SABI 2015 IOP Publishing
Journal of Physics: Conference Series **705** (2016) 012037 doi:10.1088/1742-6596/705/1/012037

A simple but efficient voice activity detection algorithm through Hilbert transform and dynamic threshold for speech pathologies

D. Ortiz P.[1], Luisa F. Villa, Carlos Salazar, and O.L. Quintero
Mathematical Modeling Research Group, GRIMMAT, School of Sciences, Universidad EAFIT, Carrera 49 NO 7 Sur-50, Medellin — Colombia.

E-mail:dpuerta1@eafit.edu.co

Abstract. A simple but efficient voice activity detector based on the Hilbert transform and a dynamic threshold is presented to be used on the pre-processing of audio signals. The algorithm to define the dynamic threshold is a modification of a convex combination found in literature. This scheme allows the detection of prosodic and silence segments on a speech in presence of non-ideal conditions like a spectral overlapped noise. The present work shows preliminary results over a database built with some political speech. The tests were performed adding artificial noise to natural noises over the audio signals, and some algorithms are compared. Results will be extrapolated to the field of adaptive filtering on monophonic signals and the analysis of speech pathologies on futures works.

1. Introduction

Many authors label the section of the speech as *voiced*, where the vocal chords vibrate and produce sound, *unvoiced*, where the vocal chords are not vibrating, and *silenced* [1] [2]. The union of these three sections is important within the tools for audio analysis because they delimit the recognition of the speech and the specific characteristics of the speaker [2]. This process of identifications of *voiced/unvoiced* and *silenced* is known as voice activity detection [3]. As this work advance, the sections of *voiced/unvoiced* will be named as speech and *silenced* sections as silences.

A silence can be defined as the absence of audible sound or as a sound with a very low intensity [4]. These silences allow identify and separate the main components inside of communication channels marking the boundaries of the prosodic units and exposes the rate at which the speaker delivers his speech. It is possible to specify the silent pauses inside of a speech as the lack of the physical perturbation of the sound wave in a medium of propagation, indicated in the audio signal as the lack of amplitude. However, the low amplitude of the silence do not imply a totally absence of sound inside of the audio signal.

It is important to provide a methodology that accomplishes to discriminate properly the silence speech sections, considering the previously mentioned about the presence of sound with low amplitude in the silent pauses. These sounds of low amplitude are known as noises, which can be described as disturbances that interfere in the signal obtained by altering their real values. From this, the following hypotheses are proposed: Is it possible to differentiate speech sections with the silent pauses in a noisy

signal? Is it possible to design an adaptive system to different types of noise that they can be present in different audios and discriminate speech and silence?

For the voice activity detection, it is common to apply different techniques that depend of the information of the obtained signal. Some features like energy, the zero-crossing rate and the coefficients of linear prediction, can be combined in such a way that the distance between them would indicate if the analysed segment is speech or silent pauses [1] or used with a threshold, fixed or dynamic, to detect the speech [5]. Other methods used probability distributions of the noise present in the silences [2] [6].

This work uses signal own features like the zero-crossing rate and the signal energy in a particular window, in order to determinate a dynamic threshold. The zero-crossing rate indicates the number of times that the signal passes, in a time gap, by the value of zero, giving a simple measure of frequency content of the signal and the signal energy represents the amplitude variations. Once this information is obtained, it is used a modification of the methodology proposed in [7] to obtain a dynamic threshold that consist of the convex combination of the maximum and minimum of each of the property calculated. Finally, a second convex combination of the two thresholds is performed. Once the threshold is obtained, it is compared with the signal coverage obtained from the Hilbert transform and determines what speech (*voiced/unvoiced*) is and what *silence* is.

This work was developed under two objectives, adaptive filtering over monophonic for the pre-processing of noisy audio signal with no reference of noise and spectral overlapping as it is shown in [8] and the analysis of speech pathologies. The second objective is planned as future work to detect speech pathologies as stuttering [9].

