HIGHER ORDER STATISTICS BASED GAUSSIANITY TEST APPLIED TO ON-LINE SPEECH PROCESING

Maria Rangoussi
Dept. of Electrical Engineering
Computer Science Division
National Technical University of Athens
Athens GR-15780, GREECE

George Carayannis
Institute for Language and
Speech Processing
22, Margari str.
Athens GR-11525, GREECE

Abstract

Detection of speech in noisy recordings becomes a challenging problem when the noise does not follow the usual whiteness, stationarity and high signal-to-noise ratio assumptions. A robust speech detector can affect significantly the performance of several speech processing tasks, such as endpoint detection, segmentation, and finally recognition, if we deal with real life data, as opposed to laboratory or controlled environment recordings. The detector proposed in this paper is based on a Gaussianity test that employs third-order cumulants of the data to decide on the binary hypotheses of noise only versus speech plus noise. Speech intervals are detected by exploiting the third-order information present in the speech signal. The detector can handle a large family of additive noises, thanks to its third-order statistics basis. The sample-adaptive and decision feedback variations proposed here, provide the detector with tracking ability both with respect to the time variations of speech and the possible non-stationarity of noise. Experiments carried out using real data, recorded in a moving car interior, show satisfactory performance of the proposed algorithms down to -6dB signal-to-noise ratio.

1 Introduction

Recordings of speech consist of alternating intervals of voice and silence. In practice, background noise is superimposed on both of them. Given a segment of such a recording, a speech detector should reliably decide whether it consists of noise only, or, of speech plus noise. If we are not interested in a more detailed decision as to the nature of the speech signal detected, then this is a binary decision problem.

Depending on the recording environment, the speech signal-to-noise ratio (SNR) can reach zero or negative dB levels, the noise can be colored, and the noise statistics may be unknown and/or time-varying. Recordings made in a moving car interior for mobile communications, in busy work rooms or in rooms with rotating machinery, provide such examples. Detection of speech in noise becomes a challenging problem when dealing with real-life noises as those just described.

Under ideal conditions, detection of speech in silence would assume the trivial form of a comparison of a given data segment with zero. In practice, however, because of estimation errors and noise, we have to resort to a binary hypotheses test to decide between silence (*null hypothesis* $\mathcal{H}_0$: absence of speech signal, noise only present) and voice (*alternative hypothesis* $\mathcal{H}_1$: both speech signal and noise present).

Conventional methods distinguish between voice and silence (noise in practice) by thresholding easily computable features such as the short time energy or the zero crossings count, [6]. Other approaches combine these features with parameters based on linear prediction coefficients (LPC), [1], or compute measures of spectral flatness, [8]. Although computationally attractive, these methods rely heavily on the assumptions that the SNR is high (around 60dB, in practice) and the noise is white and stationary.

In the present paper the speech detection problem is addressed in the domain of the *third-order cumulants* of the data, where a variety of additive noises is (at least) theoretically rejected. The detector employed is developed in [4] as a (non-)Gaussianity test in the time domain. It exploits the ability of third-order statistics to measure departure from Gaussianity by assuming theoretically zero values for Gaussian or generally symmetrically distributed (non-skewed) linear processes, and non-zero values for non-Gaussian or skewed ones. The use of this detector in the present speech detection context is prompted by the fact that third-order cumulants of voice assume significantly non-zero values, while a variety of real-life noises is non-skewed enough to be practically rejected.

Frame-adaptive and sample-adaptive algorithms to implement the test in [4] are developed here, which are appropriate for on-line applications. The proposed family of tests renders the time-domain testing of (non-)Gaussianity applicable to environments that are non-stationary with respect to both the information signal and the noise. In addition, these tests maintain an asymptotically *constant false alarms rate* because of the normal asymptotic distribution of the third-order cumulants [2], which in turn results in a central χ^2 asymptotic distribution of the test statistic. The latter allows for automatic selection of the test threshold and minimizes the heuristic elements that are inevitable in existing approaches such as [6].

