# Precise Voicing Information Extraction in Speech Signals Using the Analytic Signal

Stéphane Rossignol and Olivier Pietquin
Supélec – Metz Campus, IMS Research Group
2, rue Édouard Belin; F-57070 METZ; FRANCE
Stephane.Rossignol@supelec.fr and Olivier.Pietquin@supelec.fr

***Abstract*—** *This paper proposes a voiced – unvoiced measure based on the Analytic Signal computation. This voiced – unvoiced feature can be useful for many speech processing applications. For instance, considering speech recognition, it could be incorporated into commonly used acoustic feature vectors, such as for example the Mel Frequency Cepstral Coefficients (MFCC) and their first two derivatives, in order to improve the performance of the overall system. The evaluation of the developed measure has been performed on the TIMIT database. TIMIT has been manually segmented into phones. The voicing information can easily be derived from this segmentation. It is shown in this paper that the automatic voiced – unvoiced segmentation obtained using the method described in the next sections and the manual voiced – unvoiced segmentation provided by TIMIT are very similar.*

***Keywords*—** *Speech Processing, Signal Processing Theory and Methods*

## I. Introduction

In unit-selection-based speech synthesis (see [1]) as well as in speech recognition (see [2], [3]), having the voicing information can increase the robustness of the systems in use and the quality of their results.

Our primary motivation of developing the method presented in this paper is to obtain a reliable voiced – unvoiced segmentation. Our second goal is to obtain voiced – unvoiced segmentation marks accurately positioned in time. In [2], three measures of voicedness are described. They are based on a short time representation, in the time domain or in the frequency domain, of the signals. A 40 ms long window is used. When describing a voiced segment, it is assumed that the signal is quasi periodic. The value of 40 ms is chosen in order to get, in a time frame, a few periods of this "stationary" signal. But, unfortunately, this windowing necessarily provokes some uncertainty in the temporal position of the segmentation marks. Among the other features developed in the past, the energy, the zero crossing rate (ZCR), the autocorrelation coefficient, refined analyses of the STFT spectrum, etc. (see [4], [5], [6]) are very well known. They require to make the same stationarity assumption and thus require the same kind of frame cutting off process. The method presented in this paper is an attempt to avoid the use of such wide windows.

Few papers have addressed the analysis of shorter frames. In [7], a method based on the Gabor atomic decomposition, using 3.2 ms duration frames, has been implemented. 84 % correct classification accuracy was obtained on a database of 62 sentences, uttered by a single male speaker. In [8], short time frames, around 8 ms, are used as well. The developed features are based on the energy and on some statistical analyses of the STFT spectrum. The applications concern radio and telephone channels, for which information below 300 Hz and above 3200 Hz is usually destroyed. An error rate of 8.5 % is obtained, using around 100 phonetically balanced sentences to train and test the system. In this paper, a method based on the Analytic Signal (AS) is presented. The method provides voicing information for each of the signal samples. First, its performance are compared to the performance obtained with the short-time energy and the inverse ZCR (1-ZCR), this for a frame length varying from 2 ms to 42 ms. In the case of the Analytic Signal, the frame length concerns the size of the Hilbert filtering (see figure 1). Second, the behaviour of these three methods concerning the position of the segmentation marks and concerning the number of fast successions of voiced – unvoiced segmentation marks is analyzed.

The rest of this paper is organized as follows. In Section II, the method is derived in details. In Section III, the obtained results are presented. Finally, in Section IV, a conclusion and future works are provided.

## II. Description of the complete system

### A. Synoptic of the method

The method is presented in details below and is summarized in figure 1. The analysis of the sound is executed in four steps:

1. First, the audio signal is fed through a band-pass filter, with a $1/f^2$ decay in the band. The goal of this filtering is, in the voiced parts of the sound, to get the first partial as much prominent as possible in magnitude (see [9], where such a method, based on the first partial enhancement for pitch tracking purposes, is presented). Therefore, this signal: $s(n) \simeq a_1 \cos(2\pi f_0 n/f_s + \phi_1)$ is obtained, where $f_0$ is the fundamental frequency and $f_s$ the sampling rate. In the unvoiced parts of the sound, noise is obtained.
2. Second, the Analytic Signal is determined by using Hilbert filtering. Notice that the band-pass filtering and the Hilbert filtering are performed simultaneously, as indicated in figure 1. In the voiced parts of the sound, this signal: $X(n) \simeq a_1 \exp(2\pi j f_0 n/f_s + j\phi_1)$ is obtained.
3. Third, the modulus $A = |X(n)|$ of the Analytic Signal is estimated. In the voiced parts of the sound, it can be considered that this modulus provides an estimation

