An improved speech endpoint detection system in noisy environments by means of third-order spectra

Article *in* IEEE Signal Processing Letters · October 1999

DOI: 10.1109/97.782065 · Source: IEEE Xplore

3 authors, including:

Juan L. Navarro-Mesa
Universidad de Las Palmas de Gran Canaria
40 PUBLICATIONS **188** CITATIONS

SEE PROFILE

Eduardo Lleida
University of Zaragoza
238 PUBLICATIONS **1,888** CITATIONS

SEE PROFILE

An Improved Speech Endpoint Detection System in Noisy Environments by Means of Third-Order Spectra

J. Navarro-Mesa, *Member, IEEE,* A. Moreno-Bilbao, *Associate Member, IEEE,* and E. Lleida-Solano, *Member, IEEE*

***Abstract*—We exploit the properties of the third-order spectra in two proposals to obtain speech detection functions. One is obtained from the principal domain of the bispectrum and the other one from the integrated bispectrum. We have developed a threshold-based system in which the detection functions can be easily integrated. Experiments show the improvement on the detection scores over the energy-based function in noisy environments.**

***Index Terms*—Bispectrum, endpoint detection, integrated bispectrum, speech analysis.**

I. Introduction

THE detection of the endpoints of an utterance is required in many speech applications (e.g., coding, recognition). In the explicit methods the analysis is done over data windows from which to estimate some kind of features like short-time energy, zero-crossings, etc. [1]. The short-time energy-based detection function has been widely used because of its reliability. Once the detection function is obtained, it is easily integrated in an adaptive threshold-based endpoint detection system. When the signal is clean the detection scores are high. However, problems arise in noisy environments for low energy phonemes (e.g., some fricatives and plosives) at the endpoints. This can be overcome by taking into account the time-frequency wealth of speech and exploiting the properties of the discrete third-order spectra (TOS) in the inner (IT) [2], [4] and the outer (OT) [3], [4] triangles of the principal domain, and the integrated bispectrum (IB) [5]. The IT is sensitive to speech presence by detecting deviations from Gaussianity and the nonlinearities in the production process. The OT is sensitive to the nonstationarities and the nonlinearities associated to the speech/silence transitions. We exploit the joint properties of the IT and the OT to suggest a new detection function. The properties of the IB are a condensed version of the properties in the IT and the OT. Additionally, some real noises can be suppressed because of the ability of the TOS to be blind to symmetrically distributed noises.

Manuscript received March 1, 1999. This work was supported by the Spanish Government under Grant TIC-92-0800-C05-04. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Y. Shoham.

J. Navarro-Mesa is with the ETSI de Telecomunicación, Universidad de Las Palmas de Gran Canaria, 35017 Las Palmas de Gran Canaria, Spain (e-mail: navarro@cibeles.teleco.ulpgc.es).

A. Moreno-Bilbao is with the ETSI de Telecomunicación, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain.

E. Lleida-Solano is with the Centro Politécnico Superior, Universidad de Zaragoza, 50015 Zaragoza, Spain.

Publisher Item Identifier S 1070-9908(99)06636-5.

II. Detection Functions

Let $\{s(n)\}$ and $\{n(n)\}$ denote two (quasi) stationary, discrete-time, zero-mean, and statistically independent processes. The speech signal $s(n)$ is observed in additive noise $n(n)(\sigma_N^2$ variance) for N samples in an analysis window where speech is supposed to be stationary. We use a binary hypothesis testing framework: H_0: $y(n) = n(n)$ (null hypothesis), and H_1: $y(n) = s(n) + n(n)$, (the alternative) $\{n = 1, \cdots, N\}$. Nonvoice human sounds (e.g., clicks, breath) are considered as noise.

The short-time energy function

$$\Gamma_E = \sum_{n=1}^{N} y^2(n)/\sigma_N^2 \tag{1}$$

is optimum when speech and noise are stationary Gaussian. This situation is not realistic since they are not always stationary and Gaussian. A more realistic function comes from the detection functions obtained from the IT and the OT:

$$\Gamma_{\mathrm{IT}} = 2\Delta_N^2 N \sum_{f_j, f_k \in \mathrm{IT}} \{|B_y^{(N)}(f_j, f_k) - B_n^{(N)}(f_j, f_k)|^2 / S_y^{(N)}(f_j) S_y^{(N)}(f_k) S_y^{(N)}(f_{j+k})\} \tag{2}$$

$$\Gamma_{\mathrm{OT}} = 2\Delta_N^2 N \sum_{f_j, f_k \in \mathrm{OT}} \{|B_y^{(N)}(f_j, f_k)|^2 / c\sigma^6\} \tag{3}$$

where 'c' is a constant, Δ_N^2 is the bispectral bandwidth for the smoothing method; $S_y^{(N)}(f_j), B_y^{(N)}(f_j, f_k)$ and $B_n^{(N)}(f_j, f_k)$ are consistent estimators of the spectrum, and the bispectrum of $y(n)$ and $n(n)$, respectively. If the window is divided in K subsegments of size L, K partial bispectral estimates can be averaged. The computational demand of the bispectrum is high. This is undesirable for our purposes. The bispectral estimator we apply is based on broadband skewness measures [6]. The information in the IT and the OT is obtained directly from measures in which the averaged frequency-domain analytic signal of the K subsegments is involved. An average over K analytic signals plus only one partial bispectrum, which is less demanding, replaces the average over K bispectrum.

