Robust Voice Activity Detection Using Higher-Order Statistics in the LPC Residual Domain

Elias Nemer, *Member, IEEE*, Rafik Goubran, *Member, IEEE*, and Samy Mahmoud, *Senior Member, IEEE*

***Abstract*—This paper presents a robust algorithm for voice activity detection (VAD) based on newly established properties of the higher order statistics (HOS) of speech. Analytical expressions for the third and fourth-order cumulants of the LPC residual of short-term speech are derived assuming a sinusoidal model. The flat spectral feature of this residual results in distinct characteristics for these cumulants in terms of phase, periodicity and harmonic content and yields closed-form expressions for the skewness and kurtosis. Important properties about these cumulants and their similarity with the autocorrelation function are revealed from this exploratory part. They show that the HOS of speech are sufficiently distinct from those of Gaussian noise and can be used as a basis for speech detection. Their immunity to Gaussian noise makes them particularly useful in algorithms designed for low SNR environments. The proposed VAD algorithm combines HOS metrics with second-order measures, such as SNR and LPC prediction error, to classify speech and noise frames. The variance of the HOS estimators is quantified and used to yield a likelihood measure for noise frames. Moreover, a voicing condition for speech frames is derived based on the relation between the skewness and kurtosis of voiced speech. The performance of the algorithm is compared to the ITU-T G.729B VAD [1] in various noise conditions, and quantified using the probability of correct and false classifications. The results show that the proposed algorithm has an overall better performance than G.729B, with noticeable improvement in Gaussian-like noises, such as street and parking garage, and moderate to low SNR.**

***Index Terms*—Higher order statistics, speech analysis, voice activity detection.**

I. Introduction

VOICE activity detection (VAD) refers to the ability of distinguishing speech from noise and is an integral part of a variety of speech communication systems, such as speech coding, speech recognition, hands-free telephony, audio conferencing and echo cancellation. In the GSM-based wireless system, for instance, a VAD module [5] is used for discontinuous transmission to save battery power. Similarly, a VAD device is used in any variable bit rate codec [26] to control the average bit rate and the overall coding quality of speech. In wireless systems based on code division multiple access, this scheme is important for enhancing the system capacity by minimizing interference.

Manuscript received August 10, 1998; revised April 10, 2000. This work was completed while E. Nemer was with Nortel Networks, St. Laurent, QC, Canada. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Douglas O'Shaughnessy.

E. Nemer is with Intel Corporation, San Jose, CA 95134 USA (e-mail: enemer@ieee.org).

R. Goubran and S. Mahmoud are with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada K1S 5B6 (e-mail: goubran@sce.carleton.ca; mahmoud@sce.carleton.ca).

Publisher Item Identifier S 1063-6676(01)01324-4.

In early VAD algorithms, short-term energy, zero-crossing rate and LPC coefficients were among the common features used for speech detection [18]. Cepstral features [7], formant shape [8], and least-square periodicity measure [24] are some of the more recent metrics used in VAD designs. In the recently proposed G.729B VAD [1], a set of metrics including line spectral frequencies (LSF), low band energy, zero-crossing rate and full-band energy is used along with heuristically determined regions and boundaries to make a VAD decision for each 10 ms frame.

Higher-order statistics (HOS) have shown promising results in a number of signal processing applications, and are of particular value when dealing with a mixture of Gaussian and non-Gaussian processes and system nonlinearity [15], [16]. The application of HOS to speech processing has been primarily motivated by their inherent Gaussian suppression and phase preservation properties (e.g., [4]). Work in this area has been based on the assumptions that speech has certain HOS properties that are distinct from those of Gaussian noise, without however verifying that this is indeed the case. For this reason, it is difficult to assess the limited effectiveness reported in some of these attempts, such as [4] and [17].

While previous work in the area of speech analysis, such as detection, voicing classification or pitch estimation, have attempted to exploit some of the observed features of the HOS of speech signals, little has been done in providing an analytical framework for using these cumulants: in [25], a voiced/unvoiced detector using the bispectrum is developed and based on the observation that unvoiced phonemes are produced by a Gaussian-like excitation and thus result in a small bispectrum whereas the same is not true for voiced phonemes. In [21] a method based on Gaussianity tests for the bispectrum and the triple correlation is used to discriminate voiced and unvoiced segments. The method exploits the Gaussian blindness of HOS but not the peculiarities of the HOS of voiced speech to better classify the segments. In [3], the normalized skewness and kurtosis of short-term speech segments are used to detect transitional speech events (termed innovation), based on the observation that these two statistics take on nonzero values at the boundaries of speech segments, but no analytical ground is given to support the results. In [12] a pitch estimation method based on the periodicity of the diagonal slice of the third-order cumulant is described and yields more reliable pitch estimates than the autocorrelation, but the claim of the third-order cumulant slice having similar periodicity as the underlying speech is not clearly demonstrated.

In this paper, we present a robust VAD algorithm based on newly established HOS properties of speech. The first part of the paper is an exploratory work into the characteristics of the

third- and fourth-order cumulants of the LPC residual of speech signals. The flat spectral envelope of this residual results in distinct characteristics for these cumulants in terms of phase, periodicity and harmonic content and yields closed-form expressions for the skewness and kurtosis. It is shown, in the case of voiced speech, that these cumulants have zero-phase, a similar harmonic nature as the underlying speech and harmonic amplitudes that are a function of speech energy. The expressions for the skewness and kurtosis of voiced speech show they may be expressed in terms of speech energy and that the normalized metrics have values that are greater than zero, regardless of the speech magnitude. In addition, experimental results show that while sustained unvoiced speech has zero HOS, it is seldom the case in real-life utterances, given that unvoiced segments are short and occur at transitional speech boundaries resulting in nonzero HOS. The properties and experimental findings thus established show that the HOS of speech are in general nonzero and sufficiently distinct from those of Gaussian noise to be used as a basis for speech detection. The fact that these statistics are immune to Gaussian noise make them a set of robust metrics that are particularly effective in low SNR conditions.

The second part of the paper capitalizes on the HOS properties of speech thus established and presents a new VAD algorithm that combines HOS metrics with classical second-order measures to classify short frames as speech or noise. A necessary condition for voicing is derived based on the relation between the skewness and kurtosis of voiced speech. The practical issues related to HOS analysis such as the bias and variance of the estimators are addressed. Using the white Gaussian assumption about noise in the LPC residual, a new unbiased estimator for the kurtosis is proposed and the variances of the HOS estimators are derived and expressed in terms of the underlying process variance (i.e., the noise energy). Knowledge of these variances allows quantifying the noise likelihood of a given frame given the values of these two estimates.

The proposed algorithm is tested using a variety of noise types and different SNR levels and its performance compared to the ITU-T G.729B VAD. To quantify performance, the probability of correctly classifying speech and noise frames as well as the probability of false classification are computed by making references to truth marker files in clean speech conditions. To compute these metrics and generate the noisy speech test cases, we used the material in the TIA database [27] proposed for the evaluation of VAD algorithms. Eighty test cases were used, with each case consisting of a different combination of speech normalization level, noise type and SNR. Four SNR levels are used: ∞ dB, 18 dB, 12 dB, and 6 dB, with the SNR value computed as the ratio of the total energy of speech to that of the noise over the entire utterance, according to the procedure in [27]. The results show that the proposed algorithm performs overall better than G.729B, with noticeable improvement in Gaussian-like noises, such as street and parking garage, and moderate to low SNR. In periodic noises, such as car and train, the algorithm is marginally biased toward speech, but its probability of false classification remains lower than G.729B in all scenarios.

The paper is organized as follows: Section II describes the sinusoidal model used and the assumptions about the LPC residual. Section III details the derivations of the third- and fourth-order cumulants, and provides simulation results on typical speech signals. Section IV describes the VAD algorithm, and SectionV presents performance results under various noise conditions. Concluding remarks are given in Section VI.

II. Analytical Model for Speech

To provide an analytical framework for using higher order cumulants, it is convenient to assume a model for speech that is both mathematically manageable as well as reasonably representative of actual signals. Speech processing is one area that is dominated by linear models, in spite of the physical and experimental evidence that seems to suggest that nonlinearity needs to be accounted for [23]. It is argued in [2] that using HOS analysis of speech can reveal information about the nonlinear signal generation mechanisms, but the results in [2] were nonconclusive, even negative about the presence of this nonlinearity. In the area of coding and voice detection, models that describe speech signals as a stream of different Gaussian processes have been proposed [19], [20]. The experimental work in [6] using third- and fourth-order statistics of long-term speech showed however that these models were limited and that the speech process may be more accurately considered as mixtures of spherically invariant Gaussian distributions.

The interest of this work is in short-term speech segments, and the goal is to find robust metrics that allow distinguishing speech from noise. The intent for using higher order statistics is not to detect nonlinearity but to show that the HOS of speech are distinct from those of Gaussian noise, even in the classical paradigm of the linear production speech model. Since the secondary goal is to interpret these HOS in terms of conventional speech parameters, it seems reasonable then to use a linear model that is valid for extracting these parameters. As in any scientific approach, some model has to be assumed at first, then the results—in this case the higher order cumulants—deduced from it need to be verified by experimental data and used to show the validity of the model.

Among the sinusoidal models developed to represent speech, the simplest, and often referenced one is the so-called *zero-phase harmonic representation* [10]. The elegance of this model is in its use of the same expression for both voiced and unvoiced speech which allows for a soft decision whereby a frame may contain both types. The model is characterized by sine-wave amplitudes, a voicing probability, and a fundamental frequency. Removing (thus greatly simplifying) the phase makes it a minimal parameter set for analysis and synthesis. In this representation, a given frame is represented by a sum of harmonically related sine waves. A synthetic phase function is used such that during voiced speech, the sine waves are coherent (in phase) and during unvoiced speech they are incoherent. The speech signal over a short-term window may be expressed as

$$x(n) = \sum_{m=1}^{M} a_m \cdot \cos[(n - n_0)w_m + \psi_m + \theta_m] \qquad (1)$$

where

n_0 voice onset time;
M number of sinusoids;

a_m amplitude;

w_m excitation frequency of the mth sine wave.

