SPEECH SIGNAL DETECTION IN NOISY ENVIRONEMENT USING A LOCAL ENTROPIC CRITERION

I. Abdallah, S. Montrésor and M. Baudry

Laboratoire d'Informatique de l'Université du Maine
Email : imad@lium.univ-lemans.fr

ABSTRACT

This paper describes an original method for speech/non-speech detection in adverse conditions. Firstly, we define a time-dependent function called Local Entropic Criterion [1] based on Shannon's entropy [2]. Then we present the detection algorithm and show that at Signal to Noise Ratio (SNR) above 5 dB, it offers a segmentation comparable to the one obtained in clean conditions. We finally, describe how at very low SNR (< 0 dB) , it permits to detect speech units masked by noise.

1. INTRODUCTION

Currently, most speech recognition systems provide good performances in low-noise background. Unfortunately, these performances degrades rapidly in noisy environment. In strongly noised conditions (SNR equal to -15 dB), it becomes difficult to detect speech occurrence in the signal. Consequently, a pre-processing of the signal is required [3,4]. In order to achieve this task, we define a feature that lies on Shannon's entropy. The entropy of Shannon is considered as a cost function that measures the degree of organization of the signal.

The Local Entropic Criterion (LEC) described later in this paper, is categorized as a rupture detector used for the segmentation of signals into homogenous zones separated by transitions [1,5]. We will show that at 10 dB SNR the results approach those obtained with clean utterances and that at -20 dB SNR the LEC estimator succeed to detect speech presence in noise. In this paper, the noise is assumed to be additive white gaussian noise.

Experiment results were obtained by using an independent speaker isolated digit database in white noise conditions.

2. THE LOCAL ENTROPIC CRITERION

2.1 Shannon's Entropy

Shannon's entropy can be considered as an indicator of the signal spectrum energy concentration. It measures the degree of organization of the signal. It is defined by:

$$E(s, e_{j,k}) = -\sum_{k} \frac{\left|C_{j,k}\right|^2}{\|s\|^2} \ln \frac{\left|C_{j,k}\right|^2}{\|s\|^2}$$

where Cj,k form the set of the coefficients of the decomposition of the signal, s=$\{s_n\}_{0<=n<=N-1}$, relative to an orthonormal base $\{e_{jk}\}_{j,k}$.

2.2 THE LOCAL ENTROPIC CRITERION ALGORITHM

The LEC function derives from the Discreet Fourier Transform (DFT) coefficients computed from signal samples. It is given by:

$$LEC_W(n) = \frac{E - (E1 + E2)}{|E + E1 + E2|},$$

Where *n* is the middle of the current analysis window *W*, E is the corresponding Shanon's entropy, E1 and E2 are entropies of the left and right halves of this window relative to the (DFT) basis.

Time dependency is obtained by sliding the analysis window point by point. Decomposition basis is obtained by using a recursive DFT, which reduces considerably algorithm's cost.

In [1], we detail how the LEC function is used for continuous speech segmentation.

LEC varies with respect to spectrum energy concentration of analyzed signal. Positive LEC

values traduce dispersion of signal spectral energy while negative ones are linked to its concentration.

Segments formed by adjacent points, *n* , such as LEC(n) < = 0 (resp. LEC(n) > 0) are said *stable* (resp. *unstable*). A stable zone corresponds to the presence of a quasi-stationary part of the signal, while unstable zone can be seen as transitions between two stationary ones.

In order to localize the ruptures in the evolution of the signal, we begin by smoothing the LEC function using a low-pass filter . Ruptures will then correspond to the maximas of the smoothed LEC function relative to it's derivative.

