Improved Voice Activity Detection Using Contextual Multiple Hypothesis Testing for Robust Speech Recognition

Javier Ramírez, José C. Segura, *Senior Member, IEEE*, Juan M. Górriz, and Luz García

***Abstract*—This paper shows an improved statistical test for voice activity detection in noise adverse environments. The method is based on a revised contextual likelihood ratio test (LRT) defined over a multiple observation window. The motivations for revising the original multiple observation LRT (MO-LRT) are found in its artificially added hangover mechanism that exhibits an incorrect behavior under different signal-to-noise ratio (SNR) conditions. The new approach defines a maximum *a posteriori* (MAP) statistical test in which all the global hypotheses on the multiple observation window containing up to one speech-to-nonspeech or nonspeech-to-speech transitions are considered. Thus, the implicit hangover mechanism artificially added by the original method was not found in the revised method so its design can be further improved. With these and other innovations, the proposed method showed a higher speech/nonspeech discrimination accuracy over a wide range of SNR conditions when compared to the original MO-LRT voice activity detector (VAD). Experiments conducted on the AURORA databases and tasks showed that the revised method yields significant improvements in speech recognition performance over standardized VADs such as ITU T G.729 and ETSI AMR for discontinuous voice transmission and the ETSI AFE for distributed speech recognition (DSR), as well as over recently reported methods.**

***Index Terms*—Multiple hypothesis testing, robust speech recognition, voice activity detection (VAD).**

I. Introduction

EMERGING applications in the field of speech processing are demanding increasing levels of performance in noise adverse environments. Examples of such systems are the new voice services including discontinuous speech transmission [1], [2] or distributed speech recognition (DSR) over wireless and IP networks [3]. These systems often require a noise reduction scheme working in combination with a precise voice activity detector (VAD) in order to compensate for the harmful effect of the noise on the speech signal. 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. Sohn *et al.*. [4] proposed a robust VAD algorithm based on a statistical likelihood ratio test (LRT) involving a single observation vector. Later, Cho *et al.* [5] suggested an improvement based on a smoothed LRT. Most VADs in use today normally consider hangover algorithms based on empirical models to smooth the VAD decision. It has been shown recently that incorporating contextual information in a multiple observation LRT (MO-LRT) [6] reports benefits for speech/pause discrimination in high noise environments. This paper analyzes this method and shows a revised MO-LRT VAD that extends the number of hypotheses on the individual multiple observation window that are tested.

The rest of the manuscript is organized as follows. Section II shows a general description of a contextual multiple hypothesis testing method for voice activity detection in noisy environments. Topics such as the speech signal analysis, the definition of partial and global hypotheses that are considered in the test, and a revised maximum *a posteriori* (MAP) statistical test are presented and discussed. It is also shown that the statistical LRT proposed by Sohn *et al.* and the MO-LRT [6] are particular cases that can be derived from this more general method under several assumptions imposed on the *a priori* probability of the hypotheses. Section III analyzes the hangover mechanism artificially introduced by the MO-LRT VAD and the motivations for a revised and improved statistical test. Section IV shows the revised MO-LRT VAD in an elegant matrix form and provides examples under different SNR conditions showing the improved accuracy of the proposed method. Section V is devoted to the experimental framework including the discrimination analysis and the speech recognition performance evaluation. Finally, Section VI summarizes the conclusions of this paper.

II. Contextual Multiple Hypothesis Testing for Voice Activity Detection

It has been shown recently [6], [7] that incorporating contextual information to the decision rule yields significant improvements in speech/nonspeech discrimination in severe noise conditions. A general statistical framework is presented in this section which enables including such information in an optimum MAP test. This approach will be unveiled as a generalization of the statistical hypothesis testing method proposed by Sohn *et al.* [4] and the MO-LRT [6] previously presented.

Manuscript received October 16, 2006; revised May 31, 2007. This work was supported by the EU Sixth Framework Program under Contract IST-2002-507943 [Human Input that Works in Real Environments (HIWIRE)] and SR3-VoIP project (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. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mark Hasegawa-Johnson.

The authors are with the Department of Signal Theory, Networking, and Communications (GSTC), University of Granada, 18071 Granada, Spain.

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.2007.903937

A. Speech Signal Analysis

Let $s(n)$ be a speech signal observed in uncorrelated additive noise. The signal is decomposed into overlapped frames each of size N_w with a S_w-sample window shift and a J-point windowed DFT spectral representation is obtained on a frame by frame basis

$$S_l(k)=\frac{1}{||w||}\sum_{n=0}^{N_W-1} s(lS_w+n)w(n)e^{\frac{-j2\pi nk}{J}} \quad \forall k=0,\ldots,J-1 \tag{1}$$

where l denotes the frame index, w represents the window (typically, a Hamming window), and $||w||$ is its norm. Thus, $|S_l(k)|^2$ is an consistent estimation of the power spectral density (PSD) of the signal.

