A GEOMETRIC ALGORITHM FOR VOICE ACTIVITY DETECTION IN NONSTATIONARY GAUSSIAN NOISE

Hamza Özer and S. Gökhun Tanyer
Dept. of Electrical and Electronics Engineering, Başkent University,
Bağlıca Kampusu, Eskişehir Yolu 20.km,
Ankara 06530, TURKEY
Tel: +90 312 2341010/1236; fax: +90 312 2341051
e-mail: hamza@baskent.edu.tr

ABSTRACT

A new algorithm for voice activity detection in additive nonstationary noise is presented. The algorithm utilizes the differences of the probability distribution properties of noise and speech signal. The Magnitude Density (mdf) and the Magnitude Distribution Functions (MDF) are defined. The noise level is monitored for automatic threshold estimation. The estimate is shown to be accurate also when analysis windows do not fully contain non-speech signals and in the presence of nonstationary noise. The algorithm has been applied different type of noises (traffic, water, restaurant, ect.). The voice activity detection algorithm is shown to operate reliably in SNRs down to 0 dB and noise variance up to 10 dB/sec.

1 INTRODUCTION

The process of detecting speech in noisy acoustical environments is called the voice activity detection (VAD).Some form of VAD is required in automatic speech recognition systems. Some of those systems include; GSM-based wireless systems, multiple access schemes, such as CDMA and enhanced TDMA for wireless cellular and Personal Communications Systems and speech communication systems; speech coding, speech recognition, hands-free telephony, audio conferencing and echo cancellation. The inaccurate detection of the endpoints of speech is one of the limitations in these systems. Various VAD algorithms are developed to challenge this problem. The earlier algorithms are based on the Itakura LPC distance measure [1], on energy levels, timing, pitch and zero crossing rates [2], and periodicity measure[3]. Later, Haigh [4] developed an algorithm using cepstral features, and Yoma, McInnes and Jack used adaptive noise modeling where they assumed the noise to be reasonably stationary and correlated [5]. In parallel, those algorithms are tested on specific applications like the Pan-European digital cellular mobile telephone service [6], cellular networks [7], digital cordless telephone systems [8], and structured noise environments [9]. Most recently, El-Maleh and Kabal compared various detection algorithms for wireless personal communications systems [10].

Unfortunately, none of the present speech detection algorithms are perfect and have problems in low SNRs. Most require a noise threshold level for comparison. This threshold level is assumed to be fixed [11] or calculated in the non-speech intervals. For example, in the autoregressive analysis with the LMS algorithm, non-speech intervals are required to train the FIR filters used [5]. Similarly, third order statistics based VAD initially requires noise-only frames [10]. For problems where the signal does not have enough length of non-speech intervals and where the noise is nonstationary, the optimum threshold value should be monitored to achieve reliable detection in lower SNR levels.

In Section 2, the new technique which is developed to calculate the time-varying noise level and to guide the fixed-threshold detection algorithms is presented. Finally, numerical results and conclusions are presented in Section 3.

2 VOICE ACTIVITY DETECTION ALGORITHM

The basic function of a VAD algorithm is the comparison of some measured quantity from the input with a preset threshold. Then, voice-active or voice-inactive decision is made. VAD in nonstationary noise requires time-varying threshold value. Before describing the geometrical algorithm to monitor the noise level, let us define the Magnitude Density and Distribution functions.

2.1 The Magnitude Density Function (MDF) and The Magnitude Distribution Function (mdf)

Now, let us assume that the noise is additive so that the total signal can be written in the form

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

where $s(t)$ and $n(t)$ are the speech and noise signals, respectively. It is assumed that the speech signal $s(t)$ is band limited and noise $n(t)$ is nonstationary and Gaussian. The probability distribution function $F_x(x)$ and

the probability density function $f_x(x)$ of a random variable x are related by [12]

$$F_x(x) = \int_{-\infty}^{x} f_x(\tau)d\tau \qquad (2)$$

Let us denote the elements of the discrete-state stochastic process by $x[n]$ which are the (N+1) samples of $x(t)$ corresponding to the time instances $t = n\Delta t$ in the analysis window $[T_1 < t < T_2]$ (see Fig.4a,b,c) i.e.

