Jointly Gaussian PDF-Based Likelihood Ratio Test for Voice Activity Detection

Juan Manuel Górriz, Javier Ramírez, Elmar W. Lang, and Carlos G. Puntonet

***Abstract*—This paper presents a novel voice activity detector (VAD) for improving speech detection robustness in noisy environments and the performance of speech recognition systems in real-time applications. The algorithm is based on a generalized complex Gaussian (GCG) observation model and defines an optimal likelihood ratio test (LRT) involving multiple and correlated observations (MCO) based on jointly Gaussian probability distribution functions (jGpdf). An extensive analysis of the proposed methodology for a low dimensional observation model demonstrates 1) the improved robustness of the proposed approach by means of a clear reduction of the classification error as the number of observations is increased, and 2) the tradeoff between the number of observations and the detection performance. The proposed strategy is also compared to different VAD methods including the G.729, AMR, and AFE standards, as well as other recently reported algorithms showing a sustained advantage in speech/nonspeech detection accuracy and speech recognition performance.**

***Index Terms*—Generalized complex Gaussian (GCG) probability distribution function, robust speech recognition, voice activity detection (VAD).**

I. Introduction

THE EMERGING applications of wireless speech communication afford increasing levels of performance in noise adverse environments together with the design of high response rate speech processing systems. Examples of such systems are the new voice services including discontinuous speech transmission [1]–[3] or distributed speech recognition (DSR) over wireless and IP networks [4]–[8]. These systems often require a noise reduction scheme working in combination with a precise voice activity detector (VAD) [9] for estimating the noise spectrum during nonspeech periods in order to compensate its harmful effect on the speech signal. The nonspeech detection algorithm is an important and sensitive part of most of the existing single-microphone noise reduction schemes indeed.

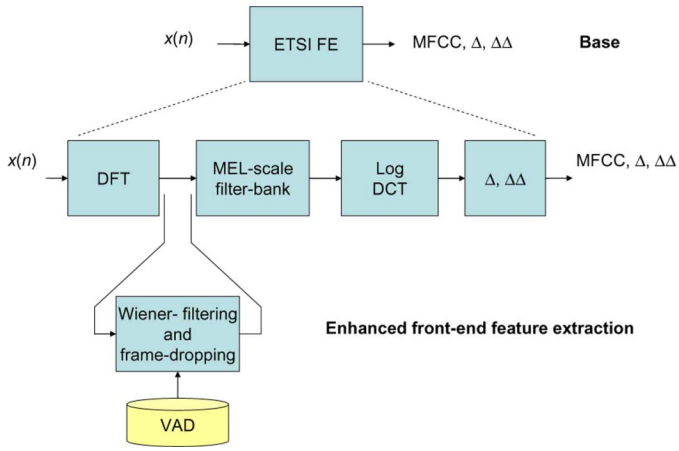

Fig. 1. ETSI standard for DSR. Front-end feature extraction.

Well-known noise suppression algorithms [10], [11] such as Wiener filtering (WF) or spectral subtraction, are widely used for robust speech recognition which is critical during VAD for attaining a high level of performance. Noneffective speech/nonspeech detection is an important source of performance degradation in automatic speech recognition (ASR) systems. There are two main motivations for that af follows.

- Noise parameters such as its spectrum are updated during nonspeech periods being the speech enhancement system is strongly influenced by the quality of the noise estimation.
- Frame-dropping (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 classification errors lead to loss of speech, thus causing irrecoverable deletion errors.

An example of such a system is the ETSI standard for DSR that incorporates noise suppression methods (see Fig. 1). The so-called advanced front-end (AFE) [4] considers an energy-based VAD to estimate the noise spectrum for Wiener filtering and a different VAD for nonspeech FD. The recognizer is usually based on hidden Markov models (HMMs), and the task consists of recognizing connected digits, which are modeled as whole word HMMs with a number of states per word, Gaussian mixtures per state, etc. First, an HMM is trained for each vocabulary word using a number of examples of that word. Second, to recognize some unknown word, the likelihood of each model generating that word is calculated and the most likely model identifies the word.

Manuscript received October 11, 2007; revised July 10, 2008. Current version published October 17, 2008. This work was supported in part by the Spanish Government under the PETRI DENCLASES (PET2006-0253), NAPOLEON (TEC2007-68030-C02-01), and TEC2008-02113 projects and the Consejería de Innovación, Ciencia y Empresa (Junta de Andalucía, Spain) under the Excellence Project (TIC-02566). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Simon King.

J. M. Górriz and J. Ramírez are with the Department of Signal Theory, Networking, and Communications, University of Granada, Granada 18071, Spain (e-mail: gorriz@ugr.es; javierrp@ugr.es).

E. W. Lang is with the Institut für Biophysik und physikalische Biochemie, University of Regensburg, Regensburg 93040, Germany (e-mail: elmar.lang@biologie.uni-regensburg.de).

C. G. Puntonet is with the Department of Computer Architecture and Technology, University of Granada, Granada 18071, Spain (e-mail: carlos@atc.ugr.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2008.2004293

During the last decade, numerous researchers have studied different strategies for detecting speech in noise and the influence of the VAD on the performance of speech processing systems [9], [12]. As an example, there is an increasing interest of studying the effect of adverse noise conditions on ASR systems in vehicle environment [13]. The in-vehicle environment provides a rich background and the big amount of noise-types significantly increase the difficulty of the problem. In this way, several techniques have been proposed towards solving the problem by performing noise tracking, environmental sniffing [14], etc. Most of the works have focussed on the development of robust algorithms with special attention on the derivation and study of noise robust features and decision rules [15]–[18]. The different approaches include those based on energy thresholds [15], pitch detection [19], spectrum analysis [17], zero-crossing rate [2], periodicity measures [20], higher order statistics [21], [22], or combinations of different features [2], [3], [23]. Sohn *et al.* [18] proposed a robust VAD algorithm based on a statistical likelihood ratio test (LRT) involving a single observation vector, and an HMM-based hang-over scheme. Later, Cho *et al.* [24] suggested an improvement based on a smoothed LRT. The statistical model proposed by Sohn was extended to generalized Gaussian distributions in [25], deriving a complex model that uses the independence assumption of each part. Most VADs in use today normally consider hangover algorithms based on empirical models to smooth the VAD decision which yields significant improvements in word ending accuracy. It has been shown recently [26]–[29] that incorporating long-term speech information to the decision rule results in benefits for speech/pause discrimination in high-noise environments also. However, an inherent delay is inevitably included, thus challenging the performance of real-time processing systems. Finally, an important assumption made on the latter works needs revision: *the independence of adjacent observations*. In any speech processing system, the input signal is usually decomposed into overlapping frames, thus a clear statistical dependence between adjacent feature vectors is introduced.

In this paper, we propose a hybrid VAD that combines the use of a hang-over mechanism and contextual information defined on the previous observations, thus avoiding the inclusion of any processing delay. In addition, the dependence between observations is also addressed by using a complex Gaussian model and the following assumption: *the observations are jointly Gaussian distributed with nonzero correlations*. Important issues that also need to be discussed are 1) the increased computational complexity mainly due to the definition of the decision rule over large data sets, and 2) the optimal criterion for the decision rule. This work represents considerable progress in the field by defining a decision rule based on an optimal statistical LRT which involves multiple and *correlated* observations. The paper is organized as follows. Section II reviews the theoretical background on the LRT statistical decision theory and proposes a generalized complex Gaussian observation model. Section III considers its application to the problem of detecting speech in a noisy signal and addresses the computation of the novel MCO-LRT for VAD. Sections IV and V analyze the proposed method for just two and three consecutively correlated speech observations, respectively, using the AURORA 3 Spanish SpeechDat-Car (SDC) database [30]. Section VII describes the experimental framework considered for the evaluation of the proposed endpoint detection, and finally, Section VIII summarizes the conclusions of this work.

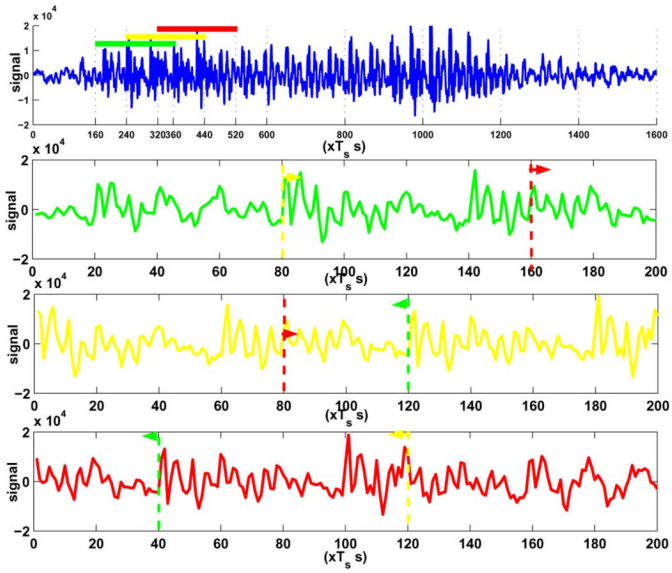

Fig. 2. Example of processing windows of an utterance in the Aurora 3 database ($T_s = 1/8000$ s). Note how the overlap between frames affects the statistical independence assumption.

II. Multiple Observation Likelihood Ratio Test

Under 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

$$L(\hat{\mathbf{y}}) = \frac{p_{\mathbf{y}|H_1}(\hat{\mathbf{y}}|H_1)}{p_{\mathbf{y}|H_0}(\hat{\mathbf{y}}|H_0)} \gtrless \frac{P[H_0]}{P[H_1]} \Rightarrow \begin{matrix} \hat{\mathbf{y}} \leftrightarrow H_1 \\ \hat{\mathbf{y}} \leftrightarrow H_0 \end{matrix} \tag{1}$$

where $P[\cdot]$ denotes the likelihood of each hypothesis, and $p_{\mathbf{y}|H_s}$ is the conditional probability of the observation $\mathbf{y}$ given the occurrence of H_s for $s = 0, 1$. In the LRT, only a single observation is considered and represented by a vector $\hat{\mathbf{y}}$. The performance of the decision procedure can be improved by incorporating more observations into the statistical test. When N measurements $\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_N$ are available in a two-class classification problem, a multiple observation likelihood ratio test (MO-LRT) [26] can be defined by

$$L_N(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_N) = \frac{p_{\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N|H_1}(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_N|H_1)}{p_{\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N|H_0}(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_N|H_0)}. \tag{2}$$

This test involves the evaluation of an Nth order LRT for which a computationally efficient method is available when the individual measurements $\hat{\mathbf{y}}_k$ are independent. However, if they are not, as for example in the VAD problem where the frames used in the computation of the observation vectors are usually overlapping in order to obtain a soft decision rule (see Fig. 2), a more appropriate model must be considered.

