with STE and ZCR.

Once again finding the value of the threshold detection is the main issue, although improvement in detection is evident. In the method, the initial value of the word is calculated using the threshold τzc increasing the size of the detection limit and thus improving efficiency of the VAD algorithm, first suggested in (3). Finally, we implement the algorithm using HOD for detection. The results obtained with this algorithm on the BEPPA battery, shows improved efficiency when the detection of unvoiced speech segment (TDD) as compared to its predecessor, when the ZCR is implemented, and in which case detection of words takes both segments into account, ie the voiced and unvoiced segments and ignoring the silent segments. In Figure 5, the above results are shown.

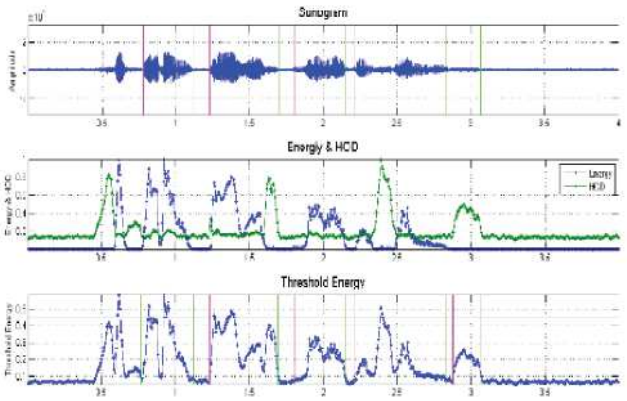

Figure. 5 VAD with STE and HOD.

The differences can still be observed and the presented thresholds used in each algorithm. The error matrix was used for the analysis of the implemented VAD ratings. To obtain the error matrix the algorithm was used with 10 sentences and for each the BEPPA battery was used, ie it had 65 words and about 40 rests. Table 1 shows the low efficiency of the algorithms that use only short-term energy to arrive at a decision, therefore the algorithm implemented in BEPPA is demonstrated to show a sensitivity of 66.2%. Since no false positive errors were detected specificity is 100%, and this percentage supports the results of (3) and (22), but shows a problem inherent in the use of the same thresholds and the corresponding classification decision. The error rate presented by this algorithm was 21.2%.

Table 1. Error matrix for the VAD with STE.

Algorithm VADSTE	Speech	Silence
False Positives	43	22
False Negatives	0	39

The results of the algorithm used with the STE and ZCR are shown in Table 2, where we can observe a slight improvement, but not a substantial change when trying to obtain a more robust algorithm. The sensitivity in this case was 75.4%. Again, the specificity is 100% effective in showing the readings with low background noise. The error rate also shows an improvement in its results to be 15.5%.

Table 2. Error matrix for the VAD with STE-ZRC

Algorithm VADSTEZRC	Speech	Silence
False Positives	49	16
False Negatives	0	38

Finally, Table 3 illustrates the results obtained by implementing the HOD using VAD detection, which shows a sensitivity of 96.9%. Demonstrating a marked improvement in the decision, this is due to the ability of the algorithm to distinguish the unvoiced speech segments. The specificity is 100% and the error rate is of 2.9%.

Table 3. Confusion matrix for the VAD with HOD

Algorithm VADHOD	Speech	Silence
False Positives	63	2
False Negatives	0	40

6. CONCLUSIONS

Due to problems encountered in empirically select the best threshold through trial and error, and the absence of bibliographic references that allow to apply several techniques for finding the most efficient threshold and, although in (16) different methods are described for VAD decision either by statistical methods, with distances and decision making by heuristics because, they were not implemented for River Plate Spanish. These results suggest that VAD cannot have universal thresholds, i.e. these cannot be implemented for any language or languages. This leads us to suggest that when the VAD threshold depends on the energy, you should implement robust decision methods for selecting optimal thresholds or using machine-learning based on statistical properties. The contribution of this research is that we can say that it is possible to use a single algorithm for different languages but it must have the ability to change the threshold adapted to the language used. Finally, it is suggested to implement algorithms comparisons with signals below freezing relations noise, that is, SNR $\leq$ - 5dB.