For a stationary periodic frame, these are harmonically related, i.e., $w_m = mw_0$, with w_0 being the fundamental frequency. The first phase term is due to the onset time n_0, defined as the time when the pitch pulse occurred relative to the beginning of the frame. The second phase component depends on a frequency cutoff w_c and a voicing probability P_v

$$\psi_m = \begin{cases} 0; & \text{for } mw_0 \le w_c(P_v) \\ U[-\pi, \pi]; & \text{for } mw_0 > w_c(P_v) \end{cases} \quad (2)$$

where $U[-\pi, \pi]$ is a uniformly distributed random variable between $-\pi$ and π. Thus, the higher the voicing probability the more sine waves are declared voiced with zero phase. Finally, the third phase component is the system phase θ_m along frequency track m. For simplicity, this component is often assumed to be zero or a linear function of frequency.

In the case of steady voiced speech, the sine waves are harmonically related and Eq (1) becomes

$$x(n) = \sum_{m=1}^{M} a_m \cdot \cos[(n - n_0)mw_0 + \psi_m + \theta_m]. \quad (3)$$

Therefore, in the framework of the sinusoidal model:

- *Steady (or stationary) voiced* speech segment is modeled as a sum of harmonically related sine waves whose frequencies are multiple of the fundamental and whose phase is determined entirely by the voice onset time n_0.
- *Nonstationary voiced* speech segment is modeled as a sum of sine waves, but only some of those may be harmonically related. For the rest, the phases are assumed deterministic but unknown.
- *Unvoiced* speech segment is modeled as a sum of incoherent sine waves whose phases are assumed random and uniformly distributed. In the more general approach, unvoiced speech is considered as a random—though not necessarily Gaussian—process.

A. LPC Residual

The LPC residual signal is the result of filtering the speech signal by the LPC prediction filter. The residual signal has a flat-envelope spectrum since, ideally, all short-term correlation is removed. Therefore, in the light of the sinusoidal model, we have the following.

- The residual of *voiced* speech may be modeled as a deterministic signal, consisting of M sinusoids with equal amplitudes, i.e., all the a_m are equal in (1). The frequencies of these sinusoids may or may not be harmonically related, depending on whether the speech is steady or nonstationary.
- The residual of *unvoiced* speech may be modeled as a harmonic process, consisting of M sinusoids with random (and equally distributed) phases. In the more general case, *unvoiced* speech may be modeled as a *non-Gaussian white* process.
- Finally, Gaussian noise at the input becomes white Gaussian in the residual. This feature is particularly interesting in the computation of the variance of the HOS estimators.

1) Effect of LPC Order: It is worth noting here that the statement about modeling speech in the residual as a sum of equal-amplitude harmonics is true, provided that the number of sinusoids contained in the original signal is greater than half the order of the LPC analysis. Otherwise, the residual signal would be zero, as the prediction filter would exactly match all the poles therein. Simulations show that for an LPC order of eight to twelve, this phenomena does not occur often, though it may occasionally happen that some voiced segments result in a near-zero residual. For this reason, a 10th-order LPC analysis is used in this experimental part of this work. This choice is also motivated by the use of this order of analysis in most speech coders.

2) Effect of Noise on the LPC Residual: The assumption about the flat spectral feature of speech and noise in the residual holds when the original signal consists of either one. If both speech and noise are present, and an autocorrelation-based method is used for LPC analysis, then the residual signal will have a flat-envelope spectrum but only in an aggregate sense. The spectral characteristics of the speech component will be highly affected by the SNR and the spectral content of the noise. For the flat-envelope nature of speech to hold, a robust method for LPC analysis is required to yield a filter that will only flatten the speech spectrum. In [17], an approach based on third-order cumulants is used and results in a robust LPC filter that is shown effective in noisy conditions. In the rest of this paper, it is assumed that such methods are used and that the flat characteristics of the speech residual hold in all conditions. The effect of a nonflat speech residual on the analytical derivations will be discussed in the appropriate section.

III. Higher-Order Cumulants of the LPC Residual

In the following section, important higher order cumulant (HOC) properties of the LPC residual of short-term speech are derived and presented in the form of theorems. The derivations are based on the sinusoidal model discussed above. The third and fourth-order cumulants are addressed separately and in each case, derived for steady, nonstationary voiced and unvoiced speech.

A. Definition and Notation

If $x(n)$, $n = 0, \pm1, \pm2, \pm3, \cdots$ is a real stationary discrete-time signal and its moments up to order p exist, then its pth-order moment function is given by

$$m_p(\tau_1, \tau_2, \cdots, \tau_{p-1}) \equiv E\{x(n)x(n+\tau_1) \cdots x(n+\tau_{p-1})\} \quad (4)$$

and depends only on the time differences $\tau_1, \tau_2, \cdots, \tau_{p-1}$ $\tau_i = 0, \pm1, \pm2, \cdots$ for all i. Here $E\{.\}$ denotes statistical expectation and for a deterministic signal, it is replaced by a time summation over all time samples (for energy signals) or time averaging (for power signals). If in addition the signal has zero

mean, then its cumulant functions (up to order four) are given by [15]

second-order cumulant:

$$C_2(\tau_1) = m_2(\tau_1), \tag{5}$$

third-order cumulant:

$$C_3(\tau_1, \tau_2) = m_3(\tau_1, \tau_2), \tag{6}$$

fourth-order cumulant:

$$\begin{aligned} C_4(\tau_1, \tau_2, \tau_3) = {} & m_4(\tau_1, \tau_2, \tau_3) - m_2(\tau_1) \cdot m_2(\tau_3 - \tau_2) \\ & - m_2(\tau_2) \cdot m_2(\tau_3 - \tau_1) - m_2(\tau_3) \\ & \cdot m_2(\tau_2 - \tau_1). \end{aligned} \tag{7}$$

By setting all the lags to zero in the above cumulant expressions, we obtain the variance, skewness and kurtosis, respectively,

$$\begin{aligned} &\text{variance:} \quad \gamma_2 \equiv C_2(0) = E\{x^2(n)\}, \\ &\text{skewness:} \quad C_3(0, 0) = E\{x^3(n)\}, \\ &\text{kurtosis:} \quad C_4(0, 0, 0) = E\{x^4(n)\} - 3[E\{x^2(n)\}]^2. \end{aligned}$$

When estimating higher-order statistics from finite data records, the variance of the estimators is reduced by normalizing the input data to have a unity variance, prior to computing the estimators. Equivalently, the third and fourth-order statistics are normalized by the appropriate powers of the data variance, thus we define the

normalized skewness:

$$\gamma_3 \equiv \frac{C_3(0, 0)}{[C_2(0)]^{1.5}} = \frac{E\{x^3(n)\}}{[E\{x^2(n)\}]^{1.5}} \tag{8}$$

normalized kurtosis:

$$\gamma_4 \equiv \frac{C_4(0, 0, 0)}{[C_2(0)]^2} = \frac{E\{x^4(n)\}}{[E\{x^2(n)\}]^2} - 3.0. \tag{9}$$

Since the third and fourth-order cumulants are multi-dimensional functions, it is customary to use only two-dimensional (2-D) slices of these, by freezing some of the lags in (6) and (7). In this paper, horizontal slices are used and defined as

third-order cumulant slice:

$$C_3[\tau] \equiv C_3(0, \tau) = m_3(0, \tau) = E[x^2(n)x(n - \tau)] \tag{10}$$

fourth-order cumulant slice:

$$C_4[\tau] \equiv C_4[0, \tau, \tau) = m_4(0, \tau, \tau) - [m_2(0)]^2 - 2[m_2(\tau)]^2$$

or

$$\begin{aligned} C_4[\tau] = {} & E[x^2(n)x^2(n + \tau)] - (E[x^2(n)])^2 \\ & - 2(E[x(n)x(n + \tau)])^2. \end{aligned} \tag{11}$$

These cumulant functions may also be normalized as

$$C_3'[\tau] = \frac{C_3[\tau]}{[C_2(0)]^{1.5}} \quad \text{and} \quad C_4'[\tau] = \frac{C_4[\tau]}{[C_2(0)]^2}. \tag{12}$$

B. Third-Order Cumulants

1) Steady Voiced Speech: Voiced speech is modeled as a sum of coherent sine waves whose frequencies are harmonically related to the fundamental. Furthermore, a linear system phase is assumed and, as a result, the phases of the sine waves in (1) are entirely determined by the onset time n_0 and a constant due to the system phase.

Theorem 1: According to the sinusoidal model, the horizontal slice $C_3[\tau]$ of the third-order cumulant of the LPC residual of steady voiced speech that is bandlimited to $f_s/4$ has M harmonics and the same periodicity as the residual itself. The amplitude of each harmonic may be written in terms of the signal energy (variance) and the number of harmonics M. Moreover, $C_3[\tau]$ has zero phase and reaches maxima at multiples of the pitch lag

$$\boxed{C_3[\tau] = 2c\left(\frac{E_s}{M}\right)^{3/2} \sum_{m=1}^{M} [2M - 1 - m] \cos(mw_0\tau)} \tag{13}$$

where $c = 2^{3/2}/8$ and E_s is the signal energy $E_s \equiv m_2(0) = M(a^2/2)$.

Proof: Consider the horizontal slice $C_3[\tau] = (1/N)\sum_n x^2(n)x(n - \tau)$ of the third-order cumulant of a deterministic signal. Using the convolutional property of the Fourier transform, it can be shown that the transform of $C_3[\tau]$ is

$$FC_3(w) = \{X(w) \otimes X(w)\}X(w) \tag{14}$$

where $X(w)$ is the transform of $x(n)$. Since the signal $x(n)$ consists of M harmonics, its spectrum is composed of M impulses on both the positive and negative frequency axes; moreover, the flat-envelope spectrum of the LPC residual implies equal magnitude impulses. Therefore, $X(w) = (a/2)e^{jkw}$ for $w = \pm(w_0, 2w_0, \cdots, Mw_0)$, and k is a constant that depends on the onset time and the system delay. The convolution of the spectrum is nonzero only at multiples of the fundamental frequency w_0. In addition, it is assumed that the signal is bandlimited to $\pi/2$ or $f_s/4$ (where f_s is the sampling frequency) and as a result, there are only $2M$ positive and $2M$ negative lags that lead to nonzero values of the autoconvolution $X(w) \otimes X(w)$. The bandlimited assumption is required to limit the convolution terms to $2M$, otherwise more nonzero terms will result on each side, due to the mirror image of the spectrum. Fig. 1 illustrates the value of the convolution for lag $w = w_0$. Due to the multiplication by $X(-w)$ in (14), only the first M lags on each of the negative and positive frequency axes are nonzero. In addition, the phase of $X(w) \otimes X(w)$ cancels out with the phase of $X(-w)$ at any lag and the resulting spectrum $FC_3(w)$ has zero phase. Table I shows the values of $FC_3(w)$ for all positive values of the lag w. Moreover, the value of $FC_3(w)$ may be expressed in terms of signal energy since $a^3 = (2E_s/M)^{3/2}$.