In this paper, a complex generalized Gaussian (CGG) model is proposed for the set of observation vectors which are assumed to be independently distributed in their components [18] and in their real and imaginary parts [25]. Thus, the MO-LRT in (2) can be rewritten as

$$L_N\left(\hat{\mathbf{y}}_1,\hat{\mathbf{y}}_2,\ldots,\hat{\mathbf{y}}_N\right)=\prod_{\mathcal{P}=\{\text{real},im\}}\prod_{\omega}\frac{p_{\mathbf{y}_\omega^{\mathcal{P}}|H_1}\left(\hat{\mathbf{y}}_\omega^{\mathcal{P}}|H_1\right)}{p_{\mathbf{y}_\omega^{\mathcal{P}}|H_0}\left(\hat{\mathbf{y}}_\omega^{\mathcal{P}}|H_0\right)} \tag{3}$$

where $\hat{\mathbf{y}}_\omega^{\mathcal{P}}=\left(\hat{y}_1^{\mathcal{P}}(\omega),\hat{y}_2^{\mathcal{P}}(\omega),\ldots,\hat{y}_N^{\mathcal{P}}(\omega)\right)$ is the vector obtained by joining the "$\mathcal{P}$" part of the component ω of each observation vector $\hat{\mathbf{y}}_k$ in the MO window, and the probability law is given by the jointly Gaussian probability density function (jGpdf)[1]

$$\left(p_{\mathbf{y}_\omega^{\mathcal{P}}|H_s}\left(\hat{\mathbf{y}}_{\boldsymbol{\omega}}^{P}|H_s\right)\right)=K_{H_s,N}\cdot\exp\left[-\frac{1}{2}\left(\hat{\mathbf{y}}_\omega^{\mathcal{P}}\right)^T\left(C^N_{\mathbf{y}_\omega,H_s}\right)^{-1}\hat{\mathbf{y}}_\omega^{\mathcal{P}}\right] \tag{4}$$

for $s=0,1$, where $K_{H_s,N}=\left(1/(2\pi)^{N/2}\left|C^N_{\mathbf{y}_{\boldsymbol{\omega}},H_s}\right|^{1/2}\right)$, $C^N_{\mathbf{y},H_s}$ is the Nth-order covariance matrix of the observation vector under hypothesis H_s, and $|\cdot|$ denotes the determinant of a matrix. As follows from (4), this novel LRT includes the dependence between adjacent observations in the covariance matrix (using the Mahalanobis distance [31]); hence, we name it MCO-LRT.

The observation model is completely defined by selecting the set of observation vectors $\mathbf{y}_k$. The model selected for the observation vector is similar to that used by Sohn *et al.* [18], consisting of the discrete Fourier transform (DFT) coefficients of the clean speech $(S(\omega))$ and the additive noise $(N(\omega))$. In the latter work, they are assumed to be asymptotically independent Gaussian random variables. Thus, the binary hypothesis is rewritten in terms of each observation vector as

$$\begin{aligned}\hat{\mathbf{y}}_k&=\mathbf{N}_k \text{ under } H_0\\ \hat{\mathbf{y}}_k&=\mathbf{S}_k+\mathbf{N}_k \text{ under } H_1\end{aligned} \tag{5}$$

where $\mathbf{N}_k=(N_k(\omega_1),\ldots,N_k(\omega_{N_{\text{DFT}}}))$ and $\mathbf{S}_k=(S_k(\omega_1),\ldots,S_k(\omega_{N_{\text{DFT}}}))$ are, respectively, the kth noise and clean signal DFT observation vector. However, in the previous works [18], [26], the covariance matrix given in (4) is either not considered at all (single observation) or assumed to be diagonal (statistical independence). Thus, we may expect obtaining a better characterization of the problem by introducing this statistical dependence. The use of the multivariate model is perfectly motivated in terms of the Mahalanobis distance [31] when the set of observations are correlated.

III. Application to VAD

The use of the MCO-LRT for voice activity detection, is mainly motivated by two factors: 1) the optimal behavior of the so defined decision rule, and 2) a multiple observation vector for classification defines a reduced variance LRT achieving clear improvements in robustness against the presence of acoustic noise in the environment. The second property is also achieved by the previous MO-LRT [26] when a large window size (N) is selected which counteracts the nonoptimality of the decision rule based on the independence assumption. Consequently, substantial improvements are expected when dealing with a small multiple and correlated observations (MCO)-window size (N). On the other hand, for real-time applications the model order cannot be as high as is usually selected [26] for speech recognition systems. Then the proposed approach is expected to improve the independent MO methodology for VAD by selecting a low model order.

The proposed MCO-LRT VAD is defined over the sliding window of observation vectors $\{\hat{\mathbf{y}}_{l-m},\ldots,\hat{\mathbf{y}}_{l-1},\hat{\mathbf{y}}_l,\hat{\mathbf{y}}_{l+1},\ldots,\hat{\mathbf{y}}_{l+m}\}$. By applying a log transformation to (3) and using the jGpdf model in (4) leads to

$$\ell_{l,N}=\sum_{\omega}\frac{1}{2}\left\{\hat{\mathbf{y}}_{\boldsymbol{\omega}}^T\Delta_N^\omega\hat{\mathbf{y}}_{\boldsymbol{\omega}}+ln\left(\frac{\left|C^N_{\hat{\mathbf{y}}_{\boldsymbol{\omega}},H_0}\right|}{\left|C^N_{\hat{\mathbf{y}}_{\boldsymbol{\omega}},H_1}\right|}\right)\right\} \tag{6}$$

where $\Delta_N^\omega=\left(C^N_{\hat{\mathbf{y}}_{\boldsymbol{\omega}},H_0}\right)^{-1}-\left(C^N_{\hat{\mathbf{y}}_{\boldsymbol{\omega}},H_1}\right)^{-1}$, $N=2m+1$ is the order of the model, l denotes the frame being classified as speech (H_1) or nonspeech (H_0), and $\hat{\mathbf{y}}_{\boldsymbol{\omega}}$ is the previously defined frequency observation vector over the sliding window. The evaluation of the LRT requires the computation of the inverse and the determinant of a matrix as shown in (6). This is not an implementation obstacle because of the reduced MCO-window size (N) and the selected model for the covariance matrix in which only the dependence between adjacent observations is considered.

A. Model Selection for the Covariance Matrix

For our purposes,[2] the covariance matrix is suitably modeled as a symmetric tridiagonal matrix, thus considering the correlation function between adjacent observations exclusively, we express it as

$$\left[C^N_{\mathbf{y}_{\boldsymbol{\omega}}}\right]_{ij}=\begin{cases}\sigma^2_{y_i}(\omega)\equiv E\left[|y_i^\omega|^2\right], & \text{if } i=j\\ r_{ij}(\omega)\equiv E\left[y_i^\omega y_j^\omega\right], & \text{if } j=i+1\\ 0, & \text{else}\end{cases} \tag{7}$$

where $1\le i\le j\le N$ and $\sigma^2_{y_i}(\omega)$ and $r_{ij}(\omega)$ are the variance and correlation frequency components of the observation vector (denoted for clarity σ_i, r_i respectively). This approach reduces the computational effort of the algorithm with additional benefits obtained from the properties of symmetric tridiagonal matrices [32].

B. Effective Computation of the LRT

Since the covariances matrices under H_0 and H_1 hypotheses are assumed to be tridiagonal symmetric matrices, the inverse matrices can be computed as [32]

$$\left[\left(C^N_{\mathbf{y}_{\boldsymbol{\omega}}}\right)^{-1}\right]_{ij}=\left[\frac{q_j}{p_j}-\frac{q_N}{p_N}\right]p_ip_j,\quad N-1\ge i\ge j\ge 0 \tag{8}$$

[1]In the following, we assume for simplicity the observation vector $\mathbf{y}_{\boldsymbol{\omega}}$ to be real. The extension of the results to a complex scenario is achieved by applying the derived expressions to the real and imaginary parts of the vector independently, i.e., using (3).

[2]The speech signal sampled at 8 kHz is usually processed on a frame by frame basis and the feature vectors $\hat{\mathbf{y}}_k$ are computed using a 25-ms frame-size and a 10-ms frame-shift.

where N is the order of the model and the set of real numbers $q_n, p_n\ (n = 1 \cdots \infty)$ satisfies the three-term recursion relation for $j \geq 1$

$$0 = r_j(q_{j-1}, p_{j-1}) + \sigma_{j+1}(q_j, p_j) + r_{j+1}(q_{j+1}, p_{j+1}) \quad (9)$$

with initial values

$$\begin{aligned} p_0 &= 1 \text{ and } p_1 = -\frac{\sigma_1}{r_1} \\ q_0 &= 0 \text{ and } q_1 = \frac{1}{r_1}. \end{aligned} \quad (10)$$

In general, this set of coefficients is defined in terms of orthogonal complex polynomials which satisfy a Wronsky-like relation [33] and have the continued-fraction representation

$$\left[\frac{q_n(z)}{p_n(z)}\right] = \frac{1}{(z-\sigma_1)-} \ominus \frac{r_1^2}{(z-\sigma_2)-} \ominus \cdots \ominus \frac{r_{n-1}^2}{(z-\sigma_n)} \quad (11)$$

where $\ominus$ denotes the continued fraction. This representation is used to compute the coefficients of the inverse matrices evaluated on $z = 0$. Since the decision rule is formulated over a sliding window consisting of $2m$+1 observation vectors, we can use the following relations to speed up the evaluation (computation of the inverse matrix) of the future decision windows:

$$\left[\frac{q_n^l(z)}{p_n^l(z)}\right] = \frac{1}{(z-\sigma_1^l)-} \ominus \left[\frac{q_{n-1}^{l+1}(z)}{p_{n-1}^{l+1}(z)}\right] \quad (12)$$

$$\left[\frac{q_n^{l+1}(z)}{p_n^{l+1}(z)}\right] = \left[\frac{q_{n-1}^{l+1}(z)}{p_{n-1}^{l+1}(z)}\right] \ominus \frac{\left(r_{n-1}^{l+1}\right)^2}{z-\sigma_n^{l+1}} \quad (13)$$

where l indexes the current sliding window. Using the Volker–Strassen formula [34] and Woodbury's identity [35], it is easy to compute the inverse matrix at frame $l+1$ as (see Appendix D)

$$\left(C_{\mathbf{y}\boldsymbol{\omega}}^{N,l+1}\right)^{-1} = (C_N)^{-1} - \frac{\left((C_N)^{-1}\mathbf{u}\right)\left(\mathbf{r_1}^T(C_N)^{-1}\right)^T}{1 + \mathbf{r_1}^T(C_N)^{-1}\mathbf{u}} \quad (14)$$

where C_N^{-1} is an Nth-order submatrix of the $(N+1)$th-order inverse matrix $\left(C_{\mathbf{y}\boldsymbol{\omega}}^{N+1,l}\right)^{-1}$ at frame l (extracting the first row and column), and $\mathbf{u}$ and $\mathbf{r_1}$ are $N \times 1$ dimensional vectors defined as $\mathbf{u} = (r_1/\sigma_1, 0, \ldots)^T$ and $\mathbf{r_1} \equiv (r_1, 0, \ldots)^T$.

