Correspondence

Voice Activity Detection in Nonstationary Noise

S. Gökhun Tanyer and Hamza Özer

***Abstract*—A new fusion method for voice activity detection in additive nonstationary noise is suggested. A performance study of the methods: fusion, the recently developed geometrically adaptive energy level, periodicity measure, and zero crossings rates, is presented. The new method is shown to operate reliably down to −5 dB SNR.**

***Index Terms*—Nonspectral analysis, speech analysis, statistical methods, voice activity detection.**

I. Introduction

The process of separating conversational speech and silence is called the *voice activity detection* (VAD). VAD was first investigated for use on TASI systems [1]. VAD is required in some speech communication applications such as speech recognition, speech coding, hands-free telephony and echo cancellation [2]. For these purposes, various types of VAD algorithms that trade off delay, sensitivity, accuracy and computational cost have been proposed. The earlier algorithms are based on the Itakura LPC distance measure [3], energy levels, timing, pitch, and zero crossing rates [4], cepstral features [5], adaptive noise modeling of voice signals [6] and the periodicity measure [7]. Some of those algorithms are implemented on specific applications like the Pan-European digital cellular mobile telephone service [8], cellular networks [9], and the digital cordless telephone systems [10].

Unfortunately, these algorithms have some problems in low SNR values, especially when the noise is nonstationary. Consistent accuracy cannot be achieved since most algorithms rely on a threshold level for comparison. This threshold level is often assumed to be fixed [11] or calculated in the silence (voice-inactive) intervals. For example, in the autoregressive analysis with the LMS algorithm, silence intervals are required to train the FIR filters [5]. Similarly, third order statistics-based VAD initially requires noise-only frames [10]. On the other hand, Yoma [6] and Haigh [5] assumed the noise to be reasonably stationary and correlated. Tucker investigated a VAD algorithm based on periodicity [7] which can operate in SNR's down to 0 dB. For some applications the speech quality can be extremely poor, and reliable operation below 0 dB SNR is required, especially when the environment is nonstationary.

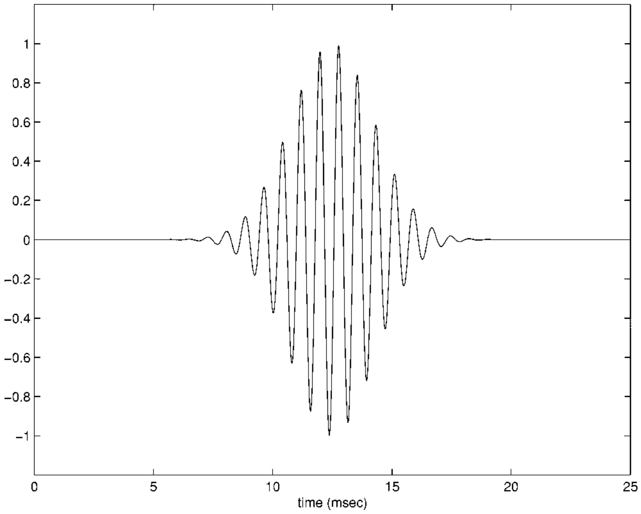

Fig. 1. Typical narrowband periodical signal.