2. Methodology

As mentioned previously, for the detection of the speech and silences sections we propose the combination of three features from the signal: the zero-crossing rate, the signal energy and the signal coverage from the Hilbert transform.

2.1. Zero-crossing rate

The zero-crossing rate is a simple measure of the frequencies in a certain signal. In speech sections, frequencies are of high amplitude and low band; therefore, the rate will be small, different to the silence [10].

$$Z_j = \sum_{i=(j-1)\cdot N+1}^{j\cdot N} |sgn[x(i)] - sgn[x(i-1)]| \tag{1}$$

N is the size of the window to measure.

2.2. Mean square error of the energy

To compute the energy is was used the mean square error of the same signal, because this gives in detail the peaks on speech and the valleys that points silences. The energy in a time window is define like

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

N is the size of the window to measure.

2.3. Signal covering

For the signal covering it was used the modulus of the analytic signal defined as

$$|\psi(t)| = (g(t)^2 + \hat{g}(t)^2)^{1/2} \tag{3}$$

Where $g(t)$ is the original signal and $\hat{g}(t)$ is the Hilbert transform of $g(t)$. The Hilbert transform is defined as

$$\hat{g}(t) = \mathcal{H}[g(t)] = g(t) * \frac{1}{\pi t} = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{g(t-\tau)}{\tau} d\tau \tag{4}$$

In figure 1 it can be seen an example of the coverage of the original signal using de modulus of the analytic signal.

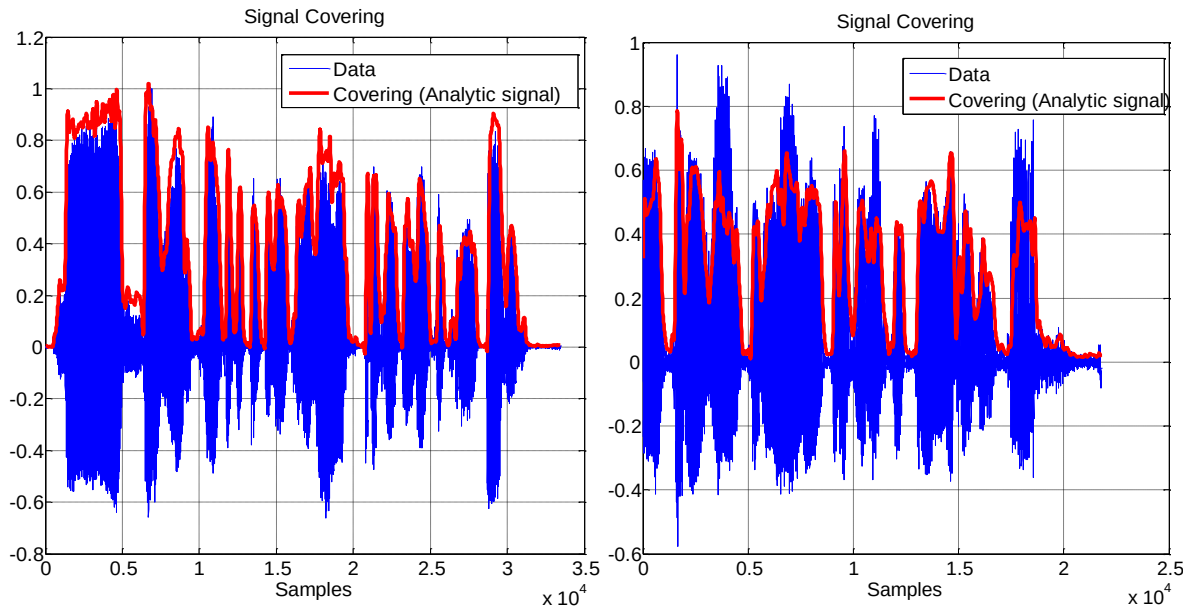

Figure 1. Signal covering for 2 different audio samples.