1058-6393/95 $4.00

2 Third-order cumulants of speech

Let us assume that the discrete-time observed signal $y(n)$ is composed of a speech component $s(n)$ plus an additive, zero-mean, random noise component $v(n)$, which is independent of $s(n)$; then $y(n) = s(n) + v(n)$. $s(n)$ is deterministic, almost periodic for voiced speech and random for unvoiced speech, thus rendering $y(n)$ a (non-)stationary, mixed spectrum process. In order to treat both cases in a common framework, we adopt the approach of [2], where cumulants are defined through *generalized* (or *time-averaged*) moments, as follows (third order case):

$$c_{3y}(\tau_1,\tau_2) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathcal{E}\{y(n)y(n+\tau_1)y(n+\tau_2)\} \quad (1)$$

where $c_{3y}(\tau_1,\tau_2)$ denotes the third-order cumulant of $y(n)$ at lag (τ_1,τ_2). Additivity in the signal domain carries over to the third-order cumulant domain, namely, under the assumption that $\lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} s(n) = 0$, it holds that $c_{3y}(\tau_1,\tau_2) = c_{3s}(\tau_1,\tau_2) + c_{3v}(\tau_1,\tau_2)$, (see [2], Section 2.2, Example 3). If $v(n)$ is Gaussian or non-skewed, it then holds that $c_{3v}(\tau_1,\tau_2) \equiv 0$, resulting in $c_{3y}(\tau_1,\tau_2) \equiv c_{3s}(\tau_1,\tau_2)$, which justifies the claim made in the introduction about noise rejection in the third-order cumulant domain.

Sample estimators of $c_{3y}(\tau_1,\tau_2)$, $0 \le \tau_2 \le \tau_1 \le N-1$ can be constructed as

$$\hat{c}_{3y}(\tau_1,\tau_2) = \frac{1}{N}\sum_{n=0}^{N-1-\tau_1} y(n)y(n+\tau_1)y(n+\tau_2) \quad (2)$$

Their mean square consistency and asymptotic normality are proved in [2] under *mixing conditions* for the signal and the noise.

Consequently, third-order statistics provide a domain where non-skewed noise superimposed on a skewed signal is rejected asymptotically in N. This property becomes useful for a noisy speech signal due to the fact that real (as opposed to synthetic) speech is skewed enough to yield significantly non-zero third-order cumulants, [7], [8]. This is mainly due to non-linearities detected in the vocal tract, [5], [10]. Those of them that follow the square law result in quadratic cross- and self-couplings of the dominant frequencies of voiced speech, - an explanation that leaves open for further research the unvoiced speech case. As an indicative example, figure 1 (top) shows the first 50 *diagonal* lags $\hat{c}_{3y}(\tau,\tau)$, $0 \le \tau \le 49$, indexed by lag τ, (i) of a noise-free voiced speech segment (/i/) of 1000 samples, (continuous line); (ii) of a real-life car noise segment of equal length (dashed line); and (iii) of their superposition at 0dB SNR (dash-dotted line). The noisy speech and the clear speech cumulant values almost coincide, thus showing that the asymptotic noise rejection in the third-order cumulant domain is obtained in practice for reasonable data lengths, even under real noise conditions.

3 The time-domain Gaussianity test

We have seen that voice can in principle be detected if one or more third-order cumulant lags are found to be non-zero. For our purposes, it has been sufficient in practice to test the set of the Q non-redundant lags closer to the origin, given collectively in the vector $\mathbf{c}_{3y} = \{ c_{3y}(\tau_1,\tau_2),\ 0 \le \tau_2 \le \tau_1 \le q \}$, where $Q = (q+1)(q+2)/2$ is on the order of 10 or 20. Then the alternative hypotheses are $\mathcal{H}_0 : \mathbf{c}_{3y} = \mathbf{0}$ and $\mathcal{H}_1 : \mathbf{c}_{3y} = \mathbf{c}_{3s} \neq \mathbf{0}$.