The computation of the determinant in the sliding window can be also computed recursively. After an initialization period of two MO windows, the determinant of the Nth-order matrix at the observation l can be evaluated according to (see Appendix C)

$$\begin{aligned} \left|C_{\mathbf{y}\boldsymbol{\omega}}^{N,l}\right| = \frac{\sigma_N^l}{\sigma_1^{l-1}}\left|C_{\mathbf{y}\boldsymbol{\omega}}^{N,l-1}\right| - \left(r_1^{l-1}\right)^2\left|C_{\mathbf{y}\boldsymbol{\omega}}^{N-2,l}\right| \\ + \frac{\left(r_{N-1}^l\right)^2}{\left(r_1^{l-2}\right)^2}\left|C_{\mathbf{y}\boldsymbol{\omega}}^{N,l-2}\right| - \left(\sigma_1^{l-2}\right)\left|C_{\mathbf{y}\boldsymbol{\omega}}^{N-1,l-2}\right| \end{aligned} \quad (15)$$

where the recursion can be easily proven using

$$\left|C_{\mathbf{y}\boldsymbol{\omega}}^{N}\right| = \sigma_N\left|C_{\mathbf{y}\boldsymbol{\omega}}^{N-1}\right| - r_{N-1}^2\left|C_{\mathbf{y}\boldsymbol{\omega}}^{N-2}\right|. \quad (16)$$

In Sections IV–VII, we present a new algorithm based on this methodology for $N = 2$ and $N = 3$. In order to obtain a connection with previous proposals, the assumption of *vanishing squared correlation functions* must be considered. Thus, we implement a novel robust speech detector with an inherent small delay that is mainly intended for real-time applications such as mobile communications. The decision function will be described in terms of the correlation and variance coefficients which constitute a correction to the previous LRT method [26] that assumes uncorrelated observation vectors.

IV. Analysis of jGpdf VAD for $N = 2$

The improvement provided by the proposed methodology is evaluated in this section by studying the case $N = 2$ (see Appendix A). In this low-dimensional problem, explicit expressions for the evaluation of the second-order MCO-LRT can be obtained and a connection with previous proposals [29] can be shown. In this case, since the covariance matrix is

$$C_{\mathbf{y}\boldsymbol{\omega}}^2 = \begin{pmatrix} \sigma_1(\omega) & r_1(\omega) \\ r_1(\omega) & \sigma_2(\omega) \end{pmatrix} \quad (17)$$

and assuming vanishing squared correlations under H_0 and H_1, the LRT can be evaluated according to

$$\begin{aligned} \ell_{l,2} = \frac{1}{2}\sum_{\omega}\frac{\gamma_1(\omega)\xi_1(\omega)}{1+\xi_1(\omega)} + \frac{\gamma_2(\omega)\xi_2(\omega)}{1+\xi_2(\omega)} \\ - \ln\left[(1+\xi_1(\omega))\right] - ln\left[(1+\xi_2(\omega))\right] \\ - 2\sqrt{\gamma_1(\omega)\gamma_2(\omega)} \\ \times\left(\rho_1^{H_0} - \frac{\rho_1^{H_1}}{\sqrt{(1+\xi_1(\omega))(1+\xi_2(\omega))}}\right) \end{aligned} \quad (18)$$

where $\rho_1^{H_1}(\omega) = \left(r_1^{H_1}(\omega)/\sqrt{\sigma_1^{H_1}(\omega)\sigma_2^{H_1}(\omega)}\right)$ and $\rho_1^{H_0}(\omega)$ are the correlation coefficients of the observations under $H_1 \& H_0$, respectively, $\xi_i(\omega) \equiv \sigma_i^s(\omega)/\sigma_i^n(\omega)$, $\gamma_i(\omega) \equiv \left(y_i^{\omega}\right)^2/\sigma_i^n(\omega)$ are the *a priori* and *a posteriori* SNRs, and l indexes the second observation. Finally, assuming that the correlation coefficients are negligible under H_0 (noise correlation coefficients) we can obtain the decision rule with the previous MO-LRT [26] as follows:

$$\ell_{l,2} = \frac{1}{2}\sum_{\omega} L_1(\omega) + L_2(\omega) + 2\sqrt{\gamma_1\gamma_2}\left(\frac{\rho_1^{H_1}}{\sqrt{(1+\xi_1)(1+\xi_2)}}\right) \quad (19)$$

where $L_i(\omega) \equiv (\gamma_i(\omega)\xi_i(\omega)/1+\xi_i(\omega)) - \ln(1+\xi_i(\omega))$ for $i = 1, 2$, are the independent LRT of the observations $\hat{\mathbf{y}}_1$, $\hat{\mathbf{y}}_2$ which are corrected with the term depending on $\rho_1^{H_1}$. At this point, ergodicity of the process in frequency space must be assumed to estimate the new model parameter $\rho_1^{H_1}$. This means that the correlation coefficients are constant in frequency (wide sense stationary) and thus, an ensemble average can be computed using the sample mean correlation of the observations $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$ included in the sliding window. Fig. 3 clarifies the motivations for using the correlation correction as shown in (19). We show the evaluation of the proposed VAD on an utterance of the AURORA 3 Spanish SpeechDat-Car database [30]. As it is indicated by black arrows, the span of the decision function over voice periods is increased significantly as the span over noise

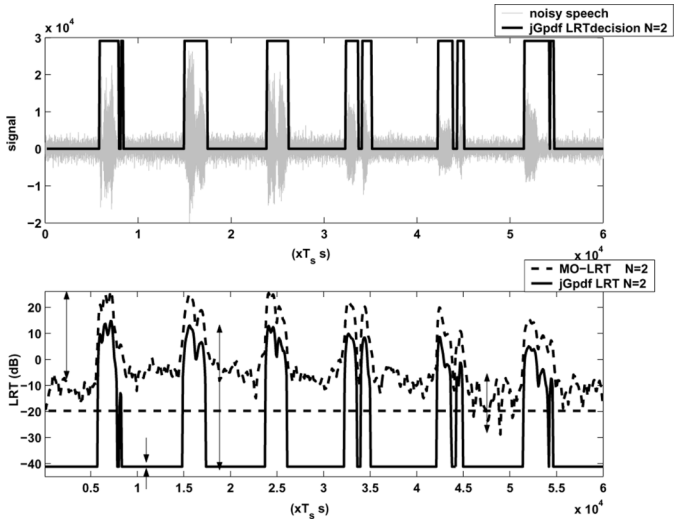

Fig. 3. Comparison between MO-LRT and MCO-LRT on an utterance of Aurora 3 database. Note how the correlation correction term provides an almost binary decision rule, decreasing the decision span in noise periods and increasing in speech ones.

periods is decreased radically. On the other hand, the decision rule of the previous MO-LRT VAD is nonstationary and noisy for a small number of observations as shown in Fig. 3.

V. Analysis of jGpdf VAD for N Observations

The improvement provided by the proposed methodology is evaluated in this section by studying the case $N = 3$ and then generalizing it to any number N of observations. In this case, the properties of a symmetric and tridiagonal matrix come into play

$$C^3_{\mathbf{y}_{\boldsymbol{\omega}}} = \begin{pmatrix} \sigma_1(\omega) & r_1(\omega) & 0 \\ r_1(\omega) & \sigma_2(\omega) & r_2(\omega) \\ 0 & r_2(\omega) & \sigma_3(\omega) \end{pmatrix}. \tag{20}$$

The likelihood ratio can be expressed as (see Appendix B)

$$\ell_{l,3} = \sum_{\omega} \ln \frac{K_{H_1,3}}{K_{H_0,3}} + \frac{1}{2}\hat{\mathbf{y}}^T_{\omega}\Delta^{\omega}_3\hat{\mathbf{y}}_{\omega} \tag{21}$$

where $K_{H_s,3}$ for $s = 0, 1$ and Δ^{ω}_3 are defined in Sections II and III, respectively. Assuming that squared correlation coefficients vanish under H_0 and H_1, the log-LRT can be evaluated as follows (for clarity we have omitted the frequency dependence of the parameters):

$$\begin{aligned} \ell_{l,3} = &\frac{1}{2}\sum_{\omega}\sum_{i=1}^{3}\left[\frac{\gamma_i\xi_i}{1+\xi_i} - \ln(1+\xi_i)\right] \\ &- 2\sqrt{\gamma_1\gamma_2}\left(\rho_1^{H_0} - \frac{\rho_1^{H_1}}{\sqrt{(1+\xi_1)(1+\xi_2)}}\right) \\ &- 2\sqrt{\gamma_2\gamma_3}\left(\rho_2^{H_0} - \frac{\rho_2^{H_1}}{\sqrt{(1+\xi_2)(1+\xi_3)}}\right) \\ &+ 2\sqrt{\gamma_1\gamma_3}\left(\rho_1^{H_0}\rho_2^{H_0} - \frac{\rho_1^{H_1}}{\sqrt{(1+\xi_1)(1+\xi_2)}}\right. \\ &\left.\times\ \frac{\rho_2^{H_1}}{\sqrt{(1+\xi_2)(1+\xi_3)}}\right). \end{aligned} \tag{22}$$

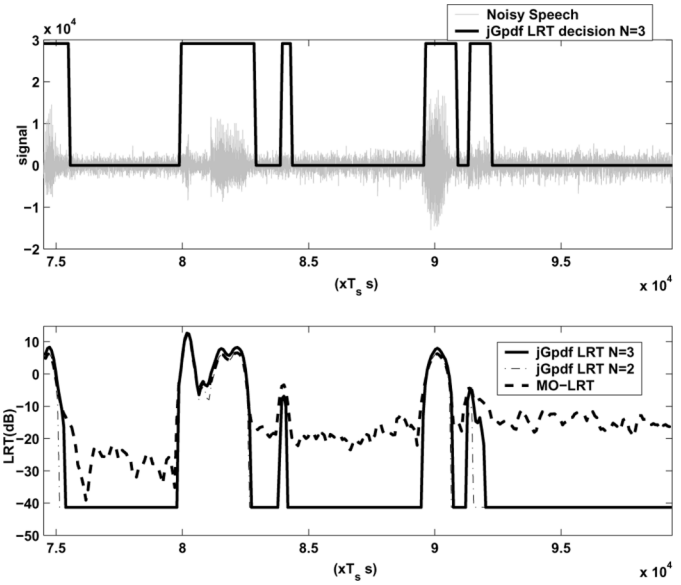

Fig. 4. Comparison among MO-LRT and second- and third-order MCO-LRTs on an utterance of Aurora 3 database. Using the same amount of information for the three algorithms, the third-order approach provides an unbiased decision which is more robust than the second approach for word ending detection.