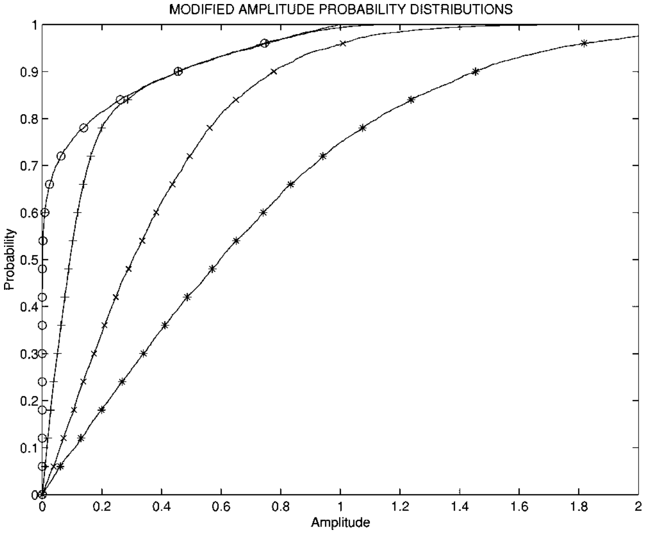

Fig. 2. The modified amplitude probability distributions for the signal given in Fig. 1 corrupted by Gaussian noise, (o) $\mathrm{SNR}_{\max} = 9$ dB, (+) $\mathrm{SNR}_{\max} = 8$ dB, (x) $\mathrm{SNR}_{\max} = 4$ dB, and (*) $\mathrm{SNR}_{\max} = -2$ dB.

II. Voice Activity Detection

The basic function of a VAD algorithm is to extract some measured features or quantities from the input signal and to compare these values with thresholds, usually extracted from the characteristics of the noise and speech signals. Then, voice-active decision is made if the measured values exceed the thresholds. VAD in nonstationary noise requires a time-varying threshold value. This value is usually calculated in the voice-inactive segments. On the other hand, for signals dominated by voice-active segments, noise can vary considerably before the next noise level re-calibration instant. We recently developed a new technique to estimate the optimum threshold for noise in the presence of speech accurately by using the amplitude probability distributions [12].

A. Classical VAD Methods

1) Energy Threshold Method: The energy threshold method is one of the earlier VAD algorithms where the energy of the signal is monitored and compared with the threshold value. The energy of the total signal in the presence of speech is assumed to be sufficiently larger than that of the background noise, and therefore the voice-active regions could be detected. The preset threshold value for a varying noise level is re-calculated for each analysis window.

Manuscript received October 6, 1998; revised August 26, 1999. This work was supported by the National Research Institute of Electronics and Cryptology, TÜBITAK-UEKAE. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Douglas O'Shaughnessy.

The author is with TÜBITAK-UEKAE, Kavaklidere, 06100 Ankara, Turkey, on leave from the Department of Electrical and Electronics Engineering, Baskent University, 06530 Ankara, Turkey (e-mail: gokhun@tubitak.gov.tr).

Publisher Item Identifier S 1063-6676(00)05180-4.

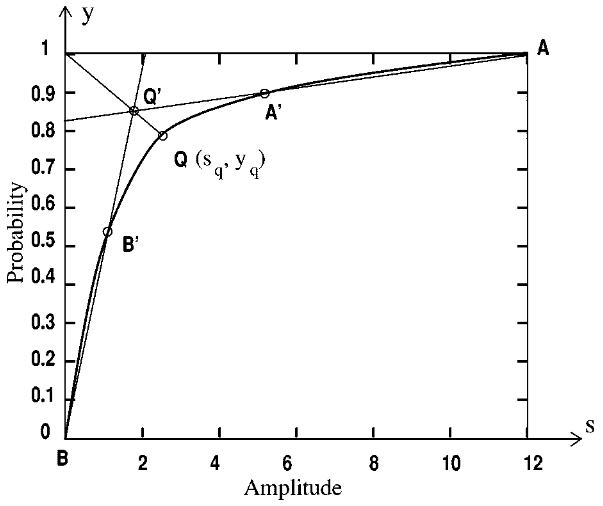

Fig. 3. Amplitude probability distribution graph and the geometrical technique to calculate the noise level.

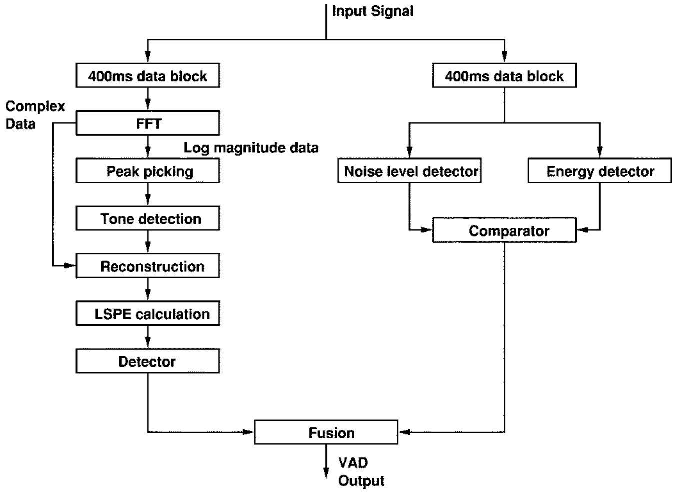

Fig. 4. Voice activity detection using the fusion of the LSPE (left side) and the geometrically adaptive energy threshold (right side) method.