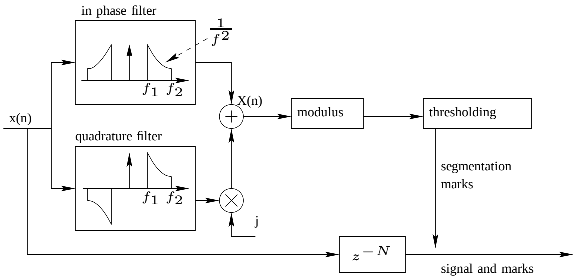


Fig. 1. Synoptic of the method

of the magnitude of the first partial ($A \simeq a_1$). In the unvoiced parts of the sound, only noise is obtained. It is thus assumed that $A$ has higher values in the voiced segments than in the unvoiced segments of the sound.

4. Fourth, this modulus $A$ is automatically thresholded (see section II-B for more details), providing the segmentation of the signal into voiced and unvoiced parts.

Notice that for the joint Hilbert and $1/f^2$ filtering, a FIR filter is used. It is thus easily possible to determine the precise position in time of the segmentation marks on the sound, as the propagation time is fixed, known and equal to $N = Tf_s/2$ samples, where $T$ is the frame length in second.

### B. Thresholding

Automatic thresholding (or binarization) methods, coming from the image processing community, are numerous (see for instance [10]). In this paper, three "ad-hoc" methods and the Otsu's method have been tested. Notice that the recording level of the analyzed sounds (see section III) vary. Thus, the thresholds need to be re-estimated for each of the analyzed sounds. This is reflected in the definitions of the three "ad-hoc" thresholds given below by the fact that the maximum value or the mean value of the signal to threshold intervene. The values $t_1$, $t_2$ and $t_3$ of these thresholds are respectively equal to:

1. $t_1 = C_1 \max[A]$; where $C_1$ is a constant used for every sound in the database.
2. $t_2 = C_2 \max[B]$; where $B$ corresponds to the 95 % smallest samples of $A$ and where $C_2$ is a constant used for every sound in the database.
3. $t_3 = C_3 \text{mean}[A]$; where $C_3$ is a constant used for every sound in the database.

A training stage is necessary to determine the values of the three constants $C_1$, $C_2$ and $C_3$.

The Otsu's method is explained in details for instance in [10]. The optimum threshold $t_O$ separating the two classes is such as their within-class variance $\sigma_w^2(t_O) = \omega_1(t_O)\sigma_1^2(t_O) + \omega_2(t_O)\sigma_2^2(t_O)$, where $\omega_1$ and $\omega_2$ are the probabilities of the two classes and $\sigma_1^2$ and $\sigma_2^2$ the variances of these classes, is minimal. In [11], appendix B, the performance of nineteen thresholding methods have been compared, considering three disruptive parameters: 1. the two classes are unbalanced; 2. the variance of one of the class is higher than the variance of the other one; 3. the two classes overlap a lot. Considering the voiced – unvoiced segmentation problem addressed in this paper, the selected techniques need to remain robust before all to the second disruptive parameter. Indeed, the variance of the Analytic Signal magnitude and of the Energy (respectively of the inverse ZCR), when going through the voiced (respectively unvoiced) parts of the signal, is high, as these features within a phone and from a phone to another one, can vary a lot; on the contrary, the AS magnitude and the Energy (respectively the inverse ZCR), when going through the unvoiced (respectively voiced) parts of the signal, vary much less within a segment and from a segment to another one. In this view, the Otsu's method proved in [11] to be one of the most robust thresholding method. This is why it has been retained in this paper.