Due to spectral symmetry, the results are the same for the negative lags; consequently, the inverse transform of $FC_3(w)$ leads to the sum of M cosine terms given by (13), from which it is evident that $C_3[\tau]$ has a maximum at $\tau = 0, P, 2P \cdots$ (where P is the pitch period, i.e., $1/w_0$). Moreover, the amplitude of the normalized cumulant slice, $C_3'[\tau]$, is only a function of M, i.e., the effect of signal energy is eliminated after normalization. This result is interesting since it draws on a similar phenomena when dealing with a random process: the normalized third and fourth-order statistics are equivalent to having an input process with a unity variance. Finally, the zero-phase characteristics of

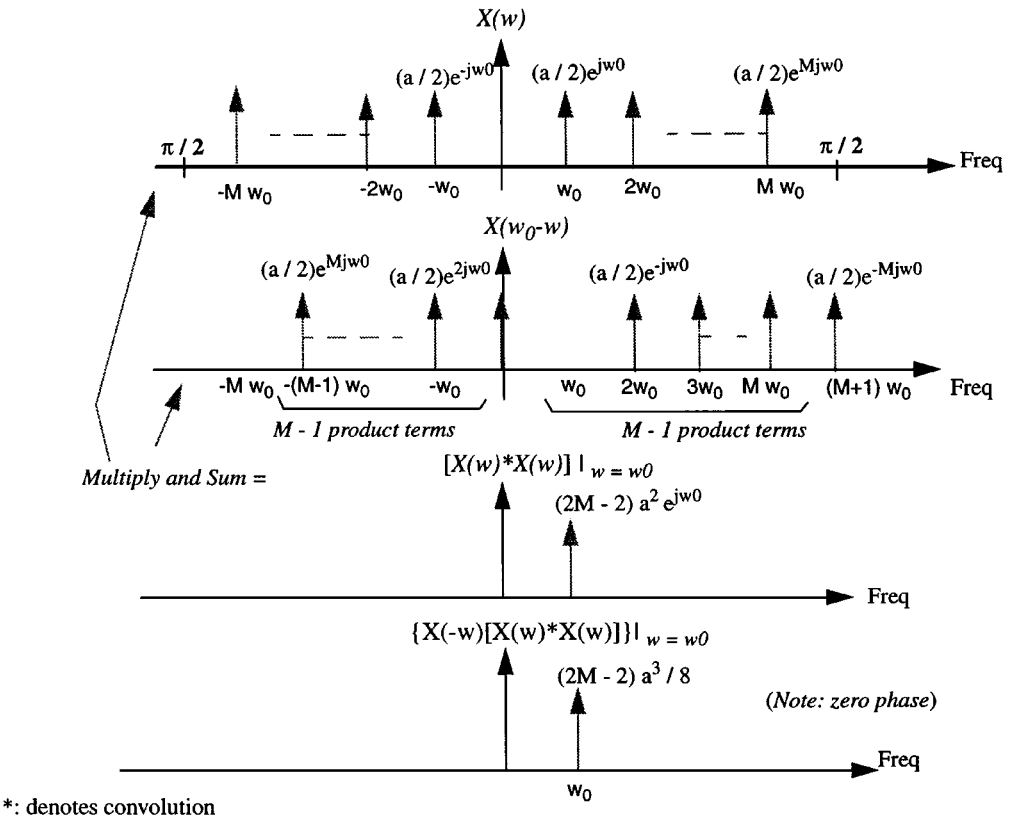

Fig. 1. Computing $FC_3(w)$ for $w = w_0$.

TABLE I
VALUE OF $FC_3(w)$ AT ALL POSITIVE LAGS

Lag (w)	X(w)*X(w)	X(-w)	FC_3(w)
0	$2Ma^2/4$	0	0
w_0	[(M -1) + (M - 1)] a^2 / 4 *Phase:* e^{jkw0}	a / 2 *Phase:* e^{-jkw0}	(2M - 2) a^3/8
$2w_0$	[(M -2) + (M - 1)] a^2 / 4 *Phase:* e^{j2kw0}	a / 2 *Phase:* e^{-j2kw0}	(2M - 3) a^3/8
$3w_0$	[(M -3) + (M - 1)] a^2 / 4 *Phase:* e^{j3kw0}	a / 2 *Phase:* e^{-j3kw0}	(2M - 4) a^3/8
.....	.....		.....
(M-1)w_0	[1 + (M - 1)] a^2 / 4 *Phase:* $e^{j(M-1)kw0}$	a / 2 *Phase:* $e^{-j(M-1)kw0}$	M a^3/8
M w_0	[0 + (M - 1)] a^2 / 4 *Phase:* e^{jMkw0}	a / 2 *Phase:* e^{-jMkw0}	(M - 1) a^3/8

the third-order cumulant is in agreement with the general property derived in [16] that third-order statistics are insensitive to time shifts.

Corollary 1: The skewness of the LPC residual of steady voiced speech may be written as a function of the energy of the residual and the number of harmonics M. The normalized skewness may be expressed in terms of M and is greater than zero for any practical values of M, namely

$$\boxed{C_3[0] = 3c(E_s)^{3/2}\left[\frac{(M-1)}{\sqrt{M}}\right]} \tag{15}$$

skewness:

normalized skewness:

$$\boxed{\gamma_3 \equiv \frac{C_3[0]}{E_s^{3/2}} = 3c\frac{(M-1)}{\sqrt{M}}}. \tag{16}$$

Proof: The skewness is found by setting $\tau = 0$ in (13)

$$\begin{aligned} C_3[0] &= 2c\left(\frac{E_s}{M}\right)^{3/2}\sum_{k=1}^{M}[2M-1-k] \\ &= 2c\left(\frac{E_s}{M}\right)^{3/2}\{M(M-1)+[1+2+3+\cdots \\ &\qquad +(M-1)]\} \\ &= 2c\left(\frac{E_s}{M}\right)^{3/2}\left\{M(M-1)+\frac{(M-1)M}{2}\right\} \end{aligned}$$

which, after rearranging terms, leads to (15).

2) Nonstationary Voiced Speech: In the case of a nonstationary voiced segment, not all harmonics may be related and as a result, the value of $C_3[\tau]$ may be zero for some of the harmonics. In the hypothetical case where no three frequencies are related, the third-order cumulant is zero at all lags. In practice, however, we expect that even for nonstationary voiced segments, some relations exist between some of the harmonics and thus $C_3[\tau]$ is rarely zero for voiced speech.

3) Unvoiced Speech: Assuming that unvoiced speech in the LPC residual is modeled as a general non-Gaussian white process, its bispectrum is 2-D-flat across all bifrequencies:

$$B(w_1, w_2) = E[X(w_1)X(w_2)X(-w_1 - w_2)] = \gamma.$$

TABLE II
VALUE OF FC_4 (w) AT ALL POSITIVE LAGS

Lag (w)	$\lvert X(w)*X(w)\rvert^2$	$m_2(0)^2\,\delta(w)$	P(w)*P(w)	$FC_4(w)$
0	$4\,M^2\,a^4/16$	$M^2\,a^4/4$	$2\,M\,a^4/16$	$-\,M\,a^4/4$
w_0	$[(M-1)+(M-1)]^2\,a^4/16$	0	$[(M-1)+(M-1)]\,a^4/16$	$[(2M-2)(2M-4)]\,a^4/16$
$2w_0$	$[(M-2)+(M-1)]^2\,a^4/16$	0	$[(M-2)+(M-1)]\,a^4/16$	$[(2M-3)(2M-5)]\,a^4/16$
$3w_0$	$[(M-3)+(M-1)]^2\,a^4/16$	0	$[(M-3)+(M-1)]\,a^4/16$	$[(2M-4)(2M-6)]\,a^4/16$
.....	.....	0	.....	.....
.....	.....	0	.....	.....
$(M-1)w_0$	$[1+(M-1)]^2\,a^4/16$	0	$[1+(M-1)]\,a^4/16$	$[M(M-2)]\,a^4/16$
$M\,w_0$	$[0+(M-1)]^2\,a^4/16$	0	$[0+(M-1)]\,a^4/16$	$[(M-1)(M-3)]\,a^4/16$
$(M+1)w_0$	$M^2\,a^4/16$	0	$M\,a^4/16$	$M(M-2)\,a^4/16$
$(M+2)w_0$	$(M-1)^2\,a^4/16$	0	$(M-1)\,a^4/16$	$(M-1)(M-3)\,a^4/16$
$(M+3)w_0$	$(M-2)^2\,a^4/16$	0	$(M-2)\,a^4/16$	$(M-2)(M-4)\,a^4/16$
.....	.....	0	.....	.....
$(2M-1)w_0$	$4\,a^4/16$	0	$2\,a^4/16$	0
$2Mw_0$	$a^4/16$	0	$a^4/16$	$-\,a^4/16$

The third-order cumulant, being the inverse Fourier transform of the bispectrum, thus consists of a 2-D delta function of amplitude γ:

$$C_3[\tau_1, \tau_2] = \gamma\delta(\tau_1, \tau_2). \tag{17}$$

C. Fourth-Order Cumulants

1) Steady Voiced Speech:

Theorem 2: According to the sinusoidal model, the horizontal slice $C_4[\tau]$ of the fourth-order cumulant of the LPC residual of steady voiced speech that is bandlimited to $f_s/4$ consists of $(2M-1)$ harmonics and has the same periodicity as the underlying signal. The value at each harmonic may be written in terms of the energy of the signal and the number of harmonics. Moreover, $C_4[\tau]$ has zero phase and maxima at multiples of the pitch lag.

