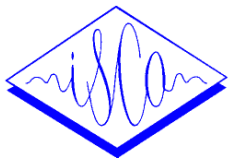

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

Skewness and Nonstationarity Measures Applied to Reliable Speech Endpoint Detection

Juan L. Navarro-Mesa & Asunción Moreno

Universitat Politècnica de Catalunya. Barcelona. Spain. email: navarro@tsc.upc.es

ABSTRACT

In this paper we are addressed to the problem of speech presence and silences detection. Speech signals possess properties to which higher order statistics are sensitive. Bispectral-based statistics have been proved a good alternative [3] to energy-based statistics in a speech presence detector scheme, but at the cost of computations. Three alternative methods are proposed for a reliable and more computationally efficient detection. Two, lag- and frequency-domain, methods for the computation of the IT and the OT contribution to total skew while keeping the ability of the statistic in [3]. And a method based in the integrated polyspectrum (IP).

1.- INTRODUCTION

The most popular methods for speech endpoint detection are based on short time energy, zero crossings, first autocorrelation lag, etc. [6]. These methods perform very well for clean speech. When the signal is not clean or the SNR is not high enough, endpoints could be obscured by noise. This situation is specially delicate for plosives and fricatives. The main problem with these sounds is their low energy and (possibly) short duration.

Detection of deviations from Gaussianity and non-stationarities are two of the problems for which third-order statistics are specially suitable. The asymmetries in the pdf of the speech signals have motivated the use of third-order statistics in speech analysis. By tracking deviations from Gaussianity and stationarities we can do a realible and accurate endpoint detection.

2. SPEECH PRESENCE AND ENDPOINTS

The problem of speech endpoint detection and, in general, speech presence detection is a complex problem where there are three possible situations: silence, speech and transitions between speech and (noisy) silence. High energy sounds are easily detectable, e.g., vowels. Problems arise when a low energy sound is at the beginning, ending or both of an utterance. In these cases the presence of additive noise becomes obscuring even when we visually try to distingish them and they could be blurred with silences. This is because local signal to noise ratio is very small.

Examples of the situation in the former paragraph can be met with plosives or fricatives (either unvoiced or voiced), and low energy vowels. Plosives are usually very short and low energy. Even though fricatives have large duration they could have low energy, too. Therefore, in a frame-based method it is difficult to distinguish them from noise. As we will show later, it is possible to discern these sounds from noise by applying third-order statistics.

3.- SPEECH DETECTION USING THE BISPECTRUM IN THE PRINCIPAL DOMAIN

Be s(n) a discret-time real, zero mean, stationary and non-Gaussian random process. The third-order cumulant and its Fourier transform, the bispectrum, are [4],

$$C_{3s}(j,k) = E\{s(n)s(n+j)s(n+k)\} \quad (1)$$

$$B(f,g) = \sum_j \sum_k C_{3s}(j,k)\, e^{-i2\pi(fi+gk)} \quad (2)$$

Sampling introduces an infinite set of parallel symmetry lines [1,2,3]. Paying attention in the principal domain of B(f,g) we can differentiate between two regions. One is the Inner Triangle (IT) where for continous-time, stationary, non-Gaussian and unaliased processes the bispectrum is non-vanishing. The other is the Outer Triangle (OT) [2] where the bispectrum will usually be nonzero if the process is either non-stationary or aliased.

In [1] the authors study the ability of bispectrum for detecting non-Gaussian signals masked by either Gaussian or non-Gaussian stationary noise. They propose a statistic function from the bicoherence evaluated in the IT region. In this function the noise effect is mitigated by extracting its bispectrum from the signal bispectrum. The statistic of this function is chi-square.

Signal presence and transitions can be detected by testing changes of stationarity [2], also. When only silence or stationary noise is present the expected value of B(f,g) in OT is zero even for non-Gaussian noise. The importance of restricting the attention to the OT triangle is that we can detect the presence of non-stationarities in a noise background of non-Gaussian random sources which makes the bispectrum nonzero in IT. In [2] the author propose a (chi-square) statistic function from the OT to detect transients by testing signals stationarity in noise.

