Statistical voice activity detection based on integrated bispectrum likelihood ratio tests for robust speech recognition

J. Ramírez,[a] J. M. Górriz, and J. C. Segura
Deptarment of Signal Theory, Networking and Communications, University of Granada, Granada, Spain

(Received 20 November 2006; revised 12 February 2007; accepted 13 February 2007)

Currently, there are technology barriers inhibiting speech processing systems that work in extremely noisy conditions from meeting the demands of modern applications. These systems often require a noise reduction system working in combination with a precise voice activity detector (VAD). This paper shows statistical likelihood ratio tests formulated in terms of the integrated bispectrum of the noisy signal. The integrated bispectrum is defined as a cross spectrum between the signal and its square, and therefore a function of a single frequency variable. It inherits the ability of higher order statistics to detect signals in noise with many other additional advantages: (i) Its computation as a cross spectrum leads to significant computational savings, and (ii) the variance of the estimator is of the same order as that of the power spectrum estimator. The proposed approach incorporates contextual information to the decision rule, a strategy that has reported significant benefits for robust speech recognition applications. The proposed VAD is compared to the G.729, adaptive multirate, and advanced front-end standards as well as recently reported algorithms showing a sustained advantage in speech/nonspeech detection accuracy and speech recognition performance.
 [DOI: 10.1121/1.2714915]

PACS number(s): 43.72.Pf, 43.72.Dv [EJS] Pages: 2946–2958

I. INTRODUCTION

The emerging applications of speech technologies (particularly in mobile communications, robust speech recognition, or digital hearing aid devices) often require a noise reduction scheme working in combination with a precise voice activity detector (VAD).[1] During the last decade numerous researchers have studied different strategies for detecting speech in noise and the influence of the VAD decision on speech processing systems.[2–8] This task can be identified as a statistical hypothesis testing problem and its purpose is the determination to which category or class a given signal belongs. The decision is made based on an observation vector, frequently called feature vector, which serves as the input to a decision rule that assigns a sample vector to one of the given classes. The classification task is often not as trivial as it appears since the increasing level of background noise degrades the classifier effectiveness and causes numerous detection errors.[9,10]

The nonspeech detection algorithm is an important and sensitive part of most of the existing single-microphone noise reduction schemes. Well-known noise suppression algorithms[11,12] such as Wiener filtering (WF) or spectral subtraction, are widely used for robust speech recognition being the VAD critical in attaining a high level of performance. These techniques estimate the noise spectrum during nonspeech periods in order to compensate for the harmful effect of the noise on the speech signal. The VAD is even more critical for nonstationary noise environments since the statistics of the background noise must be updated. An example of such a system is the ETSI standard for distributed speech recognition that incorporates noise suppression methods. The so-called advanced front-end (AFE)[13] considers an energy-based VAD in order to estimate the noise spectrum for Wiener filtering and a different VAD for nonspeech frame dropping (FD).

On the other hand, a VAD achieves silence compression in modern mobile telecommunication systems reducing the average bit rate by using the discontinuous transmission mode. Many practical applications, such as the Global System for Mobile Communications (GSM) telephony, use silence detection and comfort noise injection for higher coding efficiency. The International Telecommunication Union (ITU) adopted a toll-quality speech coding algorithm known as G.729 to work in combination with a VAD module in DTX mode. The recommendation G.729 Annex B[4] uses a feature vector consisting of the linear prediction spectrum, the full-band energy, the low-band (0–1 kHz) energy, and the zero-crossing rate. Another standard for DTX is the ETSI (adaptive multirate) AMR speech coder[3] developed by the Special Mobile Group for the GSM system. The AMR standard specifies two options for the VAD to be used within the digital cellular telecommunications system. In option 1, the signal is passed through a filterbank and the subband energies are calculated. The VAD decision depends on a measure of the signal-to-noise ratio, the output of a pitch detector, a tone detector, and the correlated complex signal analysis module. An enhanced version of the original VAD is the AMR option 2 that uses parameters of the speech encoder that are more robust to the environmental noise than AMR1 and G.729. These VADs have been used extensively in the

[a]Electronic mail: javierrp@ugr.es

 0001-4966/2007/121(5)/2946/13/$23.00

open literature as a reference for assessing the performance of new algorithms.

Most of the algorithms for detecting the presence of speech in a noisy signal only exploit the power spectral content of the signals and require knowledge of the noise power spectral density.[6,8,14,15] One of the most important disadvantages of these approaches is that no *a priori* information about the statistical properties of the signals is used. Higher order statistics methods rely on an *a priori* knowledge of the input processes and have been considered for VAD since they can distinguish between Gaussian signals (which have a vanishing bispectrum) from non-Gaussian signals. However, the main limitations of bispectrum-based techniques are that they are computationally expensive and the variance of the bispectrum estimators is much higher than that of power spectral estimators for identical data record size. These problems were addressed by Tugnait,[16,17] who showed a computationally efficient and reduced variance statistical test based on the integrated polyspectra for detecting a stationary, non-Gaussian signal in Gaussian noise. This paper shows an effective VAD based on a likelihood ratio test (LRT) defined on the integrated bispectrum of the noisy speech. The proposed approach also incorporates contextual information to the decision rule, a strategy first proposed in Ref. 18 that has reported significant benefits for different applications including robust speech recognition.[19–22] The paper includes a careful derivation of the LRT previously addressed in Refs. 23 and 24, an alternative approach based on block partitioning and averaging of the integrated bispectra, and its efficient implementation using the contextual LRT. The paper is organized as follows. Section II reviews the definition and fundamental properties of third-order cumulants and bispectrum. Section III suggests the use of integrated bispectrum for a reduced variance estimation while maintaining the benefits of higher order statistics for detection. Section IV shows the definition of the VAD based on contextual integrated bispectrum LRTs. Section V shows two different methods for integrated bispectrum estimation and LRT definition and their analysis and comparison is shown in Sec. VI. Section VII shows the receiver operating characteristic (ROC) curves and speech recognition experiments that are used to evaluate the proposed method and to compare its performance to ITU-T G.729, ETSI AMR and AFE, as well as to other recently reported VADs. Finally, Sec. VIII summarizes the conclusions of this work.

II. BISPECTRUM

The bispectrum of a deterministic, continuous-time signal $x(t)$ is defined as[25,26]

$$B(\omega_1,\omega_2)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}C_{3x}(\tau_1,\tau_2)\times\exp\{-j(\omega_1\tau_1+\omega_2\tau_2)\}d\tau_1 d\tau_2, \quad (1)$$

where

$$C_{3x}(\tau_1,\tau_2)=E\{x^*(t)x(t+\tau_1)x(t+\tau_2)\}=\int_{-\infty}^{+\infty}x^*(t)x(t+\tau_1)x(t+\tau_2)dt \quad (2)$$

is the third-order cumulant of $x(t)$, and $\omega=2\pi f$ with normalized frequency f. By the symmetry properties, the bispectrum of a real signal is uniquely defined by its values in the triangular region $0\leqslant\omega_2\leqslant\omega_1\leqslant\omega_1+\omega_2\leqslant\pi$, provided that there is no bispectral alias. In a similar fashion, the bispectrum of a discrete-time signal is defined as the two-dimensional (2D) Fourier transform:

$$B_x(\omega_1,\omega_2)=\sum_{i=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}C_{3x}(i,k)\exp\{-j(\omega_1 i+\omega_2 k)\}. \quad (3)$$

Note that from the above-presented definition the third-order cumulant can be expressed as

$$C_{3x}(i,k)=\frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}B_x(\omega_1,\omega_2)\times\exp\{j(\omega_1 i+\omega_2 k)\}d\omega_1 d\omega_2 \quad (4)$$

using the inverse Fourier transform.

Figure 1 shows the differences between cumulants and bispectrum function (magnitude and phase) for a frame containing noise only [Fig. 1(a)] and speech in noise [Fig. 1(b)]. Both data sets were extracted from an utterance of the Spanish SpeechDat-Car database[27] with uniform noise conditions so that the noise level in both data sets is the same. It can be clearly concluded that higher order statistics or polyspectra provide discriminative features for speech/nonspeech classification.[24] Even the bispectrum phase exhibits a more random behavior during nonspeech periods so that the phase entropy also provides complimentary information for VAD in noise environments.[28]

Although bispectrum methods have all the advantages of cumulants/polyspectra, their direct use has two serious limitations: (i) The computation of bispectra in the whole triangular region is huge, and (ii) the 2D template matching score in the classification is impractical. To efficiently use bispectra, integrated bispectrum methods[16,17] were proposed for different applications.[29,30]

III. INTEGRATED BISPECTRUM

Let $x(t)$ be a zero mean stationary random process. If we define $\tilde{y}(t)=x^2(t)-E\{x^2(t)\}$, the cross correlation between $\tilde{y}(t)$ and $x(t)$ is defined to be

$$r_{\tilde{y}x}(k)=E\{\tilde{y}(t)x(t+k)\}=E\{x^2(t)x(t+k)\}=C_{3x}(0,k) \quad (5)$$

so that its cross spectrum is given by

$$S_{\tilde{y}x}(\omega)=\sum_{-\infty}^{\infty}C_{3x}(0,k)\exp\{-j\omega k\} \quad (6)$$

and

$$C_{3x}(0,k)=\frac{1}{2\pi}\int_{-\pi}^{\pi}S_{\tilde{y}x}(\omega)\exp\{j(\omega_k)\}d\omega. \quad (7)$$