Proof: For a deterministic signal, $C_4[\tau]$ is given by [from (11)]

$$C_4[\tau] = \left[\frac{1}{N}\sum_n x^2(n)x^2(n+\tau)\right] - \left[\frac{1}{N}\sum_n x^2(n)\right]^2 - 2\left[\frac{1}{N}\sum_n x(n)x(n+\tau)\right]^2. \tag{18}$$

Using the convolutional property, the Fourier transform may be shown to be

$$FC_4(w) = |X(w)\otimes X(w)|^2 - [m_2(0)]^2\delta(w) - 2\{P(w)\otimes P(w)\} \tag{19}$$

where $X(w)$ is the transform of $x(n)$ and $P(w)$ is the power spectrum of $x(n)$. Since the signal $x(n)$ consists of M harmonics, its spectrum is composed of M impulses on both the positive and negative frequency axes; moreover, the flat-envelope spectrum of the LPC residual implies equal magnitude impulses. Therefore, $X(w) = (a/2)e^{jkw}$ for $w = \pm(w_0, 2w_0, \cdots, Mw_0)$ and k is a constant that depends on the onset time and the system delay. The auto-convolution of $X(w)$ is nonzero only at multiples of the fundamental frequency w_0. As in Theorem 1, it is assumed that the signal is bandlimited to $\pi/2$ or $f_s/4$ and as a result, there are only $2M$ positive and $2M$ negative lags that lead to nonzero values of the autoconvolution $X(w)\otimes X(w)$ as well as the autoconvolution $P(w)\otimes P(w)$. Therefore, $FC_4(w)$ has $2M$ nonzero values on each side of the spectrum, and only at multiples of w_0.

Clearly, the phase of $FC_4(w)$ is zero for all frequencies w since each term in Eq (19) has zero phase. Table II shows the various values of $FC_4(w)$ for all positive values of the lag w. Due to spectral symmetry, the values are the same for the negative lags. Since the signal energy is: $E_s = M(a^2/2)$, it follows that the magnitudes at the various harmonics (column five) may be expressed in terms of E_s^2. As seen from Table II, there are $2M-1$ nonzero values (due to the zero value at the next to last lag). However, depending on the value of M, some other harmonics may be zero as well.

Corollary 2: The kurtosis of the LPC residual of steady voiced speech may be expressed in terms of speech energy and the number of harmonics. The normalized kurtosis is a function of the number of harmonics only and is greater than zero for any practical value of the pitch, namely

kurtosis:

$$\boxed{C_4[0] = E_s^2\left[\frac{4}{3}M - 4 + \frac{7}{6M}\right]} \tag{20}$$

normalized kurtosis:

$$\boxed{\gamma_4 \equiv \frac{C_4[0]}{E_s^2} = \frac{4}{3}M - 4 + \frac{7}{6M}}. \tag{21}$$

Proof: The kurtosis is found by setting $\tau = 0$ in (18)

$$C_4[0] = \frac{1}{N}\sum_n x^4(n) - 3\left[\frac{1}{N}\sum_n x^2(n)\right]^2. \quad (22)$$

The value of the fourth-order moment may be determined in the frequency domain by summing the coefficients of the Fourier transform of $(1/N)\sum_n x^2(n)x^2(n+\tau)$ since

$$\frac{1}{N}\sum_n x^2(n)x^2(n+\tau) \leftrightarrow |X(w)\otimes X(w)|^2 \quad \text{Transform pair}$$

and therefore, the value at $\tau = \mathbf{0}$ is

$$\frac{1}{N}\sum_n x^4(n) = \int_{-\pi}^{\pi} |X(w)\otimes X(w)|^2\, dw.$$

The value of the Fourier coefficients $|X(w)\otimes X(w)|^2$ is given in the first column of Table II. It can be shown [13] that the sum of the coefficients across all lag values may be written in terms of the signal energy $E_s \equiv (1/N)\sum_n x^2(n)$ as

$$\frac{1}{N}\sum_n x^4(n) = \frac{E_s^2}{2}\left[\frac{8}{3}M - 2 + \frac{7}{3M}\right]. \quad (23)$$

Substituting (23) in (22) yields the value of the kurtosis given in (20).

2) Voiced Speech—General Property:

Theorem 3: If voiced speech is modeled as a deterministic harmonic signal, then the average value of the horizontal slice of the fourth-order cumulant ($C_4[\tau]$) of the LPC residual of voiced speech (both steady and nonstationary) may be written in terms of the signal energy and the number of harmonics.

Proof: The average value of $C_4[\tau]$ is the value of $FC_4(w)$ (19) at $w = 0$. First, it is observed that

$$\begin{aligned} X(f)\otimes X(f)|_{f=0} &= \int_{-\pi}^{\pi} X(\lambda)X(-\lambda)\, d\lambda \\ &= \int_{-\pi}^{\pi} |X(\lambda)|^2\, d\lambda \\ &= E_s \\ &= m_2(0). \end{aligned}$$

Therefore, setting $w = 0$ in (19)

$$\begin{aligned} FC_4(0) &= [m_2(0)]^2 - [m_2(0)]^2 - 2(P(f)\otimes P(f))|_{f=0} \\ &= -2(P(f)\otimes P(f)|_{f=0}) \\ &= -2\int_{-\pi}^{\pi} P(\lambda)P(-\lambda)\, d\lambda \\ FC_4(0) &= -2\int_{-\pi}^{\pi} |X(\lambda)|^4\, d\lambda. \end{aligned} \quad (24)$$

The spectrum $X(w)$ consists of delta functions at discrete frequencies f_m with amplitudes $a_m/2$, therefore

$$FC_4(0) = \tfrac{-1}{4}\sum_{m=1}^{M} a_m^4$$

where a_m is the amplitude of the mth sinusoid. In the LPC residual all these are equal, thus

$$E\{C_4[\tau]\} = FC_4(0) = -M(a^4/4) = \frac{-[E_s]^2}{M}. \quad (25)$$

3) Unvoiced Speech: Assuming a non-Gaussian white process, both the power spectrum and the trispectrum of the LPC residual of unvoiced speech are flat: the trispectrum is $T_e(w_1, w_2, w_3) = \gamma$ and the power spectrum $P_e(w) = \alpha$. As a result, the fourth-order moment of the residual is a delta function

$$m_4(\tau_1, \tau_2, \tau_3) = \gamma\delta(\tau_1, \tau_2, \tau_3) \quad (26)$$

and the second-order moment (i.e., the autocorrelation) is another delta function

$$m_2(\tau) = \alpha\delta(\tau). \quad (27)$$

Using (11), the kurtosis and the horizontal slice of the fourth-order cumulant of the residual signal $e(n)$ may then be written as

horizontal slice:

$$\boxed{C_4[\tau] = \gamma\delta(\tau) - \alpha^2[1 + 2\delta(\tau)]} \quad (28)$$

kurtosis:

$$\boxed{C_4[0] = \gamma - 3\alpha^2}. \quad (29)$$

D. Effect of Noise on γ_3 and γ_4

When the signal consists of both speech and noise, then $x(n) = s(n) + g(n)$. If $s(n)$ and $g(n)$ are statistically independent, then the energy of $x(n)$ is the sum of speech and noise energies: $E_x = E_s + E_N$. Second-order statistics are thus directly affected in an additive way by the presence of noise. Higher-order statistics on the other hand are immune to Gaussian noise, which has zero HOS. Since cumulants are cumulative [11], it follows that the third and fourth-order cumulants of $x(n)$ are simply those of $s(n)$. As a result, the above derivations for $C_3[0]$ and $C_4[0]$ still hold in the presence of Gaussian noise. However, when normalizing these two quantities by the signal energy E_x, the effect of the noise term in the denominator does not cancel out with the speech energy term E_s in the numerator in (15) and (20). It is easy to see that the expressions of normalized skewness (16) and kurtosis (21) of noisy speech can now be extended to include an SNR term as follows:

$$\begin{aligned} \gamma_3 &\equiv \frac{C_3[0]}{E_x^{3/2}} \\ &= 3c\left(\frac{E_s}{E_s + E_N}\right)^{3/2}\left[\frac{M-1}{\sqrt{M}}\right] \end{aligned}$$

or simply

$$\boxed{\gamma_3 = 3c\left(\frac{\text{SNR}}{\text{SNR}+1}\right)^{3/2}\left[\frac{M-1}{\sqrt{M}}\right]} \quad (30)$$

$$\gamma_4 \equiv \frac{C_4[0]}{E_x^2} = \frac{C_4[0]}{[E[s^2(n)] + E[g^2(n)]]^2} = \left(\frac{4}{3}M - 4 + \frac{7}{6M}\right)\left(\frac{E[s^2(n)]}{E[s^2(n)] + E[g^2(n)]}\right)^2$$

$$\boxed{\gamma_4 = \left(\frac{\text{SNR}}{\text{SNR}+1}\right)^2 \left[\frac{4}{3}M - 4 + \frac{7}{6M}\right]}. \quad (31)$$

Therefore, the effectiveness of these two metrics to detect voicing decreases with the SNR. Due to the squared term, the normalized kurtosis is more adversely affected by the presence of noise than the skewness.

If the noise is non-Gaussian but has a symmetrical distribution, then only its third-order statistics are zero and the above reasoning holds for the normalized skewness but not for the kurtosis. Consider the hypothetical case of Laplacian noise. Using the cumulative property of the HOS and the fact that the kurtosis of a Laplacian process may be written in terms of its energy[1] as: $C_{4g}[0] = 3(E[g^2(n)])^2$, then the normalized kurtosis of noisy speech becomes

$$\gamma_4 \equiv \frac{C_{4S}[0] + C_{4g}[0]}{E_x^2} = \frac{C_{4S}[0] + 3E[g^2(n)]}{[E[s^2(n)] + E[g^2(n)]]^2} = \left(\frac{4}{3}M - 4 + \frac{7}{6M}\right)\left(\frac{E[s^2(n)]}{E[s^2(n)] + E[g^2(n)]}\right)^2 + \left(\frac{3E[g^2(n)]}{E[s^2(n)] + E[g^2(n)]}\right)^2$$

$$\gamma_4 = \left(\frac{\text{SNR}}{\text{SNR}+1}\right)^2 \left[\frac{4}{3}M - 4 + \frac{7}{6M}\right] + 3 \cdot \left(\frac{1}{\text{SNR}+1}\right)^2.$$

The effect is similar to the case of Gaussian noise, in that the effectiveness of the kurtosis degrades with SNR. However, the problem of distinguishing speech from noise, based on this metric, needs to be reformulated since the normalized kurtosis of the noise is no longer zero, but three in this case.

E. Effect of a Nonflat LPC Residual

It was mentioned in Section II-A that noise may adversely affect the LPC analysis which will result in a residual that is flat in an aggregate sense, though the speech component itself is not flat. Even in the case where noise in not present, one would not expect the LPC residual to be perfectly flat since the LPC analysis itself is seldom perfect. It is therefore necessary to assess the implication of a nonflat residual on the properties of the HOS of speech derived so far.