2) Zero Crossings Rate Method: In this method, the zero crossings rate for each analysis window is calculated and compared with the preset threshold value [4]. The zero crossings rate of noise is assumed to be considerably larger than that of the speech signal. This assumption is accurate at high SNR values. However, it has problems at low SNR's, especially in the presence of periodic noise and speech with high zero crossing rates.

3) Least-Square Periodicity Estimator (LSPE) Method: Irwin investigated the optimum method for measuring the periodicity of speech corrupted by noise [13]. Tucker [7] designed a VAD based on periodicity. The major difficulty in designing a VAD based on periodicity is its sensitivity to any periodic signal which may well be interference or a background signal. Great care must be taken to avoid false triggering on nonspeech periodic signals. If the speech signal contains nonperiodic components, inaccurate values for endpoints of the voice-active segments can be obtained. Tucker used a preprocessor to detect and if possible remove, most of the expected types of interference. Different environments will have different interference, so the exact nature of the preprocessor will depend on the expected type of interference.

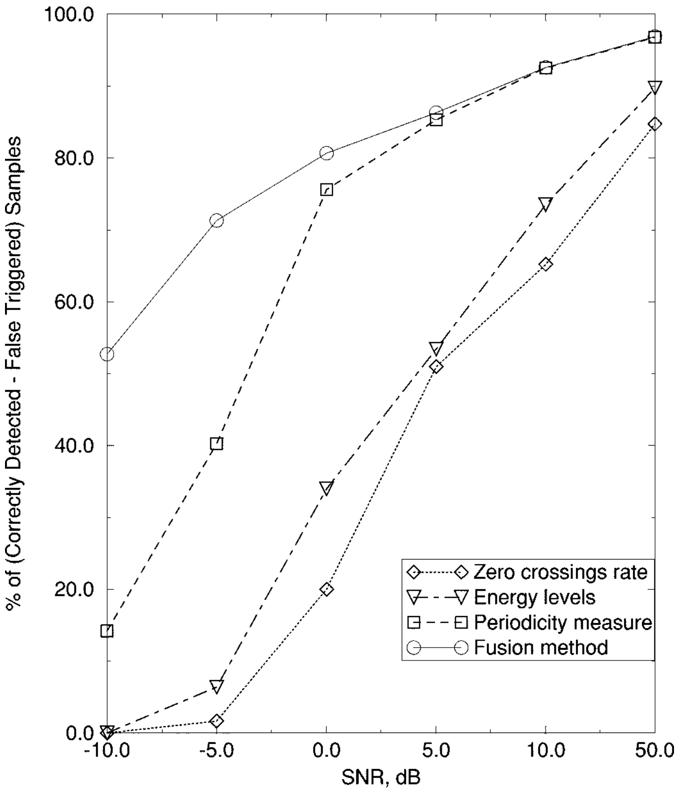

Fig. 5. The percentages of the (correctly detected-false triggered) samples corresponding to different algorithms as a function of the SNR.

B. Recently Developed Geometrically Adaptive Energy Threshold Method

In the classical energy threshold method, the threshold value is re-calculated at each voice-inactive segment. When the noise ground is nonstationary, the algorithm often 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 geometrically adaptive energy threshold (GAET) method [12] is developed to set the threshold level adaptively without the need of voice-inactive segments by using the amplitude probability distributions of the speech signal.

1) Amplitude Probability Distributions: The amplitude probability distribution (APD) is one of the most useful tools for statistical analysis of noise in communications. Theoretical and experimental studies of APD's of noise have been performed for the past 30 years or more [14]. In 1971, Spaulding [15] studied the overall environmental noise as well as noise due to specific sources. Later, in 1974 he performed noise measurements in parks and university campuses where there are less vehicular traffic, electrical equipments and other types of noise source. Graphical algorithms are developed to predict the APD's as a function of frequency, time and bandwidth of a receiver [16].