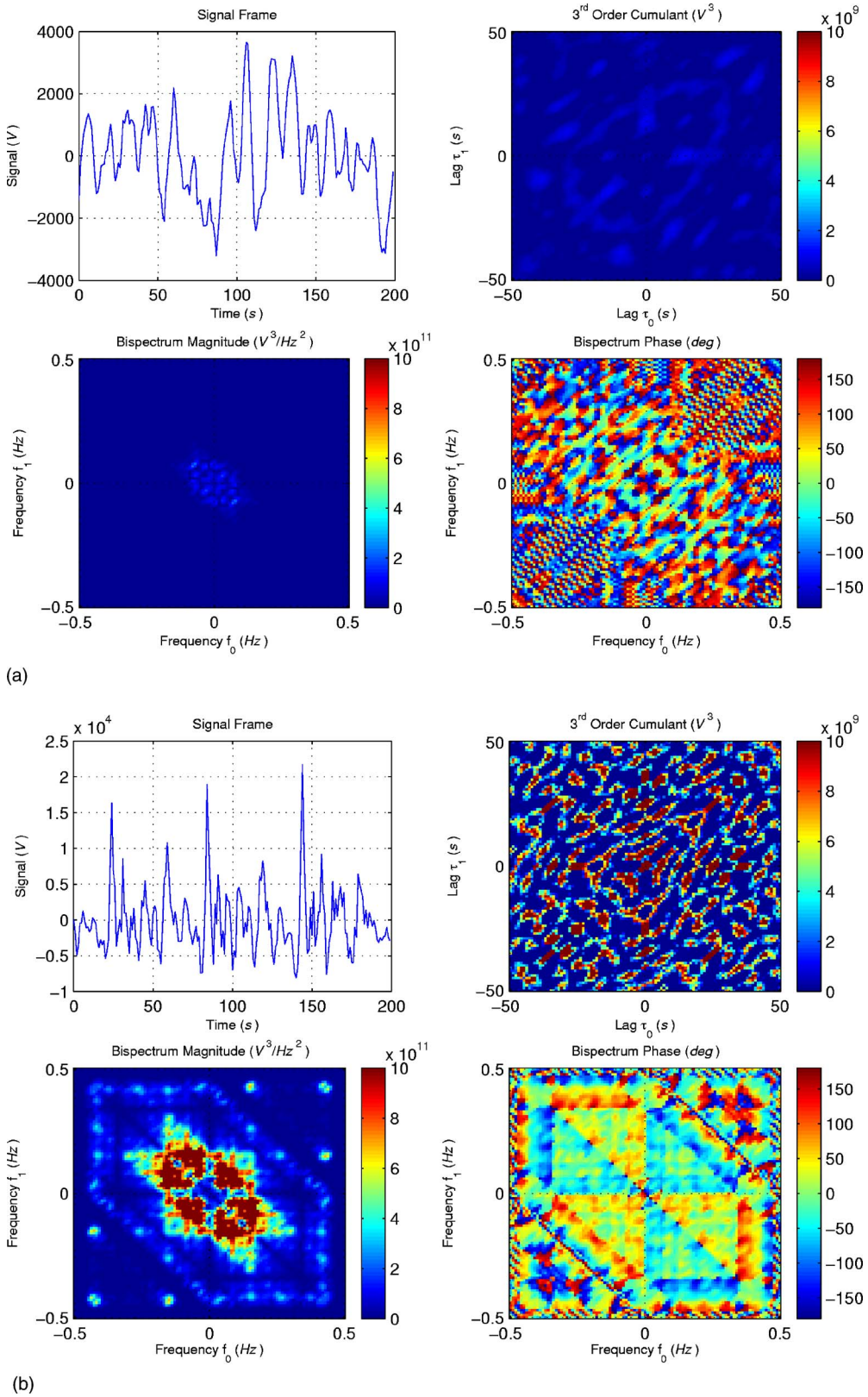


FIG. 1. Third-order statistics of: (a) a noise only signal and (b) a speech signal corrupted by car noise.

If we compare Eq. (4) with Eq. (7) we obtain

$$S_{\tilde{y}x}(\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} B_x(\omega,\omega_2)d\omega_2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} B_x(\omega_1,\omega)d\omega_1. \tag{8}$$

The integrated bispectrum is defined as a cross spectrum between the signal and its square, and therefore a function of a single frequency variable. Hence, its computation as a cross spectrum leads to significant computational savings. But more important is that the variance of the estimator is of

the same order as that of the power spectrum estimator. On the other hand, Gaussian processes have vanishing third-order moments so that the bispectrum and integrated bispectrum functions are zero as well.

Section IV shows an innovative algorithm for voice activity detection taking advantage of the statistical properties of the integrated bispectrum. The proposed method is based on a LRT that can be evaluated using a Gaussian model for the integrated bispectrum of the signal.

IV. VOICE ACTIVITY DETECTION BASED ON THE INTEGRATED BISPECTRA

This section addresses the problem of voice activity detection formulated in terms of a classical binary hypothesis testing framework:

$$H_0 \quad : \quad x(t) = n(t),$$

$$H_1 \quad : \quad x(t) = s(t) + n(t). \tag{9}$$

In a two-hypotheses test, the optimal decision rule that minimizes the error probability is the Bayes classifier. Given an observation vector $\hat{\mathbf{y}}$ to be classified, the problem is reduced to selecting the class (H_0 or H_1) with the largest posterior probability $P(H_i|\hat{\mathbf{y}})$. From the Bayes rule a statistical LRT [8] can be defined by

$$L(\hat{\mathbf{y}}) = \frac{p_{\mathbf{y}|H_1}(\hat{\mathbf{y}}|H_1)}{p_{\mathbf{y}|H_0}(\hat{\mathbf{y}}|H_0)}, \tag{10}$$

where the observation vector $\hat{\mathbf{y}}$ is classified as H_1 if $L(\hat{\mathbf{y}})$ is greater than $P(H_0)/P(H_1)$, otherwise it is classified as H_0. In Ref. 22 the LRT first proposed by Sohn [8] for VAD, which was defined on the power spectrum, is generalized and extended to the case where successive observations $\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_m$ of the noisy signal are available. The so-called multiple observation LRT (MO-LRT) reports significant improvements in robustness as the number of observations increases. This test involves evaluating the joint conditional distributions of the observations under H_0 and H_1,

$$L_m(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_m) = \frac{p_{\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_m|H_1}(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_m|H_1)}{p_{\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_m|H_0}(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_m|H_0)}, \tag{11}$$

which is easily performed if the observations are assumed to be independent.

Assuming the integrated bispectrum $\{S_{yx}(\omega):\omega\}$ as the feature vector $\hat{\mathbf{y}}$ and to be independent zero-mean Gaussian variables in the presence and absence of speech:

$$p(S_{yx}(\omega)|H_0) = \frac{1}{\pi\lambda_0(\omega)} \exp\left[- \frac{|S_{yx}(\omega)|^2}{\lambda_0(\omega)} \right],$$

$$p(S_{yx}(\omega)|H_1) = \frac{1}{\pi\lambda_1(\omega)} \exp\left[- \frac{|S_{yx}(\omega)|^2}{\lambda_1(\omega)} \right], \tag{12}$$

the evaluation of the tests defined by Eqs. (10) and (11) only requires one to estimate the integrated bispectrum $S_{yx}(\omega)$ of the noisy signal and variances λ_0 and λ_1 under absence and presence of speech in the noisy signal. Thus, taking logarithms in Eq. (10) and substituting the model defined in Eq. (12) we obtain

$$\Phi(\hat{\mathbf{y}}) = \sum_{\omega} \log\left(\frac{p(S_{yx}(\omega)|H_1)}{p(S_{yx}(\omega)|H_0)} \right) = \sum_{\omega} \left\{ \left(1 - \frac{\lambda_0(\omega)}{\lambda_1(\omega)} \right) \frac{|S_{yx}(\omega)|^2}{\lambda_0(\omega)} - \log\left(\frac{\lambda_1(\omega)}{\lambda_0(\omega)} \right) \right\}. \tag{13}$$

Finally, if we define the *a priori* and *a posteriori* variance ratios as

$$\xi(\omega) = \frac{\lambda_1(\omega)}{\lambda_0(\omega)} - 1, \quad \gamma(\omega) = \frac{|S_{yx}(\omega)|^2}{\lambda_0} \tag{14}$$

Eq. (13) can be expressed in a more compact form:

$$\Phi(\hat{\mathbf{y}}) = \sum_{\omega} \left[\left(1 - \frac{1}{1 + \xi(\omega)} \right) \gamma(\omega) - \log(1 + \xi(\omega)) \right] = \sum_{\omega} \left[\frac{\xi(\omega)\gamma(\omega)}{1 + \xi(\omega)} - \log(1 + \xi(\omega)) \right]. \tag{15}$$

Section IV A addresses two key issues in order to evaluate the proposed LRT:

(1) The estimation of the integrated bispectrum $S_{yx}(\omega)$ by means of a finite data set, and
(2) the computation of the variances $\lambda_0(\omega)$ and $\lambda_1(\omega)$ of the integrated bispectrum under H_0 and H_1 hypotheses.

A. Estimation of the integrated bispectrum $S_{yx}(\omega)$

Let $\hat{S}_{yx}(\omega)$ denote a consistent estimator of $S_{yx}(\omega)$ where $y(t) = x^2(t) - E\{x^2(t)\}$. Given a finite data set $\{x(1), x(2), \ldots, x(N)\}$ the integrated bispectrum is normally estimated by splitting the data set into blocks.[25] Thus, the data set is divided into K_B nonoverlapping blocks of data each of size N_B samples so that $N = K_B N_B$. Then, the cross periodogram of the ith block of data is given by

$$\hat{S}_{yx}^{(i)}(\omega) = \frac{1}{N_B} X^{(i)}(\omega)[Y^{(i)}(\omega)]^*, \tag{16}$$

where $X^{(i)}(\omega)$ and $Y^{(i)}(\omega)$ denote the discrete Fourier transforms of $x(t)$ and $y(t)$ for the ith block. Finally, the estimate is obtained by averaging K_B blocks,

$$\hat{S}_{yx}(\omega) = \frac{1}{K_B} \sum_{i=1}^{K_B} \hat{S}_{yx}^{(i)}(\omega). \tag{17}$$

This estimation is used to compute $\gamma(\omega)$ through Eq. (14).