In the above derivations of the third and fourth-order cumulants, the flat envelope characteristic resulted in the values of the autoconvolution of the spectrum to be expressed in terms of the signal energy. This, in turn, implied that the amplitude at each harmonic of the cumulants may be expressed in terms of the speech energy and number of harmonics, which led to more manageable expressions and to closed-form equations for the skewness and kurtosis. If the flat spectrum assumption is no longer true, then the autoconvolution can no longer be expressed in terms of the signal energy but will be a combination of the amplitude at the various signal harmonics. It is easy to see, however, that the overall behavior of these cumulants will not be significantly changed, specifically, the following.

- Zero-phase characteristic of the cumulant slices still holds and as a result the skewness and kurtosis still represent the maximum values of these functions.
- Harmonic nature of the cumulant slices remains unchanged, though the harmonic magnitudes may be different.
- Skewness and kurtosis are still positive but may not be expressed in terms of the speech energy. Instead, their value is a nonclosed-form function of the speech amplitudes at the various harmonics. As a result, the normalization by the energy (in the expressions of γ_3 and γ_4) will no longer cancel out the effect of energy in the numerator and the normalized metrics are no longer independent of signal amplitude as is the case in (16) and (21).

F. Results Using Speech Signals

The derivations presented above are verified using recorded clean speech. A tenth-order LPC analysis is performed on speech sampled at 8 kHz. The normalized third- and fourth-order cumulant functions of the LPC residual are computed using frames of 20 ms (160 samples) with a 25% overlap. The residual is low-pass filtered using a 60-tap FIR filter with a cutoff at 1.8 kHz. In order to avoid the problem of additional harmonics reported in [22], the cumulant slices are computed as

$$C_3[\tau] = \frac{1}{N-K}\sum_{n=K}^{N-1} x^2(n)x(n-\tau)$$

$$C_4[\tau] = \left[\frac{1}{N-K}\sum_{n=0}^{N-K} x^2(n)x^2(n+\tau)\right] - \left[\frac{1}{N}\sum_{n=0}^{N-1} x^2(n)\right]^2 - 2\left[\frac{1}{N-K}\sum_{n=0}^{N-K} x(n)x(n+\tau)\right]^2$$

where K is the maximum lag and N is the number of points in the frame. The idea is to keep the limit of the summation constant for all lags τ. Furthermore, the algorithm proposed in [14] may be used for fast computation of $C_3[\tau]$. The third- and fourth-order cumulants are computed for the range of lags [0, 80].

1) Voiced Speech: The waveform for the utterance "help the" spoken by a male speaker is shown in Fig. 2. Mild Gaussian noise is added (at 30 dB SNR) to avoid zero level signals. The normalized skewness and kurtosis of the LPC residual are shown in Fig. 3 for each of the frames, and it is observed that both entities are greater than zero for voiced segments, as expected. It is also worth noting that the two normalized metrics take on large positive (kurtosis) or negative (skewness) values for transient and small amplitude segments (for example at frame 30 in Fig. 3). This is mainly due to the

[1]By evaluating the moment integral of the Laplacian pdf for orders 2 and 4.

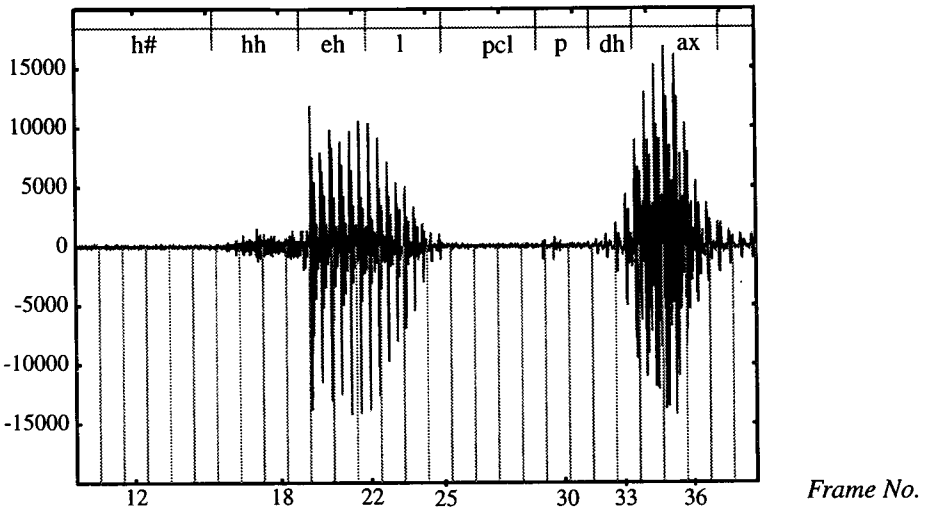

Fig. 2. Utterance "help the" by a male speaker.

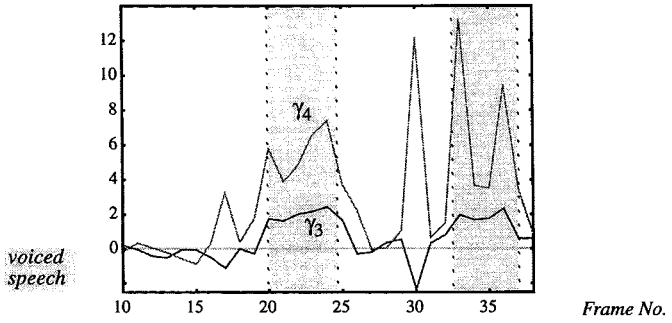

Fig. 3. Normalized skewness and kurtosis of the LPC residual.

small energy of these frames that is used for normalization, thus resulting in large normalized HOS for any transient segment, i.e., when only some of the samples are nonzero. Consequently, the normalized metrics by themselves are not sufficient for detecting voiced frames as they take on erroneously large values for small transient segments.

To better illustrate the distribution of the kurtosis, a histogram of the frame-by-frame values of the normalized kurtosis is generated for 500 frames. Another histogram is generated for the normalized kurtosis when Gaussian noise is used instead of speech prior to LPC filtering. These histograms are shown in Fig. 4 and illustrate the difference in the fourth-order statistics between speech and Gaussian noise. It is important to note here that the speech utterance contains silence periods and, thus, the kurtosis would be zero some of the time as evident from the histograms.

The second-, third-, and fourth-order cumulant slices are evaluated for a range of possible pitch lags. The third and fourth-order slices are normalized by the signal variance (12). Fig. 5 compares the normalized $C_2[\tau]$ (the autocorrelation) with the normalized $C_3[\tau]$ and $C_4[\tau]$ slices for the case of two consecutive voiced frames (20 and 21 in Fig. 2). As may be seen from Fig. 5, all three functions have a maximum at the pitch lag in addition to having zero phase (i.e., a maximum at lag 0).

2) Unvoiced Speech: To properly analyze the HOS of unvoiced speech, two sustained fricatives, namely /f/ and /h/ are recorded for a few seconds and their LPC residual used for computing the skewness and kurtosis. Histograms for the entities were then deduced and compared to those computed for the case of Gaussian noise. Comparison of the two sets of histograms suggests that the LPC residual of sustained unvoiced speech has a Gaussian-like nature since its HOS are zero. However, since in reality unvoiced speech occurs in small segments and often at transitional boundaries, it is expected that its HOS are nonzero. This phenomena is confirmed by simulation (for example frames 18 and 19 in Fig. 3) and is in agreement with the experimental findings in [3] about the nonzero normalized HOS of transitional speech segments.

IV. Voice Activity Detection and Voicing Classificaiton Using HOS

A. Rationale

It was shown in the previous sections that the skewness and kurtosis of the LPC residual of voiced speech can be expressed in terms of the number of harmonics M and signal energy and are greater than zero for any practical value of M (which is a function of pitch). The normalized statistics may be expressed in terms of M only (16), (21). This is clearly distinct from the case of Gaussian noise, where both of these entities are zero. It seems sensible then to make use of these two statistics as one way to detect voicing. The advantage of using the normalized metrics is that they are independent of the signal energy and therefore absolute thresholds may be used. However, when using normalized statistics, one has to account for the effect of noise (30), (31). Alternatively, one may consider the variance of the estimators of the skewness and kurtosis and normalize the computed entities to yield unit-variance estimators. Another point worth exploiting is the relation between the skewness and kurtosis for voiced speech and its use in identifying speech frames. These two ideas are further detailed below and are the basis of the proposed VAD algorithm.

B. Soft Detection of Noise Frames

The skewness and kurtosis of Gaussian noise are zero only in a statistical average sense. Since in practice finite length frames

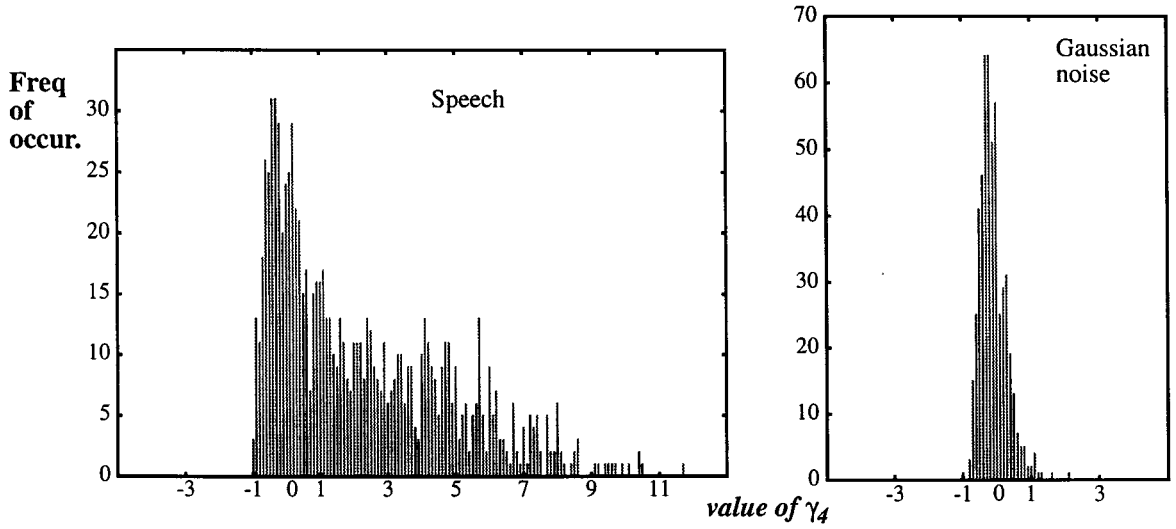

Fig. 4. Histograms of normalized kurtosis of the LPC residual (speech versus Gaussian).