Several tests have been developped from the statistics described in the former paragraphs to detect transients.

Even though endpoint detection is not the same problem as transient detection, both share some characteristics, e.g., signal arrival instants are unknown, at these instants there is a change of stationarity, noise is possibly present, etc. However, two facts make both problems clearly different. First, transients are usually deterministic but speech is a process and, second, the spectrum of transients is usually known while speech spectra is a priory unknown.

The sharing characteristics suggested us the possibility of applying the measures from IT and OT to detect endpoints. Essentially, we look for a function which detects speech presence while clearly marks endpoints. For that purpose we have proposed the joint use of both statistic functions. Let's briefly describe the method proposed in [3]. Firstly, when both functions are separately applied to detect endpoints their performance is good. However, problems with low energy sounds still persist because endpoints are not shown as clear and abrupt changes as we would like in noisy environments. We have found that this is a drawback related with normalization. Secondly, both statistics are transformed into Gaussian [2,7] (ZOT and ZIT for the statistics in OT and IT, respectively) because of the large number of data points. And thirdly, we make a quotient with the transformed statistics (ZOIT= ZOT/ZIT). The statistic from IT acts as a normalization allowing the statistic from OT to clearly point out the information about transitions as abrupt changes even when noise is present.

4.- SPEECH DETECTION USING BROAD BAND SKEWNESS MEASURES

In this section we develop two alternatives to [3] which are lag and frequency domain and computationally more efficient while showing similar or better performance detecting endpoints in noisy environments.

The start point of this section is the concept of amplitude skew and temporal skew [5]. The first one is associated with features which distinguish the original signal s(t) from the inverted signal -s(-t) and the second one is associated with features which distinguish the original signal from the reversal signal s(-t) respectively.

From the definitions in [5] we can obtain linear skew measures from the IT and the OT. However, these skew measures may not be useful for signals in which the contributions to the overall skew are of different sign and self-cancelling because they may register small values even though the signal posses large skew components. Since we are not sure about the validity of this fact for speech, we show preference for quadratic skewness measures because they can overcome this problem and have been proved [3] to perform very well.

The quadratic expressions introduced in [5] represents the cumulant energy. However, it does not distingish between amplitude and temporal skew. An appropriate broadband quadratic measures of the total amplitude (QA) (3) and temporal (QT) (4) skew is then proposed.

$$S_{QA} = \frac{1}{4}\int\int_{-\infty}^{\infty}\left[C_{3s}(u,v) + C_{3s}(-u,-v)\right]^2 du\,dv = \int\int_{-\infty}^{\infty}\left(\Re[B(f,g)]\right)^2 df\,dg \tag{3}$$

$$S_{QT} = \frac{1}{4}\int\int_{-\infty}^{\infty}\left[C_{3s}(u,v) - C_{3s}(-u,-v)\right]^2 du\,dv = \int\int_{-\infty}^{\infty}\left(\Im[B(f,g)]\right)^2 df\,dg \tag{4}$$

The measures involved in (3) and (4) can be expressed in terms of second order statistical quantities alone. Furthermore, it is possible to construct quadratic measures for sampled signals which separate out the information from the IT and the OT in the bispectral plane. The key is found to be the analytic signal S(f) and the normal and the modified 2-point autocorrelation function Ct and Gamma, respectively. The expressions for the separated contribution of the IT and OT regions to temporal and amplitude skew are the following.