The asymptotic normality result of [2] for the estimator of eq. (2) states that as $N \to \infty$,

$$\sqrt{N}\,(\hat{\mathbf{c}}_{3y} - \mathbf{c}_{3y}) \overset{distr}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \quad (3)$$

where $\mathbf{\Sigma}$ is the asymptotic covariance matrix of $\hat{\mathbf{c}}_{3y}$, $\mathbf{\Sigma} = \lim_{N\to\infty} N\,cum\{\hat{\mathbf{c}}_{3y}, \hat{\mathbf{c}}_{3y}\} = \lim_{N\to\infty} N\mathcal{E}\{(\hat{\mathbf{c}}_{3y} - \mathbf{c}_{3y})\,(\hat{\mathbf{c}}_{3y} - \mathbf{c}_{3y})^t\}$ and $(\,)^t$ denotes transposition. For N large, $\mathbf{\Sigma} \approx N\,\mathbf{C}$, where the covariance matrix $\mathbf{C} = \mathcal{E}\{(\hat{\mathbf{c}}_{3y} - \mathbf{c}_{3y})\,(\hat{\mathbf{c}}_{3y} - \mathbf{c}_{3y})^t\}$ clearly depends on the sample length N. Additionally, if $\tau = (\tau_1,\tau_2)$, $\rho = (\rho_1,\rho_2)$, then the (τ,ρ) entry of matrix $\mathbf{\Sigma}$ is given by

$$\mathbf{\Sigma}(\tau,\rho) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \sum_{\xi=-\infty}^{\infty} cum\{\, y(n)y(n+\tau_1)\, y(n+\tau_2),\ y(n+\xi)y(n+\xi+\rho_1)y(n+\xi+\rho_2)\} \quad (4)$$

and can be consistently estimated as $\hat{\mathbf{\Sigma}}(\tau,\rho)$ from a single record of data, ([2] Theorem 2.1, Section 3.3).

Consequently, the two alternative hypotheses amount to testing the (non-)zeroness of the mean between two asymptotically normal distributions:

$$\begin{aligned} \mathcal{H}_0 &: \quad \sqrt{N}\,\hat{\mathbf{c}}_{3y} \quad \overset{distr}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_0) \\ \mathcal{H}_1 &: \sqrt{N}\,(\hat{\mathbf{c}}_{3y} - \mathbf{c}_{3y}) \overset{distr}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_1), \end{aligned} \quad (5)$$

where $\mathbf{\Sigma}_i, i = 0,1$ is the asymptotic covariance under $\mathcal{H}_i, i = 0,1$.

For stationary signals, it is proved in [4] that the quadratic form

$$\hat{d} = \hat{\mathbf{c}}_{3y}^t\, \hat{\mathbf{C}}_0^{-1}\, \hat{\mathbf{c}}_{3y} = N\, \hat{\mathbf{c}}_{3y}^t\, \hat{\mathbf{\Sigma}}_0^{-1}\, \hat{\mathbf{c}}_{3y} \quad (6)$$

asymptotically follows a central χ^2 distribution with Q degrees of freedom, denoted by χ^2_Q, under $\mathcal{H}_0$. Thanks to the *generalized* asymptotic normality result of [2], quoted above, the same holds true for $\hat{d}$ computed under the present noisy speech set-up, as well. Therefore, the threshold $\mathcal{T}$ of the test is obtained from the χ^2_Q table, after fixing a *probability of false alarms*, α, and the test takes the form

$$\hat{d} \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \mathcal{T} = \chi^2_Q(\alpha). \quad (7)$$

4 Adaptive speech detection

Under the form described in the previous section, the test can assign a given segment of a recording to noise only or to speech. When applied to an open-end recording, this test should produce decisions at a desirable frequency, so as to track the speech present in the data. The simplest way to implement this is to use a fixed-length frame of, say, L samples, and slide it along the time axis on the data, either in an overlapping or in a non-overlapping way. For each frame position, the data in the analysis frame are tested and one decision is made. The overlap between two successive frame positions controls the decision frequency of the test in an obvious way: we can have one decision per L samples (no overlap) up to one decision per sample ($L-1$ samples overlap). The test statistic $\hat{d}^{(l)}$, indexed by the frame number l, is computed on the basis of (i) the asymptotic covariance matrix $\hat{\mathbf{\Sigma}}_0$ under the null hypothesis $\mathcal{H}_0$, and (ii) the third-order cumulants vector $\hat{\mathbf{c}}_{3y}$, estimated from the data in the current frame. This is the *frame-adaptive* version of the algorithm, the step of which are described below.