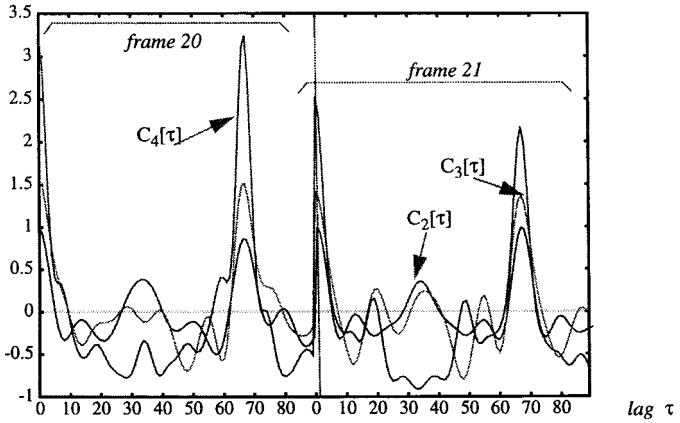

Fig. 5. Normalized $C_2[\tau]$, $C_3[\tau]$, and $C_4[\tau]$ for frames 20 and 21.

are used, the decision that a given frame is noise can only be made in a probabilistic manner with a confidence level that takes into account the variance and distribution of the estimators of the skewness and kurtosis. Given a Gaussian process $g(n)$, the estimators of the second, third, and fourth-order moments are

$$M_{kg} = \frac{1}{N}\sum_{n=0}^{N-1}[g(n)]^k \quad \text{estimator for } E[\{x(n)\}]^k \tag{32}$$

for $k = 2$, 3, and 4. In [13], it is shown that these estimators are unbiased and, for the case of white Gaussian noise, their mean and variance may be expressed in terms of the process variance, υ_g

$$E[M_{3g}] = 0 \quad \text{and} \quad E[M_{4g}] = 3\upsilon_g^2$$
$$\text{Var}[M_{3g}] = \frac{15\upsilon_g^3}{N} \quad \text{and} \quad \text{Var}[M_{4g}] = \frac{96\upsilon_g^4}{N}. \tag{33}$$

As a result, the estimator of the skewness $\widehat{SK} = M_{3g}$ is unbiased, with zero mean and known variance. Since this estimator is the sum of a large number of independent identically distributed (iid) random variables, then by the central limit theorem [9], the following normalized version of it:

$$\widehat{SK}_a = \frac{M_{3g}}{\sqrt{15\upsilon_g^3/N}} \tag{34}$$

is a Gaussian variable with zero mean and unit variance. Therefore, given the estimate of the skewness of a given frame and the corresponding scaled value denoted by "a," one can find the probability that the frame is Gaussian noise as

$$\text{Prob}[Noise] = \text{Prob}[|\widehat{SK}_a| \geq a]. \tag{35}$$

Graphically, this is equivalent to computing the area under the tail of the Gaussian curve of $\widehat{SK}_a$. Clearly when $a = 0$ the area is unity. The area under the tail of the curve can be evaluated using the erfc(x) function. For example, when $a > 0$, $\text{Prob}[Noise] = 2/\sqrt{2\pi}\int_a^\infty e^{-x^2/2}\,dx$. Thus, $\text{Prob}[Noise] = \text{erfc}(|a|)$.

It is worth noting here that while the HOS of speech are positive, a negative skewness is not necessarily an indication of noise, since transient segments can have negative HOS as was mentioned before.

Similarly, the estimator of the kurtosis is first computed from the second and fourth-order moments. To ensure an unbiased estimate, the modified estimator proposed in [13] is used

$$\widehat{KU}_U = \left(1 + \frac{2}{N}\right)M_{4g} - 3(M_{2g})^2. \tag{36}$$

This estimator is unbiased, with zero mean and known variance given in [13]. The distribution of this estimator is not straightforward, since it consists of the difference of two variables, one Gaussian and one chi-square. However, an approximation

is used here and the estimator is assumed normally distributed.[2] A unit-variance version of this zero-mean variable is defined as

$$\widehat{KU}_{Ua} = \frac{\widehat{KU}_U}{\sqrt{\frac{3\upsilon_g^4}{N}\left(104 + \frac{452}{N} + \frac{596}{N^2}\right)}}. \tag{37}$$

Therefore, given the value of the estimate of the kurtosis of a given frame and the corresponding scaled value, denoted by "b," the probability that the frame is noise is: $\text{Prob}[Noise] = \text{erfc}(|b|)$.

The discussion so far pointed out that given the estimate of the skewness and kurtosis, one can determine the probability of a frame being noise using the normalized values of these estimates and the "erfc()" function. Moreover, it is assumed that the true variance of the noise (υ_g) is known *a priori*. In reality, this is not the case but one has only a (hopefully good) estimate of the noise energy, which is estimated during frames declared nonspeech. This estimate is not equal to the true variance, but is relatively good compared to the estimates of the skewness and kurtosis, which are only deduced from a short data frame.

C. Necessary Condition for Voicing

The skewness and kurtosis of voiced speech are expressed in terms of energy and number of harmonics and may be used for detecting voiced frames. In order to eliminate the effect of energy, one may consider the normalized metrics γ_3 and γ_4, but in the presence of noise, these metrics become less effective for detecting voiced frames (30), (31). Alternatively, the ratio of the appropriate power of the skewness to that of the kurtosis may be considered as one way of eliminating the effect of signal energy in (15) and (20), while avoiding the effect of noise. Consider the ratio

$$\text{SKR} = \frac{\text{skewness}^2}{\text{kurtosis}^{1.5}} = \frac{9(M-1)^2}{8M\left[\frac{4}{3}M - 4 + \frac{7}{6M}\right]^{1.5}} \tag{38}$$

which is independent of signal energy and is only a function of M. A plot of SKR reveals that for a practical range of M $(M \geq 4)$, the ratio is confined to the range $[0 \cdots 1]$, and thus this condition is necessary for classifying a frame as a voiced frame. When only Gaussian noise is present, the ratio is undetermined since both operands are zero. In reality, this zero condition never occurs due to the variance of the estimators. However, the ratio may take on any value, including the range for voiced speech; for this reason, this is a necessary but not sufficient condition for detecting voiced frames.

D. HOS-Based VAD Algorithm

Since sustained unvoiced speech is shown to have Gaussian-like characteristics, it cannot be distinguished from Gaussian noise using HOS. However, as discussed earlier, this is not the case in reality where unvoiced speech occurs at speech transitional boundaries which have nonzero HOS. Therefore the VAD detection proposed here may be based on HOS and can be formulated as a finite two-state machine. The algorithm proposed combines the use of skewness, kurtosis, their normalized versions γ_3 and γ_4, SNR, LPC prediction error, and SKR ratio, for distinguishing speech from noise frames.

1) Data Format: Speech sampled at 8 kHz is used. A tenth-order LPC analysis is performed once every 20 ms, thus generating a 20 ms residual. Voice activity detection is carried out every 10 ms using the residual and a 20% overlap (i.e., 80 new points are combined from 20 from the past iteration).

2) HOS Computations: Every 10 ms iteration, the estimators for the second, third, and fourth-order moments are computed using (32) with $N = 100$. An autoregressive scheme is used to smooth the estimates of the moments. From these, the unbiased estimate of the kurtosis (36) is deduced. The estimate of the skewness is simply the third-order moment (32). The two metrics are then normalized by the signal energy to yield

$$\gamma_3 = \frac{\widehat{SK}}{M_{2x}^{1.5}} \quad \text{and} \quad \gamma_4 = \frac{\widehat{KU}_U}{M_{2x}^2}. \tag{39}$$

3) Noise and SNR Estimation: The noise power is estimated using frames declared as nonspeech. Moreover, it is assumed that the first three frames are nonspeech and are used to initialize the noise power estimate. Whenever a frame is declared as nonspeech, its energy is used to update the noise estimate according to an autoregressive averaging

$$\tilde{\upsilon}_g(k) = (1-\beta)\tilde{\upsilon}_g(k-1) + \beta M_{2X} \tag{40}$$

where

k	iteration index;
M_{2X}	frame energy;
$\tilde{\upsilon}_g$	estimate of the noise energy;
β	$= 0.1 \cdot \text{Prob}[Noise]$.

At every iteration, the current estimate of the noise energy is used to compute the SNR of that frame

$$\text{SNR} = Pos\left[\frac{M_{2X}}{\tilde{\upsilon}_g} - 1\right] \tag{41}$$

where $Pos[x] = x$ for $x > 0$ and 0 otherwise. Since the residual is low-pass filtered at 2 kHz, the above SNR is applicable to the lower spectrum only. Using a similar reasoning a "total SNR" metric is computed using the nonfiltered residual and the energy of the full band.

4) Probability of Noise-Only Frames: Once the skewness and kurtosis are computed, the variance of these estimates are computed using the noise energy $\tilde{\upsilon}_g$, according to (34) and (37), to yield the zero-mean, unit variance estimates $\widehat{SK}_a$ and $\widehat{KU}_{Ua}$, respectively. From these two scaled values, the probability of the frame being noise is deduced

$$\text{Prob}[Noise] = [\text{erfc}(a) + \text{erfc}(b)]/2 \tag{42}$$

where a and b are the computed values of $\widehat{SK}_a$ and $\widehat{KU}_{Ua}$, respectively.

5) SKR Ratio: The ratio is computed directly from the nonnormalized estimates of the skewness and kurtosis

$$\text{SKR} = \frac{[\widehat{SK}]^2}{[\widehat{KU}_U]^{1.5}}. \tag{43}$$

[2]This assumption is verified to be reasonable by simulation (e.g., Fig. 4).

6) LPC Prediction Error: The LPC prediction error is the inverse of the prediction gain and may be computed from the set of the reflection coefficients (r_i) generated by the LPC analysis

$$PE = \prod_{i=0}^{10}(1 - r_i^2). \tag{44}$$

The error is typically small for voiced speech and thus it is used here as an added metric, along with the SKR ratio, for detecting voiced speech segments.

7) Speech/Noise State Machine: The VAD algorithm is implemented as a two-state machine (Fig. 6). The following operations are carried out in each state.

- **Noise State**

  The noise energy is updated according to the $\mathrm{Prob}[Noise]$ (42). The SKR ratio, the Gaussian likelihood, the SNR (41) and the PE (44) values are used to determine whether the frame is speech. The occurrence of any of the following three conditions triggers a transition:

  1) $\mathrm{Prob}[Noise] < T_{Gaus}$ for two consecutive frames.
  2) SKR in voicing range and (SNR $> T_{\mathrm{SNR1}}$ or $PE < T_{PE}$) (an indication of a voiced frame).
  3) $Total$ SNR $> T_{\mathrm{SNR2}}$ (strong speech frame).