B. Computation of the variances $\lambda_0(\omega)$ and $\lambda_1(\omega)$

The statistical properties of the bispectrum estimators have been discussed in Refs. 25 and 31. Thus, the test proposed Sec. IV A and the model assumed in Eq. (12) are justified since for large N_B, the estimates $S_{yx}^{(i)}(\omega_m)$ are complex Gaussian and independently distributed of $S_{yx}^{(i)}(\omega_n)$ for

$m \neq n (m,n=1,2,\ldots,N_B/2-1)$. Moreover, their mean and variance for large values of N_B and K_B can be approximated[16] by

$$E\{\hat{S}_{yx}(\omega)\} \approx S_{yx}(\omega),$$

$$\text{var}\{\text{Re}[\hat{S}_{yx}^{(i)}(\omega)]\} \approx \frac{1}{2K_B}[S_{yy}(\omega)S_{xx}(\omega) + \text{Re}\{S_{yx}^2(\omega)\}],$$

$$\text{var}\{\text{Im}[\hat{S}_{yx}^{(i)}(\omega)]\} \approx \frac{1}{2K_B}[S_{yy}(\omega)S_{xx}(\omega) - \text{Re}\{S_{yx}^2(\omega)\}], \tag{18}$$

where Re and Im denote the real and imaginary parts of a complex number. This means that, in order to estimate the variances λ_0 and λ_1 of the integrated bispectrum under the H_0 and H_1 hypotheses, we need to obtain an expression for $S_{xx}(\omega)$ and $S_{yy}(\omega)$ when $x(t)=n(t)$ (speech absence) and $x(t)=s(t)+n(t)$ (speech presence), respectively.

1. Speech absence

Under the hypothesis H_0, $x(t)=n(t)$ and $y(t)=x^2(t)-E\{x^2(t)\}$. Thus, $S_{xx}(\omega)$ and $S_{yy}(\omega)$ are reduced to

$$S_{xx}(\omega) = S_{nn}(\omega),$$

$$S_{yy}(\omega) = S_{n^2n^2}(\omega), \tag{19}$$

where $S_{n^2n^2}$ can be expressed (see the Appendix) by

$$S_{n^2n^2}(\omega) = 2S_{nn}(\omega) * S_{nn}(\omega) + 2\pi\sigma_n^4 \tag{20}$$

and λ_0 can be estimated by evaluating $S_{nn}(\omega)$ and the variance of the noise:

$$\lambda_0(\omega) = \frac{1}{K_B}[2S_{nn}(\omega) * S_{nn}(\omega) + 2\pi\sigma_n^4\delta(\omega)]S_{nn}(\omega). \tag{21}$$

2. Speech presence

Under the hypothesis H_1, $x(t)=s(t)+n(t)$ and $S_{xx}(\omega)$ and $S_{yy}(\omega)$ require a little more computation (see the Appendix):

$$S_{xx}(\omega) = S_{ss}(\omega) + S_{nn}(\omega),$$

$$S_{yy}(\omega) = S_{s^2s^2}(\omega) + S_{n^2n^2}(\omega) + 4S_{ss}(w) * S_{nn}(\omega) - 2\pi(\sigma_s^4 + \sigma_n^4)\delta(\omega). \tag{22}$$

By using Eq. (20) for $s(t)$ and substituting it in Eq. (22) leads to

$$S_{yy}(\omega) = 2S_{ss}(\omega) * S_{ss}(\omega) + 2S_{nn}(\omega) * S_{nn}(\omega) + 4S_{ss}(\omega) * S_{nn}(\omega). \tag{23}$$

Finally, $\lambda_1(\omega)$ can be estimated in terms of $S_{ss}(\omega)$ and $S_{nn}(\omega)$ by means of

$$\lambda_1(\omega) = \frac{1}{K_B}[S_{ss}(\omega) + S_{nn}(\omega)][2S_{ss}(\omega) * S_{ss}(\omega) + 2S_{nn}(\omega) * S_{nn}(\omega) + 4S_{ss}(\omega) * S_{nn}(\omega)]. \tag{24}$$

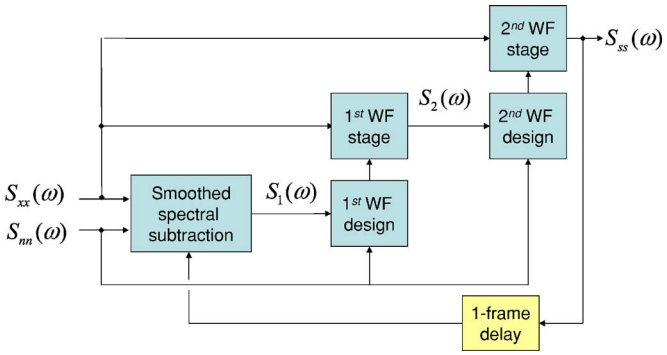


FIG. 2. Estimation of $S_{ss}(\omega)$ via smoothed spectral subtraction and Wiener filtering.

As a conclusion, a way to estimate the power spectrum of the clean signal $S_{ss}(\omega)$ is needed for the evaluation of λ_1. A method combining Wiener filtering and spectral subtraction is used in this paper to estimate $S_{ss}(\omega)$ in terms of the power spectrum $S_{xx}(\omega)$ of the noisy signal. The procedure is described as follows. During a short initialization period, the power spectrum of the residual noise $S_{nn}(\omega)$ is estimated assuming a short nonspeech period at the beginning of the utterance. Note that $S_{nn}(\omega)$ can be computed in terms of the DFT of the noisy signal $x(t)=n(t)$. After the initialization period $S_{xx}(\omega)$ is computed for each frame through Eqs. (16) and (17) and $S_{ss}(\omega)$ is then obtained by applying a denoising process. Denoising consists of a previous smoothed spectral subtraction followed by Wiener filtering. Figure 2 shows a block diagram for the estimation of the power spectrum $S_{ss}(\omega)$ of the denoised speech through the noisy signal $S_{xx}(\omega)$. It is worthwhile clarifying that $S_{nn}(\omega)$ is not only estimated during the initialization period but also updated during nonspeech frames based on the VAD decision. Thus, the denoising process consists of the following stages.

(1) Spectral subtraction:

$$S_1(\omega) = L_s S_{ss}(\omega) + (1 - L_s)\max(S_{xx}(\omega) - \alpha S_{nn}(\omega), \beta S_{xx}(\omega)). \tag{25}$$

(2) First WF design and filtering:

$$\mu_1(\omega) = S_1(\omega)/S_{nn}(\omega),$$

$$W_1(\omega) = \mu_1(\omega)/(1 + \mu_1(\omega)),$$

$$S_2(\omega) = W_1(\omega)S_{xx}(\omega). \tag{26}$$

(3) Second WF design and filtering:

$$\mu_2(\omega) = S_2(\omega)/S_{nn}(\omega),$$

$$W_2(\omega) = \max(\mu_2(\omega)/(1 + \mu_2(\omega)), \beta),$$

$$S_{ss}(\omega) = W_2(\omega)S_{xx}(\omega), \tag{27}$$

where $L_s=0.99$, $\alpha=1$, and $\beta=10^{(-22/10)}$ is selected to ensure a -22 dB maximum attenuation for the filter in order to reduce the high variance musical noise that normally appears due to rapid changes across adjacent frequency bins. The main rea-

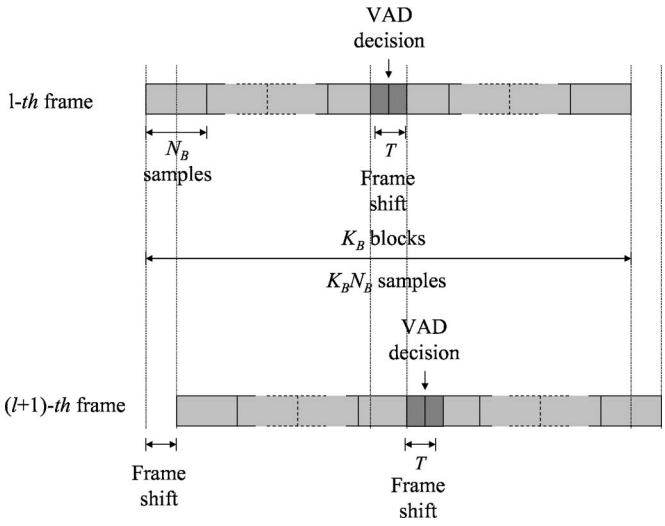


FIG. 3. Integrated bispectrum estimation by block averaging and VAD decision.

son for using a Wiener filter as noise reduction algorithm is found on its optimum performance for filtering additive noise in a noisy signal. This method is normally preferred to other conventional techniques such as spectral subtraction since the musical noise is significantly reduced as well as the variance of the residual noise. A two-stage Wiener filter configuration was used, as in the ETSI AFE standard,[13] in order to make it less sensible to the VAD decision and the noise estimation process.

V. DATA PROCESSING TECHNIQUES FOR VAD

The following shows the two different approaches for block managing that were used in this paper for the estimation of the integrated bispectrum of the input signal and its variances in the formulation of the LRT previously defined.

A. Block partitioning and averaging

The first VAD method is described as follows. The input signal $x(n)$ sampled at 8 kHz is divided into overlapping windows of size $N=K_BN_B$ samples. A typical value of the window size is about 0.2 s, which yields accurate estimations of the integrated bispectrum. The best tradeoff between block averaging (K_B) and spectral resolution (N_B) will be discussed in the next sections.