The algorithm considers several observations in the decision rule and requires the feature vectors $\mathbf{s}_l$ obtained through the analysis of each frame

$$\mathbf{s}_l=\left\{|S_l(0)|^2,|S_l(1)|^2,\ldots,|S_l(J-1)|^2\right\} \tag{2}$$

to be stored in a reindexed $(2N+1)$-observation buffer:

$$\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_{N+1},\ldots,\mathbf{x}_{2N+1}\}=\{\mathbf{s}_{l-N},\ldots,\mathbf{s}_l,\ldots,\mathbf{s}_{l+N}\}. \tag{3}$$

The content of the buffer is shifted to the left in each step of the algorithm, and the new feature vector obtained after the analysis of the current analysis window is inserted in the $(2N+1)$th position. The method now can formulate a binary decision about the presence or absence of speech in the frame stored in the central position (frame $N+1$) using the N preceding observations $\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$ and the N succeeding observations $\{\mathbf{x}_{N+1},\ldots,\mathbf{x}_{2N+1}\}$. Note that the algorithm exhibits an N-frame computational delay so that the decision over the $(N+1)$th frame of the signal is only available after the $(2N+1)$th frame has been analyzed.

B. Partial and Global Hypotheses

Our method evaluates the probability of the different individual hypotheses that can be formulated on each of the observations of the analysis buffer. The global hypothesis is defined as

$$\mathbf{h}=\{h_1,\ldots,h_{N+1},\ldots,h_{2N+1}\} \tag{4}$$

where each h_k value represents a partial binary hypothesis on the presence or absence of speech in the frame stored in the kth position of the buffer

$$h_k=0 : \text{speech absence in the } k\text{th position} : \mathbf{x}_k=\mathbf{n}$$
$$h_k=1 : \text{speech presence in } k\text{th position} : \mathbf{x}_k=\mathbf{s}+\mathbf{n}$$

where $\mathbf{s}$ and $\mathbf{n}$ denote the PSDs of the speech and noise processes, respectively.

Each possible combination of individual hypotheses defines a global hypothesis on the content of each of the frames in the buffer. Since there are $2N+1$ partial binary hypotheses, the number of possible global hypotheses is 2^{2N+1}. In order to simplify the representation, each global hypothesis is denoted by a $(2N+1)$-bit binary number so that the set of all the global hypotheses is defined as

$$\mathbf{H}=\left\{\mathbf{h}:\mathbf{h}=\sum_{k=1}^{2N+1} h_k\cdot 2^{2N+1-k};h_k\in\{0,1\}\right\}. \tag{5}$$

Finally, a partition of the set of all the possible global hypotheses is defined

$$\mathbf{H}_0=\{\mathbf{h}\in\mathbf{H}:h_{N+1}=0\} \tag{6}$$
$$\mathbf{H}_1=\{\mathbf{h}\in\mathbf{H}:h_{N+1}=1\} \tag{7}$$

depending on the hypothesis formulated on the central frame of the buffer. As an example, for $N=1$, the sets $\mathbf{H}$, $\mathbf{H}_0$, and $\mathbf{H}_1$ are given by

$$\mathbf{H}=\{0,1,2,3,4,5,6,7\} \tag{8}$$
$$\mathbf{H}_0=\{0,1,4,5\}\quad \mathbf{H}_1=\{2,3,6,7\} \tag{9}$$

or, equivalently, in binary representation as

$$\mathbf{H}=\{000,001,010,011,100,101,110,111\} \tag{10}$$
$$\mathbf{H}_0=\{000,001,100,101\} \tag{11}$$
$$\mathbf{H}_1=\{010,011,110,111\}. \tag{12}$$

C. Multiple Observation MAP Voice Activity Detection

Once all the possible hypotheses about the content of the buffer were defined, the VAD decision about presence or absence of speech can be formulated in terms of a binary hypothesis testing

G_0 : speech absence in the central frame;
G_1 : speech presence in the central frame;

that, based on the MAP optimum criterion, is expressed as

$$\Gamma\triangleq\frac{p(G_1|\mathbf{X})}{p(G_0|\mathbf{X})}=\frac{p(\mathbf{X},G_1)}{p(\mathbf{X},G_0)}\underset{G_0}{\overset{G_1}{\gtrless}}1. \tag{13}$$

