New method for epoch detection based on the Cohen's class of time frequency representations

Article *in* IEEE Signal Processing Letters · September 2001

DOI: 10.1109/97.935737 · 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

A New Method for Epoch Detection Based on the Cohen's Class of Time Frequency Representations

Juan. L. Navarro-Mesa, *Member, IEEE*, E. Lleida-Solano, *Member, IEEE*, and A. Moreno-Bilbao, *Member, IEEE*

***Abstract*—This paper presents a new method for detecting the instants of glottal closure (IGC), or epochs, in noisy environments based on the Cohen's class time–frequency representations (TFR). We define a detection function inspired in a time–frequency formulation for optimum detection and apply a morphologic closing over it to determine the epochs. It compares favorably with other methods (e.g., the SIFT-based and the Frobenius Norm [FN] function) in different levels of Gaussian noise. Experiments are carried over a data base compossed of ten speakers.**

***Index Terms*—Epoch detection, speech analysis, time–frequency representations.**

I. Introduction

MODERN speech applications (e.g., coding) tend to model the signal period-by-period. This requires period boundary determination, which can be done by detecting the instants of glottal closure (IGC). Within two epochs, the production process can be approximately modeled by a time-invariant all-pole system that imposes linear relations on the speech samples when no excitation is present. The most successful epoch determination methods rely on the detection of deviations from linearity since this fact is assumed to be associated with the excitation at the IGC. For instance, in [1], the epochs are determined from the LPC residual energy calculated from short frames. In [2], the authors develop a method that calculates the Frobenius Norm of signal matrices to detect deviations from linearity. Unfortunately, the linear model only holds approximately, and these methods fail very often.

Our method relies in the time–frequency structure of the speech during a period. When no excitation is present the characteristics of speech are assumed to be stationary. At the ICG an abrupt glottal air flow comes outside the glottis "breaking" the time–frequency structure and producing a nonstationarity shown as a wideband encrease of energy. This is a situation for which the Cohen's class of time–frequency representations is suitable.

Manuscript received October 7, 2000. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Y. Shoham.

J. L. Navarro-Mesa is with the Departmento de Señales y Comunicaciones, Universidad de Las Palmas de Gran Canaria (ULPGC), 35017 Las Palmas de Gran Canaria, Spain (e-mail: navarro@dsc.ulpgc.es).

E. Lleida-Solano is with the Centro Politécnico Superior, Universidad de Zaragoza (UZ), 50015 Zaragoza, Spain (e-mail: lleida@posta.unizar.es).

A. Moreno-Bilbao is with ETSI de Telecomunicación, Universitat Politècnica de Catalunya (UPC), 08034 Barcelona, Spain (e-mail: asuncion@gps.tsc.upc.es).

Publisher Item Identifier S 1070-9908(01)06404-5.

II. Detection Function

The expression for the Cohen's class of time–frequency (t,w) representations [3] is $C_f(t,w) = (1/2\pi)ffr(t-u,\tau)f(u+\tau/2)f^*(u-\tau/2)\mathrm{e}^{-jw\tau}dud\tau$, where $f(t)$ is the signal, and $r(t,\tau)$ is the kernel of the representation in the time-lag domain.

Our detection function arises from the time–frequency formulation for optimum detection in white Gaussian noise [4]. Let $r(t)$ be an observation and $f(t)$ a reference signal on an observation interval T. Consider the general class of receivers

$$\Lambda(t) = \int_{-\infty}^{\infty}\int_{(T)} C_r(t;t',w;\Pi(t,w))W_f(t;t',w)dt'\frac{dw}{2\pi} \quad (1)$$

where

t	given instant;
dt'	integration in the observation interval T;
$W_f(t,w)$	Wigner TFR of $f(t)$;
$C_r(t,w)$	different TFR of $r(t)$;
$\Pi(t,w)$	smoothing function (the kernel).

The interpretation of (1) is to compare the time–frequency structure of the observation to a smoothed time–frequency structure of the reference signal along the time t. In the frequency-domain, this comparison is computed over the whole frequency band.