To briefly summarize the basics, let us assume that the total signal can be written in the form

$$s(t) = c(t) + n(t) \tag{1}$$

where $c(t)$ and $n(t)$ are the clear speech and noise signals, respectively. The amplitude probability distribution (APD) function $F_s(s)$ and the

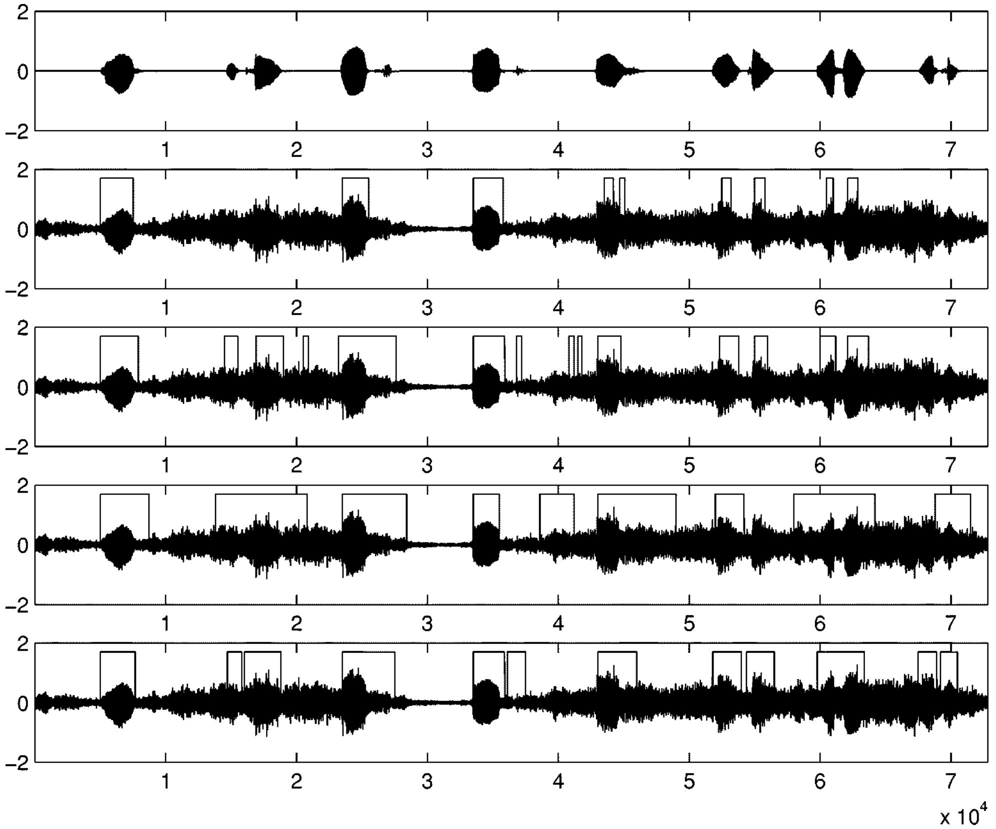

Fig. 6. The clear speech signal (words from 1 to 8 uttered in Turkish: *bir, i-ki, üç, dört, bes, al-tı, ye-di,* and *se-kiz* versus the sample number and the detected voice-active regions, from top to bottom (a) the clear speech signal, (b) the zero crossings rates method, (c) the geometrically adaptive energy threshold (GAET) method, (d) the least-Square periodicity estimator (LSPE) method, and (e) the fusion of (c) and (d), ($-5 \le \mathrm{SNR} \le 10$ dB).

TABLE I
AVERAGE* PERCENTAGES OF THE CORRECTLY DETECTED-FALSE TRIGGERRED) SAMPLES. (THE AVERAGE IS CALCULATED THROUGHOUT THE SIGNAL USED IN Fig. 6).