Again, the correlation coefficients under H_0 can be neglected, thus obtaining

$$\begin{aligned} \ell_{l,3} = &\frac{1}{2}\sum_{\omega}\sum_{i=1}^{3} L_i(\omega) + 2\sqrt{\gamma_1\gamma_2}\left(\frac{\rho_1^{H_1}}{\sqrt{(1+\xi_1)(1+\xi_2)}}\right) \\ &+ 2\sqrt{\gamma_2\gamma_3}\left(\frac{\rho_2^{H_1}}{\sqrt{(1+\xi_2)(1+\xi_3)}}\right) \\ &- 2\sqrt{\gamma_1\gamma_3}\left(\frac{\rho_1^{H_1}\rho_2^{H_1}}{\sqrt{(1+\xi_1)(1+\xi_2)^2(1+\xi_3)}}\right). \end{aligned} \tag{23}$$

A generalization of the result for N observations, assuming an Nth-order tridiagonal matrix $C^N_{\mathbf{y}_{\boldsymbol{\omega}}}$, is given by

$$\ell_{l,N} = \frac{1}{2}\sum_{\omega}\left[\sum_{i=l-m}^{l+m} L_i(\omega) + \sum_{i=l-m}^{l+m-1}\frac{2\sqrt{\gamma_i\gamma_{i+1}}\rho_i^{H_1}}{\sqrt{(1+\xi_i)(1+\xi_{i+1})}}\right] \tag{24}$$

which can be recursively computed as

$$\ell_{l+1,N} = \ell_{l,N} - \Phi_{l-m}(\omega) + \Phi_{l+(m+1)}(\omega) \tag{25}$$

where

$$\Phi_{l\pm m}(\omega) = L_{l\pm m}(\omega) + \frac{2\sqrt{\gamma_{l\pm m}\gamma_{l\pm m\mp 1}}\rho^{H_1}_{l\pm m}}{\sqrt{(1+\xi_{l\pm m})(1+\xi_{l\pm m\mp 1})}}. \tag{26}$$

The accuracy of the Nth-order MCO-LRT will be similar to the previous MO-LRT [26] since the increasing model order smooths the VAD decision. However, in a real-time scenario, the end-point detection accuracy will be increased for low orders due to the correlation correction, as shown in Fig. 4. In the latter figure, a comparison among the second- and third-order jGpdf VAD and the MO-LRT is given for an utterance of the AURORA 3 database [30]. Again, the VAD decision rule of the MO-LRT is noisy and nonstationary unlike the VAD decision rule of the proposed approach. Note how the word-ending detection rate

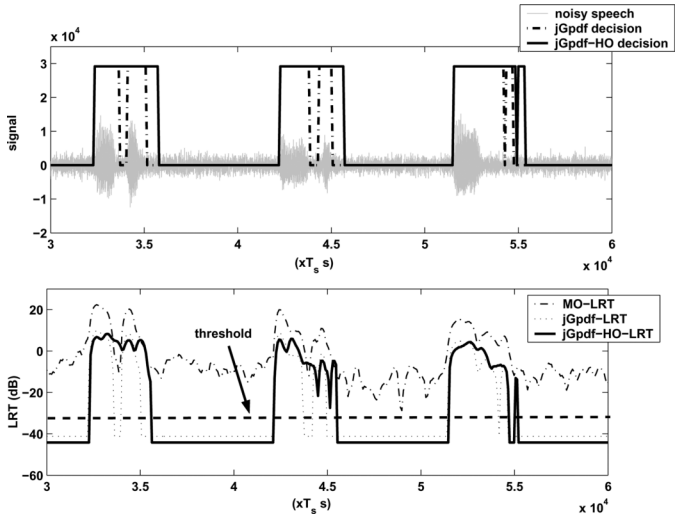

Fig. 5. Comparison among MO-LRT and second-order MCO-LRTs including the HO scheme on an utterance of Aurora 3 database. Using the same amount of information for the three algorithms, the second-order approach using a hang-over scheme provides an unbiased decision rule which is more robust than the single second-order approach for word ending detection.

of the third-order jGpdf VAD slightly outperforms the second-order approach using the same amount of data.

VI. Hang-Over Mechanism for Robust Decision

Several hang-over schemes (HO) have been proposed in order to increase the word-ending detection rate in single observation VADs [18], [25]. In this paper, we implement a very simple hang-over mechanism based on contextual information of the previous frames (if available), thus no disadvantageous delay is added to the algorithm. A smoothed LRT is defined as

$$\ell_{l,N}^{h} = \alpha \cdot \ell_{l,N} + \beta \cdot \ell_{l-l_h,N} \tag{27}$$

where the parameters l_h, α, and β are experimentally selected, i.e., $l_h = 8$, $\alpha = 0.5$, $\beta = 0.5$. Thus, we consider the same weight for both current and past decisions. This automatically improves the word ending detection since the binary nature of the decision function in the proposed approach. Using (27) leads to significant improvement in speech/nonspeech detection rate (see Fig. 5) when dealing with an almost binary and precise decision rule such as the MCO-LRT. This kind of improvement cannot be applied to the previous MO-LRT for low orders because of the changing bias of the noise periods, as shown in the same figure. Fig. 5 analyzes the effect of the hang-over scheme ($l_h = 8$, $N = 2$) on the decision function. It is clearly shown that the continuous nature of voice periods is taken into account, and again, an improvement in the detection of word endings is achieved.

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 signal-to-noise (SNR) levels [17] and the influence of the VAD decision on the performance

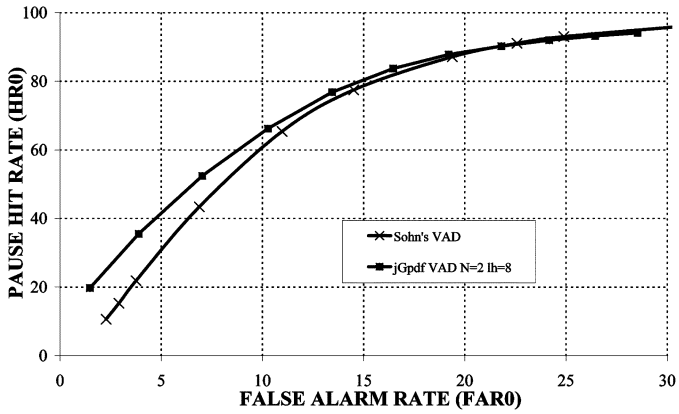

Fig. 6. ROC curves obtained in high noise conditions (5 dB) using the AURORA subset of the original Spanish SpeechDat-Car database [30].

of speech processing systems [36]. Subjective performance tests have also been considered for the evaluation of VADs working in combination with speech coders [1]. This section describes the experimental framework and the objective performance tests conducted in this paper to evaluate the proposed algorithms in the field of speech recognition.

A. Receiver Operating Characteristics (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 [37]. The AURORA subset of the original Spanish SpeechDat-Car database [30] (51.09% silence frames) was used in this analysis. This database contains 4914 recordings using close-talking and distant microphones from more than 160 speakers. The phonetic transcription of an utterance of this database is: 'tres, 'nweβe, 'θero, 'sjete, 'θinko, 'dos, 'uno, 'ot$\widehat{\int}$o, 'sejs, 'kwatro.

The files are categorized into three noise conditions: quiet, low noise, and high noise 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.

Fig. 6 compares the results of jGpdf-VAD for $N = 2$ with Sohn's VAD [18] under the same conditions (high noise condition 5 dB), i.e., using the same amount of data and a similar hang-over scheme. Note that, the ROC curve of the proposed VAD is shifted up and to the left in the ROC space. Thus, an improvement in end-point detection accuracy is achieved using a single observation, and the influence of adjacent and overlapping windows on the accuracy of the LRT based-VADs is proven. Fig. 7 shows how increasing the hang-over parameter l_h in the jGpdf-LRT VAD for $N = 3$ 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 Fig. 5 for a single utterance of the AURORA subset of the original Spanish SpeechDat-Car database [30]. Similar results are obtained for different model orders of the jGpdf-LRT VAD that exhibits a shift of the ROC

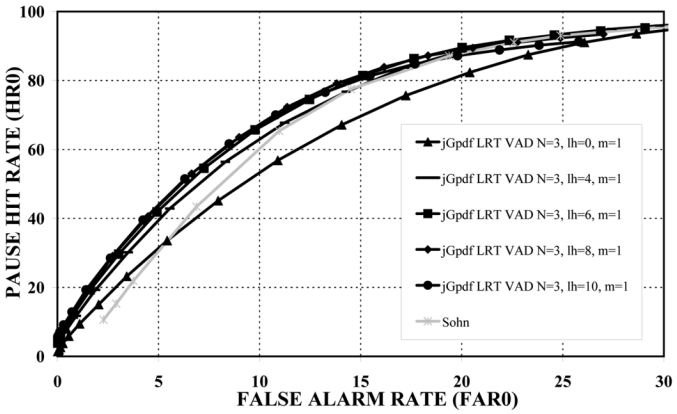

Fig. 7. Influence of the parameter l_h in the VAD detection accuracy in high noise conditions (5 dB) using the AURORA subset of the original Spanish SpeechDat-Car database [30].

curve when the hang-over parameter (l_h) increases. Finally, the benefits of contextual information (and an inherent delay) can be incorporated into our algorithm just averaging the decision rule over a set of Nth-order joint observations, i.e., $N = 2, 3$, so that, the contextual decision function is computed as

$$\tilde{\ell}_{N,m'} = \frac{1}{(2m'+1)\sum_{i=l-m'}^{l+m'} \ell_{i,N}}. \tag{28}$$

Figs. 9 and 8 show the ROC curves of the proposed VADs for different conditions (quiet, low, and high) and other frequently referred and standard algorithms [2], [3], [15]–[18] for recordings from the distant microphone. The working points of the G.729, AMR and AFE VADs are also included. A typical value of $m' = 8$ yields the best discrimination accuracy as shown in the ROC curve in Fig. 9. It is interesting to point out that 1) the proposed VAD using contextual information yields the best results over the standard and recently reported VADs, 2) the accuracy of the contextual VAD for $m' = 8$ is almost independent of the model order as shown in Fig. 9 and equivalent to the previous MO-LRT [26] (not shown for clarity), and 3) for a single observation the proposed VAD provides higher HR0 for a given false alarm FAR0 in the speech recognition working area than Sohn's VAD, which is equivalent to the MO-LRT for a single observation, and the other referred VADs, excluding Li's algorithm [16]. The latter algorithm [16] shows higher HR0 at the same FAR0 than the single observation-based approaches since it uses a prefiltering stage and *contextual information*. It is also shown in the ROC curve that the use of a third-order model provides better discrimination results than the second-order approach for single observation.