Taking into account the definition of the sets $\mathbf{H}_0$ and $\mathbf{H}_1$, the joint probabilities $p(\mathbf{X},G_1)$ and $p(\mathbf{X},G_0)$ can be obtained by

$$p(\mathbf{X},G_0)=\sum_{\mathbf{h}\in\mathbf{H}_0}p(\mathbf{X},\mathbf{h})=\sum_{\mathbf{h}\in\mathbf{H}_0}p(\mathbf{h})p(\mathbf{X}|\mathbf{h}) \tag{14}$$
$$p(\mathbf{X},G_1)=\sum_{\mathbf{h}\in\mathbf{H}_1}p(\mathbf{X},\mathbf{h})=\sum_{\mathbf{h}\in\mathbf{H}_1}p(\mathbf{h})p(\mathbf{X}|\mathbf{h}) \tag{15}$$

where $p(\mathbf{h})$ is the *a priori* probability of the global hypothesis $\mathbf{h}$. If the observations are assumed to be statistically independent, the conditional probability of each of the global hypothesis $\mathbf{h}$ can be calculated by

$$p(\mathbf{X}|\mathbf{h})=\prod_{k=1}^{2N+1}p(\mathbf{x}_k|h_k). \tag{16}$$

As a conclusion, the optimum decision criterion is defined as

$$\Gamma = \frac{\sum_{\mathbf{h}\in\mathbf{H}_1} p(\mathbf{h}) \prod_{k=1}^{2N+1} p(\mathbf{x}_k|h_k)}{\sum_{\mathbf{h}\in\mathbf{H}_0} p(\mathbf{h}) \prod_{k=1}^{2N+1} p(\mathbf{x}_k|h_k)} \underset{G_0}{\overset{G_1}{\gtrless}} 1. \tag{17}$$

Finally, a model is needed to compute the probabilities of each of the observations in the buffer given the associated hypothesis in order to complete the statistical test. Assuming the discrete Fourier transform (DFT) coefficients to be asymptotically independent Gaussian variables [8], the probabilities can be obtained as

$$p(\mathbf{x}_k|h_k) = \begin{cases} \prod_{j=0}^{J-1} \frac{1}{\pi\lambda_N(j)} \exp\left\{-\frac{|x_{j,k}|^2}{\lambda_N(j)}\right\}, & h_k = 0 \\ \prod_{j=0}^{J-1} \frac{1}{\pi(\lambda_N(j)+\lambda_S(j))} & \\ \quad \cdot \exp\left\{-\frac{|x_{j,k}|^2}{(\lambda_S(j)+\lambda_N(j))}\right\}, & h_k = 1 \end{cases} \tag{18}$$

where $x_{j,k}$ denotes the jth bin of the DFT for the kth frame in the buffer $\mathbf{X}$, and $\lambda_S(j)$ and $\lambda_N(j)$ are the PSDs of the speech and noise processes, respectively.

D. A Priori Probabilities and Contextual Information

The only way to incorporate contextual information to the decision rule—given the statical independence assumption for the observations in the buffer—is through an adequate selection of the *a priori* probabilities $p(\mathbf{h})$ of the global hypothesis. In this section, two particular cases are shown which directly lead to the statistical hypothesis testing method proposed by Sohn *et al.* [4] and the MO-LRT [6] previously presented.

1) Single Observation LRT-Based VAD: Assuming that no *a priori* information about the process is known except the probabilities of occurrence of speech frames ρ and silence frames $(1-\rho)$, the reasonable values of the *a priori* probabilities of the hypotheses are

$$p(\mathbf{h}) = \prod_{k=1}^{2N+1} \left[\rho^{h_k}(1-\rho)^{1-h_k}\right]. \tag{19}$$

Thus, taking into account the symmetry of the sets $\mathbf{H}_0$ and $\mathbf{H}_1$, the decision rule is reduced to

$$\Gamma^A = \frac{\sum_{\mathbf{h}\in\mathbf{H}_1} \prod_{k=1}^{2N+1} \left[\rho^{h_k}(1-\rho)^{1-h_k}\right] p(\mathbf{x}_k|h_k)}{\sum_{\mathbf{h}\in\mathbf{H}_0} \prod_{k=1}^{2N+1} [\rho^{h_k}(1-\rho)^{1-h_k}] p(\mathbf{x}_k|h_k)} \tag{20}$$

$$= \left[\frac{\rho p(\mathbf{x}_{N+1}|1)}{(1-\rho)p(\mathbf{x}_{N+1}|0)}\right] \underset{G_0}{\overset{G_1}{\gtrless}} 1 \tag{21}$$

or, equivalently

$$\Lambda^A = \frac{p(\mathbf{x}_{N+1}|1)}{p(\mathbf{x}_{N+1}|0)} \underset{G_0}{\overset{G_1}{\gtrless}} \frac{(1-\rho)}{\rho} \tag{22}$$

$$\log\Lambda^A = \log\left(\frac{p(\mathbf{x}_{N+1}|1)}{p(\mathbf{x}_{N+1}|0)}\right) \underset{G_0}{\overset{G_1}{\gtrless}} \log\left(\frac{1-\rho}{\rho}\right). \tag{23}$$