Unfortunately, the denominator in (2) leads to an erratic function that produces many false alarms [6], [7], which can be reduced by replacing $S_y^{(N)}(f_j)$ by $c\sigma^6$. This is possible

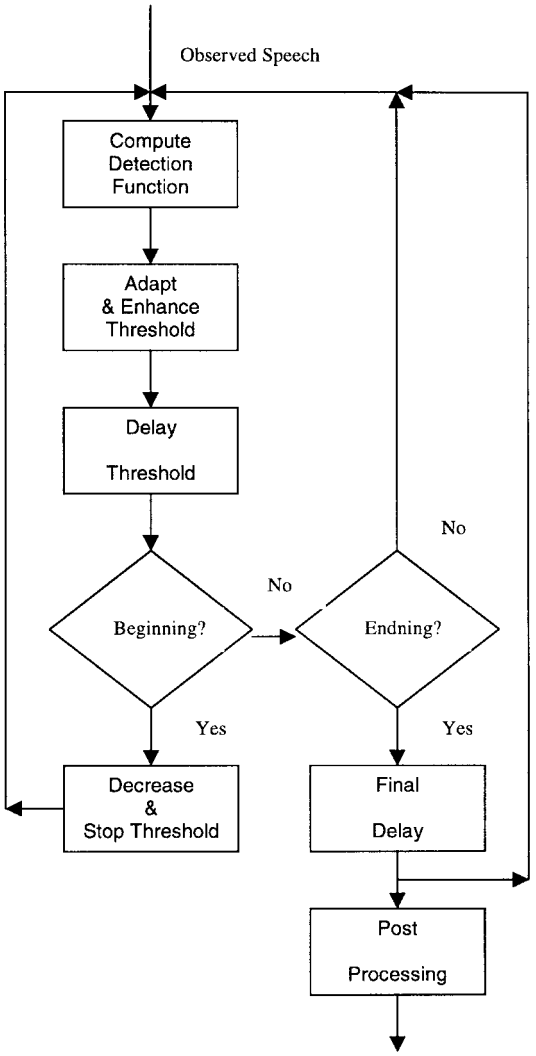

Fig. 1. Flow diagram of the endpoint detection system.

since we are more interested in detecting speech presence than in differentiating the different information to which a bicoherence function like the one in (2) is sensitive to. Also, in our experience, $B_n^{(N)}(f_j, f_k)$, while desirable, is not extrictly necessary. Therefore, it can be neglected allowing computational savings. Thus, the remaining Γ_{IT} is similar to (3) except for the limits of the summation.

We have applied both (2) and (3) to speech endpoint detection with success. Anyway, we have found that human noises cause many false alarms. That is why we propose to apply $\Gamma_{\mathrm{OIT}} = \mathrm{abs}(\Gamma_{\mathrm{IT}} - \Gamma_{OT})$. The effect of Γ_{OT} over Γ_{IT} is a subtraction that during silences and short bursts smoothes Γ_{IT} facilitating a reliable detection. Fortunately, the computation of Γ_{OIT} needs only one bispectrum because Γ_{IT} and Γ_{OT} can be computed at the same time.

The discrete IB is a cross spectrum between $Y(k)$, the DFT of $y(n)$, and $R(k)$, the DFT of $r(n) = y^2(n) - E\{y^2(n)\}$. It is a function of a single frequency and it represents significant computational savings with respect to the classical bispectrum. We derive a detection function inspired in the coherence function from the IB [5], [7]:

$$\Gamma_{\mathrm{IB}} = \sum_{k=1}^{N_B/2-1} |S_{yr}(k)|^2 \tag{4}$$

where N_B is the number of frequency points, '$*$' means complex conjugate and $S_{yr}(k) = [Y(k) - Y_n(k)][R^*(k) - R_n^*(k)]$ is an IB subtraction of noise. $S_{yr}(k)$ is computed by averaging the partial estimates $S^{(i)}y_r(k)$ form K subsegments. Expression (4) seems a suggesting choice for endpoint detection because it concentrates several properties of (2) and (3).

III. Threshold-Based Endpoint Detector

The observed signal is preemphasized to reduce the low frequency components of noise and to increase the energy of the upper frequencies of speech (see Fig. 1). Both sizes N and the overlap between windows are a compromise between time precision and consistency of the estimations. In all cases $N = 37'5$ ms. (300 samples, $F_s = 8$ KHz) and the overlap between windows is 90%. For the computation of Γ_{IB} the noise integrated bispectrum is estimated during the first 2 s of the observation. The adaptation of the threshold depends on the mean and the variance of the detection function in some preceding instants, about ten is enough. The threshold is enhanced to reduce the effect of false alarms. When the signal-to-noise ratio (SNR) is low the increase of the detection function is so small that the threshold evolves in parallel to it making difficult the detection. For facilitating detection the threshold adaptation is delayed some instants so that the increase in the detection function becomes bigger. When a beginning is detected the threshold adaptation is stopped, if not adaptation continues. If an ending is detected adaptation starts again after a delay of a few instants so as to avoid the effect of speech in the thresholds during silence. Finally, a postprocessing is applied to reject too short utterances and to keep intrasyllabic silences.