BIBLIOGRAPHY

1. Kondoz A. M. . Digital Speech: Coding for Low Bit Rate Communication Systems. 1st ed. UK. *John Wiley & Sons*; 2006
2. Campanella S. J. Advanced Digital Comunications Systems and Signal Processing Techniques. 1st ed. Engliwood Cliffs NJ. Prentice Hall; 1987
3. Rabiner L. R. and Sambur M. R. . An algorithm for determinig the endpoints of isolated utterances. *The bell system Technical*.1975;1(0):297-301
4. Rabiner L. R. and Sambur M. R.. Voiced-unvoiced-silence detection using the itakura lpc distance measure. *IEEE International Conference on Acoustics, Speech, and Signal Processing*.1977;1(1):323-325
5. Irwin M. J. Periodicity estimation in the presence of noise. *in Acoustics Conference.* 1980.
6. Tucker R. Voice activity detection using a periodicity measure in *IEEE Proceedings Communications*. 1992; vol. 139:377–380.
7. Haigh J. A. and Mason J. S., Robust voice activity detection using cepstral features- *in IEEE TENCON'93*; 1993;vol. 3:321–324.
8. Srinivasan K and Gersho A. Voice activity detection for cellular networks. in *IEEE in workshop on speech coding for telecommunications*. 1993: 85–86.
9. Huan Y. S. Benyassine A. Shlomot E. and Yuen E.. A robust low complexity voice activity detection algortihm for speech communication systems. *in Workshop on Speech Coding for Telecommunications.*1997;1():97-98
10. ITU-T. A silence compression scheme for g.729 optimized for terminals conforming to itu-t v.70. *Technical report, ITU-T* Rec. G.729 Annex B, November 1996.
11. Stegmann J. and Shröder G. Robust voice activity detection based on the wavelet transform. *in IEEE Workshop in Speech Coding, Pennsylvania*. USA. September 1997:99–100
12. Renevey P. and Drygajlo A. Entropy based voice activity detection in very noise conditions.

in EUROSPEECH, European Conference on Speech Communication and Technology. September 2001
13. Gökhun S. and Özer H. Voice activity detection in nonstationary noise. *in IEEE Transactions on Speech and Audio Processing.* 2000:478 – 482.
14. Bing F W and Kun C W. Voice activity detection based on autocorrela function using wavelet transform and teager energy operato. *Computational Linguistics and Chinese Language Processing*.2006;11(1):87-100
15. Kim N. S., Chang J-H. and Mitra S. K. Voice activity detection based on multiple statistical models *in IEEE signal pricessing* June 2006:1965–1976.
16. Gorriz J.M. Nuevos avances en la detección de actividad de voz mediante estadistica de alto orden y estrategias de optimización. *Tesis doctoral. Departamento de arquitectura y tecnología de computadores - Universidad de Granada*. Mayo 2006
17. Hansen J.H Deller J.R and Proakis J.G. Discrete time processing of speech signals. *John Wiley & Sons*. New York. 1993.
18. Rabiner L.R. and Schafer R.W. Digital processing of speech signals. *Prentice-Hall signal processing series*. Prentice-Hall. 1978.
19. Jyh-Shing R. J. . *Audio Signal Processing and Recognition*. http://mirlab.org/jang/books/audioSignalProcessing/ (accessed 9 de Junio 2010).
20. Tato J.M. Características acústicas de nuestro idioma. *Revista Otolaringológica*. 1949;1:17–34.
21. Rufiner H. L. Sigura A.D. Ossela, E. and M. E. and Torres. Sistema de administración para batería de ensayos para pacientes con prótesis auditivas. *in XVII Congreso Argentino de Bioingeniería - SABI'09*. 2009:94–98.
22. Kondoz A.M. Digital speech Coding for Low Bit Rate Communication Systems. *John Wiley & Sons*. USA. 2004