## III. Experiments

### A. The TIMIT database

The evaluations were performed on the TIMIT database (see [12]). TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. 69.5 % of the speakers are males. The considered phones are: 8 stops (/b/, /d/, /g/, /p/, /t/, /k/, /dx/, /q/); 6 closures (/bcl/, /dcl/, /gcl/, /pcl/, /tcl/, /kcl/); 2 affricates (/jh/ and /ch/); 8 fricatives (/s/, /sh/, /z/, /zh/, /f/, /th/, /v/, /dh/); 7 nasals (/m/, /n/, /ng/, /em/, /en/, /eng/, /nx/); 7 semivowels and glides (/l/, /r/, /w/, /y/, /hh/, /hv/, /el/); 20 vowels (/iy/, /ih/, /eh/, /ey/, /ae/, /aa/, /aw/, /ay/, /ah/, /ao/, /oy/, /ow/, /uh/, /uw/, /ux/, /er/, /ax/, /ix/, /axr/, /ax-h/); and three other symbols: the pause, the epenthetic silence and the begin/end marker (non-speech events). This leads to a total of 61 different segmentation labels, from which the TIMIT voicing information is derived. The voiced – unvoiced segmentation problem is quite balanced. Indeed, around 56 % of the samples, over a few hundred millions, are labeled as voiced, and 44 % as unvoiced. Notice that silences and non-speech events are considered in this paper as unvoiced segments.

### B. Training stage for the constants $C_1$, $C_2$, $C_3$ and behaviour of the Otsu's method

Experimentally, it appears that the three "ad-hoc" thresholds have the same behaviour, as table 1 shows. These thresholds are almost equally robust. The Otsu's method provides systematically over estimated thresholding values. Modifying the original method, a multiplicative constant $C_\text{O}$ can be thus added, leading to the threshold $t_O^M = t_O C_O$. A training stage is necessary to determine the value of $C_\text{O}$. In table 1, obtained results with the modified Otsu's method are shown. They are very similar to the results obtained using the three "ad-hoc" thresholds.

|  | $C_1$ | DR (%) | $C_2$ | DR (%) |
|---|---|---|---|---|
| AS | 5.3933e-2 | 92.173 | 1.0311e-1 | 92.596 |
| E | 4.3775e-3 | 81.54 | 7.9306e-3 | 81.681 |
| ZCR | 6.8927e-1 | 89.14 | 7.2904e-1 | 89.061 |

|  | $C_3$ | DR (%) | $C_O$ | DR (%) |
|---|---|---|---|---|
| AS | 3.4162e-1 | 92.599 | 2.1949e-1 | 92.089 |
| E | 4.7992e-2 | 81.767 | 1.2366e-2 | 81.416 |
| ZCR | 9.9290e-1 | 89.066 | 1.1212e0 | 88.476 |

Table 1. Trained Constants for the 4 thresholds and Detection Rates (DR) obtained on the TIMIT database using the Analytic Signal (AS) compared to results obtained using the Energy (E) and the Zero Crossing Rate (ZCR); $T = 14\ ms$

As the best voiced – unvoiced segmentation performance are obtained for a 14 ms window length (see figure 2) for the AS-based and the Energy-based methods and as the performance of the ZCR-based method for this window length is close to the best performance obtained using this method, it has been decided to show in this section the training results for the threshold constants obtained considering this window length only. When the Analytic Signal method is used, a Detection Rate comprised between 92.089 % (with the modified Otsu's method) and 92.599 % (with the third threshold) is obtained. When the Energy-based method is used a Detection Rate comprised between 81.416 % (with the modified Otsu's method) and 81.767 % (with the third threshold) is obtained. When the ZCR-based method is used a Detection Rate comprised between 88.476 % (with the modified Otsu's method) and 89.066 % (with the third threshold) is obtained. The original Otsu's method provides Detection Rates of only 74.602 % (with the AS-based method) and 54.924 % (with the Energy-based method). With the ZCR-based method, the Detection Rate is less damaged, as it is still equal to 88.409 %.

In the rest of the paper, the third "ad-hoc" threshold has been retained, for the AS-based, the Energy-based and the ZCR-based methods.