$$S_{QAIT} = \sum_{t=0}^{N-1} C_t^2 C_t^* + \Gamma_t^2 \Gamma_t^* = \frac{32}{N^3}\sum_{f,g\in IT}\left(\Re[B(f,g)]\right)^2 \tag{5}$$

$$S_{QTIT} = \sum_{t=0}^{N-1} C_t^2 C_t^* - \Gamma_t^2 \Gamma_t^* = \frac{32}{N^3}\sum_{f,g\in IT}\left(\Im[B(f,g)]\right)^2 \tag{6}$$

$$S_{QAOT} = \sum_{t=0}^{N-1} C_t^3 + \Gamma_t^3 = \frac{96}{N^3}\sum_{f,g\in OT}\left(\Re[B(f,g)]\right)^2 \tag{7}$$

$$S_{QAOT} = \sum_{t=0}^{N-1} C_t^3 - \Gamma_t^3 = \frac{96}{N^3}\sum_{f,g\in OT}\left(\Im[B(f,g)]\right)^2 \tag{8}$$

Thus, (5) and (6) are quadratic measures of amplitud skew (QA) and temporal skew (QT) based in the IT, and, (7) and (8) are quadratic measures of amplitud and temporal stationarity based in the OT. Notice that lag-domain measures substitute frequency-domain in [1,2,3].

In the last part of this section we want to show how the total skew in the IT and the OT can also be obtained from frequency-domain measures without calculating the whole bispectrum. We can demonstrate that the following frequency-domain expressions hold for the IT and OT contribution to skew.

$$S_{OT} = \sum_{t=0}^{N-1} C_t^3 = \frac{96}{N}\sum_{f,g\in OT}|B(f,g)|^2 = \frac{8}{N^5}\sum_{f,g\in OT}|S(f)S(g)S^*(f+g)|^2 \tag{9}$$

$$S_{IT} = \sum_{t=0}^{N-1} C_t^3 C_t^* = \frac{32}{N} \sum_{f,g \in IT} |B(f,g)|^2 = \frac{8}{N^5} \sum_{f,g \in IT} |S(f)S(g)S^*(f+g)|^2 \quad (10)$$

where S(f) is the frequency-domain analytic signal. We propose to evaluate the right-hand side by first averaging S(f) over several records and in a second step evaluate the summatory.

Both the lag- and the frequency-domain skew measures from the IT and the OT are used to obtain statistic functions like in section 3 and [3]. The lag- and the frequency-domain statistics will be called ZOIT_T and ZOIT_F, respectivly.

5.- SPEECH DETECTION USING THE INTEGRATED POLYSPECTRUM

The problem of detecting unknown, random, stationary, non-Gaussian signal in Gaussian noise of unknown correlation structure can be attacked using the Integrated Polyspectrum (IP) [7]. In this section we introduce the concept of IP and propose the statistic in [7] to detect speech endpoints.

Suppose that a given frame of length N is divided into K nonoverlapping records of size Nb samples so that N=KNb. Let $S^{(i)}(w)$, $Y^{(i)}(w)$ and $R^{(i)}(w)$ denote the DFT of the data, squared data and

$$r(t) = s^3(t) - 3s(t)E\{s^2(t)\} - E\{s^3(t)\} \quad (11)$$

of the i-th record respectively. Then, the expressions for the integrated bispectrum and trispectrum at each frequency (w_m) are,

$$T_b(\omega_m) = \frac{1}{K} \sum_{i=1}^{K} \left\{ \frac{1}{Nb} S^{(i)}(\omega_m) \left[Y^{(i)}(\omega_m) \right]^* \right\} \quad (12)$$

$$T_t(\omega_m) = \frac{1}{K} \sum_{i=1}^{K} \left\{ \frac{1}{Nb} S^{(i)}(\omega_m) \left[R^{(i)}(\omega_m) \right]^* \right\} \quad (13)$$

where m=1,2,...,Nb/2-1. Notice that (12) and (13) are the cross spectrum between the signal and its square and cubic values. Another interpretation is that both represent the integration over one or two frequencies in the bi- or triplane, respectively.