METHOD	**Average Percent**
Zero Crossings Rate	34.2
Geometrically Adaptive Energy Threshold	72.2
LSPE	71.4
Fusion	96.9

amplitude probability density (apd) function $f_s(s)$ of a continuous time random variable $s(t)$ are related by

$$F_s(s) = \int_{-\infty}^{s} f_s(\xi)\,d\xi = \int_{0}^{s} f_s(\xi)\,d\xi. \qquad (2)$$

Let us denote the elements of the stochastic process on discrete time signal by $s[k]$ which are the $(N+1)$ samples of $s(t)$ at $t = k\Delta t$ in the analysis window $(T_1 \le t \le T_2)$, i.e.,

$$s[k] = s(T_1 + k\Delta t) \qquad (3)$$

for $k = 0, 1, \ldots, N$ and where

$$\Delta t = \frac{(T_2 - T_1)}{N}. \qquad (4)$$

Note that the sampling rate does not need to obey the Nyquist criteria and N is assumed to be sufficiently large. The APD, $F_s[m]$, and apd, $f_s[m]$ of discrete time random variable $s[k]$ can be defined as

$$F_s[m] = \sum_{i=0}^{m} f_s[i] \qquad (5)$$

and where $f_s[i]$ is the number of samples of $s[k]$ satisfying

$$i\Delta s \le |s[k]| < (i+1)\Delta s \qquad (6)$$

normalized by the total number of samples N, and where Δs is the resolution parameter. Note that for $\Delta t \to 0$, $\Delta s \to 0$, $(T_1 \to -\infty)$, $(T_2 \to \infty)$ and $(N \to \infty)$, $F_s[m]$ and $f_s[m]$ converge to $F_s(s)$ and $f_s(s)$, respectively. If the APD of the signal and noise, $F_s(s)$ and $F_n(n)$, are different then, $F_s[m]$ and $F_n[m]$ are expected to be different. For a corrupted signal, the signal and noise will partially occupy different regions on the APD.

2) Modified Amplitude Probability Distribution (MAPD) Function: In this work, the modified amplitude distribution (MAPD) function $R_s[m]$ is defined. It can implicitly be obtained by setting the x and y-axis by $y = k/N$ and $x = \mathrm{sort}(s[k])$ respectively where sorting is done in ascending order, and any sorting algorithm can be used [12]. A narrowband signal shown in Fig. 1 is corrupted by Gaussian noise, and the corresponding MAPD functions for various SNR's are given in Fig. 2. It can be shown that $R_s[m]$ is equivalent to $F_s[m]$ and is numerically more accurate (more resemblance to $F_s(t)$). It requires comparatively few data points (samples). From now on,