Our method [5] treats the speech signal as a succession of time-varying spectra where there is not any known reference signal. In order to adapt (1) for epoch detection, we consider that the reference and observation signals are obtained from the speech signal in adjacent intervals. Instead of applying (1) directly, our experiments show that the best detection scores are obtained when $W_f(t,w)$ is replaced by $C_f(t,w;\Pi)$. The analysis is performed sample–by–sample over a sliding window (smaller than a period) and two consecutive windows are considered to obtain $C_f(t-1,w;\Pi)$ and $C_f(t,w;\Pi)$, respectively. Then, our detection function becomes

$$\Lambda(t) = \int_{-\infty}^{\infty} C_f(t;t'-1,w;\Pi(t,w))C_r(t;t',w;\Pi(t,w))dw. \quad (2)$$

This particular formulation takes the form of a spectral density correlator [4] and from a theoretic point of view it is a suboptimum detector because the formulation is not exactly as expressed in (1). In the stationary portions of the period $\Lambda(t)$ is smooth while at the IGC it becomes sharp because of the nonstationarity and the encrease of energy associated to them. Thus, the glottal closures take the form of sharp peaks in $\Lambda(t)$.

Finally, the choice of $\Pi(t,w)$ is important since it determines the properties of the TFR. A proper formulation for nonsta-

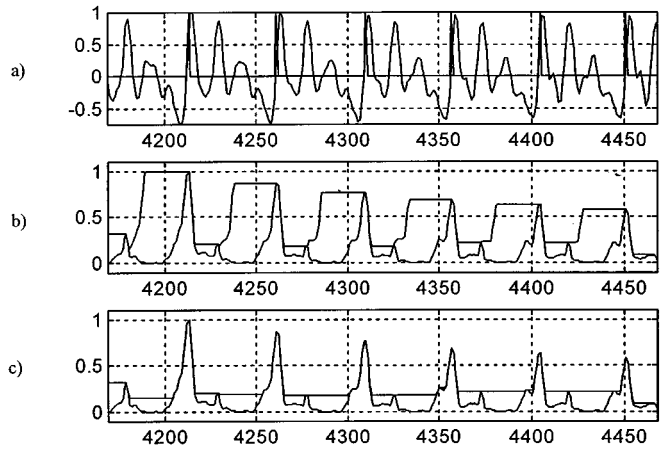

Fig. 1. (a) Speech signal and corresponding ICG marks, (b) dilation, and (c) closing over $\Lambda(t)$.

tionary processes and a good time–frequency resolution are desirable properties. Our studies and experiments show that the best performance is achieved with Born–Jordan representation (BJ–TFR) among many others (e.g., cone-shaped kernel).

III. IGC Extraction by Means of Morphological Closing

The morphological closing [6] is a filter based on dilation (maxima) and erosion (minima) operations over a set of samples, the structurant element. The closing filter keeps unaltered the positive peaks in $\Lambda(t)$, whereas it flattens the rest of the function thus allowing the detection of the IGC [Fig. 1(c)]. The size of the structurant element must be smaller than a period in order to filter small peaks between IGC maxima [Fig. 1(b)]. Maxima detection is automatically performed using a difference filter plus a zero-crossing to detect positive/negative alternances. In Fig. 1, there is an example of ICG extracted [Fig. 1(a)] after morpholigical closing.

IV. Description of the Algorithm

In our study, we assume that voiced/unvoiced detection and pitch estimation have been done by any of the classical methods prior to the experiments. In Fig. 2, we have the block diagram of our method. The analysis window and the structurant element sizes, and the detection function are adjusted with regard to the pitch period. We must remark that a rough estimation to the pitch value is enough for a proper performance. At any instant, the analysis is made sample by sample and epoch marks are given as soon as they appear.

V. Experiments and Results

We have used a speech database with their corresponding laryngogram of five men and five women (about 40 s long). The sampling frequency is 20 kHz. Prior to the expriments, we have used the laryngogram signals to extract correct IGC (29292 were obtained), voiced/unvoiced intervals and pitch. The noise is zero-mean white Gaussian and the SNRs during the voiced intervals vary from clean speech to 0 dB.

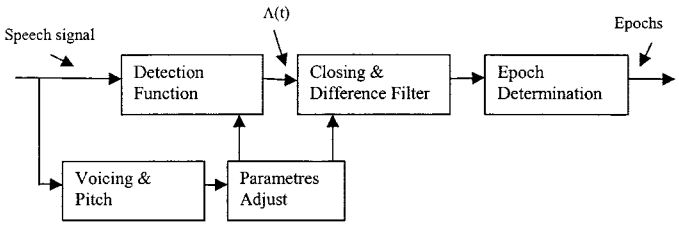

Fig. 2. Block diagram of our ICG detection scheme.