### C. Effect of the window length on the performance

Figure 2 shows that the AS-based feature developed in this paper provides much better results than the Energy-based and the ZCR-based features, which are two commonly used features for the voiced – unvoiced detection problem. This figure shows as well that the method remains robust even for very short windows. Considering the AS-based method, the best result is obtained for a 14 ms frame length (a Detection Rate of 92.599 % is obtained). For a 42 ms frame length, the Detection Rate drops to 92.056 %; for a 2 ms frame length, the Detection Rate drops to 90.908 %, which is still acceptable. Considering the Energy-based method, the best result is obtained for a 14 ms frame length as well (a Detection Rate of 81.767 % is obtained). Considering a 42 ms window length a Detection Rate of 80.880 % is obtained; for a 2 ms frame length, the Detection Rate drops to 79.922 %. Considering the ZCR-based method, the best result is obtained for a 42 ms frame length (a Detection Rate of 89.456 % is obtained). Considering a 2 ms window length, the Detection Rate drops to 85.699 %; for a 14 ms frame length, a Detection Rate of 89.066 % is obtained.

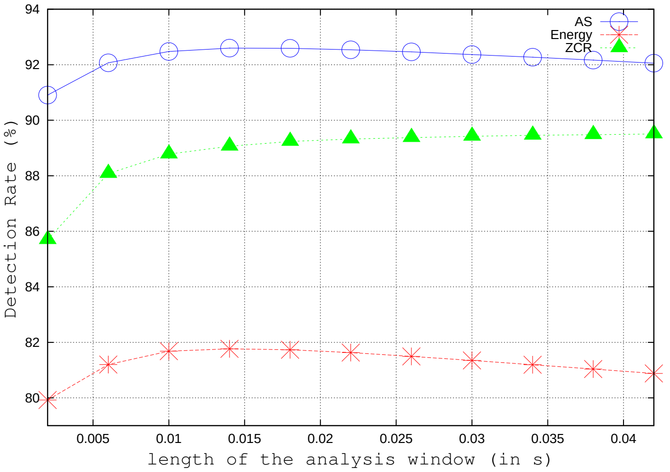


Fig. 2. Detection Rate obtained for different window lengths (AS method: circles+solid line; Energy method: stars+dashdot line; 1-ZCR: triangles+dashed line); the third "ad-hoc" threshold is used

### D. Timing aspects – precision in the position of the found segmentation marks

In this section, the precision in position of the found ruptures between the voiced (respectively unvoiced) segments and unvoiced (respectively voiced) segments, compared to the positions provided in TIMIT, is studied. A segmentation mark is positioned on the sample $(i)$ if for this sample it was decided in the previous stages of the analysis that the signal is voiced (respectively unvoiced) and if for the previous sample $(i-1)$ it was decided that the signal is unvoiced (respectively voiced).

Notice that, in natural speech, successions of several voiced phones or of several unvoiced phones are possible. The method described in this paper has not been designed to determine the boundaries between successive voiced phones or between successive unvoiced phones. Therefore, it is not possible to allocate to every TIMIT segmentation mark one of the found segmentation marks.

Instead, one of the TIMIT segmentation marks is allocated to each of the found segmentation marks. The closest TIMIT segmentation mark is selected. The distance $d$ in ms between a found segmentation mark and its corresponding TIMIT segmentation mark characterizes the precision in position of the found segmentation marks. However, false alarms could occur. That is to say: a transition is detected within a voiced or an unvoiced segment. In that case, the distance $d$ is too big and the allocation has to be rejected. It is assumed in this paper that the speech rate in terms of phones per second in English is around 12 (see [13]), leading to a duration of around 80 ms for each phone. It is thus decided in this paper that if

$d$ is bigger than 30 % of this duration, the allocation must be rejected.

Figure 3 shows that, for longer windows, as expected, $d$ increases as the window length increases. However, for shorter windows, $d$ increases as the window length decreases. This is due to the fact that for shorter windows, the analysis becomes less stable, giving rapid successions of segmentation marks, that is to say very short voiced (respectively unvoiced) followed by very short unvoiced (respectively voiced) segments. Several of the found segmentation marks have thus the same TIMIT segmentation mark as allocated mark (this is shown in details in section III-E), leading to a blurring effect. The best performance are obtained, considering respectively the AS-based, the Energy-based and the ZCR-based methods, for a window length equal to 18 ms, 14 ms and 26 ms. Notice that, considering shorter windows, the AS-based method provides more precisely positioned segmentation marks than both the Energy-based and the ZCR-based methods.