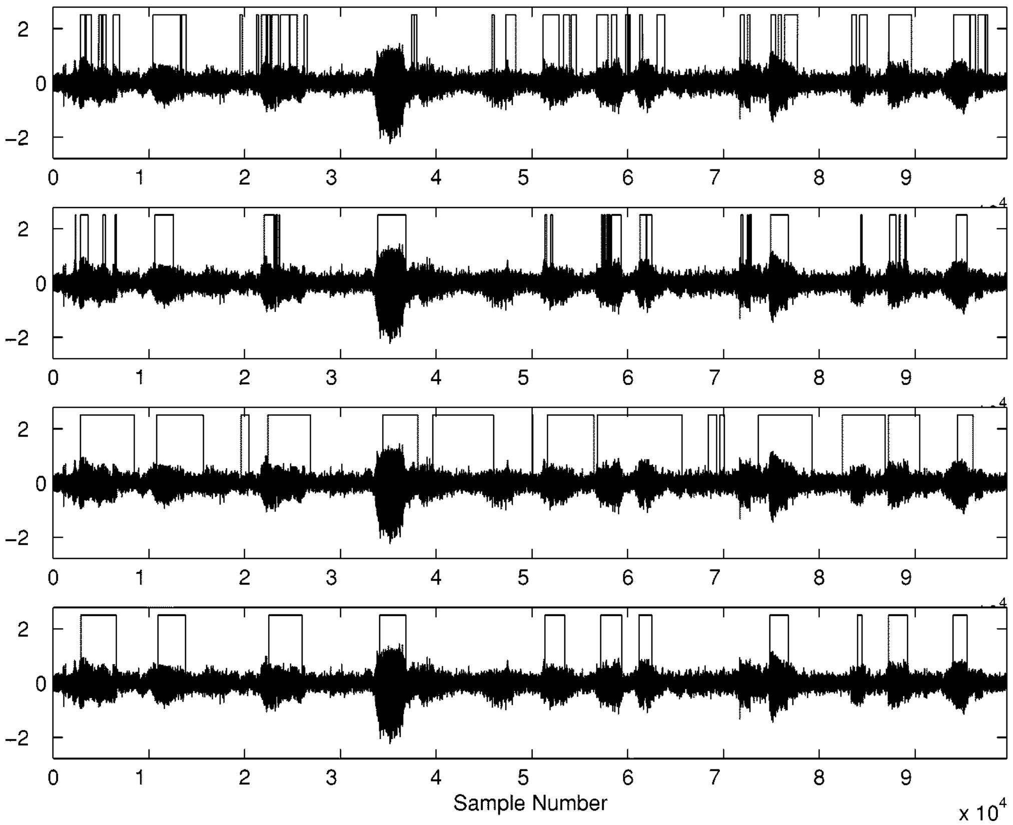

Fig. 7. Voice activity detection results for the corrupted speech signal where words from 1 to 9 uttered in Turkish: *bir, i-ki, üç, dört, bes, al-tı, ye-di, se-kiz*, and *do-kuz*, from top to bottom (a) the zero crossings rates method, (b) the geometrically adaptive energy threshold (GAET) method, (c) the least-Square periodicity estimator (LSPE) method, and (d) the fusion of (b) and (c).

$R_s[m]$ will be used in place of $F_s[m]$. $R_c[m]$, $F_c[m]$, $f_c[m]$ and $R_n[m]$, $F_n[m]$, $f_n[m]$ can similarly be defined using the samples $c[k]$ and $n[k]$, respectively.

3) Geometrical Technique to Calculate the Noise Level: It is observed on the $R_s[m]$ plot that the samples $s[k]$ and $n[k]$ are partially separated, and the amplitude of the zero mean Gaussian noise samples $n[k]$ locate closer to the origin whereas, the clear signal samples $c[k]$ dominate the higher values as illustrated in Fig. 2. It is interesting to note that the noise level approximately corresponds to the bending point, and that it shifts to a higher amplitude level as the increase in noise decreases the SNR from 9 dB to 8 dB. However, this is not very obvious for $\mathrm{SNR} = 4$ and -2 dB. A geometrical technique can heuristically be used to find the bending point on the MAPD graph which represents the noise level (see Fig. 3). The point Q' can be found by intersecting the two tangent lines passing through the points A-A' and B-B' respectively. Then, a third line passing through the top left corner and the point Q' crosses the MAPD graph at the "optimum" point, $Q(s_q, y_q)$. For robustness to various window sizes and different applications, the intersection point Q' is calculated using three different tangent lines, choosing $y = 1/5, 1/10$ and $1/20$ for point B' and $y = 4/5, 9/10$ and $19/20$ for point A' respectively where these values are observed to be not very critical. The average value for the point $Q(s_q, y_q)$ is used to calculate s_q. The noise level s_q can further be multiplied by a safety coefficient α $(0.8 < \alpha < 1.2)$ which is a constant throughout the detection process. The SNR estimation error is calculated for Gaussian, water, public restaurant and the traffic noise, and is observed to be less than 3 dB for -15 dB $< \mathrm{SNR} < 20$ dB [17].

III. Proposed Fusion Method

The energy threshold method has problems in nonstationary noise and low SNR, and the zero crossings rate method provides some improvement. The LSPE method is accurate down to 0 dB SNR if false triggerings could be avoided. The GAET method is robust to nonstationary noise but false triggering is often observed when noise has short bursts. Above all the other combinations, the fusion of the GAET and the LSPE methods are observed to yield more accuracy and reliability (see Fig. 4). The GAET method keeps track of the nonstationary background noise while the LSPE analyzes the periodical content of the incoming signal.

In the GAET branch of the fusion method, the speech signal is processed in 1 200 ms analysis blocks, and the noise level is recalculated for each block. Each block is divided into 32 ms long 75 frames, overlapping 16 ms. Voice-active decision for a frame is made if speech signal is above the threshold more than 50% of the time, and the branch returns 1; otherwise, it returns 0.