2.4. Dynamic threshold

For the calculation and implementation of the dynamic threshold, zero crossings and energy as dynamic features of the signal are used. First, both of them are extracted using overlapped time windows so non-stationary changes can be measured correctly, then, these data vectors are normalized, so the maximum value will be 1 and they can be compared with the signal information.

Once the data is normalized, a modification of the method proposed in [7] is used. This method consists in a convex combination of the maximum and minimum levels of the characteristic in each window. The zero-crossing rate and the energy threshold is defined by

$$\begin{aligned} E_{th}(j) &= (1-\lambda_E)\cdot E_{max} + \lambda_E \cdot E_{min} \\ Z_{th}(j) &= (1-\lambda_Z)\cdot Z_{max} + \lambda_Z \cdot Z_{min} \end{aligned} \tag{5}$$

Where λ is a scaling factor that control the process of estimation and j indicates the window. For diferents types of signals this value may vary depending of its characteristics [7], then, a scaling factor that depend directly of the signal

$$\lambda_E = \frac{E_{max} - E_{min}}{E_{max}} \qquad \lambda_Z = \frac{Z_{max} - Z_{min}}{Z_{max}} \tag{6}$$

It's possible that the minimum values of these two features can change until to find a value almost zero. In this case, the thresholds don't adapt properly to the signal changes, i.e., if it finds a value close to zero

(that is the minimum in all the information of the signal), the threshold for the energy and the zero-crossing rate, will be kept constant and low, which will give incorrect information in the case that there are silent pauses with noise of amplitude and high frequencies. To avoid this, the minimum value of (6) is increased slightly and is defined by

$$\begin{aligned} E_{min}(j) &= E_{min}(j-1) \cdot \Delta_E(j) \\ Z_{min}(j) &= Z_{min}(j-1) \cdot \Delta_Z(j) \end{aligned} \tag{7}$$

The parameter Δ is define as

$$\Delta(j) = \Delta(j-1) \cdot \alpha \tag{8}$$

where α is a growth factor. Once this threshold is obtained for the energy and the zero-crossing rate, it is defined the global threshold for discriminate the silent pauses in a speech like a convex combination of the two previous thresholds

$$TH(j) = (1-p) \cdot E_{th}(j) + p \cdot Z_{th}(j) \tag{9}$$

where p is the scaling factor of the convex combination. Once the dynamic threshold is obtained of the signal it is possible to compare the coverage of the same signal obtained from (3). If the coverage is below of the limit, the audio section is considered a silence, if it is above, is considered a speech section.

3. Silence detection procedure

Once calculated the features of the signal (Zero Crossing rate, energy and signal covering) and obtained the dynamic threshold, the procedure for detecting the silences sections by comparing the threshold and coverage was established.

1. First, signal data is normalized, followed by a pre-filtering band-pass with cut frequency 100-3200 Hz.
2. For the dynamic threshold, first the maximum and minimum variables for the energy and the zero-crossing rate are determined. For the energy, the maximum will be the average of the data and the minimum variable will be the minimum value. For the zero-crossing rate, if the first value is equal to zero, it will be taken the average as maximum value, if is different to zero, will be the first value of the data. For the minimum variable, will be the minimum zero-crossing rate of the data, if these is equal to zero, this variable is taken as an epsilon $\epsilon > 0$ (a low number closed to 0).
3. Once the maximum and minimum are determined, it follows to determine the threshold for the energy and the zero-crossing rate for each overlapped window. In this case, the overlapping was set in 90% of the size of the window. Later, the total threshold of the window is calculated using (9).
4. The complete signal coverage is determined from of the analytic signal by using the Hilbert transform, then, a decimation over the analytic signal is made to smooth the covering.
5. Finally, the dynamic threshold is compared with the coverage obtained in step 4. If the threshold is above of the coverage this audio section is taken as silent pauses; if the threshold is below, is taken as speech.