FRAME-ADAPTIVE SPEECH DETECTOR

Initialization:

step I.1: Select the values of Q, α, and $\mathcal{T} = \chi^2_Q(\alpha)$.

step I.2: Use the $\mathcal{N}_0$ initial noise-only samples of the data, $\{y(n)\}_{n=0}^{\mathcal{N}_0-1}$, to estimate the covariance matrix of $\hat{\mathbf{c}}_{3y}$ under $\mathcal{H}_0$, $\hat{\mathbf{C}}_0$.

step I.3: Compute the (pseudo-) inverse $\hat{\mathbf{P}}_0 = \hat{\mathbf{C}}_0^{-1}$.

Repeat for $l = 1, 2, 3, \ldots$:

step R.1: Estimate $\hat{\mathbf{c}}_{3y}^{(l)}$ on the basis of the data in the l-th frame $\mathbf{y}^{(l)} = [\, y(\mathcal{N}_0 + (l-1)(L/\mathcal{L}) + 1) \cdots y(\mathcal{N}_0 + (l-1)(L/\mathcal{L}) + L)\,]$, where factor $\mathcal{L}$ controls the overlap ratio between two successive frames (e.g., no overlap for $\mathcal{L} = 1$).

step R.2: Compute the quadratic form $\hat{d}^{(l)} = [\, \hat{\mathbf{c}}_{3y}^{(l)} \,]^t \; \hat{\mathbf{P}}_0 \; [\, \hat{\mathbf{c}}_{3y}^{(l)} \,]$.

step R.3: Set $b(l) = 1$ if $\hat{d}^{(l)}$ exceeds $\mathcal{T}$ or $b(l) = 0$ otherwise.

This algorithm produces the binary decision sequence $b(l)$, $l = 1, 2, \ldots$, indexed by the frame number l.

In step **I.2**, the estimation of $\hat{\mathbf{C}}_0$ requires an initial stretch of noise-only data, $\{y(n)\}_{n=0}^{\mathcal{N}_0-1}$, which are segmented into, say, R non-overlapping records of length L each; then $\mathcal{N}_0 = R \cdot L$ and

$$\hat{\mathbf{C}}_0 = (1/R) \sum_{r=1}^{R} (\hat{\mathbf{c}}_{3y}^{(r)} - \bar{\mathbf{c}}_{3y}) (\hat{\mathbf{c}}_{3y}^{(r)} - \bar{\mathbf{c}}_{3y})^t, \quad (8)$$

where $\bar{\mathbf{c}}_{3y} = (1/R) \sum_{r=1}^{R} \hat{\mathbf{c}}_{3y}^{(r)}$. Alternatively, $\hat{\mathbf{C}}_0$ can be consistently estimated from the single data record $\{y(n)\}_{n=0}^{\mathcal{N}_0-1}$, as described in [3], or [2], Section 3.3.

The use of frames that slide on the data in a more or less overlapping fashion is a somewhat awkward attempt to track the time-varying nature of the speech signal. A much more "natural" approach would be to update $\hat{\mathbf{c}}_{3y}$ with every new sample $y(n)$, while using an appropriate multiplicative frame on the data to smoothly weight out the past in favor of more recent observations.

This idea prompts the use of the following sample-adaptive estimator proposed in [9],

$$\hat{\mathbf{c}}_{3y}^{(n)} = \hat{\mathbf{c}}_{3y}^{(n-1)} + (1-\lambda) \, [\, \mathbf{y}_3^{(n)} - \hat{\mathbf{c}}_{3y}^{(n-1)} \,]. \quad (9)$$

The entries of the $Q \times 1$ vector $\mathbf{y}_3^{(n)}$ are of the form $y_3^{(n)}(\tau_1, \tau_2)$, $0 \le \tau_2 \le \tau_1 \le q$, where the triple product $y_3^{(n)}(\tau_1, \tau_2) = y(n-q)y(n-q+\tau_1)y(n-q+\tau_2)$ is an instantaneous approximation of $c_{3y}(\tau_1, \tau_2)$ at time n. If we insert this iteration in eq. (6), we obtain

$$\hat{d}^{(n)} = [\lambda \hat{\mathbf{c}}_{3y}^{(n-1)} + (1-\lambda)\mathbf{y}_3^{(n)}]^t \hat{\mathbf{C}}_0^{-1} [\lambda \hat{\mathbf{c}}_{3y}^{(n-1)} + (1-\lambda)\mathbf{y}_3^{(n)}]. \quad (10)$$

This iteration forms the basis of the sample-adaptive detection algorithm, outlined below.