$$x[n] = x(T_1 + n\Delta t) \qquad \text{for } n = 0, 1, 2, 3, ..., N$$

and where

$$\Delta t = \frac{T_2 - T_1}{N}$$

and $[T_1 < t < T_2]$ is the region where the signal is analyzed. The Magnitude Distribution Function (MDF) $A_x[m]$ and the Magnitude Density Function (mdf) $a_x[m]$ can be defined as

$$A_x[m] = \sum_{k=0}^{m} a_x[k] \qquad (3)$$

and where $a_x[m]$ is the number of samples of $x[m]$ satisfying

$$m\Delta x \leq \mid x[n] \mid < (m+1)\Delta x \qquad (4)$$

where Δx is the resolution parameter. $A_s[m]$ ($a_s[m]$) and $A_n[m]$ ($a_n[m]$) can similarly be defined using the samples $s[n]$ and $n[n]$ of $s(t)$ and $n(t)$, respectively. Note that (MDF) and (mdf) converges to $F_{|x|}(x)$ and the the probability density function $f_{|x|}(x)$ for $\Delta x \to 0, \Delta t \to 0$, $(T_1 \to -\infty)$ and $(T_2 \to \infty)$, respectively (see Figs.1,2).

Examining the two MDF's, $A_x[m]$ and $A_s[m]$, it can be seen that the signal and noise occupies different regions on the MDF and can partially be separated. This is true if the speech and noise have different expected values. Before describing the algorithm, let us further define function $S_x[m]$. $S_x[m]$ is obtained by interchanging $y = sort(x[m])$ function with respect to the $y = x$ line and normalizing it to 1 (see Fig.3). There are various types of sorting techniques any of which can be used. It can be shown that $S_x[m]$ is equivalent to ADF of $x[n]$, has finer resolution and requires less computer time to obtain compared to ADF. From now on, $S_x[m]$ will be used in place of MDF.

2.2 The Geometric Algorithm To Estimate The Noise Level

It is observed that in the $S_x[m]$ (ADF), the samples $s[n]$ and $n[n]$ are partially separated. Zero mean Gaussian noise samples $n[n]$ locate close to the origin whereas the speech samples $s[n]$ dominates the end region. A geometrical technique can heuristically be used to find the point on the MDF graph which represents the upper threshold for noise (see Fig.3). This point can be found using the following procedure; the intersection point of the two lines which are tangent to the start and end points of the ADF, respectively. A third line passing through the top left corner and the intersection point crosses the ADF at the 'optimum' point. The value found at this optimum point can be used as the threshold value or multiplied by a safety coefficient α $(1 < \alpha < 1.5)$ which is constant throughout the detection process.

3 RESULTS AND CONCLUSIONS

The geometric algorithm for estimation of the noise level is illustrated in Fig.3. It is observed that signal and noise are partially separated by calculating the ADF of $x[n]$, since the (pdf) of the signal and noise are different. Later, the noise level was geometrically found by intersecting the three lines shown in Fig.3. The reconstructed envelope of the noise is shown in Fig.4.d. Robustness to errors occur for insufficient data (short analysis window) is obtained by averaging the multiple values obtained.

A new algorithm for voice activity detection in additive nonstationary noise is presented. The probabilistic properties of a signal is examined by the Magnitude Density (mdf) and the Magnitude Distribution Functions (MDF). The geometrical technique based on MDF is illustrated in the problem of detection of the words in nonstationary Gaussian noise (see Fig.4). The algorithm is applied to the traffic, water, and restaurant noises. The results show that the algorithm is successful for all these noises. The noise is chosen to vary at rates up to -5 to 5dB/sec. The VAD algorithm based on energy levels along with the new geometrical technique is observed to operate reliably in SNRs down to 0 dB. Better performance is expected with the use of more recent algorithms[3-11].

References

[1] Rabiner L. R. and Sambur M. R.," Voiced-unvoiced-silence detection using the Itakura LPC distance measure," *Proc. Intl. Conf. Acoust., Sp., and Sig. Proc.*, pp. 323-326, May 1977.

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

[3] Tucker R., " Voice activity detection using a periodicity measure," *IEE Proceedings-I*, vol. 139, pp. 377-380, August 1992.