4. Results and analysis

4.1. Test

For the test, it was made a database with different political speeches published on the internet. These speeches were recorded in noisy environments that can disturb the voice activity detection and the noise

has spectral overlapping with the real signal of the speech. The data base has a sample rate of 8 KHz, and to analyse the correct voice activity detection, the speech and silence section where identified manually by an expert operator as shown in figure 2. To test the robustness of the algorithm it was added artificial Gaussian white noise with SNR of 5 dB, 15dB and 20dB measure using the energy of the noise and the signal, and the result was compared with a benchmark algorithm found in [11]. An error measure was used to calculate the performance of the algorithm. This measure was made by comparing the samples in the signal identified by the expert as speech or silence and results of the algorithm. Another measure used was the number of silences identified by the algorithm.

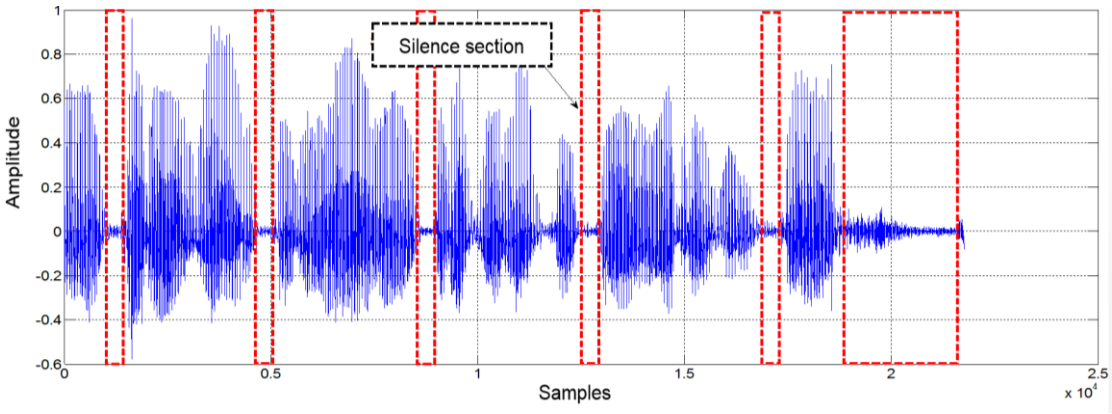

Figure 2. Silence section identified by an expert for the test.

To obtain the analytic signal it was established a decimation factor of 10 from different tests, watching that this describe in a good way the signal coverage. For the extraction of the features, which define the dynamic threshold, were used small windows of 12.5 ms or 100 samples over the sample rate (8 KHz, 8000 samples per second) with an overlapping of the 90%. Were used small windows with the objective of that abrupt changes do not alter the measure and the overlapping allows following precisely the characteristics behavior. The growth factor α for the minimum energy was settled in 1.0001, so the minimum energy grows up in a low rate, and the scaling factor p in 0.1 to prioritize the measure of the energy threshold.

4.2. Results and analysis

For the first test, signals where used without adding synthetic noise, as it was mentioned before, audio signals has their natural noise. Results can be seen on table 1 where the percentage of error and the number of silences identified by both algorithms is presented.

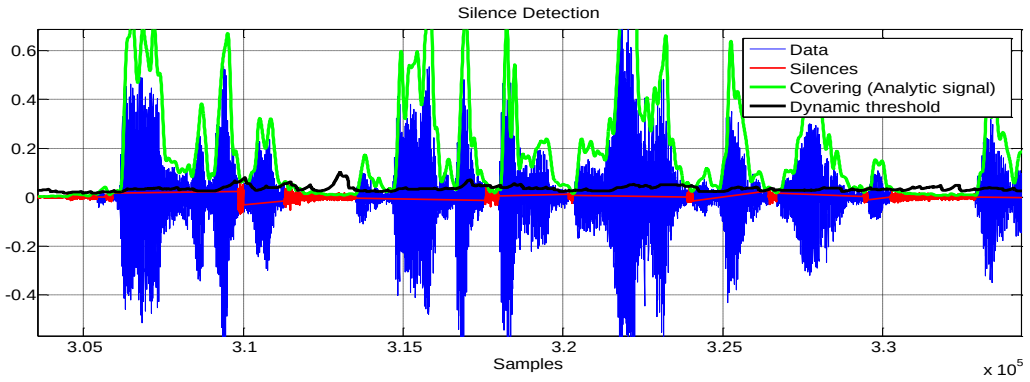

Figure 3. Behaviour of the proposed algorithm. On green the coverage of the signal is shown. On black, the dynamic threshold. The speech section can be find on blue, and silences on red.