An important question is how to connect the IP with the statistics in sections 3 and 4. First, we must take into account that the initial assumption for the signal include non-Gaussianity and stationarity. As we know non-zero values in the OT appear when the stationarity assumption does not hold. Therefore, in (12) and (13) stationarity information will always be mixed with the Guasianity information. The fact of integrating make impossible to distinguish or separate the information in the IT and the OT regions. Thus, any statistic funtion based in the IP extract both informations at the same time and it can be seen as a joint test for stationarity and Gaussianity. We propose a statistic from the IP based on the results in [2].

$$T_? = \sum_{m=1}^{Nb/2-1} \frac{|T_?(\omega_m)|^2}{\sigma^6} \quad (14)$$

where sigma is the noise variance and ? means bispectral or trispectral measure. T? is a chi-square statistic with Nb/2-1 degrees of freedom.

6.- EXPERIMENTS AND RESULTS

Database is composed by speech signals sampled at 16 KHz. The frame length must be long enough to have good estimates but short enough to assume local stationarity in it. In our experiments we use 320-points frames. The overlapping between frames is of 300 samples and the averages to estimate the statistics are made with 128-points records with a 90% overlap. Added noises added are either stationary white Gaussian or Exponencial. The experiments we have made are addressed to demonstrate that the functions proposed perform reliably and accurately in those cases where speech presence detection is really problematic.

The reference method is the energy-based and we compare it to the three proposed methods; time-domain (ZOIT_T), frequency-domain (ZOIT_F) and IP-based Tb (bispectral) and Tt (trispectral). The performance of all methods is evaluated with low energy voiced and unvoiced sounds.

We show experiments with low energy consonants, such as plosives /p/, /t/, /k/, fricatives /s/, /f/ or voiced /r/. The sentences were uttered by three male speakers (figs. 1,2 and 4) and a female speaker (fig. 3). On the top of the figures there are the pronunced phonemes, their SNR and the kind of noise. Notice that, the SNR used are enough to obscure the low energy sounds.

Figure 1 shows three transitions, silence-/t/, /o/-silence and silence-/k/ at the samples 1200,3100 and 3590, respectively. In figure 3 there is a transition silence-/f/ at sample 1600. The conclusion is straigth forward, the time- and the frequency-domain third-order measures show the endpoints more clearly and accurately than the energy. The transitions are shown as very abrupt changes letting better to discern the endpoints.

Figure 2 shows two transitions, silence-/p/ and /r/-silence at samples 500 and 2500, respectively. Figure 4 shows two transitions, /s/-silence and silence-/p/ at samples 1550 and 2450, respectively. The third-order measures are based in the IP. Again, the third-order measures (IP-based) show a good performance and making comparisons we can see that changes are shown as more abrupt jumps than with the energy-based detector.

7.- CONCLUSIONS.

All experiments show that the higher-order based statistic functions we propose are good alternatives to the energy-based detector which can be replaced by our statistics in a speech presence and endpoint detection scheme. Further work must be done to test our methods in a data base to make comparisons with other methods.

8.- ACKNOWLEDGEMENT.

This work has been granted by TIC-92-0800-C05-04.

9.- REFERENCIES.

[1] M. J. Hinich and G. R. Wilson. " Detection of Non-Gaussian Signals in Non-Gaussian Noise Using the Bispectrum ". IEEE Trans. on ASSP and Signal Processing, VOL. 38, No. 7, July 1990.

[2] M. J. Hinich . " Detecting a Transient Signal by Bispectral Analysis ". IEEE Trans. on ASSP and Signal Processing, VOL. 38, No. 7, July 1990.

[3] J.L. Navarro, A. Moreno and E. Lleida. " Bispectral-Based Statistics Applied to Speech Endpoint Detection ". Proc. of the IEEE Signal Processing ATHOS Workshop on Higher-Order Statistics. Girona, Spain. 1995.

[4] C. L. Nikias and M. R. Raghuveer. " Bispectrum Estimation: A Digital Processing Framework ". Proceedings of the IEEE, VOL. 75, No. 7, July 1987.

[5] A.T. Parsons and M.L.Williams. " The Construction of Broadband Higher Order Spectral Measures ". Higher Order Statistics. Elsevier Science Publushers B.V. 1992.

[6] L. R. Rabiner and R. W. Schafer. " Digital Processing of Speech Signals ". Prentice-Hall Inc., 1978.