In the LSPE branch, the size of the analysis block and frame lengths are the same as in the GAET branch. The algorithm returns 1 for a frame if peaks are present at least 20% of its FFT frame, otherwise it returns 0 as given in [7]. Hanning windowing is used for both branches.

In the fusion algorithm, the digital output for both branches are monitored. The weighted sum of the two outputs using the weights α and $\beta = (1 - \alpha)$, is compared to 0.5 for the overall voice-active decision. One algorithm can be dominated by the other with proper adjustment of the weights. It is interesting to note that the false triggering rate increases when the weight for the LSPE branch β is increased, whereas, the rate of miss increases when α is increased. This suggests that the miss/false triggering rate can be controlled by proper setting of those

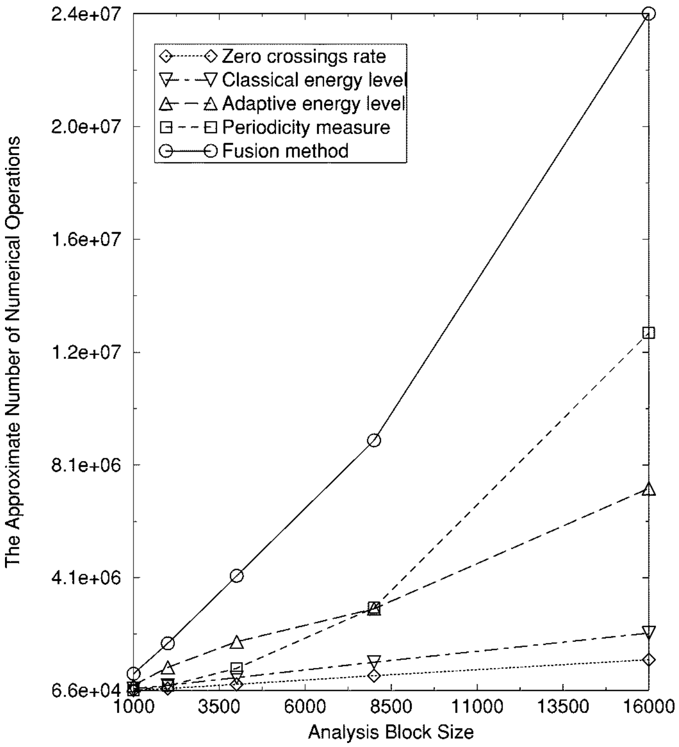

Fig. 8. Approximate number of numerical operations required by the algorithms as a function of the analysis block size.

weights. The length of the analysis block for the GAET can be decreased for highly nonstationary backgrounds, but this would raise the threshold artificially when SNR and word length are both above 20 dB and 1200 ms, respectively.

IV. Results

As an objective measure to evaluate the algorithms, the percentages of the (correctly detected-false triggered) samples of speech at different SNR's are shown in Fig. 5 where the SNR is measured over a set of eight words (words from 1 to 8 uttered in Turkish: bir, i-ki, üç, dört, bes, al-tı, ye-di, se-kiz), and noise is assumed to be additive and stationary. Later, the same speech signal is corrupted by the nonstationary noise as shown in Fig. 6. Maximum rate of change in noise level, around 0 dB SNR, measures approximately 5 dB/s. It is observed that the zero crossings rate method misses the two syllable words "'two" and "eight" in low SNR whereas the geometrically adaptive energy threshold method misses only "eight". The LSPE method clearly has the problem of false triggering in high noise. The fusion method is observed to detect all the words accurately, and is shown to detect the endpoints of the voice active regions accuately even in the presence of nonstationary noise. The short jitters can be minimized by postprocessing. The average percentages of the (correctly captured-false triggered) samples for the example given in Fig. 6 are also calculated (Table I). The algorithm is also tested on 25 recordings of women, men and children recorded in various noise backgrounds: traffic, restaurant, school corridor and water fountain noise as opposed to clean speech with the post-hoc addition of noise as in Fig. 6. A typical result using the speech of a woman in traffic is given in Fig. 7. It is observed that the proposed method is accurate except for the first syllable of 7 and misses the second syllable of 9 where the SNR is estimated to be below -10 for those segments. A listening comprehension test is done among 20 Turkish speaking engineering students using the detected segments of the words. The results show that those students recognized more than 80% of the words correctly. Finally, the computational complexity for each algorithm and for various block sizes is calculated and compared in Fig. 8.