Figure 3 illustrates the way the signal is processed and the block of data the decision is made for. Note that the decision is made for a T-sample data block around the midpoint of the analysis window where T is the "frame-shift." Thus, a large data set is used to estimate the integrated bispectrum $S_{yx}(\omega)$ by averaging K_B successive blocks of data while the decision is made for a shorter data set. As in most of the standardized VADs,[3,4,13] the frame-shift is 80 samples so that the VAD frame rate is 100 Hz.

After having estimated the power spectrum $S_{ss}(\omega)$ of the clean signal through the denoising process shown earlier, the variances $\lambda_0(\omega)$ and $\lambda_1(\omega)$ of the integrated bispectrum under speech absence and speech presence are computed by evaluating the convolution operations required by Eqs. (21) and (24), respectively. Then, the *a priori* and *a posteriori* variance ratios $\xi(\omega)$ and $\gamma(\omega)$, as defined in Eq. (14), can be estimated and the VAD decision is obtained by comparing the LRT defined in Eq. (15) to a given threshold η. If the LRT is greater than the threshold η the frame is classified as speech, otherwise it is classified as nonspeech. Finally, in order to track nonstationary noisy environments the power spectrum estimate of the noise is updated based on the current observation of the noisy signal:

$$S_{nn}(\omega) = L_n S_{nn}(\omega) + (1 - L_n) S_{xx}(\omega) \tag{28}$$

with $L_n=0.98$ when the VAD detects a nonspeech observation.

This method ensures a reduced variance estimation of the integrated bispectrum of the noisy signal by block averaging. However, the window shift T is typically much smaller than the block size and, therefore, the method is computationally expensive. In order to solve this problem, a computationally efficient but also effective method is developed in Sec. V B.

B. Contextual likelihood ratio test

Most VADs in use today normally consider hang-over algorithms based on empirical models to smooth the VAD decision. It has been shown recently[21,22] that incorporating long-term speech information to the decision rule reports benefits speech/pause discrimination in high noise environments, thus making unnecessary the use of hang-over mechanisms based on hand-tuned rules. The VAD previously proposed addresses this problem by formulating a smoothed decision based on a large data set. However, an optimum statistical test involving multiple and independent observations of the input signal can be also defined as in Ref. 22 over the integrated bispectrum of the noisy signal using Eq. (11).

The proposed MO-LRT formulates the decision for the central frame of a $(2m+1)$-observation buffer $\{\hat{\mathbf{y}}_{l-m}, \ldots, \hat{\mathbf{y}}_{l-1}, \hat{\mathbf{y}}_l, \hat{\mathbf{y}}_{l+1}, \ldots, \hat{\mathbf{y}}_{l+m}\}$:

$$L_{l,m}(\hat{\mathbf{y}}_{l-m}, \ldots, \cdot\, \hat{\mathbf{y}}_{l+m}) = \frac{p_{\mathbf{y}_{l-m},\ldots,\mathbf{y}_{l+m}|H_1}(\hat{\mathbf{y}}_{l-m}, \ldots, \hat{\mathbf{y}}_{l+m}|H_1)}{p_{\mathbf{y}_{l-m},\ldots,\mathbf{y}_{l+m}|H_0}(\hat{\mathbf{y}}_{l-m}, \ldots, \hat{\mathbf{y}}_{l+m}|H_0)}, \tag{29}$$

where l denotes the frame being classified as speech (H_1) or nonspeech (H_0). Note that, assuming statistical independence between the successive observation vectors, the corresponding log-LRT:

$$\ell_{l,m} = \sum_{k=l-m}^{l+m} \ln \frac{p_{\mathbf{y}_k|H_1}(\hat{\mathbf{y}}_k|H_1)}{p_{\mathbf{y}_k|H_0}(\hat{\mathbf{y}}_k|H_0)} \tag{30}$$

is recursive in nature, and if the Φ function is defined as

$$\Phi(k) = \ln \frac{p_{\mathbf{y}_k|H_1}(\hat{\mathbf{y}}_k|H_1)}{p_{\mathbf{y}_k|H_0}(\hat{\mathbf{y}}_k|H_0)}, \tag{31}$$

Eq. (30) can be calculated as

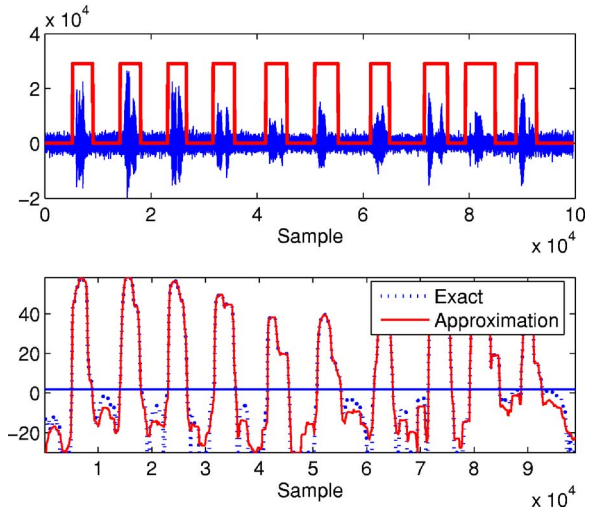


FIG. 4. Operation of the MO-LRT VAD defined on the integrated bispectrum.

$$\ell_{l+1,m} = \ell_{l,m} - \Phi(l-m) + \Phi(l+m+1). \tag{32}$$

Now, if the integrated bispectrum of the noisy signal is considered as the feature vector, Eq. (31) is reduced to be

$$\Phi(k) = \sum_{\omega} \left[\frac{\xi_k(\omega)\gamma_k(\omega)}{1+\xi_k(\omega)} - \log(1+\xi_k(\omega)) \right], \tag{33}$$

where $\xi_k(\omega)$ and $\gamma_k(\omega)$ denote the *a priori* and *a posteriori* variance ratios for the kth frame as defined in Eq. (14).

As a conclusion, the decision rule is formulated over a sliding window consisting of $(2m+1)$ observation vectors around the frame the decision is made for. This fact imposes an m-frame delay to the algorithm that, for several applications, including robust speech recognition, is not a serious implementation obstacle.

Figure 4 shows an example of the operation of the contextual MO-LRT VAD on an utterance of the Spanish SpeechDat-Car database.[27] Figure 4 shows the decision variables for the tests defined by Eq. (33) and, alternatively, for the test with the second log-term in Eq. (33) suppressed (*approximation*) when compared to a fixed threshold $\eta=1.5$. For this example, $N_B=256$ and $m=8$. This approximation reduces the variance during nonspeech periods. It can be shown that using an eight-frame window reduces the variability of the decision variable yielding to a reduced noise variance and better speech/nonspeech discrimination. On the other hand, the inherent anticipation of the VAD decision contributes to reduce the number of speech clipping errors. The use of this test reports quantifiable benefits in speech/nonspeech detection as will be shown in Sec. VII.

C. Comparison in terms of computational cost

It is interesting to compare the two methods proposed for voice activity detection based on single and multiple-observation LRTs. Both methods exhibit the advantages of VADs employing contextual information for formulating the decision rule since the decision variable is built on a long-term data set. The first method decomposes a large analysis window into K_B blocks each of size N_B samples and the integrated bispectrum $S_{yx}(\omega)$ is estimated by averaging K_B blocks using Eq. (16) for each frame-shift T. This can be computationally expensive since the frame-shift is usually much lower than the window size $(N=K_BN_B)$. The second method based on the MO-LRT is more efficient since it just requires one to compute the integrated bispectrum of the current frame (N_B samples). The test is then built on a moving average fashion by means of Eq. (32). This is clearly more efficient in terms of complexity since a single bispectrum computation is performed for each frame-shift T instead of the K_B bispectrum estimations required by the first method. On the other hand, both methods exhibit high speech/nonspeech discrimination accuracy in noisy environments for equivalent delay configurations as will be shown in the following.