This test is only based on the central observation in the buffer and discards any contextual information. By substituting (18) in the previous equation finally leads to

$$\log\Lambda^A = \sum_{j=0}^{J-1} \left(\frac{\gamma_{j,(N+1)}\xi_j}{1+\xi_j} - \log(1+\xi_j)\right) \underset{G_0}{\overset{G_1}{\gtrless}} \log\left(\frac{1-\rho}{\rho}\right) \tag{24}$$

$$\xi_j = \frac{\lambda_S(j)}{\lambda_N(j)} \quad \gamma_{j,(N+1)} = \frac{\left|\mathbf{x}_{j,(N+1)}\right|^2}{\lambda_N(j)} \tag{25}$$

where ξ_j is the *a priori* signal-to-noise ratio (SNR) in the jth bin, and $\gamma_{j,(N+1)}$ is the *a posteriori* SNR of the central frame in the buffer. In practice, an average criterion with a range independent of the number of bins J is defined as

$$\log\Lambda^{SO} = \frac{1}{J}\sum_{j=0}^{J-1} \left(\frac{\gamma_{j,(N+1)}\xi_j}{1+\xi_j} - \log(1+\xi_j)\right) \underset{G_0}{\overset{G_1}{\gtrless}} \eta. \tag{26}$$

Note that this test is essentially the statistical LRT proposed by Sohn *et al.* [4].

2) Multiple Observation LRT-Based VAD: Another option for the selection of the *a priori* probability of the hypotheses is related to the fact that speech and nonspeech frames do not appear in an isolated way but on consecutive speech segments. Thus, the set of probabilities can be defined

$$p(\mathbf{h}) = \begin{cases} (1-\rho), & \mathbf{h} = 0 \\ \rho, & \mathbf{h} = 2^{2N+1} - 1 \\ 0, & \text{otherwise} \end{cases} \tag{27}$$

where ρ is the probability of occurrence of speech frames. This assumption is correct if the the analysis window is shorter than the minimum length of speech and nonspeech periods except for the transitions. With this model, the statistical test is reduced to

$$\Gamma^B = \frac{\rho\prod_{k=1}^{2N+1} p(\mathbf{x}_k|1)}{(1-\rho)\prod_{k=1}^{2N+1} p(\mathbf{x}_k|0)} \underset{G_0}{\overset{G_1}{\gtrless}} 1 \tag{28}$$

or, equivalently

$$\Lambda^B = \frac{\prod_{k=1}^{2N+1} p(\mathbf{x}_k|1)}{\prod_{k=1}^{2N+1} p(\mathbf{x}_k|0)} \underset{G_0}{\overset{G_1}{\gtrless}} \frac{(1-\rho)}{\rho} \tag{29}$$

$$\log\Lambda^B = \log\left(\frac{\prod_{k=1}^{2N+1} p(\mathbf{x}_k|1)}{\prod_{k=1}^{2N+1} p(\mathbf{x}_k|0)}\right) \underset{G_0}{\overset{G_1}{\gtrless}} \log\left(\frac{1-\rho}{\rho}\right). \tag{30}$$

By substituting (18) in the previous equation, the decision rule is finally defined as

$$\log\Lambda^B = \sum_{k=1}^{2N+1}\sum_{j=0}^{J-1} \left(\frac{\gamma_{j,k}\xi_j}{1+\xi_j} - \log(1+\xi_j)\right) \underset{G_0}{\overset{G_1}{\gtrless}} \log\left(\frac{1-\rho}{\rho}\right) \tag{31}$$

$$\xi_j = \frac{\lambda_S(j)}{\lambda_N(j)} \quad \gamma_{j,k} = \frac{|\mathbf{x}_{j,k}|^2}{\lambda_N(j)} \tag{32}$$

where ξ_j is the *a priori* SNR for the jth band, and $\gamma_{j,k}$ denotes the *a posteriori* SNR for the jth band at the kth frame of the