V. Conclusion

The commonly used VAD algorithms, including the recently introduced geometrically adaptive energy threshold (GAET) method, are shown to have problems in the presence of nonstationary noise, specially below 0 dB SNR. The estimate for noise level is shown to be accurate even when the analysis windows do not fully contain voice-inactive signals in the presence of nonstationary noise. This new geometrical algorithm is observed to improve the energy threshold method approximately 5 dB. The fusion of the GAET and the LSPE algorithms is observed to operate reliably down to -5 dB, most speech can be detected above -10 dB and tested to be accurate for both men and women on various noise backgrounds. The computation complexity of the GAET and the LSPE algorithms are observed to be approximately equal, and the new fusion method's complexity to be approximately equal to the total of both algorithms, which is acceptable for most applications.

References

[1] K. Bullington and J. M. Fraser, "Engineering aspects of TASI," *Bell Syst. Tech. J.*, pp. 353–364, Mar. 1959.

[2] O. Tanrikulu, B. Baykal, A. G. Constantinides, and J. A. Chambers, "Critically sampled sub-band acoustic echo cancellers based on IIR and FIR filter banks," *IEEE Trans. Signal Proccessing*, vol. 45, pp. 901–912, Apr. 1997.

[3] L. R. Rabiner and M. R. Sambur, "Voiced-unvoiced-silence detection using the Itakura LPC distance measure," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, May 1977, pp. 323–326.

[4] J. C. Junqua, B. Reaves, and B. Mak, "A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognize," in *Proc. Eurospeech'91*, 1991, pp. 1371–1374.

[5] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstral features," in *Proc. IEEE TENCON*, China, 1993, pp. 321–324.

[6] N. B. Yoma, F. McInnes, and M. Jack, "Robust speech pulse-detection using adaptive noise modeling," *Electron. Lett.*, vol. 32, July 1996.

[7] R. Tucker, "Voice activity detection using a periodicity measure," *Proc. Inst. Elect. Eng.*, vol. 139, pp. 377–380, Aug. 1992.

[8] 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 *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Glasgow, U.K., May 1989, pp. 369–372.

[9] K. Srinivasan and A. Gersho, "Voice activity detection for cellular networks," in *Proc. IEEE Speech Coding Workshop*, Oct. 1993, pp. 85–86.

[10] S. Sasaki and I. Matsumoto, "Voice activity detection and transmission error control for digital cordless telephone system," *IEICE Trans. Commun.*, vol. E77B, no. 7, pp. 948–955, 1994.

[11] D. R. Halverson, "Robust estimation and signal detection with dependent nonstationary data," *Circuits Syst. Signal Process.*, vol. 14, no. 4, pp. 465–472, 1995.

[12] H. Özer and S. G. Tanyer, "A geometric algorithm for voice activity detection in nonstationary Gaussian noise," in *Proc. EUSIPCO'98*, Rhodes, Greece, Sept. 1998.

[13] M. J. Irwin, "Periodicity estimation in the presence of noise," in *Proc. Acoust. Conf.'79*, Windemere, U.K., 1980.

[14] F. Horner, *Advance in Radio Research*, New York: Academic, 1964, vol. 2, pp. 121–204.

[15] A. D. Spaulding, W. H. Ahlbeck, and L. R. Espeland, "Urban residential man-made radio analysis and predictions," U.S. Govt. Printing Office, Washington, DC, OT Telecommun. Res. Eng. Rep., 1971.

[16] T. Nakai and H. Ohba, "On the graphical method of drawing APD's for atmospheric radio noise," *IEEE Trans. Electromagn. Compat.*, vol. EMC-26, pp. 71–78, May 1984.

[17] H. Özer, "Signal detection and estimation in nonstationary background," M.S. thesis, Dept. Elect. Electron. Eng., Başkent Univ., Ankara, Turkey, Aug. 1998.