Thus, among all the VAD schemes 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 Li's algorithm [16], that is based on an optimum linear filter for edge detection. The proposed VAD also improves Marzinzik VAD [17] that tracks the power spectral envelopes, and the Sohn VAD [18] that formulates

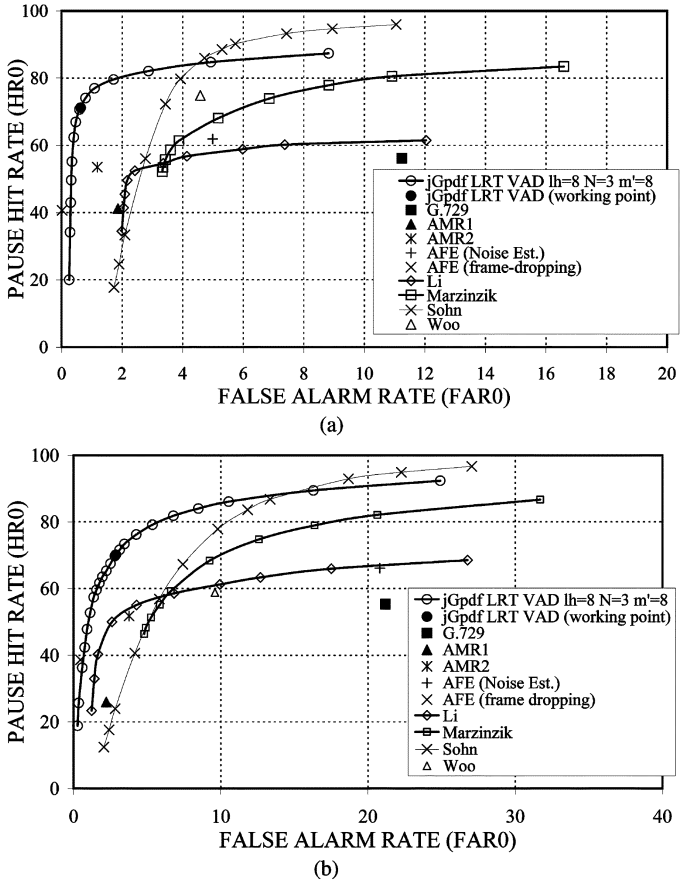

Fig. 8. ROC curves for standard and recently reported VADs obtained using the AURORA subset of the original Spanish SpeechDat-Car database (a) in quiet noise condition (25 dB) and (b) in low noise condition (15 dB).

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 experiments described above yields a first measure of the performance of the VAD. Other measures of VAD performance that have been reported are the clipping errors [1]. 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.

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 [15] since non-efficient speech/nonspeech classification is an important source of performance degradation for speech recognition systems working in noisy environments [38]. This section evaluates the

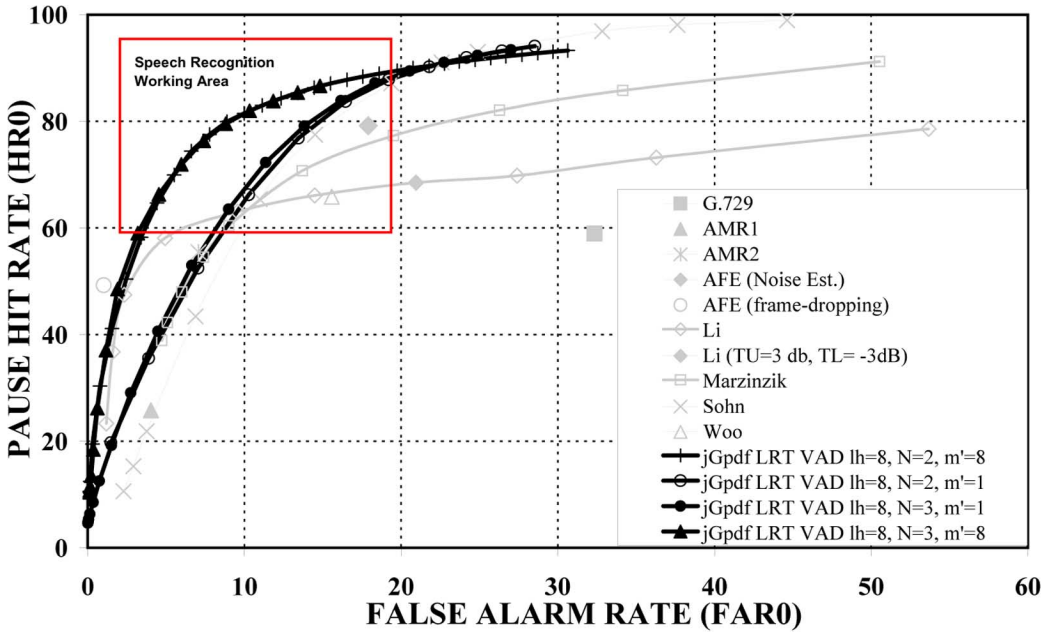

Fig. 9. ROC curves for standard and recently reported VADs obtained in high noise condition (5 dB) using the AURORA subset of the original Spanish SpeechDat-Car database [30].

VAD according to the objective it was developed for, that is, by assessing the influence of the VAD in a speech recognition system.

Fig. 1 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 [39]. An enhanced feature extraction scheme incorporating a noise reduction algorithm and nonspeech frame-dropping (FD) was built on the base system [39]. The noise reduction algorithm has been implemented as a single Wiener filtering (WF) stage as described in the AFE standard [4] 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.

The recognizer is based on the HTK (Hidden Markov Model Toolkit) software package [40]. The task consists of recognizing connected digits which are modeled as whole word HMMs 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. In speech recognition experiments, the selection of the VAD threshold is based on the results obtained in detection experiments (working points in ROCs for all conditions). The working point (selected threshold) should correspond with the best tradeoff between the hit rate and false alarm rate, then the threshold is adaptively chosen depending on the noisy condition. The algorithm for fixing the threshold is similar to that used in the AMR1 standard [3] and other previous works [28], [41]. It is assumed that the system will work at different noisy conditions and that an optimal threshold can be determined for the system working in the cleanest and noisiest conditions [27], [41].

Finally, recognition performance is assessed in terms of the word accuracy (WAcc) which takes into account the number of substitution errors (S), deletion errors (D), and insertion errors (I)

$$\text{WAcc}(\%) = \frac{N - D - S - I}{N} \times 100\% \tag{29}$$

where N is the total number of words in the testing database.

1) Results for the Aurora 3 Database: 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 the 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.

Table I shows the recognition performance for the Spanish SpeechDat-Car database when WF and FD are performed on the base system [39]. Again, the jGpdf-VAD outperforms all the algorithms used for reference yielding relevant improvements in speech recognition as the MO-LRT algorithm [26]. In addition, not only the delayed LRT version averaging over MCO for $m' = 8$ and $N = 3$ obtains the best recognition results, but the single observation vector clearly improves the recognition word accuracy of the previous LRT by Sohn (or MO-LRT for single observation) [18] and other standardized VADs used in the WF and FD stages. Note how significant the improvement is on Li's VAD since the working point of the latter algorithm causes the insertion of many pause frames in the recognition

TABLE I
AVERAGE WORD ACCURACY (%) FOR THE SPANISH SDC DATABASES. ($*$: $N = 3$, $m = 1$; $**$: $N = 3$, $m' = 8$). COMPARISON TO: (a) STANDARD VADS AND (b) RECENTLY PUBLISHED VAD METHODS

(a)

	G.729	AMR1	AMR2	AFE	Base	**jGpdf-LRT***	**jGpdf-LRT ****
WM	88.62	94.65	95.67	95.28	92.94	**96.20**	**96.29**
MM	72.84	80.59	90.91	90.23	83.31	**91.73**	**91.88**
HM	65.50	62.41	85.77	77.53	51.55	**84.62**	**87.01**
Avg.	75.65	74.33	90.78	87.68	75.93	**90.85**	**91.73**

(b)

	Woo	Li	Marzinzik	Sohn	MO-LRT	**jGpdf-LRT***	**jGpdf-LRT ****
WM	95.35	91.82	94.29	96.07	96.33	**96.20**	**96.29**
MM	89.30	77.45	89.81	91.64	91.61	**91.73**	**91.88**
HM	83.64	78.52	79.43	84.03	87.43	**84.62**	**87.01**
Avg.	89.43	82.60	87.84	90.58	91.79	**90.85**	**91.73**

TABLE II
INSERTIONS, DELETIONS, AND SUBSTITUTIONS FOR THE PROPOSED MO-JGPDF VAD IN WF AND WF + FD OPERATION

	# Deletions			# Substitutions			# Insertions		
	WM	MM	HM	WM	MM	HM	WM	MM	HM
WF	127	136	80	454	166	233	440	401	379
WF + FD	114	135	75	205	144	108	98	104	103

TABLE III
AVERAGE WORD ACCURACY FOR CLEAN AND MULTICONDITION AURORA-2 TRAINING/TESTING EXPERIMENTS. COMPARISON TO: (a) STANDARD VADS AND (b) RECENTLY PUBLISHED VAD METHODS

(a)

	G.729	AMR1	AMR2	AFE	**jGpdf-LRT***	**Hand-labelling**
Base + WF	66.19	74.97	83.37	81.57	**83.94**	**84.69**
Base + WF+ FD	70.32	74.29	82.89	83.29	**85.31**	**86.86**

(b)

	Woo	Li	Marzinzik	Sohn	**jGpdf-LRT****	**MO-LRT**
Base + WF	83.64	77.43	84.02	83.89	**84.30**	**84.33**
Base + WF+ FD	81.09	82.11	85.23	83.80	**86.21**	**86.14**

training. Thus, the jGpdf VAD for $m = 1$ and $N = 3$ is the choice to be taken when dealing with real-time application since it provides the best tradeoff between computational delay and speech recognition and detection rate.