SAMPLE-ADAPTIVE SPEECH DETECTOR

Initialization :

steps I.1 to I.3 : As in the frame adaptive version.

step I.4 : Initialize $\hat{\mathbf{c}}_{3y}^{(q)} = \mathbf{0}$ and $\hat{d}^{(q)} = 0$, set $\lambda < 1$.

Repeat for $n = q+1, q+2, q+3, \ldots$:

step R.1: Update $\hat{\mathbf{c}}_{3y}^{(n-1)}$ into $\hat{\mathbf{c}}_{3y}^{(n)}$, as in eq. (9).

step R.2: Update $\hat{d}^{(n-1)}$ into $\hat{d}^{(n)}$, as in eq. (10).

step R.3: Set $b(n) = 1$ if $\hat{d}^{(n)}$ exceeds $\mathcal{T}$ or $b(n) = 0$ otherwise.

This algorithm produces the decision sequence $b(n)$, $n = q+1, q+2, \ldots$, indexed by the sample number n.

The factor λ employed in eq. (9) is in fact a *forgetting factor* that limits the memory of the adaptation, to gain a certain tracking ability. Its value must be set close to but less than 1. Eq. (9) that employs the forgetting factor λ is the constant gain form of an analogous update that uses decreasing gains, and specifically $(1/n)$ in lieu of $(1-\lambda)$. The latter converges in the mean square sense to the ensemble vector $\mathbf{c}_{3y}$, because for each n it is algebraically equal to the corresponding batch estimator of eq. (2), applied to the data set $[\, y(1) \ldots y(n) \,]$, which is known to be m.s.s. consistent, [2]. Instead of converging, the constant gain version will exhibit a random residual error, which accounts for its tracking ability when applied on non-stationary data, as it is the case here. Due to the quadratic form of eq. (10), the convergence / tracking ability comments made for $\hat{\mathbf{c}}_{3y}^{(n)}$ carry over smoothly to $\hat{d}^{(n)}$, provided that the inverse covariance matrix $\mathbf{P}_0$ is known and remains constant.

5 Adaptation with decision feedback

The adaptive algorithms proposed in the previous section make use of the inverse covariance matrix of the noise, $\mathbf{P}_0 = [\mathbf{C}_0]^{-1}$, which is estimated from an initial stretch of noise-only data and used as a constant thereafter. This procedure is computationally advantageous; however, (i) the performance of the test thereafter depends on the quality of this estimate and (ii) the applicability of the test is limited to stationary noise conditions. If the noise is non-stationary, the initial estimate of $[\mathbf{C}_0]^{-1}$ will soon become "out of date", and the performance will drop, because the asymptotic χ^2 result of eq. (6) will become invalid. It is therefore meaningful to try to improve this initial estimate by exploiting noise-only stretches that occur in the data after the initial part used for training, through a decision feedback scheme. This modification is possible both for the frame- and the sample-adaptive forms of the detector and it offers the (non-)Gaussianity test of [4] the novelty of being applicable to non-stationary environments. Indeed, in the latter case, and if the time variation of the noise statistics is slow relative to the algorithm's time constant, decision feedback can significantly improve the performance of the algorithms in section 4, at the cost of increased computations.