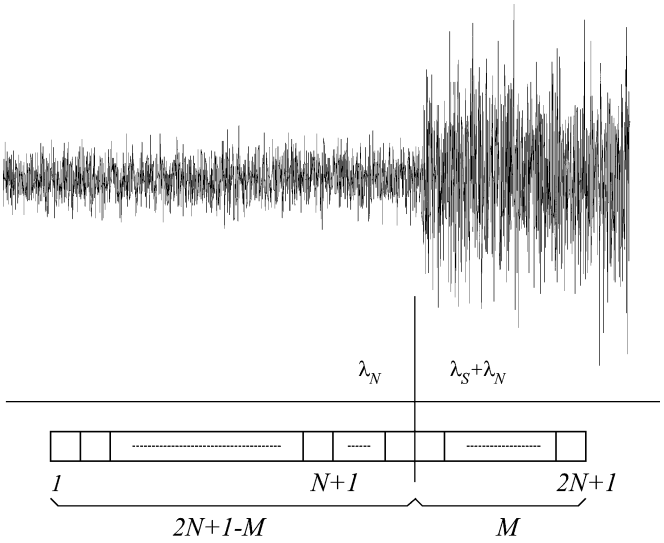

Fig. 1. Detection of a step change in the variance of the signal using a multiple observation window.

buffer. As in the previous example, a scaled decision rule independent of N and J

$$\log \Lambda^{MO} = \frac{1}{J(2N+1)} \sum_{k=1}^{2N+1} \sum_{j=0}^{J-1} \left(\frac{\gamma_{j,k}\xi_j}{1+\xi_j} - \log(1+\xi_j) \right) \underset{G_0}{\overset{G_1}{\gtrless}} \eta \tag{33}$$

is preferred. This statistical test is essentially the MO-LRT proposed in [6] and can be understood as an average of the decision criterion Λ^{SO} over the frames present in the buffer.

III. Hangover Analysis of the MO-LRT Method

The MO-LRT yields significant improvements in speech detection accuracy as well as speech recognition performance in severe noise conditions [6] when compared to the noncontextual LRT proposed by Sohn *et al.* [4]. However, a degradation of the ability of the method to detect nonspeech frames is shown under high SNR conditions due to the implicit hangover present in the MO-LRT VAD. Although a hangover scheme is adequate for improving voice activity detection performance and recovering low energy beginnings and endings of speech segments, the desired behavior is to increase the hangover length under low SNRs while reducing it for high SNRs. This section analyzes the hangover of the MO-LRT VAD and shows an opposite behavior to the desired one which results prejudicial under high SNR conditions.

For the analysis of the hangover duration, the speech-to-nonspeech transition in Fig. 1 is studied. This corresponds to a situation where the first $2N+1-M$ observations in the buffer are nonspeech frames while the remaining M observations are speech frames. In such a situation, the expected value of the MO-LRT decision rule can be calculated as follows

$$E[\log \Lambda^{MO}] = \frac{1}{J(2N+1)} \sum_{k=1}^{2N+1} \sum_{j=0}^{J-1} \left(\frac{E[\gamma_{j,k}]\xi_j}{1+\xi_j} - \log(1+\xi_j) \right) \tag{34}$$

where the expected value of the *a posteriori* SNR of each frame depends on the presence or absence of speech

$$E[\gamma_{j,k}] = \begin{cases} (1+\xi_j); & 2N+2-M \le k \le 2N+1 \\ 1; & 1 \le k \le 2N+1-M \end{cases}. \tag{35}$$

By substituting the expected value of the SNR in (34), the expected value of the decision rule is reduced to

$$E[\log \Lambda^{MO}] = \frac{1}{J(2N+1)} \left[M \sum_{j=0}^{J-1} \left(\frac{\xi_j^2}{1+\xi_j} \right) + (2N+1) \sum_{j=0}^{J-1} \left(\frac{\xi_j}{1+\xi_j} - \log(1+\xi_j) \right) \right]. \tag{36}$$

Then, by solving $E[\log \Lambda^{MO}] = \eta$, the number of observations M that originates a VAD transition is given by

$$M = (2N+1) \left[\frac{\eta - \frac{1}{J} \sum_{j=0}^{J-1} \left(\frac{\xi_j}{1+\xi_j} - \log(1+\xi_j) \right)}{\frac{1}{J} \sum_{j=0}^{J-1} \left(\frac{\xi_j^2}{1+\xi_j} \right)} \right]. \tag{37}$$

As a conclusion, this equation shows the expected value of M for which $E[\log \Lambda^{MO}]$ exceeds the decision threshold η. Thus, the hangover length can be calculated as $\Delta = (N+1) - M$. An analysis of (37) shows that M decreases with the increasing ξ_j, so that the hangover Δ increases with the *a priori* SNR. This behavior reduces the utility of MO-LRT under a wide range of SNR conditions and constitutes the motivation for a revised method.