Note that these particular databases used in the AURORA 3 experiments have prolonged nonspeech periods unlike the AURORA 2 database [42], and then the effectiveness of the VAD results turn even more important for the speech recognition system. This fact can be clearly shown when comparing the performance of the proposed VAD to Marzinzik VAD [17]. The word accuracy of both VADs is quite similar for the AURORA 2 task (see Table I in Section VIII). However, the proposed VAD yields a significant performance improvement over Marzinzik VAD [17] for the AURORA 3 database as shown in Table I.

The number of deletions, substitutions, and insertions for AURORA 3 when the proposed MO-jGpdf VAD is used within the speech recognition system for noise filtering, and optionally, for frame dropping is shown in Table II. We clearly observe the reduction on the number of substitutions and insertions in the WF-FD operation.

2) Results for the Aurora 2 Database: The original AURORA-2 database [42] consists of sequences of up to seven connected digits spoken by American English talkers which used as source speech (clean TIdigits database with 43.72% silence frames), and a selection of eight different real-world noises which are artificially added at SNRs from 20 to −5 dB. These noisy signals represent the most probable application scenarios for telecommunication terminals (suburban train, babble, car, exhibition hall, restaurant, street, airport, and train station). Two training modes are defined for the experiments conducted on the AURORA-2 database: 1) training on clean data only (Clean Training), and 2) training on clean and noisy data (Multicondition Training).

Table III shows the recognition performance achieved by the different VADs that were compared. These results are averaged over the three test sets of the AURORA-2 recognition experiments and SNRs between 20 and 0 dBs. Note that, for the recognition experiments based on the AFE VADs, the same configuration of the standard [4], which considers different VADs for WF and FD, was used. The proposed VAD, using contextual information, yields the same recognition accuracy than the previous version [26], and 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. The improvements shown in Tables I and III are mainly due to:

- a reduction of the number of substitution errors when the VAD is only used for WF-based speech enhancement;
- a significant reduction of the number of insertion errors (especially when the HMM models are trained using clean speech) when the VAD is additionally used for non-speech frame-dropping (this reduction is just slightly prejudiced by a corresponding increase in the number of deletions so that the overall ASR performance is significantly improved).

VIII. Conclusion

This paper presented a novel VAD for improving speech detection robustness in noisy environments. The proposed method is developed on the basis of previous proposals which incorporate long-term speech information into the decision rule. However, the assumption of independence between observations was alleviated since this hypothesis is not realistic at all. The proposed algorithm defined an optimal likelihood ratio test based on multiple and correlated observation vectors, which avoids the need of smoothing the VAD decision, in order to offer significant benefits for speech/pause detection in noisy environments. The algorithm had an optional inherent delay when the average of the decision function over MO-windows was considered but for several applications, including robust speech recognition, does not represent a serious implementation obstacle. However, for real-time applications this computational delay poses a major problem since the reduction of the number of observations leads to a significant degradation in the accuracy of speech processing systems. To avoid this deficiency a more suitable model for this scenario was introduced and its computational burden was also discussed. Several experiments were conducted to evaluate this approach for $N = 2$ and $N = 3$. An analysis based on the ROC curves revealed a clear reduction of the classification error for two observations and also when the number of observations is increased as expected. In this way, the proposed jGpdf-VAD outperformed, *under the same conditions*, Sohn's VAD which assumes a single observation in the decision rule and uses a HMM-based hangover, but also other methods including the standardized G.729, AMR, and AFE VADs, and other recently reported VAD methods in both speech/nonspeech detection performance.

In the Appendix A, complete recursion for the computation of a jGpdf over a sliding MO window is derived. This recursion is based on the recursive computation of the inverse and determinant of any symmetric tridiagonal matrix which has many advantages in real-time applications.

Appendix

Computation of the LRT for $N = 2$: In this section, a complete derivation of the MCO-LRT is derived from the GCG observation model for $N = 2$. From (4) and considering the statistical dependence between adjacent observations, we have that the MCO-LRT can be expressed as

$$\ell_{l,2} = \sum_{\omega} \ln \frac{K_{H_1,2}}{K_{H_0,2}} + \frac{1}{2}\hat{\mathbf{y}}_{\omega}^{T}\Delta_2^{\omega}\hat{\mathbf{y}}_{\omega} \tag{30}$$

where (for clarity we have omitted the frequency dependence of the parameters)

$$\ln \frac{K_{H_1,2}}{K_{H_0,2}} = \frac{1}{2} ln \left(\frac{\left|C^N_{\mathbf{y}_{\boldsymbol{\omega}},H_0}\right|}{\left|C^N_{\mathbf{y}_{\boldsymbol{\omega}},H_1}\right|} \right) = \frac{1}{2} \frac{\sigma_1^{H_0}\sigma_2^{H_0} - \left(r_1^{H_0}\right)^2}{\sigma_1^{H_1}\sigma_2^{H_1} - \left(r_1^{H_1}\right)^2} \tag{31}$$

and $C_{\mathbf{y}_{\boldsymbol{\omega}}}$ is defined as in (20). If we assume that the voice signal is observed in independent additive noise, that is, for each observation $i = 1, 2$

$$\begin{aligned} H_1: &\quad \sigma_i^{H_1} = \sigma_i^n + \sigma_i^s \\ H_0: &\quad \sigma_i^{H_0} = \sigma_i^n \end{aligned} \tag{32}$$

then, the second term in (6) can be expressed as

$$\ln \frac{K_{H_1,2}}{K_{H_0,2}} = \frac{1}{2}\left[\ln \left(\frac{1 - \left(\rho_1^{H_0}\right)^2}{1 - \left(\rho_1^{H_1}\right)^2} \right) - \ln \left(\prod_{i=1}^{2} (1 + \xi_i) \right) \right] \tag{33}$$

where $\rho_1^{H_s} \equiv \left(r_1^{H_1} / \sqrt{\sigma_1^{H_1}\sigma_2^{H_1}} \right)$ is the correlation coefficient under H_1, and $\xi_i(\omega) \equiv \left(\sigma_i^s(\omega)/\sigma_i^n(\omega)\right)$ is the *a priori* SNR.

On the other hand, the inverse of the covariance matrix is expressed in terms of the orthogonal complex polynomials $q_k(z)$, $p_k(z)$ as

$$\left(C^2_{\mathbf{y}_{\boldsymbol{\omega}},H_s}\right)^{-1} = \begin{pmatrix} \left[\frac{q_0}{p_0} - \frac{q_2}{p_2}\right] p_0 p_0 & \left[\frac{q_1}{p_1} - \frac{q_2}{p_2}\right] p_0 p_1 \\ \left[\frac{q_1}{p_1} - \frac{q_2}{p_2}\right] p_0 p_1 & \left[\frac{q_1}{p_1} - \frac{q_2}{p_2}\right] p_1 p_1 \end{pmatrix}_{H_s} \tag{34}$$

where $p_0 = 1$, $q_0 = 0$, $p_1 = -\sigma_1/r_1$, and $q_2/p_2 = \sigma_2/\left(r_1^2 - \sigma_1\sigma_2\right)$ under hypothesis H_s. Thus, the second term of (30) can be expressed as

$$\hat{\mathbf{y}}_{\omega}^{T}\Delta_2^{\omega}\hat{\mathbf{y}}_{\omega} = \left(y_1^{\omega}\right)^2 \left(\Delta_2^{\omega}\right)_{00} + \left(y_2^{\omega}\right)^2 \left(\Delta_2^{\omega}\right)_{11} + 2y_1^{\omega}y_2^{\omega}\left(\Delta_2^{\omega}\right)_{01} \tag{35}$$

where

$$\left(\Delta_2^{\omega}\right)_{00} = \left(\sigma_2^{H_0}/\sigma_2^{H_0}\sigma_1^{H_0} - \left(r_1^{H_0}\right)^2\right) - \left(\sigma_2^{H_1}/\sigma_2^{H_1}\sigma_1^{H_1} - \left(r_1^{H_1}\right)^2\right)$$

$$\left(\Delta_2^{\omega}\right)_{11} = \left(\sigma_1^{H_0}/\sigma_2^{H_0}\sigma_1^{H_0} - \left(r_1^{H_0}\right)^2\right) - \left(\sigma_1^{H_1}/\sigma_2^{H_1}\sigma_1^{H_1} - \left(r_1^{H_1}\right)^2\right)$$

and

$$\left(\Delta_2^{\omega}\right)_{01} = \left(r_1^{H_0}/\left(r_1^{H_0}\right)^2 - \sigma_2^{H_0}\sigma_1^{H_0}\right) - \left(r_1^{H_1}/\left(r_1^{H_0}\right)^2 - \sigma_2^{H_0}\sigma_1^{H_0}\right).$$

Finally, if we define the *a posteriori* SNR $\gamma_i(\omega) \equiv \left(y_i^{\omega}\right)^2/\sigma_i^n(\omega)$ and neglect the squared correlation functions under both hypotheses, (18) is obtained.