Table 1. This table shows the results for the VAD of 10 audio signals of the data base. Method 1 is the algorithm presented in this work. Method 2 is the benchmark method found on literature [11]. Here the percentage of accuracy is presented and the number of silences identified

	Natural noise				
	Method 1		**Method 2**		
Audios	%	N. Sil.	%	N. Sil.	N. Silence Real
Audio 1	16.30	3	16.30	3	4
Audio 2	8.30	6	23.30	3	7
Audio 3	13.50	6	31.00	2	8
Audio 4	6.34	9	27.34	3	10
Audio 5	31.25	5	42.91	1	12
Audio 6	16.37	6	38.24	1	8
Audio 7	3.52	7	28.52	2	7
Audio 8	5.325	11	22.82	5	12
Audio 9	9.45	7	21.12	3	6
Audio 10	1.71	10	1.92	5	10

As it is shown, the performance of the proposed method is much better than the algorithm proposed on [11]. The percentage of accuracy has an average of 11.21% and the number of silences identified are closed to the values to those founded by the expert. In figure 3 can be seen the behavior of the proposed algorithm, showing the combination of the dynamic threshold and the covering of the signal to identify the silence section.

One of the objectives of using a dynamic threshold is that it can adapt to the spectral characteristics over the signal. As it can be seen on figure 3 dynamic threshold can change over time and by different kind of spectral overlapping with natural noise in this case.

The algorithm was also tested contaminating the audio signals with Gaussian white noise with SNR of 5 dB, 15 dB and 20 dB. Results can be observed on table 2 to 4.

Table 2. This table shows the results for the VAD of 10 audio signals contaminated with white Gaussian noise with SNR of 20 dB. Method 1 refers the proposed algorithm. Method 2 is the benchmark method found on literature [11].

	Gaussian Noise SNR 20 dB				
	Method 1		**Method 2**		
Audios	%	N. Sil.	%	N. Sil.	N. Silence Real
Audio 1	18.70	5	18.70	3	4
Audio 2	12.35	9	22.35	3	7
Audio 3	36.25	14	36.25	2	8
Audio 4	36.25	18	32.75	3	10
Audio 5	5.12	12	37.21	1	12
Audio 6	8.11	9	34.36	1	8
Audio 7	3.83	7	28.83	2	7
Audio 8	11.55	15	23.22	5	12
Audio 9	18.96	8	24.79	3	6
Audio 10	15.58	13	22.58	5	10

Table 3. This table shows the results for the VAD of 10 audio signals contaminated with white Gaussian noise with SNR of 15 dB. Method 1 refers the proposed algorithm. Method 2 is the benchmark method found on literature [11].

	Gaussian Noise SNR 15 dB				
	Method 1		**Method 2**		
Audios	%	N. Sil.	%	N. Sil.	N. Silence Real
Audio 1	55.49	8	29.24	3	4
Audio 2	07.27	8	22.27	3	7
Audio 3	35.35	13	39.73	2	8
Audio 4	49.37	21	35.37	3	10
Audio 5	7.23	13	36.40	1	12
Audio 6	2.26	8	32.89	1	8
Audio 7	16.79	9	31.79	2	7
Audio 8	26.24	18	23.32	7	12
Audio 9	20.86	8	26.70	3	6
Audio 10	17.51	13	28.01	4	10

Table 4. This table shows the results for the VAD of 10 audio signals contaminated with white Gaussian noise with SNR of 5 dB. Method 1 refers the proposed algorithm. Method 2 is the benchmark method found on literature [11].