In order to introduce decision feedback in the frame-adaptive detector, $\hat{\mathbf{C}}_0$ and $\bar{\mathbf{c}}_{3y}$ obtained as in eq. (8) on the basis of the first R frames of L samples each, and denoted as $\hat{\mathbf{C}}_0^{(R)}$ and $\bar{\mathbf{c}}_{3y}^{(R)}$ hereafter, are used to initialize an update recursion for $\hat{\mathbf{C}}_0$. This recursion is triggered every time a frame is assigned to noise by the test, i.e., every time $b(l) = 0$, to produce a fresh estimate $\hat{\mathbf{C}}_0^{(m)}, m = R+1, R+2, \ldots$. It uses the data in this frame to update first the third-order cumulant vector $\hat{\mathbf{c}}_{3y}^{(m)}$ and then $\bar{\mathbf{c}}_{3y}^{(m)} = (m-1)/m\, \bar{\mathbf{c}}_{3y}^{(m-1)} + 1/m\, \hat{\mathbf{c}}_{3y}^{(m)}$, and use the latter in

$$\hat{\mathbf{C}}_0^{(m)} = \tfrac{m-1}{m}\hat{\mathbf{C}}_0^{(m-1)} + \tfrac{1}{(m-1)}[\hat{\mathbf{c}}_{3y}^{(m)} - \bar{\mathbf{c}}_{3y}^{(m)}]\,[\hat{\mathbf{c}}_{3y}^{(m)} - \bar{\mathbf{c}}_{3y}^{(m)}]^t. \tag{11}$$

Consistency in the mean square sense can be established for these adaptations under stationary conditions, by relating them to their "batch" counterparts.

As it is $\hat{\mathbf{P}}_0$ rather than $\hat{\mathbf{C}}_0$ that is used to compute the quadratic form $\hat{d}$, a matrix (pseudo-) inversion should follow each update, to obtain $\hat{\mathbf{P}}_0^{(m)} = [\hat{\mathbf{C}}_0^{(m)}]^{-1}$. This increase of the computational cost can be avoided if we exploit the rank-1 update nature of eq. (11) along with the matrix inversion lemma, through which we can directly update $\hat{\mathbf{P}}_0^{(m)}$, instead of first updating $\hat{\mathbf{C}}_0^{(m)}$ and then inverting. This is possible through the recursion

$$\hat{\mathbf{P}}_0^{(m)} = \tfrac{m}{m-1}\hat{\mathbf{P}}_0^{(m-1)} - \frac{\frac{m}{m-1}\hat{\mathbf{P}}_0^{(m-1)}[\hat{\mathbf{c}}_{3y}^{(m)} - \bar{\mathbf{c}}_{3y}^{(m)}][\hat{\mathbf{c}}_{3y}^{(m)} - \bar{\mathbf{c}}_{3y}^{(m)}]^t\hat{\mathbf{P}}_0^{(m-1)}}{(\frac{m-1}{m})^2 + [\hat{\mathbf{c}}_{3y}^{(m)} - \bar{\mathbf{c}}_{3y}^{(m)}]^t\hat{\mathbf{P}}_0^{(m-1)}[\hat{\mathbf{c}}_{3y}^{(m)} - \bar{\mathbf{c}}_{3y}^{(m)}]} \tag{12}$$

initialized by $\hat{\mathbf{P}}_0^{(0)} = [\hat{\mathbf{C}}_0^{(0)}]^{-1}$, where $\bar{\mathbf{c}}_{3y}^{(m)}$ is as above. Again here the adaptations can be made more alert to changes if the decreasing gain $(m-1)/m$ is replaced by a constant gain λ, set to a value less than but close to 1. A practical algorithm that incorporates the decision feedback modification into the frame-adaptive detector is described below.

DECISION FEEDBACK DETECTOR

Initialization :

steps I.1 to I.3: As in the frame-adaptive detector.

step I.4 : Initialize $\bar{\mathbf{c}}_{3y}^{(0)} = \bar{\mathbf{c}}_{3y}$, as in eq. (8), set $\lambda < 1, m = 0$.

Repeat for $l = 1, 2, 3, \ldots$:

step R.1: As in the frame-adaptive detector.

step R.2: Compute the quadratic form $\hat{d}^{(l)} = [\,\hat{\mathbf{c}}_{3y}^{(l)}\,]^t\ \hat{\mathbf{P}}_0^{(m)}\ [\,\hat{\mathbf{c}}_{3y}^{(l)}\,]$.

step R.3: Set $b(l) = 1$ if $\hat{d}^{(l)}$ exceeds $\mathcal{T}$ or $b(l) = 0$ otherwise.

step R.4: If $b(l) = 0$ then increase m by 1 and update $\hat{\mathbf{P}}_0^{(m)}$ using the current estimate $\hat{\mathbf{c}}_{3y}^{(l)}$, as in eq. (12) (decreasing gain $(m-1)/m$ or constant gain λ).