In order to illustrate the dependence of Δ with the SNR, we assume that both the signal and the noise are white Gaussian processes. Thus, the *a priori* SNR is constant for all the frequency bands ($\xi_j = \xi$ for $0 \le j \le (J-1)$), and (37) is reduced to

$$M = (2N+1) \left(\frac{1+\xi}{\xi^2} \right) \left[\eta + \log(1+\xi) - \frac{\xi}{1+\xi} \right]. \tag{38}$$

Fig. 2 shows the increasing hangover length Δ with the *a priori* SNR. This undesired effect is what prejudices the accuracy of the MO-LRT VAD under high SNR conditions. This effect can be also shown in Fig. 3, where the evolution of the decision rules Λ^{SO} and Λ^{MO} is compared for a white signal observed in noncorrelated additive white noise. Note that Λ^{MO} exhibits a smoother behavior since it is essentially an averaged version of Λ^{SO} over $2N+1$ consecutive frames. On the other hand, an advanced detection of the word beginnings and a delayed detection of the word endings is observed. The hangover is symmetric and its length is consistent with the expected value of six frames shown in Fig. 2.

IV. Revised MO-LRT

The undesired behavior of the hangover is mainly due to not considering more than just the two global hypotheses consisting of all the frames belonging to the same class. Note that, the most

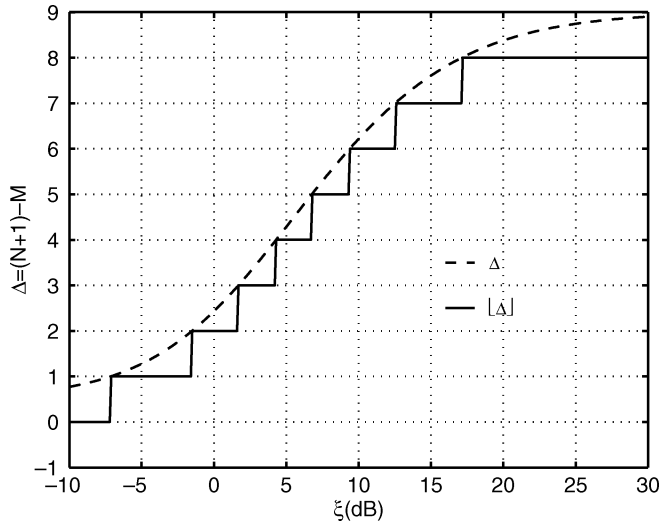

Fig. 2. Hangover length as a function of the *a priori* SNR (ξ) for $N = 8$ and $\eta = 0$.

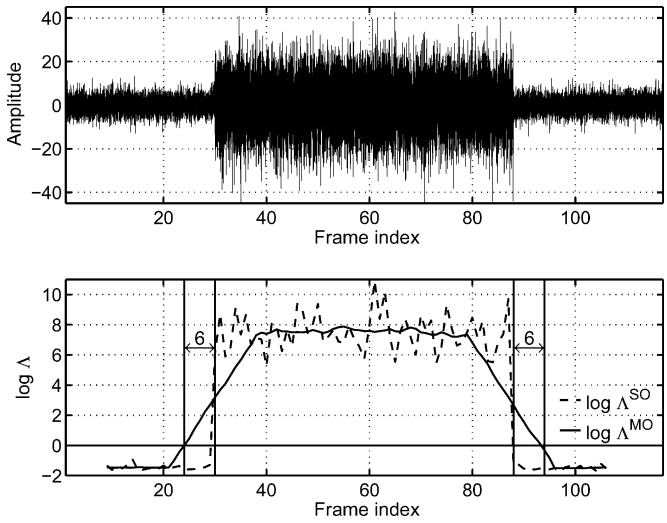

Fig. 3. Evolution of Λ^{SO} and Λ^{MO} for a white signal in additive uncorrelated white noise with $\xi = 10$ dB and $N = 8$.

probable hypotheses may differ from these two global hypothesis especially when detecting a speech-to-nonspeech transition or a nonspeech-to-speech transition. For such a situation, the most probable hypothesis will typically contain both speech and nonspeech partial hypotheses over the individual observations in the buffer. This section shows a revised version of the MO-LRT VAD that extends the number of global hypotheses that are evaluated in order to improve its performance on the transitions.

For the selection of the *a priori* probabilities of the global hypotheses, we assume that the length of the buffer is reduced enough not to occur more than just one speech-to-nonspeech or nonspeech-to-speech transition within the buffer. Thus, it can be shown that the number of hypotheses to be considered is reduced to $2(2N + 1)$ instead of 2^{2N+1}. As an example, for $N = 1$, the hypotheses considered are

$$\{000, 001, 011, 100, 110, 111\}. \tag{39}$$

More generally, the subset $\mathbf{H}^*$ of $\mathbf{H}$ containing no null probability hypotheses can be described in matrix form so that each row identifies a global hypothesis. These hypotheses can be expressed in matrix form resulting the Hankel matrix

$$\mathbf{K} = \begin{bmatrix} 0 & 0 & \cdots & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & \cdots & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \cdots & 1 & \cdots & 1 & 1 \\ 1 & 1 & \cdots & 1 & \cdots & 1 & 1 \\ 1 & 1 & \cdots & 1 & \cdots & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & \cdots & 0 & 0 \end{bmatrix}. \tag{40}$$