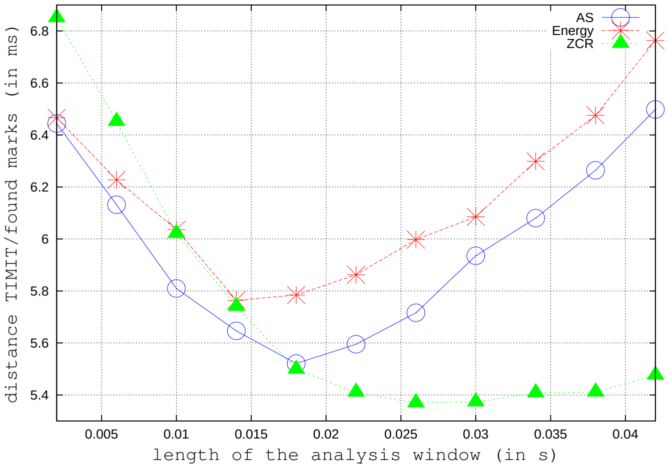


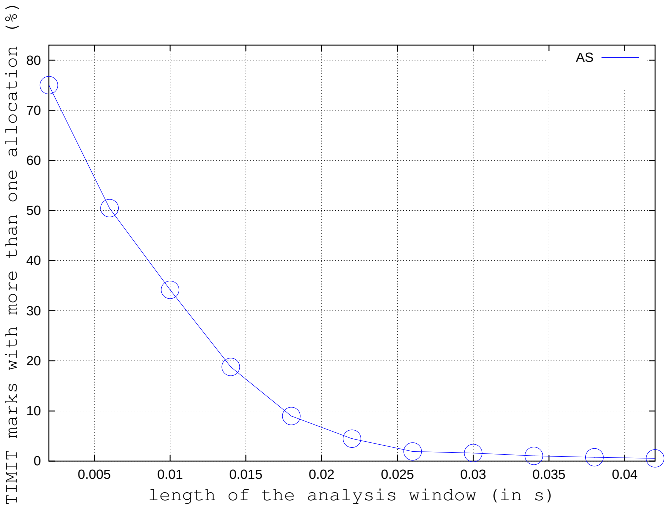
Fig. 3. Precision in the time position (in ms) of the found segmentation marks, for different window lengths (AS method: circles+solid line; Energy method: stars+dashdot line; 1-ZCR: triangles+dashed line)

### E. *Timing aspects – number of rapid successions of segmentation marks*

In order to get more insight concerning the performance of the method, two additional measures have been extracted:

- The ratio of voiced and unvoiced segments shorter than $S$ ms, $S$ being small, relative to the total number of found segments, is computed. $S$ has been chosen equal to 5 ms. The results are shown in figure 4. As expected, for shorter window lengths, much more short segments are obtained. The AS-based, Energy-based and ZCR-based methods provide similar results, the AS-based method being slightly better.
- The mean number of found segmentation marks having the same TIMIT segmentation mark as allocated mark is computed. The results are shown in figure 5. For shorter window lengths, more possible allocations are accepted. The AS-based method shows a slightly more stable behaviour than the Energy-based and the ZCR-based methods.

Results show that some post processing stage of the voiced – unvoiced measure is definitely requested. Techniques to reduce the number of short segments need to be implemented. It could be envisioned for instance to use more than only one measure, that is to say to fusion the results provided by the AS-based method described in this paper and other robust methods. The goal of this fusion stage would be to overcome the problems mentioned in the previous section and in this one, this in order to improve the precision in the time location of the found segmentation marks when using very short window lengths.