[7] J. K. Tugnait. " Detection of Non-Gaussian Signals Using Integrated Polyspectrum ". IEEE Trans. on Signal Processing. Vol. 42, No. 11. November 1994.

Fig. 1 .../f/ /o/ /k/ /a/... SNR=30 dB

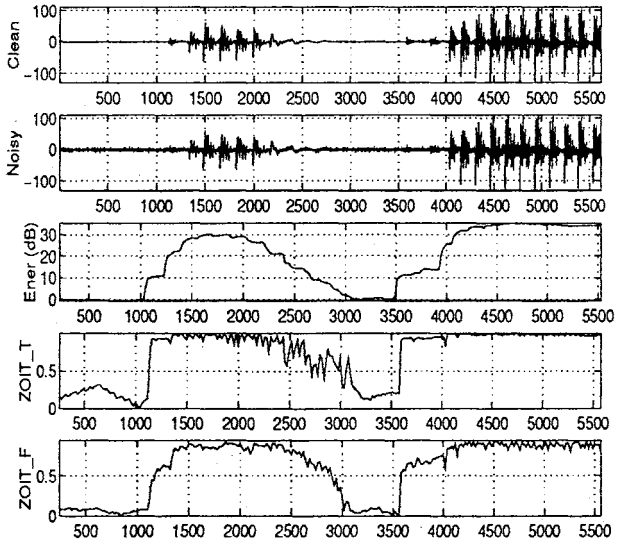

Fig. 2 .../p/ /o/ /r/... Gaussian Noise SNR=20 dB

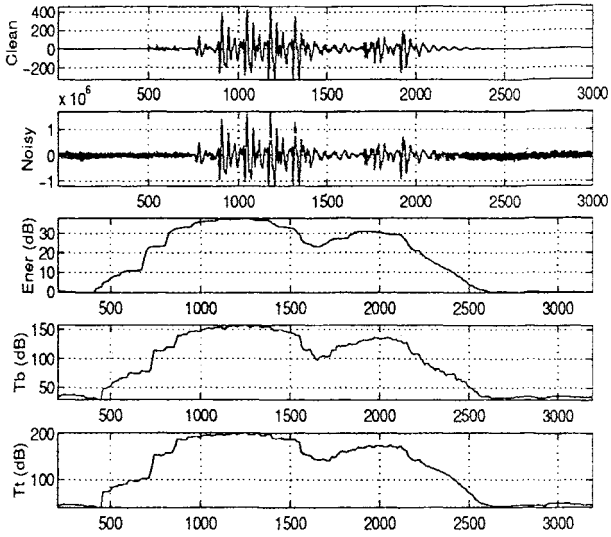

Fig. 3 .../f/ /u/ /e/... Exponential Noise SNR=30 dB

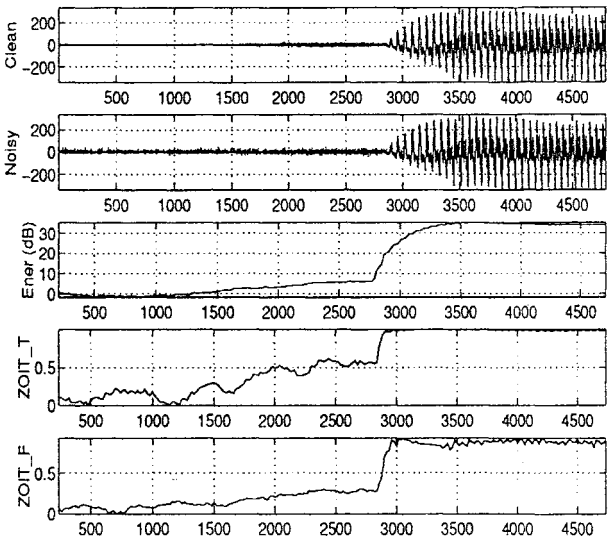

Fig. 4 .../i//o//s/ /p//o/... Exponential Noise SNR=20 dB