	Gaussian Noise SNR 5 dB				
	Method 1		**Method 2**		
Audios	%	N. Sil.	%	N. Sil.	N. Silence Real
Audio 1	98.24	11	45.74	3	4
Audio 2	13.77	9	23.77	3	7
Audio 3	46.65	15	42.27	2	8
Audio 4	56.43	22	38.93	3	10
Audio 5	17.04	15	40.37	1	12
Audio 6	3.44	8	34.07	1	8
Audio 7	40.62	12	40.62	2	7
Audio 8	35.69	20	26.94	7	12
Audio 9	50.44	12	32.94	3	6
Audio 10	22.83	8	33.33	5	10

Table 5. Average Error comparison of the two methods

	Average error %			
	Natural Noise	**SNR 20 dB**	**SNR 15 dB**	**SNR 5 dB**
Method 1	11.21	16.67	23.84	38.51
Method 2	27.08	28.10	30.57	35.90

Table 6. Standard deviation Error comparison of the two methods

	Standard deviation error %			
	Natural Noise	**SNR 20 dB**	**SNR 15 dB**	**SNR 5 dB**
Method 1	8.68	11.52	17.97	27.20
Method 2	11.45	6.65	5.70	6.95

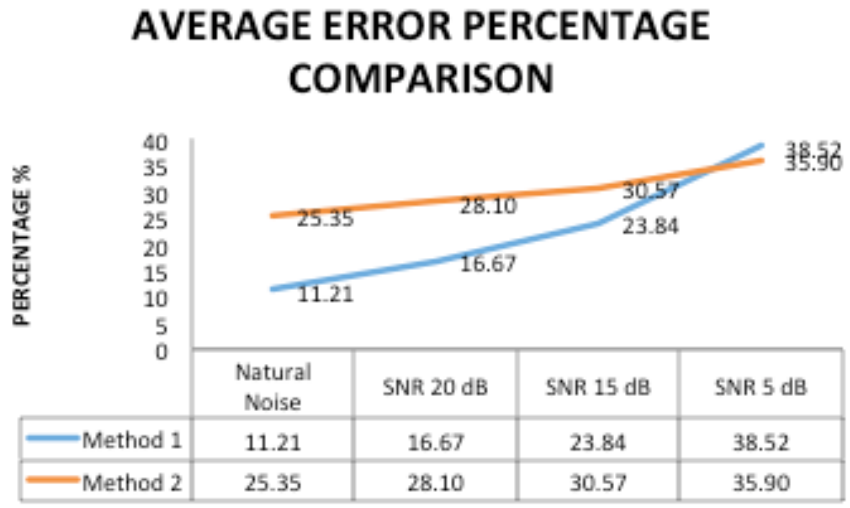

Figure 4. Average percentage error comparison between the two methods.

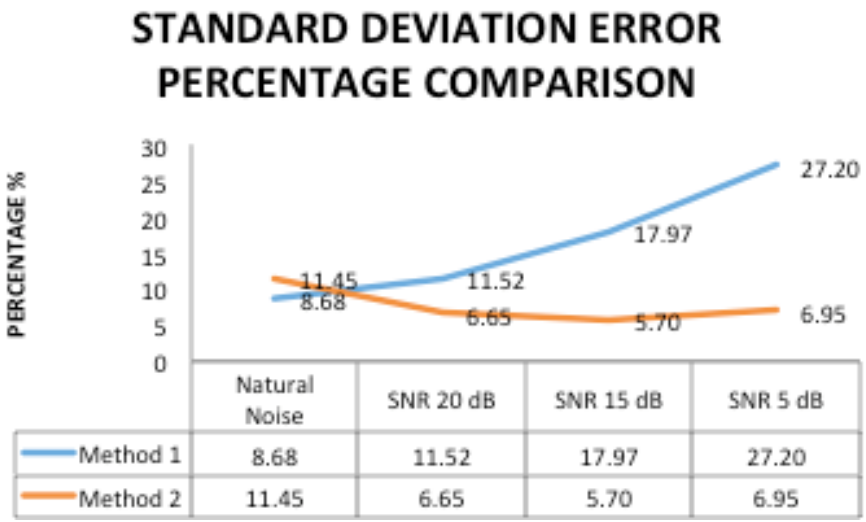

Figure 5. Standard deviation percentage error comparison between the two methods.