IV. Experiments and Results

The experiments have been made over 1078 isolated English digits (49 men and 49 women) from the Texas Instruments database ($F_s = 8$ KHz). Each speaker pronounced digits one to nine and two versions of zero. All of them were manually labeled before the experiments. The additive noises are; Gaussian, exponential, internal/external telephone line, air conditioning, car engine, keyboard, and fan. The SNR ranges from 25 to 0 dB. We consider that an endpoint is lost if the error is higher than 100 ms.

In Fig. 2, we present the averaged (all noises) detection scores in terms of the SNR. Detections (reliability) are given in percent. The mean errors (accuracy) are given in milliseconds over the detected endpoints. As it can be seen, the amount of beginnings [Fig. 2(a)] and endings [Fig. 2(c)] for Γ_{IB} and Γ_{OIT} clearly outperform Γ_E, especially below 15 dB. Furthermore, the bispectral measures show a higher robustness to the noise conditions. Below 15 dB the energy function becomes highly unreliable while Γ_{IB} and Γ_{OIT} show a good performance. Between 10–20 dB the amount of detections

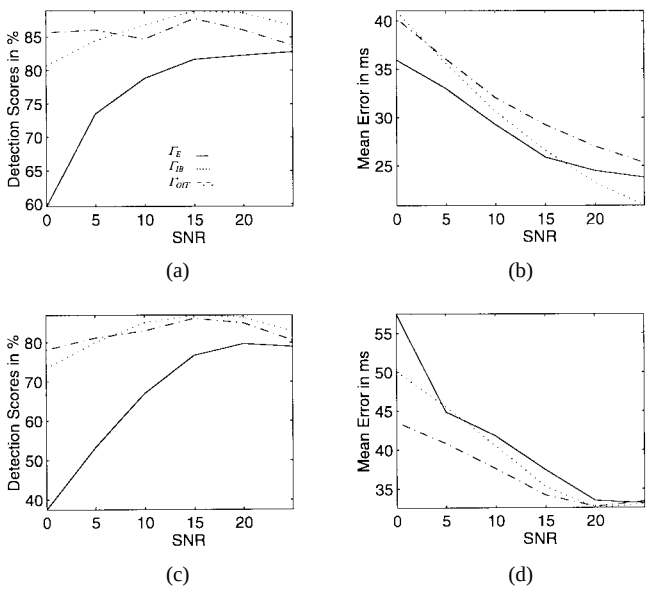

Fig. 2. Detection scores in terms of the SNR, 0–25 dB. (a) Detected beginnings. (b) Mean error, beginnings (ms). (c) Detected endings. (d) Mean error, endings (ms).

seems better than for higher SNR. This is because below 20 dB the human noises start to be obscured by the additive noise thus becoming negligible for the detection functions. As result, false alarms are reduced. The mean errors for the beginnings [Fig. 2(b)] and the endings Fig. 2(d)] also increase with noise but, in any case, they are bounded to one or two analysis windows. From an accuracy point of view, any function performs significantly better than the others do.

V. Conclusion

We have proven, in different noise backgrounds, that the bispectral-based detection functions are admissible candidates to substitute the energy function in a threshold-based endpoint detection system. The bispectral measures suffer from higher computational demand than the energy one. Our bispectral detection functions are significantly less demanding than the original ones. At present, our work is addressed to the application of our detection system to speech coding and recognition in noisy environments.

References

[1] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[2] M. Hinich and G. Wilson. "Detection of non-Gaussian signals in non-Gaussian noise using the bispectrum," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1126–1131, July 1990.

[3] M. Hinich, "Detecting a transient signal by bispectral analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1277–1283, July 1990.

[4] M. Hinich and H. Messer, "On the principal domain of the discrete bispectrum of a stationary signal," *IEEE Trans. Signal Processing*, vol. 43, pp. 2130–2134, Sept. 1995.

[5] J. Tugnait, "Detection of non-Gaussian signals using integrated polyspectrum," *IEEE Trans. Signal Processing*, vol. 42, pp. 3137–3149, Nov. 1994.

[6] J. Navarro-Mesa and A. Moreno-Bilbao, "Skewness and nonstationarities measures applied to reliable speech endpoint detection," in *Proc. 4th ESCA Eurospeech*, Madrid, Spain, vol. 1, pp. 1423–1426, Sept. 1995.

[7] J. Navarro-Mesa, A. Moreno-Bilbao, and E. Lleida-Solano, "Bispectral-based statistics applied to speech endpoint detection," in *Proc. IEEE Signal Processing ATHOS Workshop on Higher-Order Statistics*, Girona, Spain, 1995, vol. 1, pp. 280–283.