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

[5] Yoma N. B., McInnes F., and Jack M., " Robust speech pulse-detection using adaptive noise modelling," *Electron. Lett.*, vol. 32, no. 15, July 1996.

[6] Freeman D. K., Cosier G., Southcott C. B., and Boyd I., " The voice activity detector for the Pan-European digital cellular mobile telephone service," *Proc. Intl. Conf. Acoust., Sp., and Sig. Proc.*, pp. 369-372, Glasgow, May 1989.

[7] Srinivasan K. and Gersho A.," Voice activity detection for cellular networks," *Proc. of the IEEE Speech Coding Workshop*, pp. 85-86, October 1993.

[8] Sasaki S. and Matsumoto I.," Voice activity detection and transmission error control for digital cordless telephone system," *IEICE Trans. on Communications*, vol. E77B, Iss 7, pp. 948-955, 1994.

[9] Hoyt J. D. and Wechsler H.," Detection of human speech in structured noise," *Proc. Intl. Conf. Acoust., Sp., and Sig. Proc.*, pp. II-237-II-240, Australia, May 1994.

[10] El-Maleh K. and Kabal P.," Comparison of voice activity detection algorithms for wireless personal communications systems," *IEEE Proc. Canadian Conf. Elect. and Comp. Eng.*, pp. 470-473, May 1997.

[11] Halverson D. R.," Robust estimation and signal detection with dependent nonstationary data," *Circuits Systems and Signal Processing* , vol. 14, Iss 4, pp. 465- 472, 1995.

[12] Papoulis A., Probability, Random Variables and Stochastic Processes. Tokyo: McGraw Hill, 1984.

4 FIGURES

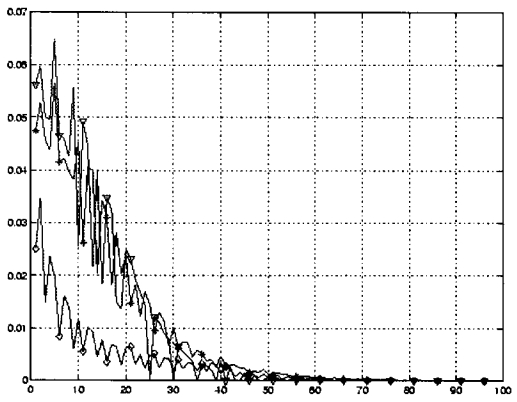

Fig. 1. The amplitude probability density (adf) functions
(SNR is oscillating in (-5,5) dB)
(X) $a_x(x)$, ($\diamond$)$a_s(s)$, (V)$a_n(n)$.

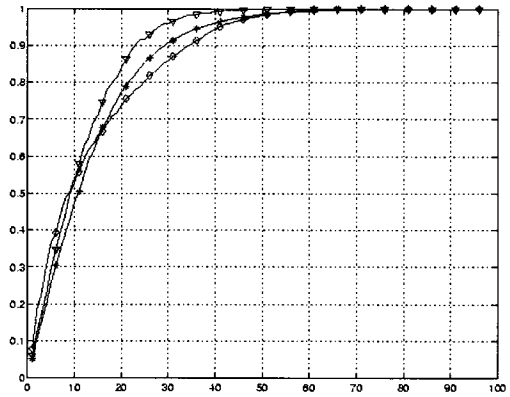

Fig. 2. The amplitude probability distribution (ADF) functions
(SNR is oscillating in (-5,5) dB)
(X) $a_x(x)$, ($\diamond$)$a_s(s)$, (V)$a_n(n)$.

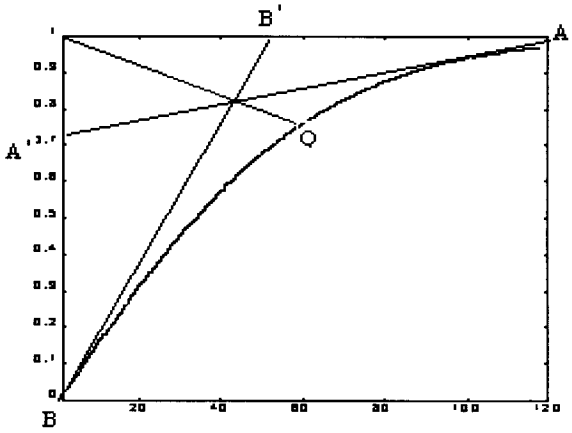

Fig. 3. The amplitude probability distribution (ADF) function $S_x(x)$ and the geometrical technique to estimate the optimum noise level.