A. Computation of the LRT for $N = 3$

It is easy to prove (22) just following the methodology proposed in the previous section for $N = 3$. We note that in this case

$$\ln \frac{K_{H_1,3}}{K_{H_0,3}} = \frac{1}{2}\left[\ln\left(\frac{1-\left(\rho_1^2+\rho_2^2\right)^{H_0}}{1-\left(\rho_1^2+\rho_2^2\right)^{H_1}}\right) - \ln\left(\prod_{i=1}^{3}(1+\xi_i)\right)\right] \quad (36)$$

and Δ_3^{ω} is computed using the following expression under hypotheses $H_0 \& H_1$:

$$\begin{aligned}\hat{\mathbf{y}}_{\omega}^T \left(C^3_{\mathbf{y}_{\omega},H_s}\right)^{-1} \hat{\mathbf{y}}_{\omega} &= \frac{1}{1-(\rho_1^2+\rho_2^2)} \\ &\times \left[\left(\frac{1-\rho_2^2}{\sigma_1}\left(y_1^{\omega}\right)^2\right) + \frac{\left(y_2^{\omega}\right)^2}{\sigma_2} \cdots \right. \\ &+ \left(\frac{1-\rho_1^2}{\sigma_3}\left(y_3^{\omega}\right)^2\right) - 2\rho_1 \frac{y_1^{\omega} y_2^{\omega}}{\sqrt{\sigma_1\sigma_2}} \\ &\left. - 2\rho_2 \frac{y_2^{\omega} y_3^{\omega}}{\sqrt{\sigma_2\sigma_3}} + 2\rho_1\rho_2 \frac{y_1^{\omega} y_3^{\omega}}{\sqrt{\sigma_1\sigma_3}}\right]. \end{aligned} \quad (37)$$

The selection of a tridiagonal symmetric covariance matrix provides the generalization for the Nth order. In this case, we have to include all the cross-correlation terms of the observations as in (22), which provides additional benefits in the observation model. However, the improvement on the old MO-LRT [26] for higher order is minor due to the reduced variance in the decision rule of both approaches.

B. Proof for the Recursive Computation of the Determinant

Given an Nth-order symmetric tridiagonal matrix $C^N_{\mathbf{y}_{\boldsymbol{\omega}}}$

$$C^N_{\mathbf{y}_{\boldsymbol{\omega}}} = \begin{pmatrix} \sigma_1 & r_1 & 0 & \cdots & \\ r_1 & \sigma_2 & r_2 & 0 & \cdots \\ 0 & r_2 & \sigma_3 & r_3 & \cdots \\ \vdots & & & & \\ 0 & \cdots & & r_{N-1} & \sigma_N \end{pmatrix} \quad (38)$$

its determinant function satisfies (16). Equivalently

$$\left|C^N_{\mathbf{y}_{\boldsymbol{\omega}}}\right| = \sigma_1 \left|\hat{C}^{N-1}_{\mathbf{y}_{\boldsymbol{\omega}}}\right| - r_1^2 \left|\hat{C}^{N-2}_{\mathbf{y}_{\boldsymbol{\omega}}}\right| \quad (39)$$

where $\hat{}$ denotes the determinant of the matrix that consist of high-index elements. Let us consider three consecutive MCO-windows and denote by $A_{N,j}$ for $j = l-2, l-1, l$ their determinant functions (see Fig. 10). They satisfy the following relations:

$$\begin{aligned} A_{N,l} &= \sigma_N^l A_{N-1,l} + (r_{N-1}^l)^2 A_{N-2,l} \\ A_{N,l-1} &= \sigma_1^{l-1} A_{N-1,l} + (r_1^{l-1})^2 A_{N-2,l} \\ A_{N,l-2} &= \sigma_1^{l-2} A_{N-1,l-2} + (r_1^{l-2})^2 A_{N-2,l} \end{aligned} \quad (40)$$

where the relations $\hat{A}_{N-2,l-1} = A_{N-2,l}$ and $\hat{A}_{N-2,l-2} = A_{N-2,l}$ must be considered. Thus, introducing the second and third relation in the expression for $A_{N,l}$ we derive the recursive relation for the computation of the determinant function at current frame l in terms of frames *l-1* and *l-2* as shown in (15).

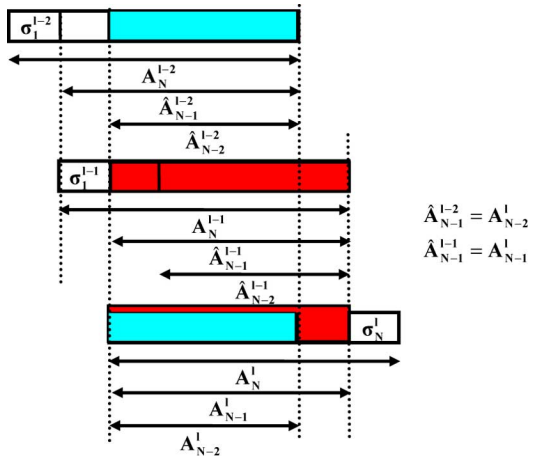

Fig. 10. Connection among the determinants of adjacent symmetric tridiagonal matrices.

C. Recursive Computation of the Inverse of a Symmetric Tridiagonal Matrix

In the computation of the (8), the term which requires the most computational effort is the polynomial fraction [32]. Thus our discussion is mainly focussed on deriving a recursive implementation for this fraction of orthogonal polynomials. Given an Nth-order tridiagonal symmetric matrix at time l

$$D_N^l = \begin{pmatrix} a_0 & b_0 & 0 & 0 & \cdots & 0 \\ b_0 & a_1 & b_1 & 0 & \cdots & 0 \\ 0 & b_1 & a_2 & b_2 & \cdots & 0 \\ \vdots & & & & a_{N-2} & b_{N-2} \\ 0 & \cdots & & 0 & b_{N-2} & a_{N-1} \end{pmatrix} \quad (41)$$

and its inverse $\left(D_N^l\right)^{-1}$ (that is the fractions q_k/p_k, for $k = 1, \ldots, N$ are already known) we have to derive D_N^{l+1} in terms of the previous state l. Let $\tilde{D}_N^l$ denotes the transpose matrix with respect the main inverse diagonal, that is

$$\tilde{D}_N^l = \begin{pmatrix} a_{N-1} & b_{N-2} & 0 & 0 & \cdots & 0 \\ b_{N-2} & a_{N-2} & b_{N-3} & 0 & \cdots & 0 \\ 0 & b_{N-3} & a_{N-3} & b_{N-4} & \cdots & 0 \\ \vdots & & & & a_1 & b_0 \\ 0 & \cdots & & 0 & b_0 & a_0 \end{pmatrix}. \quad (42)$$

The inverse matrix of $\tilde{D}_N^l$ is also known since $\left(\tilde{D}_N^l\right)^{-1} = \widetilde{\left[\left(D_N^l\right)^{-1}\right]}$. From this new defined matrix, it is easy to compute $(\tilde{D}_{N+1}^l)^{-1}$ using the following expression (Method 2):

$$\left[\frac{q_{N+1}^l(z)}{p_{N+1}^l(z)}\right] = \frac{1}{(z-a_N)-} \ominus \left[\frac{q_N^l(z)}{p_N^l(z)}\right] b_{N-1} \quad (43)$$

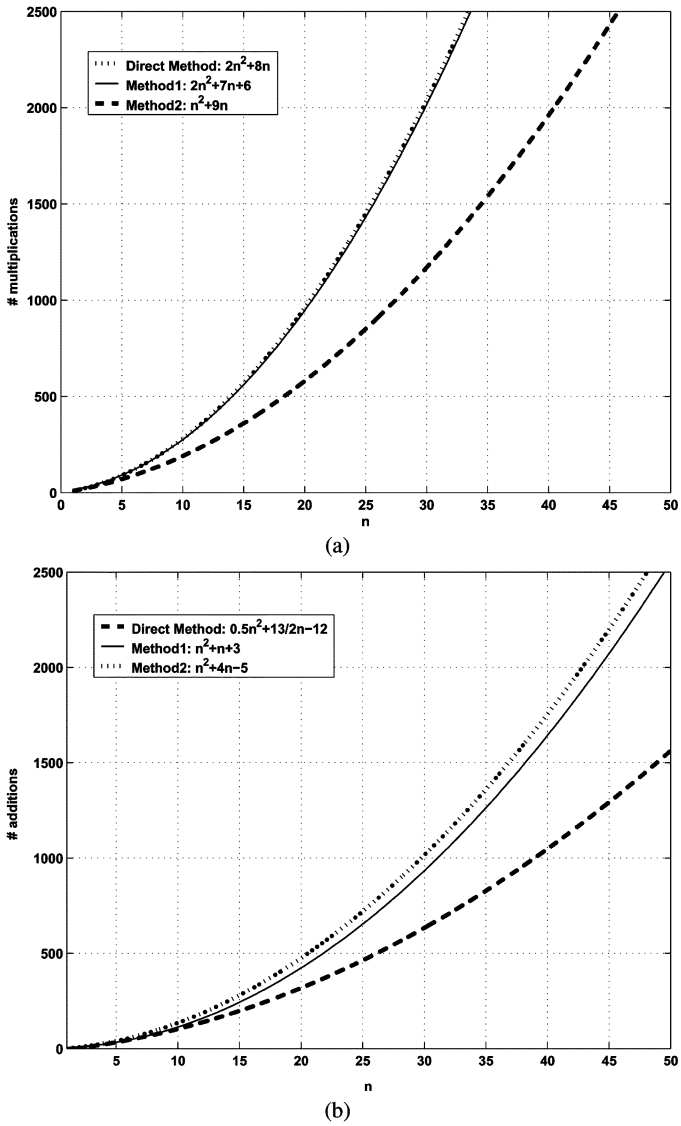

Fig. 11. Computational complexity reduction in the computation of the inverse matrix. (a) Number of multiplications needed in the Direct method and methods 1 and 2. (b) Number of additions needed in the Direct method and methods 1 and 2.