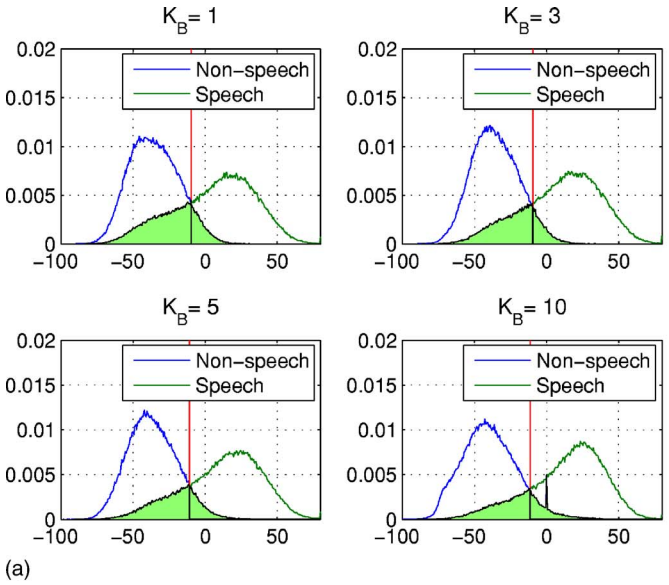
VI. ANALYSIS OF THE PROPOSED METHODS

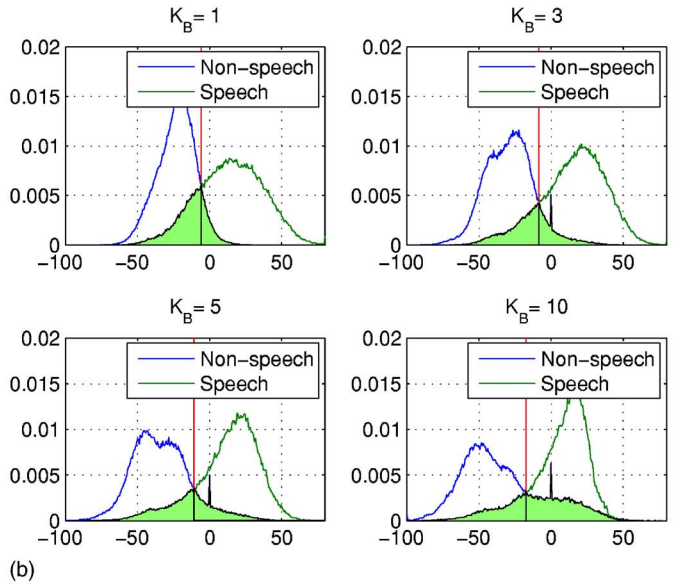
In a Bayes classifier, the overlap between the distributions of the decision variable represents the VAD error rate.[21] In order to clarify the motivations for the proposed algorithms, the distributions of the LRTs defined by Eqs. (15) and (30) were studied as a function of the design parameters K_B, N_B, and m. A hand-labeled version of the Spanish SpeechDat-Car database[27] was used in the analysis. This database contains recordings from close-talking and distant microphones at different driving conditions: (a) Stopped car, motor running, (b) town traffic, low speed, rough road, and (c) high speed, good road. The most unfavorable noise environment (i.e., high speed, good road, and distant microphone) with an average SNR of about 5 dB was selected for the experiments. Thus, the decision variables defined by Eqs. (15) and (30) were measured during speech and nonspeech periods for the whole database, and the histogram and probability distributions were built. Separate results for each of the data processing techniques discussed previously are shown in the following.

A. VAD based on block averaging

Figure 5 shows the distributions of speech and noise for different values of K_B and N_B. It is clearly shown that the distributions of speech and noise are better separated when increasing the number of blocks (K_B). When K_B increases the noise variance decreases and the speech distribution is shifted to the right being more separated from the nonspeech distribution. Thus, the distributions of speech and nonspeech are less overlapped and consequently, the error probability is reduced.

The reduction of the distribution overlap yields improvements in speech/pause discrimination. This fact can be shown by calculating the misclassification errors of speech and noise for an optimal Bayes classifier. Figure 5 also shows the areas representing the probabilities of incorrectly detecting speech and nonspeech and the optimal decision threshold. Figure 6 shows the independent decision errors for speech and nonspeech and the global error rate as a function of K_B for $N_B=256$. The error rates were obtained by computing the areas under the class distributions of the decision variable shown in Fig. 5. The global error rate represents the





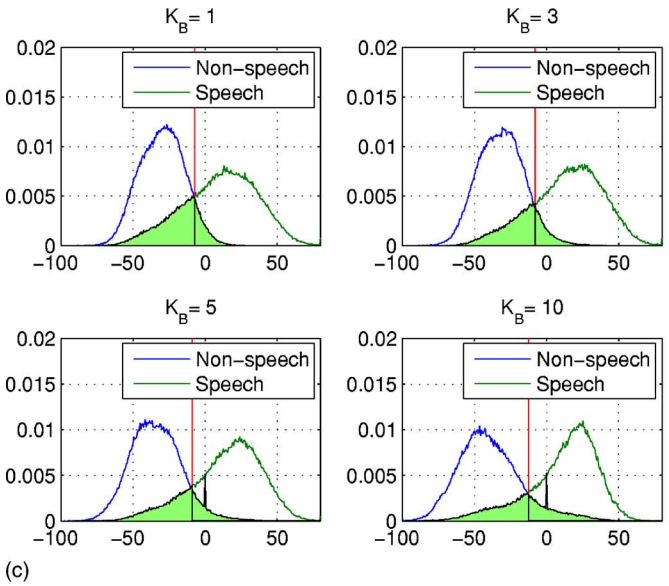


FIG. 5. Distributions of the decision variable for the VAD based on block averaging. (a) N_B=64. (b) N_B=128. (c) N_B=256.

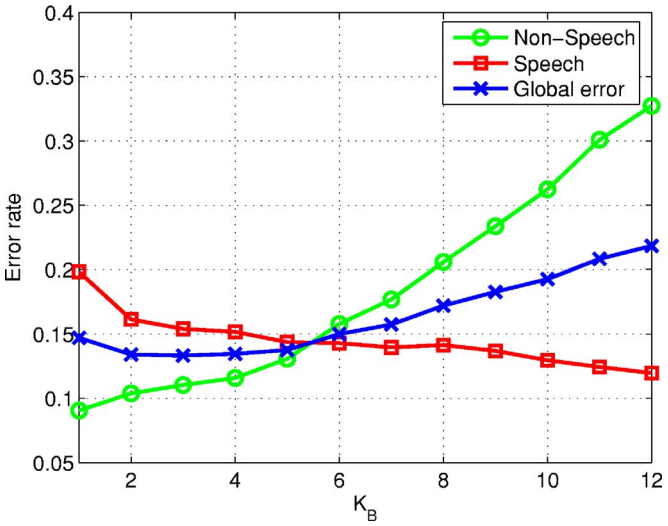


FIG. 6. Probability of error as a function of K_B for N_B=256.

total overlapped area while the speech and nonspeech error rate represent the areas below and above the optimum decision threshold, respectively. The speech detection error is clearly reduced when increasing the length of the window (K_BN_B) while the increased robustness is only damaged by a moderate increase in the nonspeech detection error. These improvements are achieved by reducing the overlap between the distributions when K_B is increased as shown in Fig. 5. It is interesting to show that the optimal values of the parameters K_B and N_B are conditioned to a fixed window size (K_BN_B). This fact is shown in Fig. 7 where the minimum value of the global error rate depends on both N_B and K_B and is obtained for a typical window size of about 640–800 samples (80–100 ms).

B. Integrated bispectrum MO-LRT VAD

Similar results were obtained for the MO-LRT VAD based on the integrated bispectrum of the noisy signal. Figure 8 shows the distributions of speech and noise for different values of K_B and N_B=256. Again the noise variance decreases with m and the speech distribution is shifted to the

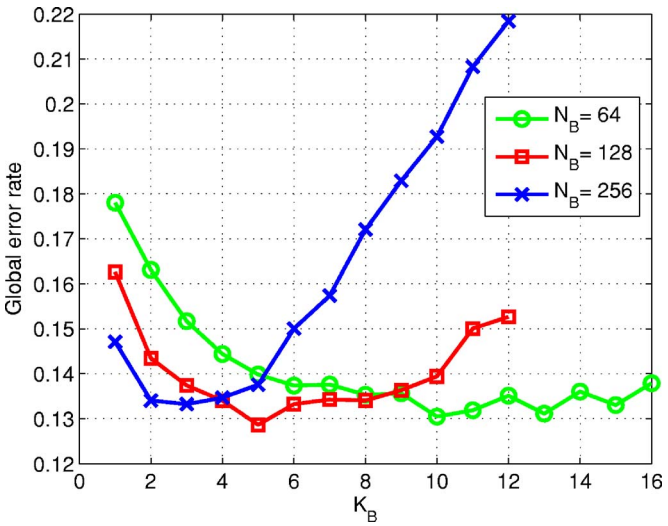


FIG. 7. Global error rates as a function of K_B for N_B=64, 128, and 256.

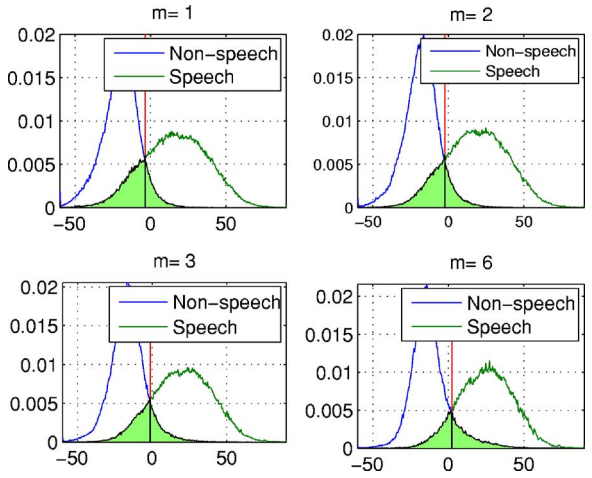


FIG. 8. Distributions of the decision variable for the VAD based on MO-LRT.

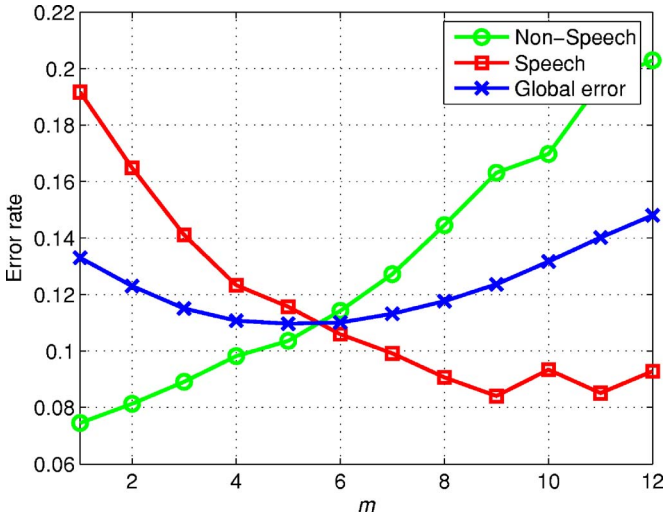


FIG. 9. Probability of error as a function of m for N_B=256.

right. Figure 9 shows the independent decision errors for speech and nonspeech and the global error rate as a function of m for N_B=256. The total error is reduced with the increasing length of the window (m) and exhibits a minimum value for a fixed order. According to Fig. 9, the optimal value of the order of the VAD is m=6. Thus, increasing the length of the window is beneficial in high noise environments since the VAD introduces an artificial "hang-over" period which reduces front and rear-end clipping errors. This saving period is the reason for the increase of the nonspeech detection error shown in Fig. 9.