In the same way, the hypotheses with no null probability in the subsets $\mathbf{H}_0$ and $\mathbf{H}_1$ can be represented by matrices $\mathbf{K}_0$ and $\mathbf{K}_1$, that are built by selecting the rows of $\mathbf{K}$ whose central element is 0 or 1, respectively. As an example, for $N = 1$, $\mathbf{K}$, $\mathbf{K}_0$, and $\mathbf{K}_1$ are

$$\mathbf{K} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

$$\mathbf{K}_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \quad \mathbf{K}_1 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}. \tag{41}$$

The subsets containing hypotheses with nonzero *a priori* probability will be denoted as $\mathbf{H}^*$, $\mathbf{H}_0^*$, and $\mathbf{H}_1^*$ in the rest of the paper. These sets consists of the hypotheses with up to one speech-to-nonspeech or nonspeech-to-speech transition and are defined by matrices $\mathbf{K}$, $\mathbf{K}_0$ and $\mathbf{K}_1$, respectively.

A. A Priori Probabilities

The *a priori* probabilities of the hypotheses can be calculated by taking into account that there exist $4N$ hypotheses that correspond to nonspeech-to-speech and speech-to-nonspeech transitions in the set $\mathbf{H}^*$. Among them, $2N$ assume a 0 hypothesis for the central frame while the remaining $2N$ assume a 1 hypothesis. For the computation of the *a priori* probabilities of the hypotheses $\mathbf{h}$, a T-frame signal containing L speech frames $(L < T)$ grouped in K speech segments (speech blocks larger than the length of the analysis window) is considered. For large T, the probability of the hypotheses $\mathbf{h}$ with a single speech-to-nonspeech transition or nonspeech-to-speech transition is easily obtained as the number K of speech segments in the data sample divided by the total number of frames T, that is, the probability $\phi = K/T$ of speech blocks. For the hypothesis $\mathbf{h} = 2^{2N+1} - 1$, the number of cases is equal to the number of frames L minus the number of all the situations corresponding to transitional hypotheses assuming a 1 hypothesis for the central frame in the analysis window, that is, $p(\mathbf{h}) = (L - 2NK)/T = \rho - 2N\phi$, where ρ denotes the *a*

TABLE I
PROBABILITY OF SPEECH FRAMES ρ AND SPEECH SEGMENTS ϕ FOR THE AURORA3 SPANISH DATABASE ("B": SEQUENCE OF 10 ISOLATED DIGITS, "C": SEQUENCE OF FOUR OR MORE CONNECTED DIGITS, "I": ONE DIGIT PER UTTERANCE)

	'b'	'i'	'c'
ρ	0,394	0,296	0,619
ϕ	0,010	0,008	0,013

priori probability of speech frames. In a similar way, for $\mathbf{h}=0$, $p(\mathbf{h})=[(T-L)-2NK]/T=1-\rho-2N\phi$. As a conclusion

$$p(\mathbf{h})=\begin{cases}\phi, & \mathbf{h}\in\mathbf{H}^*\wedge\mathbf{h}\neq 0\wedge\mathbf{h}\neq 2^{2N+1}-1\\ \rho-2N\phi, & \mathbf{h}=2^{2N+1}-1\\ (1-\rho)-2N\phi, & \mathbf{h}=0\\ 0, & \mathbf{h}\notin\mathbf{H}^*.\end{cases} \tag{42}$$

Table I shows the probabilities ρ and ϕ calculated for the AURORA3 Spanish SpeechDat-Car database for the different kinds of recordings.

B. Revised Statistical Test: RMO-LRT

Once the values of the *a priori* probabilities of the hypotheses have been obtained, a revised statistical test can be defined. As a starting point, the general statistical test defined in (17) but particularized for the no null probability hypothesis in $\mathbf{H}_0^*$ and $\mathbf{H}_1^*$ is considered

$$\Gamma=\frac{\sum\limits_{\mathbf{h}\in\mathbf{H}_1^*}p(\mathbf{h})\prod\limits_{k=1}^{2N+1}p(\mathbf{x}_k|h_k)}{\sum\limits_{\mathbf{h}\in\mathbf{H}_0^*}p(\mathbf{h})\prod\limits_{k=1}^{2N+1}p(\mathbf{x}_k|h_k)}\underset{G_0}{\overset{G_1}{\gtrless}}1. \tag{43}$$