- **Speech State**

  The noise likelihood (42) along with the values of γ_3 and γ_4 (39) are used to determine whether the frame is Gaussian. After a hangover period (two to three frames), transition to the noise state occurs

$$\text{If } \{\mathrm{Prob}[Noise] > T_{Gaus} \quad \text{and} \quad \gamma_3 < T_{\gamma_3} \quad \text{and} \quad \gamma_4 < T_{\gamma_4}\}.$$

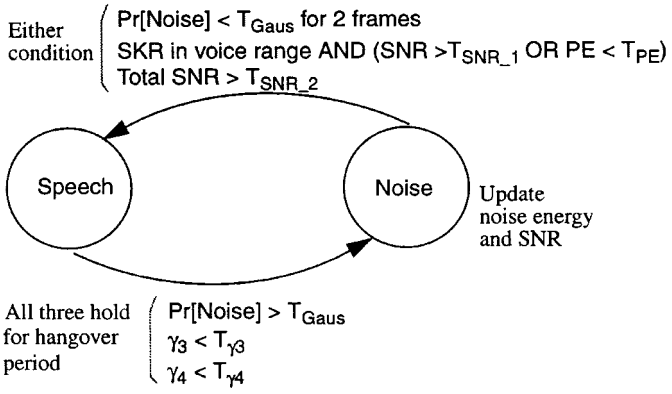

Fig. 6. HOS-based VAD state machine.

V. Experimental Results

To evaluate the effectiveness of the HOS-based VAD, we calculated the probability of correct and false detection for a number of noisy speech scenarios. To obtain these two metrics and to generate the noisy speech, we used the TIA database [27] recommended for the evaluation of VAD algorithms in cellular systems using discontinuous transmission. The database contains clean speech files, the corresponding truth marker files, and various noise files. The speech material consists of ten data files spoken by five male and five female speakers. Each speech file is the recording of one side of a two-way phone conversation lasting about 11 min. The noise material consists of four noise recordings from a commuter train, a parking garage, a street corner, and a car interior. Both speech and noise files have been preprocessed by a modified-IRS filtering to emulate the transfer function of the telephone network. Both speech and noise files have been normalized to -26 dBov, defined as the level relative to a 16-bit saturation.

The database also contains a utility to generate eighty test cases by various mixing of the speech and noise files. Each case is a different combination of the speech normalization level, the noise type and the SNR. Four SNR levels are used: ∞ dB, 18 dB, 12 dB, and 6 dB. The SNR value is computed as the ratio of the total energy of speech to that of the noise over the entire utterance. To assess the performance of the proposed algorithm, we defined the following metrics.

- Pc_{Speech}: Probability of correctly detecting speech frames. Computed as the ratio of correct speech detections to the total number of hand-labeled speech frames.
- Pc_{Noise}: Probability of correctly detecting noise frames. Computed as the ratio of correct noise detections to the total number of hand-labeled noise frames.
- Pf: Probability of false detection. Computed as the ratio of incorrectly classified speech or noise frames to the total number of frames.

In addition to computing the above metrics, marker files are generated and compared to the truth markers in order to visually inspect the performance. For each of the 80 test cases, the Pc's and Pf's of the proposed VAD are computed and compared to those of G.729B [1]. The results from all the cases involving the same noise type and SNR level are averaged and summarized in one row of Table III.

The results show that the proposed VAD has an overall better performance than G.729B in all SNRs and all the noise types used here, as evident by a lower probability of false detection (Pf) and a higher probability of correct speech detection (Pc_{Speech}). The difference in performance is particularly noticeable in moderate to low SNR (12 dB and 6 dB) and in street and garage noise. These noise types are likely to be Gaussian, being the aggregate of many independent sources. In these types, the probability of correct noise detection (Pc_{Noise}) is also higher at all SNRs for the HOS-based VAD, demonstrating that the metrics used, such as the normalized third and fourth-order statistics, are effective noise discriminators regardless of the noise energy levels. The case of street noise at 12 dB and a female speaker is shown in Fig. 7. The difference in classification is particularly noted during the nonspeech segment where G.729B has a rather erratic behavior and results in incorrect oscillations between states.

In all noise types and SNRs, including clean speech, the probability of correct speech detection (Pc_{Speech}) is higher for the proposed VAD. The improvement ranges from two to ten percentage points. This fact demonstrates that the combined second- and higher-order metrics used for speech detection, such as the SKR ratio, the Gaussian likelihood, the low-band and full-band SNR, and the LPC prediction error, are effective discriminators of speech in all noise types, even in

TABLE III
Pc'S AND Pf'S FOR THE HOS-BASED AND G.729B VAD

Noise Environment		Pc Speech (%)		Pc Noise (%)		Pf (%)	
Type	SNR	HOS VAD	G.729B	HOS VAD	G.729B	HOS VAD	G.729B
Street	18 dB	95.5	88.8	81.4	74.9	12.5	19.3
	12 dB	88.5	80.9	86.8	77.9	12.7	21.1
	6 dB	85.3	74.8	90.8	79.3	12.5	22.5
Garage	18 dB	96.8	91.4	87.2	81.6	8.8	14.2
	12 dB	85.6	77.8	90.1	80.9	11.8	20.4
	6 dB	81.1	68.7	90.4	79.8	13.5	24.9
Car	18 dB	93.6	87.7	85.9	90.0	9.8	11.2
	12 dB	88.4	79.8	90.9	92.4	9.8	11.5
	6 dB	79.9	76.5	94.3	92.5	9.9	11.9
Train	18 dB	91.4	87.3	89.4	90.0	9.7	11.1
	12 dB	88.6	84.3	90.9	91.4	9.8	10.6
	6 dB	66.1	61.9	90.2	91.9	22.3	23.6
None	inf dB	99.9	99.9	53.9	46.3	22.2	25.8

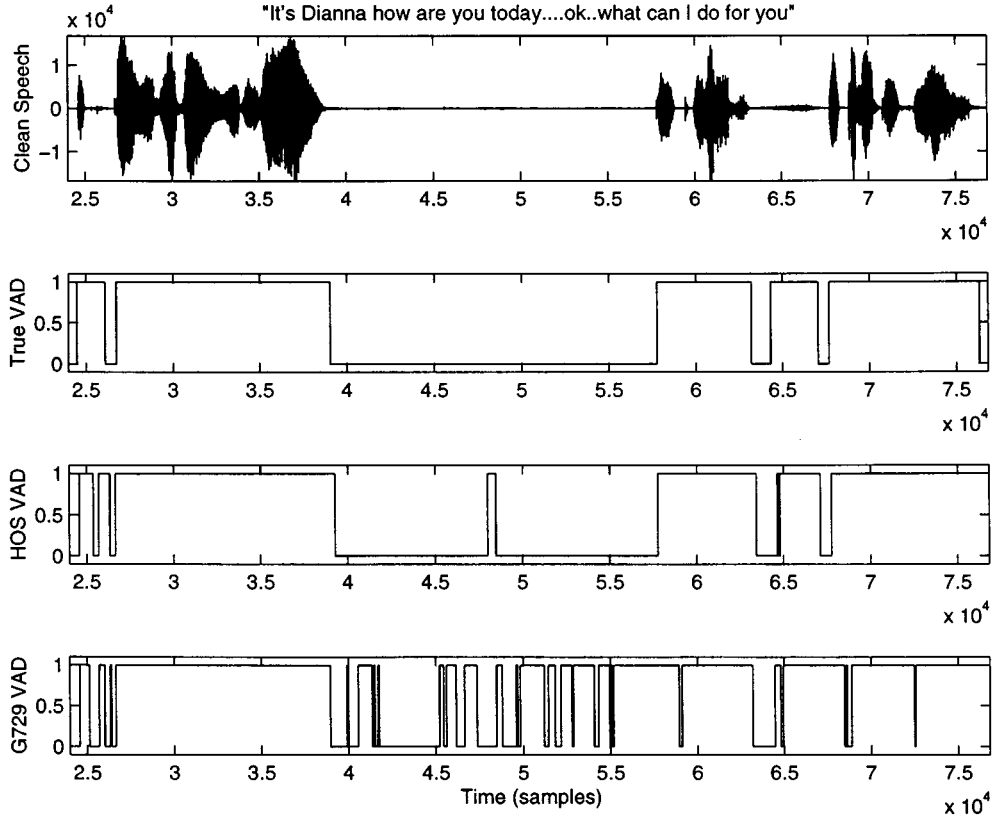

Fig. 7. HOS-based and G.729B VAD in street noise conditions (12 dB).

non-Gaussian noise or in no noise at all. This result further validates the HOS properties of speech derived in Sections III-B1, III-C1, and IV-C.

In periodic noise types, such as car and train noise, the probability of correct noise detection (Pc_{Noise}) is somewhat lower, though marginally, for the HOS-based VAD. This is due to the fact that the noise in this case has a dominant periodic component and thus a nonzero HOS. The reason for the HOS-based decisions being biased toward speech in non-Gaussian conditions may be explained by considering the state machine transitions (Fig. 6): a non-Gaussian noise will result in a low noise likelihood measure which will delay or inhibit the transition from the speech to the noise states, resulting in more speech classifications. The consequences of this behavior however are not as

severe as if the problem occurred on the other transition, since then it will result in falsely classifying speech as noise and this has more detrimental consequences.

VI. Conclusion

The objective of this paper is to exploit the properties of higher-order statistics for the sake of finding a robust algorithm for voice activity detection in the presence of noise. To this end, we studied the third and fourth-order cumulants of speech in the LPC residual domain, and unveiled important properties about these cumulants whose relevance goes beyond the goal of a VAD application.

- Horizontal slices of the third and fourth-order cumulants of the LPC residual of voiced speech have the same periodicity as the underlying speech. Their harmonic amplitudes may be expressed in terms of the speech energy and the number of harmonics.
- These cumulant slices have zero phase and reach maxima at multiples of the pitch lag and, as such, may be used for pitch estimation in a similar manner as the autocorrelation function.
- Skewness and kurtosis of voiced speech are nonzero and may be used as a basis for speech detection or voicing classification. When normalized by the appropriate power of the signal energy, these metrics are independent of signal levels, which makes them convenient as detectors since absolute thresholds may be used.
- Average value of the horizontal slice of the fourth-order cumulant of the LPC residual of voiced speech may be written in terms of the signal energy and the number of harmonics.
- Ratio of the appropriate powers of the skewness to that of the kurtosis of voiced speech is independent of signal energy and is confined to a small range for any practical range of the pitch.
- Unvoiced speech in the LPC residual may not be modeled as a harmonic process but rather as a general white process.