C. Concluding remarks

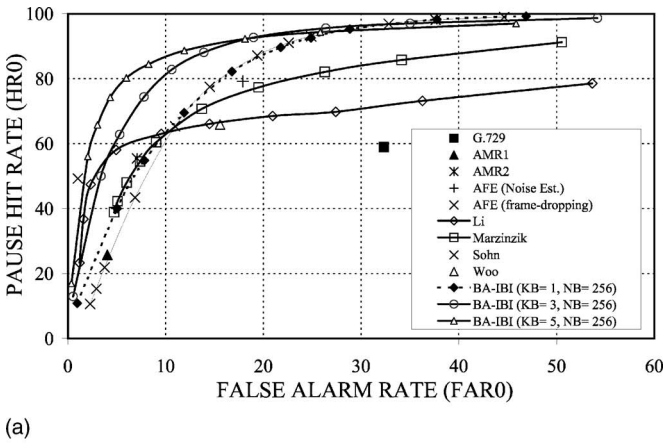
The two alternative methods for data processing and definition of the decision rule yield high discrimination accuracy and minimum global classification error rate for a given length of the analysis window. It can be concluded from Figs. 7 and 9 that the best results are obtained for a typical value of the window size of about 80–100 ms. Thus, both methods benefit from using contextual information for the formulation of the decision rule. The VAD module then exhibits a delay of about half the length of the analysis window that for several applications including real-time speech transmission can be a serious implementation obstacle. However, for other applications including robust speech recognition, a delay of about 50–80 ms in the VAD module does not represent a problem and can be accepted.

VII. EXPERIMENTAL FRAMEWORK

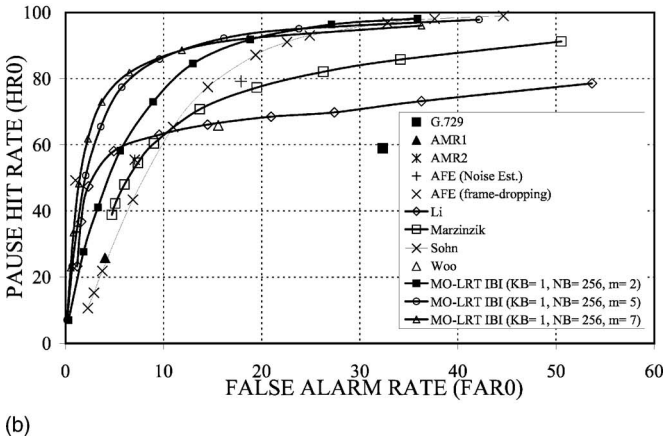
Several experiments are commonly conducted in order to evaluate the performance of VAD algorithms. The analysis is mainly focused on the determination of the error probabilities or classification errors at different SNR levels[6] and the influence of the VAD decision on the performance of speech processing systems.[1] Subjective performance tests have also been considered for the evaluation of VADs working in combination with speech coders.[32] The follwing describes the experimental framework and the objective performance tests conducted in this paper to evaluate the proposed algorithms.

A. ROC curves

The ROC curves are frequently used to completely describe the VAD error rate. They show the tradeoff between speech and nonspeech detection accuracy as the decision threshold varies.[21] The AURORA subset of the original Spanish SpeechDat-Car database[27] was used in this analysis. This database contains 4914 recordings using close-talking and distant microphones from more than 160 speakers. The



(a)



(b)

FIG. 10. ROC curves obtained in high noise conditions. (a) Block based integrated bispectrum LRT VAD. (b) Integrated bispectrum MO-LRT VAD.

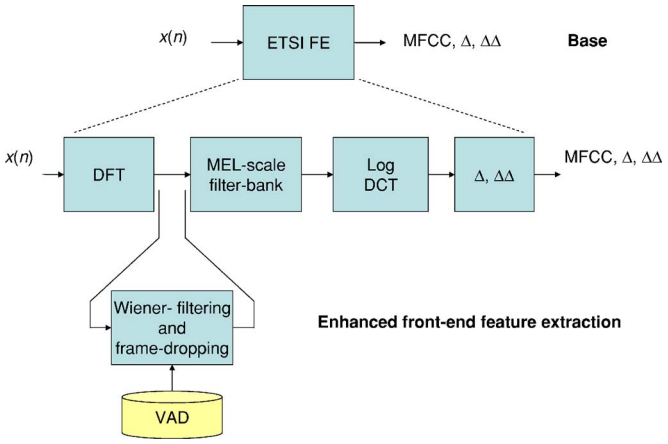


FIG. 11. Speech recognition experiments. Front-end feature extraction.

files are categorized into three noisy conditions: Quiet, low noisy, and highly noisy conditions, which represent different driving conditions with average SNR values between 25 and 5 dB. The nonspeech hit rate (HR0) and the false alarm rate (FAR0=100-HR1) were determined for each noise condition being the actual speech frames and actual speech pauses determined by hand-labeling the database on the close-talking microphone.

Figure 10 shows the ROC curves of the proposed VADs and other frequently referred algorithms[6,8,14,15] for recordings from the distant microphone in high noisy conditions. The working points of the G.729, AMR and AFE VADs are also included. Figure 10(a) shows how increasing the number of blocks (K_B) in the block averaging integrated bispectrum (BA-IBI) LRT VAD leads to a shift-up and to the left of the ROC curve in the ROC space. This result is consistent with the analysis conducted in Figs. 5 and 7 that predicts a minimum error rate for K_B close to five blocks. Similar results are obtained for the efficient MO-LRT IBI VAD that exhibits a shift of the ROC curve when the number of observations (m) increases as shown in Fig. 10(b). Again, the results are consistent with our preliminary experiments and the results shown in Figs. 8 and 9 that expect a minimum error rate for m close to eight frames. Both methods show clear improvements in detection accuracy over standardized VADs and over a representative set of recently published VAD algorithms.[6,8,14,15]

Thus, among all the VADs examined, our VAD yields the lowest false alarm rate for a fixed nonspeech hit rate, and also the highest nonspeech hit rate for a given false alarm rate. The benefits are especially important over G.729, which is used along with a speech codec for discontinuous transmission, and over the algorithm of Li *et al.* (Ref. 15), that is based on an optimum linear filter for edge detection. The proposed VAD also improves Marzinzik VAD[6] that tracks the power spectral envelopes, and the Sohn VAD[8] that formulates the decision rule by means of a statistical likelihood ratio test defined on the power spectrum of the noisy signal.

It is worthwhile mentioning that the above-described experiments yield a first measure of the performance of the VAD. Other measures of VAD performance that have been reported are the clipping errors.[32] These measures provide valuable information about the performance of the VAD and can be used for optimizing its operation. Our analysis does not consider or analyze the position of the frames within the word and assesses the hit rates and false alarm rates for a first performance evaluation of the proposed VAD. On the other hand, the speech recognition experiments conducted later on the AURORA databases will be a direct measure of the quality of the VAD and the application it was designed for. Clipping errors are indirectly evaluated by the speech recognition system since there is a high probability of a deletion error to occur when part of the word is lost after frame-dropping.

TABLE I. Average word accuracy (%) for the AURORA 2 for clean and multicondition training experiments. Results are averaged for all the noises and SNRs ranging from 20 to 0 dB.

	G.729	AMR1	AMR2	AFE	MO-LRT IBI
WF	66.19	74.97	83.37	81.57	84.15
WF+FD	70.32	74.29	82.89	83.29	85.71
	Woo	Li	Marzinzik	Sohn	Hand-labeled
WF	83.64	77.43	84.02	83.89	84.69
WF+FD	81.09	82.11	85.23	83.80	86.86

B. Speech recognition experiments

Although the ROC curves are effective for VAD evaluation, the influence of the VAD in a speech recognition system was also studied. Many authors claim that VADs are well compared by evaluating speech recognition performance[14] since nonefficient speech/nonspeech classification is an important source of the degradation of recognition performance in noisy environments.[2] There are two clear motivations for that: (i) Noise parameters such as its spectrum are updated during nonspeech periods being the speech enhancement system strongly influenced by the quality of the noise estimation, and (ii) FD, a frequently used technique in speech recognition to reduce the number of insertion errors caused by the noise, is based on the VAD decision and speech misclassification errors lead to loss of speech, thus causing irrecoverable deletion errors. This section evaluates the VAD according to the objective it was developed for, that is, by assessing the influence of the VAD in a speech recognition system.

TABLE II. Average word accuracy (%) for the Spanish SDC databases.