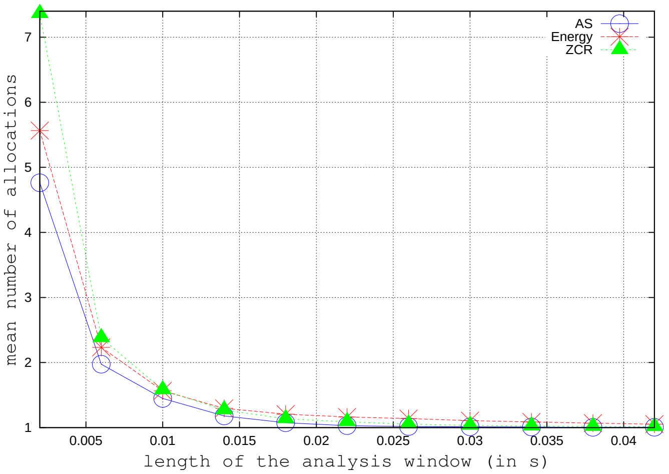


Fig. 5. Mean number of found marks having the same TIMIT mark as allocated mark, for different window lengths (AS: circles+solid line; Energy: stars+dashdot line; 1-ZCR: triangles+dashed line)

Rate dramatically drops to 79.603 % for a SNR equal to 9 dB and 74.023 % for a SNR equal to 5 dB.

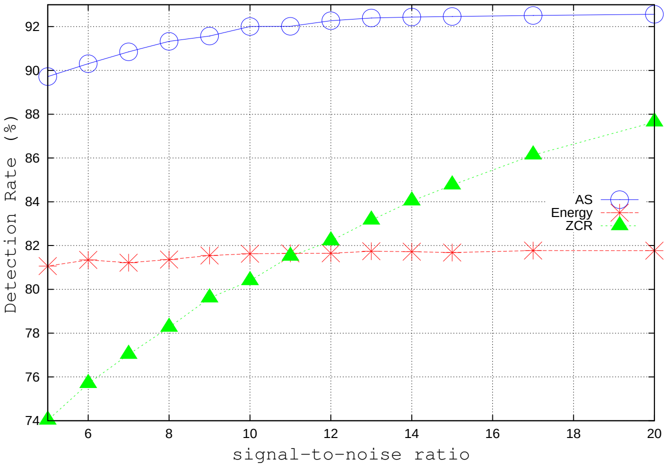


Fig. 6. Detection rate, for different SNR; $T = 14\ ms$ (AS: circles+solid line; Energy: stars+dashdot line; 1-ZCR: triangles+dashed line)

Furthermore, it can be noticed in figure 7 that considering the AS-based method, the $C_3$ value obtained after the training stage does not depend so much on the SNR. On the contrary, the $C_3$ value obtained after the training stage considering the Energy-based method is more than tripled when the SNR drops to 9 dB, and more than fivefold increased when the SNR drops to 5 dB. This indicates that the AS-based method does not absolutely require to get an estimate of the SNR, and that it is definitely not the case for the Energy-based method. Figure 8 shown this. $C_3$ is taken constant and equal to its value when there is no additive noise. It can be seen that, considering the Energy-based method, the Detection Rate dramatically drops from 79.266 % to 55.208 % when the SNR decreases only from 13 dB to 10 dB, and that it is much less the case considering the AS-based method.

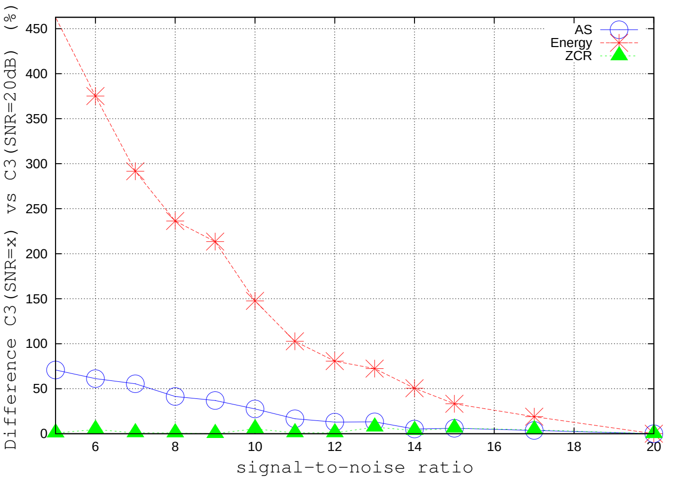


Fig. 7. Difference in % of the value of the trained $C_3$ constant for different SNR, when the value obtained for a SNR equal to 20 dB is taken as reference; $T = 14\ ms$ (AS: circles+solid line; Energy: stars+dashdot line; 1-ZCR: triangles+dashed line)