By the results is clear that the proposed algorithm is robust under the different test performed. On tables 2 and 3, results shows that the algorithm is consistent with the first test where no noise was added. The low percentage of error shows that the algorithm is robust against the noise with low and middle energy. Also the number of silence section detected keeps closed to the real.

Although in test with SNR of 20 dB and 15 dB shows good result, is important to note that as energy of the signal increases, the percentage of error increases too. As shown in table 5, the performance of the proposed algorithm gets worse as the energy of the noise increase, but it remains at a low percentage.

It can be observe in Figure 5, the value of the standard deviation of error increases as the noise level in the audio, which means that the method is prone to failure in the presence of noisy signals unlike the other method in which the standard deviation of error remains constant.

5. Conclusions

Considering that noise is a natural phenomenon when getting the information, is important to build tools that can adapt to this noises without inconvenient. Comparing this test with real life, different kind of noise can be found when getting the information to analyze such other voices, short circuits and others.

Voice activity detection take an important place in issues such as emotion detection in patients with diseases or emotional disorders, in remote monitoring of these patients, in pathologies of the vocal tract, and others. From the analysis carried out, it can be said that although the algorithm proposed has a simple structure, it is robust and consistent against noise of different energies so it can be implemented in different applications for the detection of pathologies related to speech.

These results could be used for stablish relationships of the presence and frequency of these segments in a speech with the objective to detect deception, emotional states in social interaction, shortcomings of affective disorder or pathologies associated with speech like the stuttering.

6. References

[1] B. S. Atal and L. R. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Aplications to speech Recognition," *IEEE*, pp. 201-212, 1976.

[2] G. Saha, S. Chakroborty and S. Senapati, "A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications," Indian Institute of Technology, Khragpur.

[3] F. G. Germain, D. L. Sun and G. J. Mysore, "Speaker and Noise Independent Voice Activity Detection," Proceedings of Interspeech, Lyon, 2013.

[4] R. V. Prasad, A. Sangwan, H. S. Jamadagni, C. M. C., R. Sah and V. G. , "Comparison of Voice Activity Detection Algorithms for VoIP," in *Proceedings of the Seventh International Symposium on Computers and Communications (ISCC'02)*, 2002.

[5] E. Verteletskaya and B. Simak, "Voice Activity Detection for Speech Enhancement Applications," Acta Polytechnica, Praha, 2010.

[6] S. G. Tanyer and H. Özer, "Voice Activity Detection in Nonstationary Noise," *IEEE*, pp. 478-482, 2000.

[7] K. Skahnov, E. Verteletskaya and B. Simak, "Dynamical Energy-Based Speech/Silence Detector for Speech Enhancement Applications," in *Proceedings of the World Congress on Engineering*, Londres, 2009.

[8] D. Ortiz and O. L. Quintero, "Una aproximación al filtrado adaptativo para la cancelación de ruidos en señales de voz monofónicas," in *XVI Congreso Latinoamericano de Control Automático, CLCA 2014*, Cancún, 2014.

[9] P. Pichot, Diagnostic and Statistical Manual of Mental Disorders, Washington, D.C.: American Psychiatric Association, 1994.

[10] R. G. Bachu, S. Kopparthi, B. Adapa and B. D. Barkana, "Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal".

[11] Z. H. Tan and B. Lindberg, "Low-Complexity Variable Frame Rate Analysis for Speech Recognition and Voice Activity Detection," IEEE Journal of Selected Topics in Signal Processing, vol. 4, p. 5, January 2010.