This algorithm produces the sequence of decisions $b(l)$, $l = 1, 2, 3, \ldots$, indexed by the frame number l.

The introduction of the decision feedback scheme in the sample-adaptive detector is obtained in an analogous way and will be omitted due to lack of space, with the comment that (i) $\hat{\mathbf{c}}_{3y}^{(n)}$ and $\hat{d}^{(n)}$ are updated with every new sample as in eqs. (9) and (10), while (ii) the noise statistics update, analogous to eq. (12) with constant gain, is triggered only when noise-only data are detected. Different gains λ, μ are therefore possible for the two adaptations.

6 Experimental results

The performance of the proposed algorithms is tested on real data, recorded in a moving car interior. In figure 1 (bottom) we show the noisy speech recording at SNR = -6 dB. The results of the frame-adaptive detector, when applied on the data of figure 1 with $Q = 10$, $L = 256$ and overlap ratio $\mathcal{L} = 7/8$, are given in figure 2. The upper part shows the test statistic $d^{(l)}$, indexed by the frame number l along with the automatically chosen threshold for $\alpha = 1\%$, while the lower part shows the binary decision result $b(l)$, superimposed on the noise-free speech signal, for better visualization of the results. Figure 3 shows the corresponding results for the sample-adaptive detector with $\lambda = 0.999$, indexed by the sample number n. It can be seen that both the proposed algorithms correctly discern speech from noise-only areas in the noisy data. Finally, figure 4 shows the effect of decision feedback on the frame-adaptive detector, applied on another noisy speech signal at SNR = 3 dB (top: clear speech). The spikes on the test statistic that produce false alarms (middle) are clearly reduced when decision feedback is used (bottom).

7 Conclusion

A family of speech detection algorithms based on the third-order statistics of speech is proposed in this paper. They employ a form of the time-domain Gaussianity test of [4] to test the binary hypotheses of noise-only versus speech plus noise. They maintain an asymptotically constant false alarms rate, allowing for automatic threshold setting, while they maintain satisfactory performance down to SNR = -6 dB. The adaptive forms developed, along with the decision feedback scheme employed, render the proposed algorithms appropriate for on-line operation and for non-stationary environments.

Acknowledgement

The authors wish to thank MATRA Communication, France, for the real car data used in the experiments; these are the property of MATRA within the Freetel/ESPRIT project.

References

[1] B.S.Atal, L.R.Rabiner, "A pattern recognition approach to V-U-S classification with applications to speech recognition," *IEEE Trans. on ASSP*, vol. 24, No 3, pp. 201-212, June 1976.

[2] A.V.Dandawaté, G.B.Giannakis, "Asymptotic theory of mixed time averages and k-th order cyclic moment and cumulant statistics," *IEEE Trans. Info Theory*, 1995, (to appear).

[3] B.Friedlander, B. Porat, "Asymptotically optimal estimation of MA and ARMA parameters of non-Gaussian processes from high-order moments," *IEEE Trans. on Automatic Control*, vol. 35, No 1, pp. 27-35, Jan. 1990.

[4] G.B.Giannakis, M.K.Tsatsanis, "Time-domain tests for Gaussianity and time reversibility," *IEEE Trans. on Signal Processing*, Dec. 1994 (to appear).

[5] P.Maragos, J.F.Kaiser, and T.F.Quatieri, "Energy separation in Signal Modulations with application to speech analysis," *IEEE Trans. on Signal Proc.*, vol. 41, No 10, pp. 3024-3051, Oct. 1993.

[6] L.R.Rabiner, M.R.Sambur, "Voiced - Unvoiced - Silence detection using the Itakura LPC distance measure," *IEEE Proc. ICASSP*, pp. 323-326, May 1977.