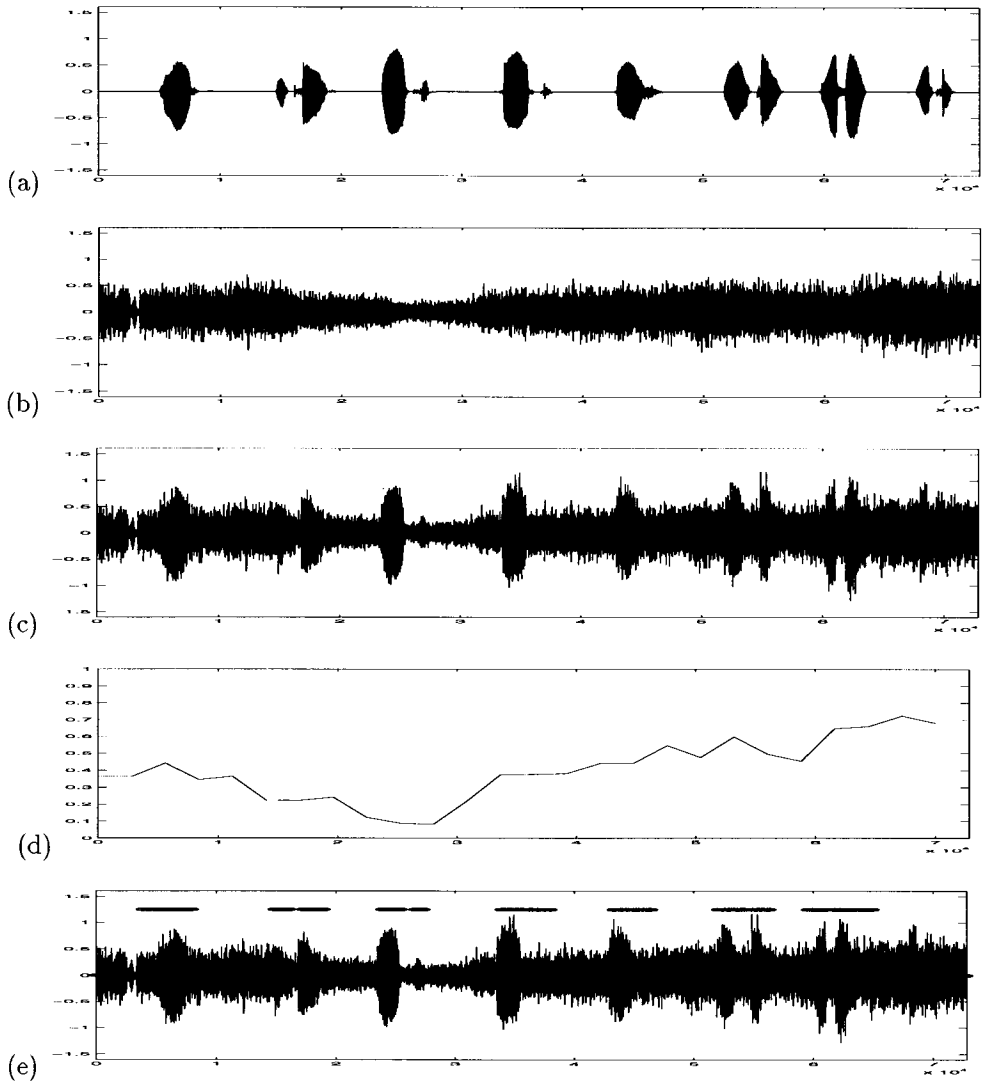

Fig. 4. Detection of the voiced-unvoiced region. (a) voice, (b) traffic noise, (c) voice in traffic noise, (d) reconstructed envelope of noise, (e) detected voiced regions (SNR is in (-5,5)dB).