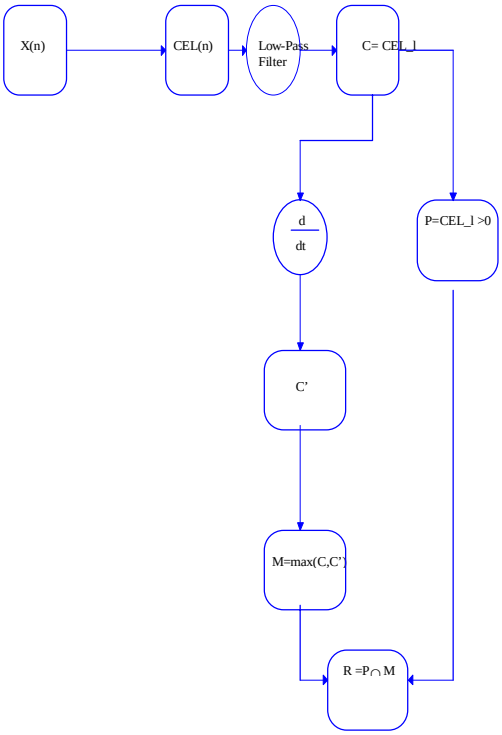


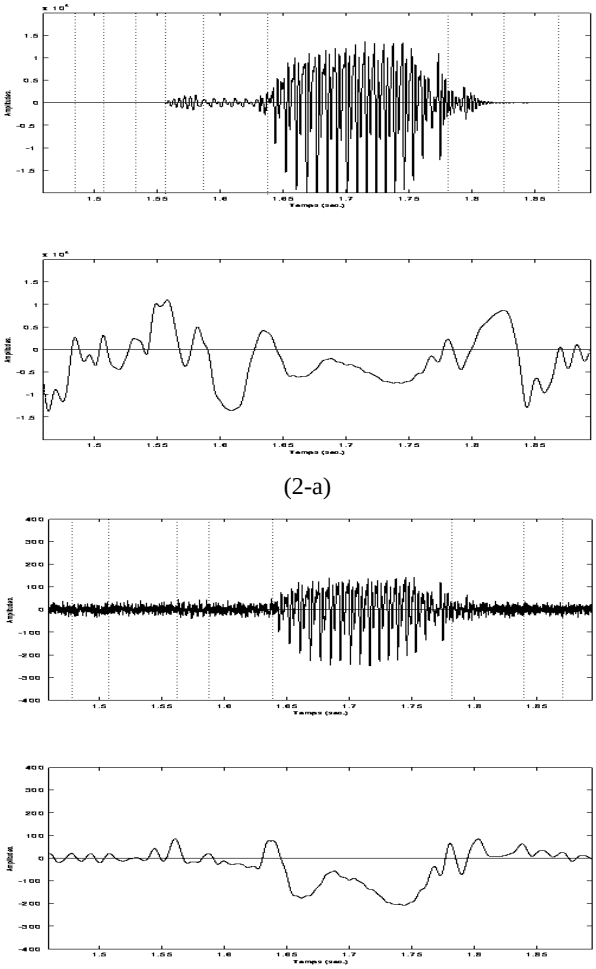
figure 1 : Rupture detection algorithm using LEC function.

For speech signals, the presence of voiced units is traduced by the creation of stable zones relative to the LEC, while unvoiced ones such as plosives are considered as unstable zones relative to the LEC.

2.3 Detection of the Vocalic Segments in Noisy Environment

Many approaches has been used in speech enhancement in adverse conditions including spectral subtraction [6], subspace decomposition [7,8] and minimum mean square error [MMSE] estimator [9]. The LEC estimator can be seen as a spectral estimator. It allows us to detect signal organisation by focusing on the energy concentration in the frequency domain.

We have tested the performance of the LEC's algorithm on a clean signal and on the same one with additive noise with SNR equal to 10 dB. The results are comparable with those obtained in clean conditions. A comparison between a segmentation obtained in clean and noisy conditions is described in figure 3.



(2-a)

(2-b)

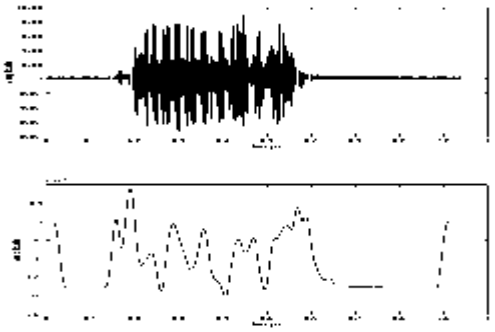
figure 2: Segmentation of the French word «deux» using LEC function (2-a). The vertical lines corresponds to the ruptures defined by LEC algorithm. In (2-b) we added white gaussian noise to the same sentence with an SNR equal to 10 dB. Notice that the segmentation obtained by the LEC estimator of the speech sentence is the same under clean or noisy conditions.