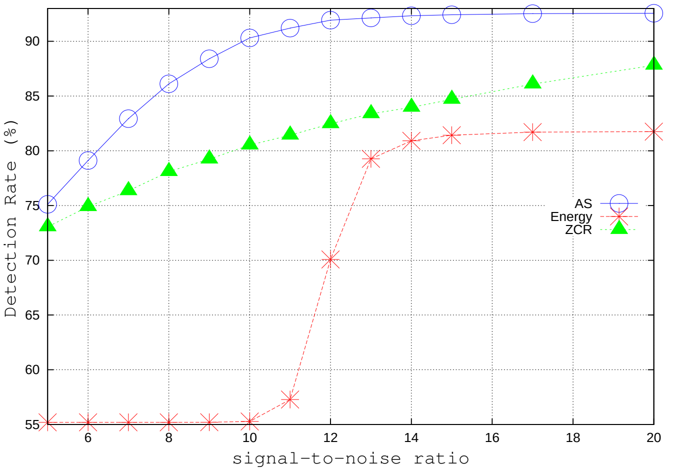


Fig. 8. Detection rate, for different SNR; $T = 14\ ms$; $C_3$ not adjusted to the SNR (AS: circles+solid line; Energy: stars+dashdot line; 1-ZCR: triangles+dashed line)

The $C_3$ value, considering the ZCR-based method, does not depend at all on the SNR. This reflects the fact that the method is less robust to noise than the other two methods. However, if $C_3$ is taken constant and equal to its value when there is no additive noise, the ZCR-based method shows a quite robust behaviour, and becomes almost competitive with the AS-based method when the SNR is very low.

Similar results are obtained when a longer window length, that is to say 42 ms, is chosen. This is shown in figures 9 and 10. The three methods show slightly better behaviours.

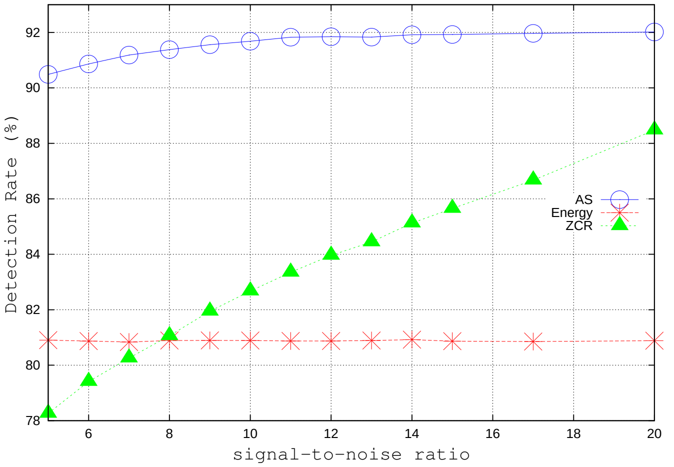


Fig. 9. Detection rate, for different SNR; $T = 42\ ms$ (AS: circles+solid line; Energy: stars+dashdot line; 1-ZCR: triangles+dashed line)

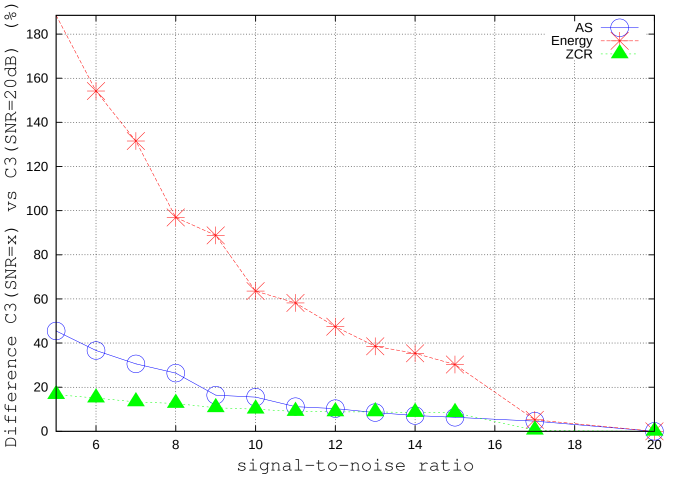


Fig. 10. Difference in % of the value of the trained $C_3$ constant for different SNR, when the value obtained for a SNR equal to 20 dB is taken as reference; $T = 42\ ms$ (AS: circles+solid line; Energy: stars+dashdot line; 1-ZCR: triangles+dashed line)