(or using the original matrix and the recursive relations for the polynomials q_j and p_j given in Section III, Method 1), where the polynomial fractions $\left(q_j^l(z)/p_j^l(z)\right), j=0,\ldots,N-1$ are already computed at frame l using (9) with initial values

$$p_0 = 1 \text{ and } p_1(z) = \frac{z - a_{N-1}}{b_{N-1}}$$
$$q_0 = 0 \text{ and } q_1(z) = \frac{1}{b_{N-1}}. \tag{44}$$

Once the inverse of the matrix for the $N+1$th order at frame l $\left(D_{N+1}^l\right)^{-1}$ is obtained, the computation of the Nth order matrix at frame $l+1$ $\left(D_N^{l+1}\right)^{-1}$ is straight forward. Since the previous matrix $\tilde{D}_{N+1}^l$ (equivalently D_{N+1}^l) is related to $\tilde{D}_N^{l+1}$ (resp. to D_N^{l+1}) as

$$\tilde{D}_{N+1}^l = \begin{pmatrix} & & \vdots \\ & \tilde{D}_N^{l+1} & 0 \\ & & b_0 \\ \cdots & 0 \quad b_0 & a_0 \end{pmatrix}$$
$$D_{N+1}^l = \begin{pmatrix} a_0 & b_0 & 0 & \cdots \\ b_0 & & & \\ 0 & & D_N^{l+1} & \\ \vdots & & & \end{pmatrix}. \tag{45}$$

If we define the N-dimensional vector $\mathbf{b_0} \equiv (b_0, 0, \ldots)^T$, we can use the Volker–Strassen formula [34] to connect the inverse of the matrices as

$$\left(D_{N+1}^l\right)^{-1} = \begin{pmatrix} a_0 & \mathbf{b_0}^T \\ \mathbf{b_0} & D_N^{l+1} \end{pmatrix}^{-1} = \begin{pmatrix} a_0^{-1} + \frac{\mathbf{b_0}^T (C_N)^{-1} \mathbf{b_0}}{a_0^2} & -a_0^{-1}\mathbf{b_0}^T (C_N)^{-1} \\ -(C_N)^{-1}\mathbf{b_0} a_0^{-1} & (C_N)^{-1} \end{pmatrix} \tag{46}$$

where $C_N = D_N^{l+1} - \mathbf{b_0} a_0^{-1} \mathbf{b_0}^T$. Finally, we define $\mathbf{u} = (bo/a_0, 0, \ldots)^T$ and use the Woodbury's identity [35] to establish

$$\left(D_N^{l+1}\right)^{-1} = (C_N)^{-1} - \frac{\left((C_N)^{-1}\mathbf{u}\right)\left(\mathbf{b_0}^T (C_N)^{-1}\right)^T}{1 + \mathbf{b_0}^T (C_N)^{-1}\mathbf{u}}. \tag{47}$$

Using this methodology, we reduce the number of multiplications needed to compute the inverse of a symmetric tridiagonal matrix unlike the number of additions as shown in Fig. 11. Anyway, this procedure means a clear reduction in computational complexity and a novel methodology for continuous real-time matrix inversion.

ACKNOWLEDGMENT

The authors would like to thank Mrs. M. E. C. Lara for help in writing this paper in English.

REFERENCES

[1] 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.*, vol. 35, no. 9, pp. 64–73, Sep. 1997.

[2] "A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70," ITU, ITU-T Rec. G.729-Annex B, 1996.

[3] "Voice activity detector (VAD) for adaptive multi-rate (AMR) speech traffic channels," ETSI, ETSI EN 301 708 Rec., 1999.

[4] "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," ETSI, ETSI ES 202 050 Rec., 2002.

[5] R. M. Gray, "The 1974 origins of VoIP," *IEEE Signal Process. Mag.*, vol. 22, no. 4, pp. 87–90, Jul. 2005.

[6] L. Ding, A. Radwan, M. El-Hennawey, and R. Goubran, "Measurement of the effects of temporal clipping on speech quality," *IEEE Trans. Instrum. Meas.*, vol. 55, no. 4, pp. 1197–1203, Aug. 2005.

[7] C. Wang, K. Sohraby, R. Jana, J. Lusheng, and M. Daneshmand, "Voice communications over ZigBee networks," *IEEE Commun. Mag.*, vol. 46, no. 1, pp. 121–127, Jan. 2008.

[8] L. Sun and E. Ifeachor, "Voice quality prediction models and their application in VoIP networks," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 809–820, Aug. 2006.
[9] R. L. Bouquin-Jeannes and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech Commun.*, vol. 16, pp. 245–254, 1995.
[10] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1979, pp. 208–211.
[11] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.
[12] R. L. Bouquin-Jeannes and G. Faucon, "Proposal of a voice activity detector for noise reduction," *Electron. Lett.*, vol. 30, no. 12, pp. 930–932, 1994.
[13] J. H. L. Hansen, X. X. Zhang, M. Akbacak, U. H. Yapanel, and B. Pellom, "Cu-move: Advanced in-vehicle speech systems for route navigation," in *DSP for In-Vehicle and Mobile Systems*. New York: Springer-Verlag, 2004.
[14] M. Akbacak and J. H. L. Hansen, "Environmental sniffing: Noise knowledge estimation for robust speech systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 465–477, Feb. 2007.
[15] K. Woo, T. Yang, K. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electron. Lett.*, vol. 36, no. 2, pp. 180–181, 2000.
[16] 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.*, vol. 10, no. 3, pp. 146–157, Mar. 2002.
[17] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 2, pp. 341–351, Feb. 2002.
[18] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 16, no. 1, pp. 1–3, Jan. 1999.
[19] R. Chengalvarayan, "Robust energy normalization using speech/nonspeech discriminator for German connected digit recognition," in *Proc. Eurospeech*, Budapest, Hungary, Sep. 1999, pp. 61–64.
[20] R. Tucker, "Voice activity detection using a periodicity measure," *IEE Proc. Commun., Speech, Vision*, vol. 139, no. 4, pp. 377–380, 1992.
[21] J. M. Górriz, J. Ramírez, C. G. Puntonet, and J. C. Segura, "Generalized LRT-based voice activity detector," *IEEE Signal Process. Lett.*, vol. 13, no. 10, pp. 636–639, Oct. 2006.
[22] 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.*, vol. 13, no. 8, pp. 497–500, Aug. 2006.
[23] S. G. Tanyer and H.Özer, "Voice activity detection in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 478–482, Jul. 2000.
[24] Y. D. Cho, K. Al-Naimi, and A. Kondoz, "Improved voice activity detection based on a smoothed statistical likelihood ratio," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2001, vol. 2, pp. 737–740.
[25] J.-H. Chang, J. W. Shin, and N. S. Kim, "Voice activity detector employing generalised Gaussian distribution," *Electron. Lett.*, vol. 40, no. 24, pp. 1561–1563, 2004.
[26] 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.*, vol. 12, no. 10, pp. 837–844, Oct. 2001.
[27] J. M. Górriz, J. Ramírez, J. C. Segura, and C. G. Puntonet, "An effective cluster-based model for robust speech detection and speech recognition in noisy environments," *J. Acoust. Soc. Amer.*, vol. 120, no. 470, pp. 470–481, 2006.
[28] J. M. Górriz, J. Ramírez, E. W. Lang, and C. G. Puntonet, "Hard c-means clustering for voice activity detection," *Speech Commun.*, vol. 44, pp. 1638–1649, 2006.
[29] J. M. Górriz, J. Ramirez, J. C. Segura, and C. G. Puntonet, "An improved MO-LRT VAD based on a bispectra Gaussian model," *Electron. Lett.*, vol. 41, no. 15, pp. 877–879, 2005.
[30] 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 *Proc. II LREC Conf.*, 2000, CD-ROM.
[31] B. Manly, *Multivariate Statistical Methods: A Primer*. London, New York: Chapman & Hall, 1986.
[32] H. Yamani and M. Abdelmonem, "The analytic inversion of any finite symmetric tridiagonal matrix," *J. Phys. A: Math. Gen.*, vol. 30, pp. 2889–2893, 1997.
[33] N. Akhiezer, *The Classical Moment Problem*. Edinburgh, U.K.: Oliver and Boyd, 1965.
[34] C. L. T. H. Cormen, R. Rivest, and C. Stein, *Introduction to Algorithms*. New York: MIT Press and McGraw-Hill, 2001.
[35] S. Kay, *Modern Spectrum Estimation: Theory and Applications*. Englewood Cliffs: Prentice-Hall, 1988.
[36] R. L. Bouquin-Jeannes and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech Commun.*, vol. 16, pp. 245–254, 1995.
[37] J. Ramírez, J. C. Segura, M. C. Benítez, A. d. l. Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, vol. 42, no. 3-4, pp. 271–287, 2004.
[38] L. Karray and A. Martin, "Towards improving speech detection robustness for speech recognition in adverse environments," *Speech Commun.*, no. 3, pp. 261–276, 2003.
[39] "Speech processing, transmission, and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithms," ETSI, ETSI ES 201 108 Rec., 2000.
[40] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge, U.K.: Cambridge Univ. Press, 1997.
[41] J. Ramírez, J. C. Segura, C. Benítez, A. d. l. Torre, and A. Rubio, "An effective subband OSF-based VAD with noise reduction for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 6, pp. 1119–1129, Nov. 2005.
[42] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions," in *Proc. ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France, Sep. 2000, CD-ROM.

Juan Manuel Górriz received the B.Sc. degree in physics and electronic engineering from the University of Granada, Granada, Spain, and the Ph.D. degree from the Universities of Cádiz and Granada, Spain, in 2000, 2001, 2003, and 2006, respectively.

He is currently an Associate Professor with the Department of Signal Theory, Networking, and Communications at the University of Granada. He has coauthored more than 140 technical journals and conference papers in these areas and has served as Editor in Chief for the *Open Acoustics Journal*, Bentham, since 2007. His present interests lie in the field of statistical signal processing and its application to speech and image processing.

Javier Ramírez received the M.A.Sc. and Ph.D. degrees in electronic engineering from the University of Granada, Granada, Spain, in 1998 and 2001, respectively.

Since 2001, he has been an Associate Professor in the Department of Signal Theory Networking, and Communications, University of Granada. His research interest includes robust speech recognition, speech enhancement, voice activity detection and design, seismic signal processing, and implementation of high-performance digital signal processing systems. He has coauthored more than 100 technical journal and conference papers in these areas. He has served as reviewer for several international journals and conferences.

Elmar W. Lang received the physics diploma in 1977, the Ph.D. degree in physics (summa cum laude) in 1980, and habilitated in biophysics in 1988 from the University of Regensburg, Regensburg, Germany.

He is Adjunct Professor of biophysics at the University of Regensburg, where he is heading the Neuro- and Bioinformatics Group. Currently, he serves as an Associate Editor of *Neurocomputing and Neural Information Processing-Letters and Reviews*. His current research interests include biomedical signal and image processing, independent component analysis and blind source separation, neural networks for classification and pattern recognition, and stochastic process limits in queuing applications.

Carlos G. Puntonet was born in Barcelona, Spain, on August 11, 1960. He received the B.Sc., M.Sc., and PhD. degrees in electronics physics from the University of Granada, Granada, Spain, in 1982, 1986, and 1994, respectively.

He has been a Visiting Researcher at the Laboratorie de Traitement d'Images et Reconnaissance de Formes (INPG, Grenoble, France), at the Institute of Biophysics (Regensburg, Germany), and at the Institute of Physical and Chemical Research (RIKEN, Nagoya, Japan). Currently, he is an Associate Professor in the Department of Architecture and Computer Technology, University of Granada. His research interests lie in the fields of signal and image processing, independent component analysis and blind separation of sources, artificial neural networks, optimization methods, and biomedical applications.