Unlike other reported work in the area of HOS for speech, a more fundamental approach is taken here whereby analytical derivations were first deduced based on a speech model, thus providing a basis for justifying or refuting the experimental findings. The rationale for considering the LPC residual is its flat spectral envelope which makes the HOC derivations for speech more tractable and allows quantifying the bias and variance of the HOS estimators for Gaussian noise.

Simulations using actual speech signals demonstrated that the derivations and the underlying speech model are valid for voiced speech, but that sustained unvoiced speech has a Gaussian-like nature unlike the hypothesis of the sinusoidal model. The skewness and kurtosis of voiced speech are shown analytically and experimentally to be nonzero. Moreover, the nature of unvoiced speech, occurring in short segments at transitional boundaries, results in nonzero HOS in practical situations. These are the pivotal concepts on which an HOS-based VAD algorithm is developed. The relation between the two metrics is used as a condition for an improved detection of speech frames. Moreover, the variance of the HOS estimators is derived for the case of Gaussian noise and is used to quantify the likelihood of a given frame being noise. The resulting algorithm combines HOS metrics with second-order measures, such as low-band and full-band SNR and the LPC prediction error, to classify frames into one of the two states.

Compared to G.729B, the proposed algorithm is based on a more analytical framework, is conceptually simpler and uses a smaller parameter set, making it is easier to tune. It is also appropriate for use in conjunction with speech coders where such parameters as the LSFs are not available.

The performance in noise of the two algorithms shows the HOS-based VAD has superior performance to G.729B in terms of a lower probability of false classification and a higher probability of correct speech classification. This fact suggests that HOS-based methods have promising potential in yielding VAD algorithms that would significantly advance the current state of the art. The work however does not claim these statistics to be superior in and by themselves to second-order approaches, but rather that they provide additional information about the signal that is immune to the presence of noise, and that makes them particularly effective in low SNR applications. Clearly, successful algorithms are those that can combine the two approaches and harness the advantages of both.

Future work in this area includes investigating the combination of more metrics and tuning the algorithm with speech recorded in more diverse noise environments. Other applications of the theoretical findings, particularly pitch estimation, are also under consideration.

Acknowledgment

The authors are grateful to Dr. S. Coulombe, Prof. D. O'Shaughnessy, and all the reviewers for valuable suggestions on this work.

References

[1] A. Benyassine, E. Shlomot, and H. Su, "ITU-T recommendation G.729, annex B, a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data spplications," *IEEE Commun. Mag.*, pp. 64–72, Sept. 1997.

[2] J. Fackrell, "Bispectral analysis of speech signals," Ph.D. dissertation, Univ. Edinburgh, Edinburgh, U.K., 1996.

[3] A. Falaschi and I. Tidei, "Speech innovation characterization by higher-order moments," in *Visual Representation of Speech Signals*, M. Cooke, S. Beet, and M. Crawford, Eds. New York: Wiley, 1993.

[4] R. Fulchiero and A. Spanias, "Speech enhancement using the bispectrum," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 4, 1993, pp. 488–491.

[5] 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. Acoustics, Speech, Signal Processing*, May 1989, pp. 369–372.

[6] G. Gabor and Z. Gyorfi, "On the higher order distributions of speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 602–603, Apr. 1988.

[7] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstral features," in *IEEE TEN-CON*, 1993, pp. 321–324.

[8] J. D. Hoyt and H. Wechsler, "Detection of human speech in structured noise," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, May 1994, pp. 237–240.

[9] A. Leon-Garcia, *Probability and Random Process for Electrical Engineering.* Reading, MA: Addison-Wesley, 1989.

[10] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, p. 744, Aug. 1986.
[11] J. Mendel, "Tutorial on higher-order statistics in signal processing and system theory: Theoretical results and some applications," *Proc. IEEE*, vol. 79, pp. 278–305, Mar. 1991.
[12] A. Moreno and J. Fonollosa, "Pitch determination of noisy speech using higher order statistics," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, 1992, pp. 133–136.
[13] E. Nemer, "Speech analysis and quality enhancement using higher-order cumulants," Ph.D. dissertation, Carleton Univ., Ottawa, ON, Canada, 1999.
[14] E. Nemer, R. Goubran, and S. Mahmoud, "An efficient algorithm for computing the triple correlation," in *IEEE Canadian Conf. Electrical Computer Engineering*, vol. 1, May 1997, pp. 19–22.
[15] C. Nikias and J. Mendel, "Signal processing with higher-order statistics," *IEEE Trans. Signal Processing*, vol. 41, pp. 10–38, Jan. 1993.
[16] C. Nikias and M. Raghuveer, "Bispectrum estimation: A digital signal processing framework," *Proc. IEEE*, vol. 75, pp. 869–891, July 1987.
[17] K. Paliwal and M. Sondhi, "Recognition of noisy speech using cumulant-based linear prediction analysis," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1991, pp. 429–432.
[18] L. R. Rabiner and M. R. Sambur, "Voiced–unvoiced-silence detection using the Itakura LPC distance measure," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, May 1977, pp. 323–326.
[19] V. Ramamoorty and T. Ericson, "Speech coding based on a composite-Gaussian source model," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1979, pp. 534–537.
[20] V. Ramamoorty, "Voice/unvoiced detection based on a composite-Gaussian source model," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1980, pp. 57–60.
[21] M. Rangoussi and G. Carayannis, "Higher-order statistics based Gaussianity test applied to on-line speech processing," in *Proc. IEEE Asilomar Conf*, 1995, p. 303.
[22] D. Ruiz, M. Carrion, A. Gallego, and A. Medouri, "Parameter estimation of exponentially damped sinusoids using a higher correlation-based method," *IEEE Trans. Signal Processing*, vol. 43, pp. 2665–2667, Nov. 1995.
[23] H. Teager and S. Teager, "Evidence for nonlinear sound production in the vocal tract," in *Speech Production and Speech Modeling*, W. J. Hardcastle and A. Marchal, Eds. Norwell, MA: Kluwer, 1990, pp. 241–261.
[24] R. Tucker, "Voice activity detection using a periodicity measure," *Proc. Inst. Elect. Eng.*, vol. 139, no. 4, pp. 377–380, Aug. 1992.
[25] B. Wells, "Voiced/unvoiced decision based on the bispectrum," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1985, pp. 1589–1592.
[26] *Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital systems*, TIA doc. PN-3292, Jan. 1996.
[27] *TDMA minimum performance standards for discontinuous transmission operation of mobile stations*, TIA doc. and database IS-727, June 1998.

Elias Nemer (S'91–M'95) received the B.Eng (EE), M. Eng (EE), and M.B.A. degrees from McGill University, Montreal, QC, Canada, in 1985, 1989, and 1993, respectively, and the Ph.D. degree in systems and computer engineering from Carleton University, Ottawa, ON, Canada, in 1999.

From 1986 to 1990, he was with CAE Electronics, Montreal, as a hardware DSP Designer for real-time flight simulators. From 1993 to 1995, he was involved in the system design of wireless data networks at Bell Northern Research. From 1996 to 2000, he was with the speech research group at Nortel Networks, Montreal, where he lead various research projects in speech processing for cellular telephony. Since 2000, he has been a Senior Member of Staff with the Networking Division, Intel, San Jose, CA, and a System Architect for advanced broadband cable communications. His research interests include digital signal processing, speech enhancement, speech coding, voice over IP, and wideband communications.

Rafik A. Goubran (S'85–M'85) was born in Cairo, Egypt, in 1955. He received the B.Sc. and M.Sc. degrees in electrical engineering from the Department of Electronics and Communication Engineering, Cairo University, in 1978 and 1981, respectively, and the Ph.D. degree in electrical engineering from the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada, in 1986.

In January 1987, he joined the Department of Systems and Computer Engineering, Carleton University, where he is now Associate Professor and Chair. He has acted as a Consultant to several industrial and government organizations, including Nortel, Mitel, Bell Canada, Vienna Systems (Nokia), Revenue Canada, the Department of National Defense (DND), Bota Teleconferencing, Matcom, the National Research Council of Canada (NRC), and Data Measurement Corporation. His research interests include digital signal processing (DSP) and its applications in acoustics, speech processing, communications, and analytical chemistry, voice transmission over IP (VoIP), and ATM networks. His current research projects deal with audio quality improvement in telephony, audio teleconferencing, acoustic echo and noise cancellation, adaptive filter structures, beamforming using microphone arrays, and narcotics detection using ion mobility spectrometry. Other interests include mobile communications, digital systems design, DSP hardware, multiprocessor architectures for DSP, and computer architecture.

Dr. Goubran is a member of the Association of Professional Engineers of Ontario.

Samy Mahmoud (M'83–SM'83) received the the M.Eng and Ph.D degrees in electrical engineering from Carleton University, Ottawa, ON, Canada, in 1971 and 1974, respectively.

He joined the Faculty of Engineering, Carleton University, in 1975 where he served as the Chair of the Department of Systems and Computer Engineering from 1988 to 1996. At present, he is the Dean of the Faculty of Engineering and Design. His main academic and professional interests are in the areas of mobile and personal communication systems, broadband networks, and distributed computing systems. He has directed several large R&D projects in these areas, involving joint university/industry collaborations. He has recently led a major initiative to establish the National Capital Institute of Telecommunications (NCIT), a joint research organization involving several large international companies in the telecommunications and computer industries, leading university researchers, and scientists and engineers from two major Canadian Government research laboratories (CRC and NRC). The new Institute has been launched with major financial contributions form the Government of Canada, the Province of Ontario, and the industrial partners. In the past 20 years, he has he published more than 150 archival and conference papers in the areas of wireless communications and broadband networks and supervised over 70 masters theses and doctoral dissertations. He holds senior consulting appointments with major international companies in the field of telecommunications.

Dr. Mahmoud was a co-guest editor of two issues of the IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS in the area of wireless and satellite communications. He has won several recognition awards in recent years for pioneering research work leading to university/industry technology transfer. He is a co-recipient of the Stentor 1993 National Telecommunications Award, and co-recipient of IEEE TRANSACTIONS ON VEHICULAR COMMUNICATIONS Best Paper Award in 1994.