	Base	Woo	Li	Marz.	Sohn	G.729	AMR1	AMR2	AFE	**MO-LRT IBI**
WM	92.94	95.35	91.82	94.29	96.07	88.62	94.65	95.67	95.28	**96.39**
MM	83.31	89.30	77.45	89.81	91.64	72.84	80.59	90.91	90.23	**91.75**
HM	51.55	83.64	78.52	79.43	84.03	65.50	62.41	85.77	77.53	**86.65**
Avg.	75.93	89.43	82.60	87.84	90.58	75.65	74.33	90.78	87.68	**91.60**

Figure 11 shows a block diagram of the speech recognition experiments conducted to evaluate the proposed VAD. The reference framework (base) considered for these experiments is the ETSI AURORA project for distributed speech recognition.[33] The recognizer is based on the HTK (Hidden Markov Model Toolkit) software package.[34] The task consists of recognizing connected digits which are modeled as whole word HMMs (Hidden Markov Models) with 16 states per word, simple left-to-right models, and three Gaussian mixtures per state (diagonal covariance matrix). Speech pause models consist of three states with a mixture of six Gaussians per state. The 39-parameter feature vector consists of 12 cepstral coefficients (without the zero-order coefficient), the logarithmic frame energy plus the corresponding delta and acceleration coefficients. Two training modes are defined for the experiments conducted on the AURORA-2 database: (i) Training on clean data only (Clean Training), and (ii) training on clean and noisy data (Multi-Condition Training). For the AURORA-3 SpeechDat-Car databases, the so called well-matched (WM), medium-mismatch (MM), and high-mismatch (HM) conditions are used. These databases contain recordings from the close-talking and distant microphones. In WM condition, both close-talking and hands-free microphones are used for training and testing. In MM condition, both training and testing are performed using the hands-free microphone recordings. In HM condition, training is done using close-talking microphone recordings from all the driving conditions while testing is done using the hands-free microphone at low and high noise driving conditions. Finally, recognition performance is assessed in terms of the word accuracy (WAcc) that considers deletion, substitution, and insertion errors.

An enhanced feature extraction scheme incorporating a noise reduction algorithm and nonspeech FD was built on the base system.[33] The noise reduction algorithm has been implemented as a single WF stage as described in the AFE standard[13] but without mel-scale warping. No other mismatch reduction techniques already present in the AFE standard have been considered since they are not affected by the VAD decision and can mask the impact of the VAD precision on the overall system performance.

Table I shows the recognition performance achieved by the different VADs that were compared. These results are averaged over the three test sets (A, B, and C) of the AURORA-2 recognition experiments[35] and SNRs between 20 and 0 dB. Note that, for the recognition experiments based on the AFE VADs, the same configuration of the standard,[13] which considers different VADs for WF and FD, was used. The proposed integrated bispectrum MO-LRT VAD outperforms the standard G.729, AMR1, AMR2, and AFE VADs in both clean and multicondition training/testing experiments. When compared to recently reported VAD algorithms, the proposed one yields better results being the one that is closer to the "ideal" hand-labeled speech recognition performance.

Table II shows the recognition performance for the Spanish SpeechDat-Car database when WF and FD are performed on the base system.[33] Again, the VAD outperforms all the algorithms used for reference yielding relevant improvements in speech recognition. Note that these particular databases used in the AURORA 3 experiments have longer nonspeech periods than the AURORA 2 database, and then the effectiveness of the VAD results are more important for the speech recognition system. This fact can be clearly shown when comparing the performance of the proposed VAD to Marzinzik VAD.[6] The word accuracy of both VADs is quite similar for the AURORA 2 task. However, the proposed VAD yields a significant performance improvement over Marzinzik VAD[6] for the AURORA 3 database.

VIII. CONCLUSIONS

This paper showed two different schemes for improving speech detection robustness and the performance of speech recognition systems working in noisy environments. Both methods are based on statistical likelihood ratio tests defined on the integrated bispectrum of the speech signal which is defined as a cross spectrum between the signal and its square, and inherits the ability of higher order statistics to detect signals in noise with many other additional advantages: (i) Its computation as a cross spectrum leads to significant computational savings, and (ii) the variance of the estimator is of the same order of power spectrum estimators. The proposed methods incorporate contextual information to the decision rule, a strategy that has reported significant improvements in speech detection accuracy and robust speech recognition applications. They differ in the way the signal is processed in order to obtain precise estimations of the integrated bispectrum and its variance. The optimal window size was determined by analyzing the overlap between the distributions of the decision variable and the error rate of an optimum Bayes classifier. The experimental analysis conducted on the well-known AURORA databases has reported significant improvements over standardized techniques such as ITU G.729, AMR1, AMR2, and ESTI AFE VADs, as well as over recently published VADs. The analysis assessed: (i) The speech/nonspeech detection accuracy by means of the ROC curves, with the proposed VAD yielding improved hit rates and reduced false alarms when compared to all the reference algorithms, and (ii) the recognition rate when the VAD is considered as part of a complete speech recognition system, showing a sustained advantage in speech recognition performance.

ACKNOWLEDGMENTS

This work has received research funding from the EU 6th Framework Programme, under Contract No. IST-2002-507943 (HIWIRE, Human Input that Works in Real Environments) and SESIBONN and SR3-VoIP projects (Nos. TEC2004-06096-C03-00, TEC2004-03829/TCM) from the Spanish government. The views expressed here are those of the authors only. The Community is not liable for any use that may be made of the information contained therein.

APPENDIX: VARIANCES OF THE INTEGRATED BISPECTRUM FUNCTION

This appendix demonstrates Eqs. (20) and (22) that were used in Sec. IV B for the computation of the variances $\lambda_0(\omega)$

and $\lambda_1(\omega)$ of the integrated bispectrum function under speech absence and speech presence, respectively.

Let us assume the clean signal $s(t)$ and the noise $n(t)$ to be stationary, zero mean ($E[s(t)]=E[n(t)]=0$), statistically independent processes so that

$$r_x(k) = r_s(k) + r_n(k) \Rightarrow S_{xx}(\omega) = S_{ss}(\omega) + S_{nn}(\omega), \tag{A1}$$

where $r_x(k)$ denotes the autocorrelation function of the signal corrupted by additive noise.

In order to derive the variances of the integrated bispectrum under the hypotheses H_0 and H_1, it is first needed to evaluate the autocorrelation function of the sequence $y(t)=x^2(t)-E[x^2(t)]$, which is defined as

$$\begin{aligned} r_{yy}(k) &= E[y(t)y(t+k)] \\ &= E[(x^2(t) - E[x^2(t)])(x^2(t+k) - E[x^2(t+k)])]. \end{aligned} \tag{A2}$$

If the variance of the signal is defined as $\sigma_x^2=E[x^2(t)]=E[x^2(t+k)]$ then

$$r_{yy}(k) = E[(x^2(t)x^2\ (t+k)] - \sigma_x^4. \tag{A3}$$

Moreover, if the clean signal and the noise are assumed to be noncorrelated ($E[s(t)n(t)]=E[s(t)]E[n(t)]=0$) the second term on the right-hand side of Eq. (A3) can be expressed as

$$\begin{aligned} \sigma_x^4 = E^2[(s(t)+n(t))^2] &= (\sigma_s^2 + \sigma_n^2 + \overbrace{2E[s(t)n(t)]}^{0})^2 = \sigma_s^4 \\ &+ \sigma_n^4 + 2\sigma_s^2\sigma_n^2. \end{aligned} \tag{A4}$$

By defining $\bar{y}(t) \equiv x^2(t)$, Eq. (A3) can be expressed as

$$\begin{aligned} r_{\bar{y}\bar{y}}(k) &\equiv E[(s(t)+n(t))^2(s(t+k)+n(t+k))^2] \\ &= E[s^2(t)s^2(t+k)] + E[n^2(t)n^2(t+k)] \\ &+ E[s^2(t)n^2(t+k)] + 2E[s^2(t)n(t+k)s(t+k)] \\ &+ E[n^2(t)s^2(t+k)] + 2E[n^2(t)n(t+k)s(t+k)] \\ &+ 2E[n(t)s(t)s^2(t+k)] + 2E[n(t)s(t)n^2(t+k)] \\ &+ 4E[n(t)s(t)n(t+k)s(t+k)], \end{aligned} \tag{A5}$$

where $r_{s^2s^2}=E[s^2(t)s^2(t+k)]$ and $r_{n^2n^2}=E[n^2(t)n^2(t+k)]$.

Using the relation between fourth-order moments and cumulants given in Ref. 36 for a set or random variables:

$$\begin{aligned} C_{x_1,\ldots,x_n} = &\sum_{p_1,\ldots,p_m} (-1)^{m-1}(m-1)! \\ &\times E\left[\prod_{i\in p_1} X_i\right] \cdots E\left[\prod_{i\in p_m} X_i\right], \end{aligned} \tag{A6}$$

where $\{p_1,\ldots,p_m\}$ are all the partitions with $m=1,\ldots,n$ of the set of integers $\{1,\ldots,r\}$. In particular, for $p=4$:

$$\begin{aligned} C_{x_1,x_2,x_3,x_4} &= (-1)^0(0!)E[x_1x_2x_3x_4] + (-1)^1(1!) \\ &\times [E[x_1x_2]E[x_3x_4] + E[x_1x_4]E[x_2x_3] \\ &+ E[x_1x_3]E[x_2x_4] + E[x_1]E[x_2x_3x_4] \\ &+ E[x_2]E[x_1x_3x_4] + E[x_3]E[x_1x_2x_4] \\ &+ E[x_4]E[x_1x_2x_3]] + (-1)^2(2!) \\ &\times [\times E[x_1]E[x_2]E[x_3x_4] + E[x_1]E[x_3]E[x_2x_4] \\ &+ E[x_1]E[x_4]E[x_2x_3] + E[x_2]E[x_3]E[x_1x_4] \\ &+ E[x_2]E[x_4]E[x_1x_3] + E[x_3]E[x_4]E[x_1x_2]] \\ &+ (-1)^3(3!)E[x_1]E[x_2]E[x_3]E[x_4]. \end{aligned} \tag{A7}$$

Note that Eq. (A7) can be significantly reduced if the random variables are assumed to be zero mean:

$$\begin{aligned} C_{x_1,x_2,x_3,x_4} &= E[x_1x_2x_3x_4] - [E[x_1x_2]E[x_3x_4] \\ &+ E[x_1x_4]E[x_2x_3] + E[x_1x_3]E[x_2x_4]]. \end{aligned} \tag{A8}$$