At very low SNR (< -15 dB), it becomes difficult to detect speech signal presence. In this case LEC algorithm is used in order to localize speech activity masked by noise. LEC being sensible to the spectral energy concentration, it appears to perform poorly when there is virtually no significant speech activity while it seems to be a robust estimator for the detection of parts of speech signal as vocalic segments kernels drowned in noise.

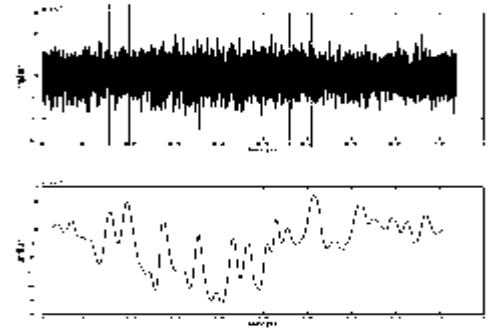
Vocalic kernels can be seen as concatenation of quasi-stationary signals with spectral energy concentration around their harmonic components. In the spectral domain they are less affected by the noise presence than other speech components as fricatives or plosives.

In a stable zone, minimas of the smoothed LEC function relative to it's derivative, indicates the area were the spectral energy concentration is very high. When noise is added to quasi-stationary zones, it can be traduced by a dispersion of the spectral energy. In spite of this energy dispersion, presence of quasi-stationary zones masked by noise is traduced by the existence of negative minimas relative to the LEC function. Maximas of the LEC function will then correspond to the beginning and the end of stable zones masked by noise.

Figure 3 shows a comparison between two LEC functions corresponding to vowel /a/ in clean conditions and with added noise (SNR = -10dB).



(3-a)



(3-b)

figure 4 : (3-a) vowel /a/ and its corresponding LEC function. (3-b) the same vowel /a/ masked by noise with SNR equal to 10 dB and the corresponding LEC function. Notice that vowel presence is indicated by the presence of stable zones were the LEC values are minimal.

In a detection approach, we are confronted to choose between two simple hypothesis H and K :

H : x = noise alone

K : x = signal + noise .

In order to decide if we are under the hypothesis H or K, the LEC estimator will take the following decisions :

if x is a stable zone, we calculate local minima's value (minl) of LEC function relative to x. Then we calculate the global minimum MING value of LEC function. Then we calculate local and global maximum (maxl and MAXG)of LEC function.

In a stable zone:

1- if minl < λ.MING and maxl > λ.MAXG, were λ is a fixed threshold with 0 <= λ <=1, x is considered as speech unit masked by noise.

2- if minl >= λ.MING and maxl <= λ.MAXG, x is considered as noise. Let us notice that in this case, x can corresponds to noise or to the presence of unvoiced speech unit.

3- if x corresponds to an unstable zone, it will be considered as noise.

3. EXPERIMENTATION

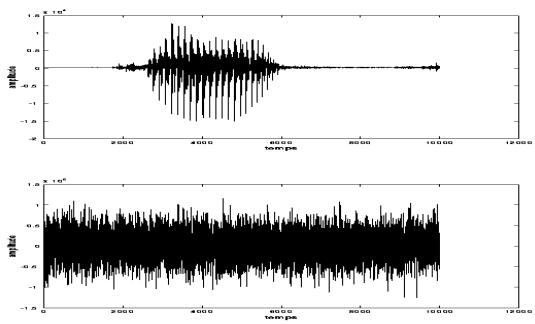
The above described algorithm has been tested on isolated noisy speech utterances with Signal to Noise Ratios (SNR) varying from 6 to -24 dB.

As an illustration, figure 4 shows LEC curve compared with the energy's curve of a part of a speech signal with additive noise with a SNR equal to -20 dB.

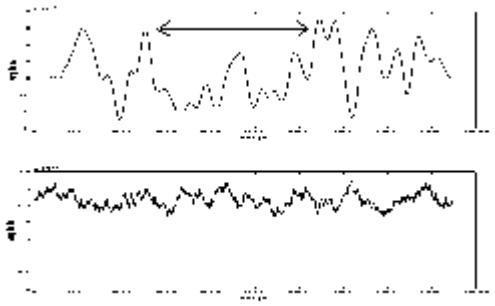
As a comparison, we can clearly see that our algorithm yield a good detection of such a signal while energy's curve is unable to.