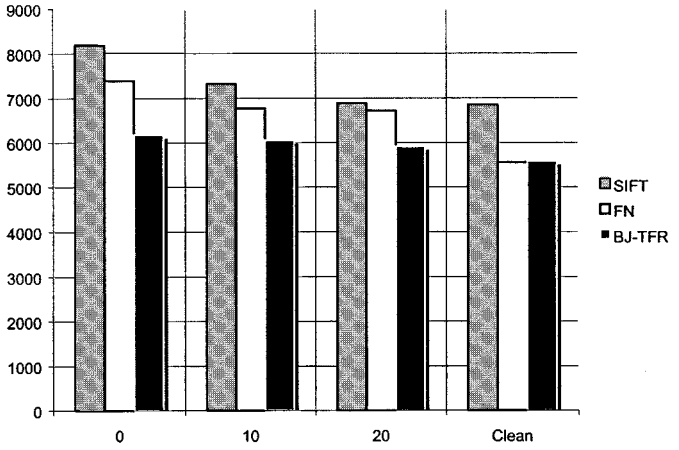

Fig. 3. Number of omissions per method (reliability).

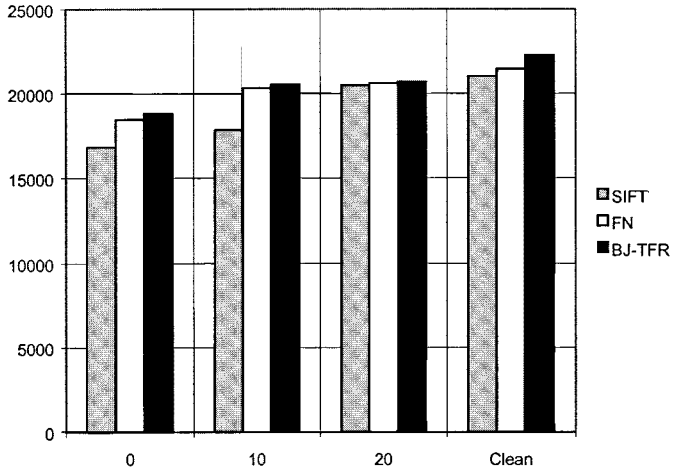

Fig. 4. Amount of correct plus fine detections (accuracy).

At a given instant, for both the TFR-based and the FN-based functions, the analysis window length is a 40% of the pitch period. For the SIFT-based detection function, a 12-coeffcients LPC analysis is performed over frames of 30 ms long. The structurant element size is a 50% of the period.

For evaluation criteria, the accuracy is defined from detections as follows. Correct (error less than a 6% of the period), fine (less than a 25%), and gross if it is higher. The criterion for reliability is based on the amount of omissions, that is, the ICG that have not been detected.

The number of omissions are presented in Fig. 3. It can be seen that the reliability is dependent on the SNR. The best scores in all cases are obtained with the BJ–TFR, which is also less sensitive to the noise conditions. In Fig. 4, we present the amount of

correct plus fine detections. Again, the BJ–TFR shows the best performance.

VI. Conclusion

We have presented a new epoch detection method that exploits the time–frequency structure of speech during the ICG. It compares favorably with respect to classical methods like the SIFT-based and the FN-based. Future work must be addressed to study the influence of voiced/unvoiced decision and pitch estimation errors. Also, other types of noise (not only Gaussian) must be experimented with.

Acknowledgment

The authors would like to thank the Department of Communication and Neuroscience, Keele University, U.K., for supplying the speech data base. They would also like to thank the reviewers for their useful comments.

References

[1] F. Plante, G. Meyer, and W. Ainsworth, "Pitch Detection: Auditory versus Inverse Filtering," *Proc. Inst. Acoust.*, pt. 5, vol. 16, pp. 81–88, 1996.

[2] C. Ma, Y. Kamp, and L. F. Willens, "A Frobenius Norm Approach to Glottal Closure Detection from the Speech signal," *IEEE Trans. Speech Audio Processing*, vol. 2, Apr. 1994.

[3] L. Cohen, *Time–Frequency Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[4] P. Flandrin, "A Time–Frequency Formulation of Optimum Detection," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1377–1384, Sept. 1988.

[5] J. L. Navarro-Mesa and E. Lleida-Solano, "Representaciones Tiempo Frecuencia no Paramétricas de Voz y Aplicaciones," Ph.D. dissertation, Politèc. Catalunya, Barcelona, Spain, 1997.

[6] J. Serra and P. Soille, "Mathematical Morphology and its Applications to Signal Processing," in *Computational Imaging and Vision*. Norwell, MA: Kluwer, 1994.