[7] M.Rangoussi, A.Delopoulos, M.K.Tsatsanis, "On the use of higher order statistics for robust endpoint detection of speech," *IEEE Proc. Workshop on HOS*, pp. 56-60, CA, USA, June 1993.

[8] M.Rangoussi, S.Bakamidis, G.Carayannis, "On the use of SVD and high-order statistics for robust endpoint detection of speech," in *Levels in Speech Communication: interactions and relations*, Ed. J.Schoentgen et al., Elsevier, Brussels, 1994.

[9] A. Swami, J. Mendel, "Adaptive system identification using cumulants," *IEEE Proc. ICASSP'88*, vol. 4, pp. 2248-2251.

[10] H.M.Teager and S.M.Teager, "Evidence of nonlinear sound production mechanisms in the vocal tract," In *Speech production and speech modeling*, NATO ASI, Ser. D, vol. 55, Bonas, France, July 1989; Kluwer Academic, 1990.

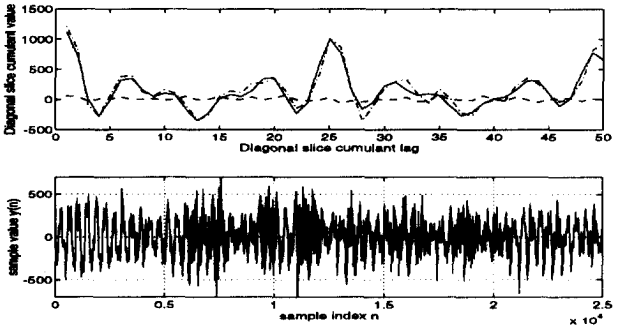

Figure 1: *Top:* Third-order cumulants of speech. *Bottom:* Noisy recorded signal y(n), SNR = -6dB.

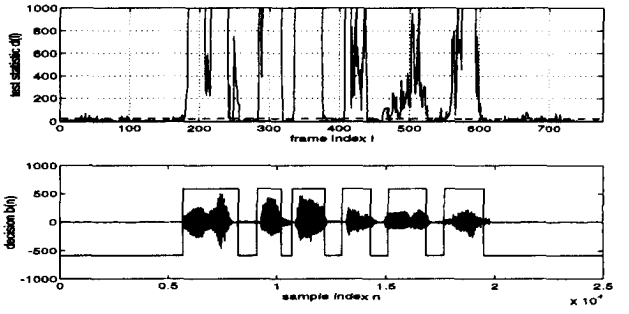

Figure 2: Frame-adaptive detector results.

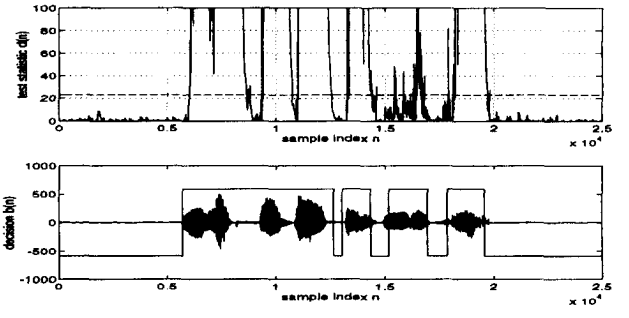

Figure 3: Sample-adaptive detector results.

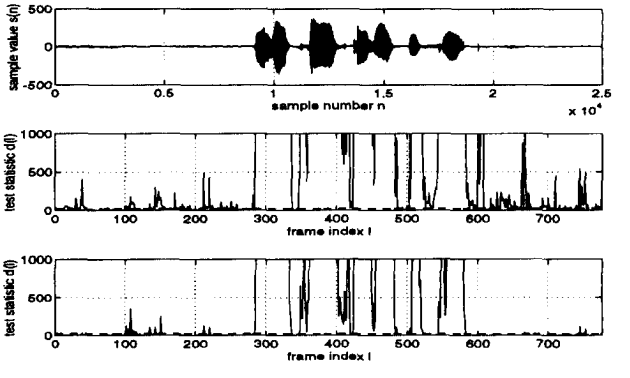

Figure 4: Frame-adaptive detector, clear speech (top), test statistic without (middle) and with (bottom) decision feedback.