For each SNR, we calculate the rate of false alarm detection and the Correct Detection rate (CDR). The choice of the threshold λ is important. When λ is low we reduce the risk of false alarms detection (FAD); in the same time, presence of some speech units can be omitted. When taking a high threshold λ, we increase the correct detection rate but FAD rate too. Notice here that we fixed a minimum length (40 ms) fore stable zones to be considered as the indicator of speech presence.

Figure 5 shows the ratio of (FAD) and (CD) for SNR varying from 10 to -20 dB and λ=0.5. For an example at SNR =-20dB, CDR is equal to 79.99 % and FAD rate is equal to 49.56 %. For SNR above 0 dB CDR = 100 %.and FAD =0 %.



(4-a)



(4-b)

figure 4. Speech/non-speech detection using LEC. (5-a) : a part of a speech signal (up) noised with an SNR equal to -20 dB (down). (5-b) : Detection using the LEC function. (up) and Energy's curve of the noised signal (down).

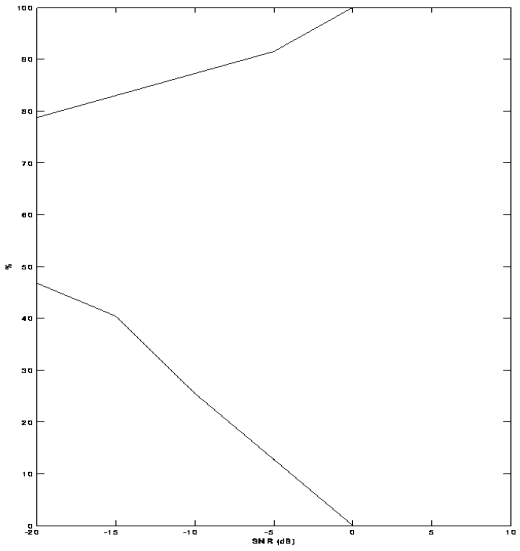


figure 5 : Correct detection rate (upper curve) and False alarm detection rate (lower curve), relative to SNR with λ=0.5.

4. CONCLUSION

In this paper we presented a new algorithm for speech/non-speech detection based on a Local Entropic Criterion. At SNR above 5 dB, it offers a segmentation comparable to the one obtained in clean conditions. At very low SNR (< 0 dB) , it permits to detect speech units masked by noise. We are currently working on further tests and verifications especially for signals noised with different models of noise [5,6] such as crowd noise. In adverse conditions, we think that the algorithm might advantageously be integrated in speech recognition systems.

5. REFERENCES

[1] I. Abdallah, S ; Montresor and M. Baudry: «*Un Al*gorithme Recursif pour la Segmentation des Signaux de Parole Basé sur un Critère Entropique Local». 4th congress of the French Acoustical Society (SFA), April 1997.

[2] R. Coifmann and V. Wickerhauser: «Entropy-Based Algorithms for Best-Basis Selection». IEEE transactions on info. theory, vol 38, n°2, pp 713-718, Mars 1992.

[3] R. Andre-Obrecht and J.B Puel: «détection des débuts et fin de parole en environnement difficile». quatorzième Colloque GRETSI, pp 157-160**,** September 1993.

[4] L. Mauuary, J. Monné : «Speech/Non-speech Detection for Voice Response Systems». 3rd European Conference on Speech Communications and Technology. September 1993.

[5] R. Andre-Obrecht: «A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals». IEEE transactions on Acoustics, Speech, and Signal Processing, n°36, pp 29-40, January 1988.

[6] Y. Malca and D. Wulich : « Improved Spectral Subtraction for Speech Enhancenment ». EUSIPCO, 10-13th September 1996, Trieste-Italy.

[7] P.S. Hansen, P.C Hansen, S. Hansen and J. Sorensen:« Noise Reduction of Speech Signals Using the Rank-Revealing Uliv Decomposition ». EUSIPCO, 10-13th September 1996, Trieste-Italy.

[8] D. Darlington and D. Campbell :« Sub-band Adaptive Filtering Applied to Speech Enhancement ». EUSIPCO, 10-13th September 1996, Trieste-Italy.

[9] P. Scalart, J. Vieira Filho and J. Chiquito:« On Speech enhancement Algorithms Based on MMSE Estimation ». EUSIPCO, 10-13th September 1996, Trieste-Italy