## IV. Conclusion and Future Works

In this paper, a method for automatically obtain the voiced – unvoiced segmentation of monophonic speech signals is presented. This method is based on the Analytic Signal. In terms of voiced – unvoiced segmentation, it is shown that the method described in this paper provides reliable detection results. Similarly, a great precision in terms of time location of the segmentation marks is obtained. Using 14 ms duration frames provides the most reliable voicing information: a Detection Rate of 92.599 % is obtained. Concerning the time location of the segmentation marks, the best results are obtained for a 18 ms window length. Notice that in that case, our voicing measure remains reliable, as the Detection Rate is still equal to 92.593 %. Furthermore, the AS-based method is much more robust to noise than the Energy-based and ZCR-based methods.

Improvements of the technique are envisioned, concerning mainly the post-processing of the shortest segments. This will allow to get the full potential of the method concerning the use of very short window lengths in order to get more precisely positioned in time segmentation marks. The periodic structure of the voiced signals is not taken into account at the moment. The method could provide an estimate of the pitch $f_0$, and thus could be adapted in order to take advantage of this. Its performance would be probably even better. Using the precise voiced – unvoiced information obtained with the method described in this paper, for instance in existing speech recognition systems and speech alignment systems, could improve their performance and could improve their speed.

## References

[1] D. Meen, T. Svendsen, and J. E. Natvig, “Improving Phone Label Alignment Accuracy by Utilizing Voicing Information,” in *Proc. ICSLP Speccom*, 2005.

[2] A. Zolnay, R. Schlüter, and H. Ney, “Robust Speech Recognition Using a Voiced-Unvoiced Feature,” in *ICSLP*, 2002.

[3] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, “The HTK Book (for HTK version 3.2),” Tech. Rep., 2002.

[4] B. S. Atal and L. R. Rabiner, “A Pattern Recognition Approach to Voiced – Unvoiced – Silence Classification with Applications to Speech Recognition,” in *IEEE Transaction on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, 1976, pp. 201–212.

[5] M. Greenwood and A. Kinghorn, “SUVing: Automatic Silence/Voiced/Unvoiced Classification of Speech,” in *Undergraduate Coursework – Department of Computer Science, The University of Sheffield, UK*, 1999.

[6] S. Rossignol, X. Rodet, J. Soumagne, J.-L. Collette, and P. Depalle, “Feature Extraction and Temporal Segmentation of Acoustic Signals,” in *Proceedings of the ICMC*, 1998.

[7] A. P. Lobo and P. C. Loizou, “Voiced/Unvoiced Speech Discrimination in Noise Using Gabor Atomic Decomposition,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003, pp. 820–823.

[8] A. M. C. R. Borzino, D. G. da Silva, and J. A. Apolinário Jr., “Voiced/unvoiced classification for short time frames and its application to frequency-domain cryptanalysis,” in *International Workshop on Telecommunications (IWT)*, 2007.

[9] W. Hess, *Pitch Determination of Speech Signals*. Springer-Verlag, 1983.

[10] P. K. Sahoo, S. Soltani, and K. C. Wong, “A Survey of Thresholding Techniques,” in *Computer Vision, Graphics, and Image Processing*, vol. 41, 1988, pp. 233 – 260.

[11] S. Rossignol, “Segmentation et indexation de signaux sonores musicaux,” Ph.D. dissertation, University of Paris VI – Jussieu, July 2000.

[12] Linguistic Data Consortium, “The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT),” NIST Speech CD 1-1.1, 1990.

[13] J.-L. Rouas, J. Farinas, and F. Pellegrino, “Évaluation automatique du débit de la parole sur des données multilingues spontanées,” in *Journées d’Étude sur la Parole (JEP)*, 2004.