Under the statistical independency assumption, the cross cumulants are null and the cross terms in Eq. (A5) are reduced to

$$\begin{aligned} &E[s^2(t)n^2(t+k)] = \overbrace{C_{s,s,n,n}}^{0} + 2E[s(t)n(t+k)]E[s(t)n(t+k)] \\ &\qquad + E[s^2(t)]E[n^2(t+k)], \\ &2E[s^2(t)n(t+k)s(t+k)] \\ &= 2(\overbrace{C_{s,s,n,s}}^{0} + 2E[s(t)n(t+k)]E[s(t)s(t+k)] \\ &+ E[s^2(t)]E[s(t+k)n(t+k)]), \\ &E[n^2(t)s^2(t+k)] = \overbrace{C_{n,n,s,s}}^{0} + 2E[n(t)s(t+k)]E[n(t)s(t+k)] \\ &\qquad + E[n^2(t)]E[s^2(t+k)], \\ &2E[n^2(t)n(t+k)s(t+k)] \\ &= 2(\overbrace{C_{n,n,s}}^{0} + 2E[n(t)n(t+k)]\ E[n(t)s(t+k)]) + E[n^2(t)]E[s(t \\ &+ k)n(t+k)]), \\ &2E[n(t)s(t)s^2(t+k)] \\ &= 2(\overbrace{C_{n,s,s,s}}^{0} + 2E[n(t)s(t+k)]E[s(t)s(t+k)] \\ &+ E[n(t)s(t)]E[s^2(t\ +k)]), \\ &2E[n(t)s(t)n^2(t+k)] \\ &= 2(\overbrace{C_{n,s,n,n}}^{0} + 2E[n(t)n(t+k)]E[s(t)n(+k)] \\ &+ E[n(t)s(t)]E[n^2(t+k)]), \\ &4E[n(t)s(t)n(t+k)s(t+k)] \\ &= 4(\overbrace{C_{n,s,n,s}}^{0} + E[n(t)n(t+k)]E[s(t)s(t+k)] \\ &\qquad + E[n(t)s(t+k)]E[s(t)n(t+k)] \\ &\qquad + E[s(t)s(t+k)]E[n(t)n(t+k)]). \end{aligned} \tag{A9}$$

Using these expressions, Eq. (A5) is reduced to

$$r_{\bar{y}\bar{y}}(k) = r_{s^2s^2}(k) + r_{n^2n^2}(k) + 2\sigma_s^2\sigma_n^2 + r_s(k)r_n(k) \tag{A10}$$

and Eq. (A3) can be expressed as

$$r_{yy}(k) = r_{s^2s^2}(k) + r_{n^2n^2}(k) + 4r_s(k)r_n(k) - (\sigma_s^4 + \sigma_n^4). \tag{A11}$$

Finally, the power spectrum of the squared centered signal $y(t)$ is obtained after computing the discrete Fourier transform (DFT) on Eq. (A11):

$$S_{yy}(\omega) = S_{s^2s^2}(\omega) + S_{n^2n^2}(\omega) + 4S_{ss}(\omega) * S_{nn}(\omega) - 2\pi(\sigma_s^4 + \sigma_n^4)\delta(\omega), \tag{A12}$$

where the asterisk "$*$" denotes convolution in the frequency domain.

Finally, under speech absence $x(t)=n(t)$ and assuming $n(t)$ to be a Gaussian process:

$$E[n^2(t)n^2(t+k)] = \overbrace{C_{n,n,n,n}}^{0} + 2\overbrace{E[n(t)n(t+k)]E[n(t)n(t+k)]}^{r_n(k)r_n(k)} + \overbrace{E[n^2(t)]E[n^2(t+k)]}^{\sigma_n^4} \tag{A13}$$

or equivalently:

$$S_{n^2n^2}(\omega) = 2S_{nn}(\omega) * S_{nn}(\omega) + 2\pi\sigma_n^4\delta(\omega). \tag{A14}$$

Note that, using the above-derived equations, the variances of the integrated bispectrum function under H_1 and H_0 hypotheses can be computed and the statistical tests used in the proposed VAD methods are completely described and justified.

[1] R. L. Bouquin-Jeannes and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," Speech Commun. **16**, 245–254 (1995).

[2] L. Karray and A. Martin, "Towards improving speech detection robustness for speech recognition in adverse environments," Speech Commun. 261–276 (2003).

[3] ETSI, "Voice activity detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels," ETSI EN 301 708 Recommendation, 1999 (European Telecommunications Standards Inst., France).

[4] ITU, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," ITU-T Recommendation G.729-Annex B, 1996 (International Telecomm. Union, Geneva).

[5] A. Sangwan, M. C. Chiranth, H. S. Jamadagni, R. Sah, R. V. Prasad, and V. Gaurav, "VAD techniques for real-time speech transmission on the Internet," in IEEE International Conference on High-Speed Networks and Multimedia Communications, 2002, pp. 46–50.

[6] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," IEEE Trans. Speech Audio Process. **10**, 341–351 (2002).

[7] 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 Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 1989, pp. 369–372.

[8] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," IEEE Signal Process. Lett. **16**, 1–3 (1999).

[9] I. Potamitis and E. Fishler, "Speech activity detection and enhancement of a moving speaker based on the wideband generalized likelihood ratio and microphone arrays," J. Acoust. Soc. Am. **116**, 2406–2415 (2004).

[10] J. Górriz, J. Ramírez, J. C. Segura, and C. Puntonet, "An effective cluster-based model for robust speech detection and speech recognition in noisy environments," J. Acoust. Soc. Am. **120**, 470–481 (2006).

[11] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 1979, pp. 208–211.

[12] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, Signal Process. **27**, 113–120 (1979).

[13] ETSI, "Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; advanced front-end feature extraction algorithm; Compression algorithms," ETSI ES 202 050 Recommendation, 2002 (European Telecommunications Standards Inst., France).

[14] K. Woo, T. Yang, K. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," Electron. Lett. **36**, 180–181 (2000).

[15] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," IEEE Trans. Speech Audio Process. **10**, 146–157 (2002).

[16] J. K. Tugnait, "Detection of non-Gaussian signals using integrated polyspectrum," IEEE Trans. Signal Process. **42**, 3137–3149 (1994).

[17] J. K. Tugnait, "Corrections to detection of non-Gaussian signals using integrated polyspectrum," IEEE Trans. Signal Process. **43**, 2792–2793 (1995).

[18] J. Ramírez, J. C. Segura, M. C. Benítez, A. de la Torre, and A. Rubio, "A new adaptive long-term spectral estimation voice activity detector," in Proceedings of EUROSPEECH 2003, Geneva, Switzerland, pp. 3041–3044.

[19] A. Sangwan, W. Zhu, and M. Ahmad, "Improved voice activity detection via contextual information and noise suppression," in IEEE International Symposium on Circuits and Systems (ISCAS), 2005, pp. 868–871.

[20] J. Ramírez, J. C. Segura, C. Benítez, A. de la Torre, and A. Rubio, "An effective subband osf-based vad with noise reduction for robust speech recognition," IEEE Trans. Speech Audio Process. **13**, 1119–1129 (2005).

[21] J. Ramírez, J. C. Segura, M. C. Benítez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," Speech Commun. **42**, 271–287 (2004).

[22] J. Ramírez, J. C. Segura, C. Benítez, L. García, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," IEEE Signal Process. Lett. **12**, 689–692 (2005).

[23] J. Górriz, J. Ramírez, J. Segura, and C. Puntonet, "Improved MO-LRT VAD based on bispectra Gaussian model," Electron. Lett. **41**, 877–879 (2005).

[24] J. Ramírez, J. M. Górriz, J. C. Segura, C. G. Puntonet, and A. Rubio, "Speech/non-speech discrimination based on contextual information integrated bispectrum LRT," IEEE Signal Process. Lett. **13** (2006).

[25] D. R. Brillinger and M. Rosenblatt, *Spectral Analysis of Time Series* (Wiley, New York, 1968).

[26] C. Nikias and M. Raghuveer, "Bispectrum estimation: A digital signal processing framework," Proc. IEEE **75**, 869–891 (1987).

[27] A. Moreno, L. Borge, D. Christoph, R. Gael, C. Khalid, E. Stephan, and A. Jeffrey, "SpeechDat-Car: A large speech database for automotive environments," in Proceedings of the II LREC Conference 2000.

[28] J. M. Górriz, J. Ramírez, C. G. Puntonet, and J. Segura, "An efficient bispectrum phase entropy-based algorithm for VAD," in Interspeech 2006, pp. 2322–2325.

[29] X. Zhang, Y. Shi, and Z. Bao, "A new feature vector using selected bispectra for signal classification with application in radar target recognition," IEEE Trans. Signal Process. **49**, 1875–1885 (2001).

[30] X. Liao and Z. Bao, "Circularly integrated bispectra: Novel shift invariant features for high-resolution radar target recognition," Electron. Lett. **34**, 1879–1880 (1998).

[31] D. Brillinger, *Time Series Data Analysis and Theory* (Holt, Rinehart and Winston, New York, 1975).

[32] A. Benyassine, E. Shlomot, H. Su, D. Massaloux, C. Lamblin, and J. Petit, "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 applications," IEEE Commun. Mag. **35**, 64–73 (1997).

[33] ETSI, "Speech processing, transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms," ETSI ES 201 108 Recommendation, 2000 (European Telecommunications Standards Inst., France).

[34] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book* (Cambridge University Press, New York, 1997).

[35] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions," in ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium, Paris, France, 2000 (Intl. Speech Communication Assn.).

[36] C. Nikias and A. Petropulu, *Higher Order Spectra Analysis: a Non-linear Signal Processing Framework* (Prentice Hall, Englewood Cliffs, NJ, 1993).