Although (16) and (18) can be used to evaluate the statistical test defined by (43), this test requires to evaluate the $2(2N+1)$ probabilities of the different global hypotheses (i.e. $2N+1$ of the set $\mathbf{H}_1^*$ and $2N+1$ of $\mathbf{H}_0^*$). If (43) is approximated by taking the maximum value of the hypotheses, a revised statistical test can be defined as

$$\Gamma^*=\frac{\max\limits_{\mathbf{h}\in\mathbf{H}_1^*}\left\{p(\mathbf{h})\prod\limits_{k=1}^{2N+1}p(\mathbf{x}_k|h_k)\right\}}{\max\limits_{\mathbf{h}\in\mathbf{H}_0^*}\left\{p(\mathbf{h})\prod\limits_{k=1}^{2N+1}p(\mathbf{x}_k|h_k)\right\}}\underset{G_0}{\overset{G_1}{\gtrless}}1 \tag{44}$$

and taking logarithms leads to

$$\log\Gamma^*=\max_{\mathbf{h}\in\mathbf{H}_1^*}\left\{\log p(\mathbf{h})+\sum_{k=1}^{2N+1}\log p(\mathbf{x}_k|h_k)\right\}-\max_{\mathbf{h}\in\mathbf{H}_0^*}\left\{\log p(\mathbf{h})+\sum_{k=1}^{2N+1}\log p(\mathbf{x}_k|h_k)\right\}\underset{G_0}{\overset{G_1}{\gtrless}}0. \tag{45}$$

The properties of the algorithm are easier to analyze if the decision criterion is defined in this way while this approximation yields similar speech/nonspeech classification performance.

This test can be expressed in matrix form as follows. First, we define the vectors consisting of the probabilities of the observations under the "0" and "1" hypothesis

$$\mathbf{B}_0=[\log p(\mathbf{x}_1|0)\log p(\mathbf{x}_2|0)\ldots\log p(\mathbf{x}_{2N+1}|0)]^{\mathrm{T}} \tag{46}$$

$$\mathbf{B}_1=[\log p(\mathbf{x}_1|1)\log p(\mathbf{x}_2|1)\ldots\log p(\mathbf{x}_{2N+1}|1)]^{\mathrm{T}} \tag{47}$$

and the column vectors $\mathbf{P}_1$ and $\mathbf{P}_0$ consisting of the logarithmic *a priori* probabilities of the hypotheses in $\mathbf{K}_1$ and $\mathbf{K}_0$, respectively, are calculated using (42). Then, the column vectors $\mathbf{L}_1$ and $\mathbf{L}_0$ defined by

$$\mathbf{L}_1=\mathbf{K}_1\mathbf{B}_1+(\mathbf{I}-\mathbf{K}_1)\mathbf{B}_0 \tag{48}$$

$$\mathbf{L}_0=\mathbf{K}_0\mathbf{B}_1+(\mathbf{I}-\mathbf{K}_0)\mathbf{B}_0 \tag{49}$$

are computed and the test is reduced to

$$\log\Gamma^*=\max(\mathbf{L}_1+\mathbf{P}_1)-\max(\mathbf{L}_0+\mathbf{P}_0)\underset{G_0}{\overset{G_1}{\gtrless}}0. \tag{50}$$

As an example, for $N=1$, the different matrices and vectors are

$$\mathbf{B}_0=[\log p(\mathbf{x}_1|0)\log p(\mathbf{x}_2|0)\log p(\mathbf{x}_3|0)]^{\mathrm{T}} \tag{51}$$

$$\mathbf{B}_1=[\log p(\mathbf{x}_1|1)\log p(\mathbf{x}_2|1)\log p(\mathbf{x}_3|1)]^{\mathrm{T}} \tag{52}$$

$$\mathbf{K}_0=\begin{bmatrix}0&0&0\\0&0&1\\1&0&0\end{bmatrix}\quad\mathbf{K}_1=\begin{bmatrix}1&1&1\\1&1&0\\0&1&1\end{bmatrix} \tag{53}$$

$$\mathbf{P}_0=\begin{bmatrix}\log((1-\rho)-2N\phi)\\ \log\phi\\ \log\phi\end{bmatrix} \tag{54}$$

$$\mathbf{P}_1=\begin{bmatrix}\log(\rho-2N\phi)\\ \log\phi\\ \log\phi\end{bmatrix} \tag{55}$$

$$\mathbf{L}_0=\begin{bmatrix}\log p(\mathbf{x}_1|0)+\log p(\mathbf{x}_2|0)+\log p(\mathbf{x}_3|0)\\ \log p(\mathbf{x}_1|0)+\log p(\mathbf{x}_2|0)+\log p(\mathbf{x}_3|1)\\ \log p(\mathbf{x}_1|1)+\log p(\mathbf{x}_2|0)+\log p(\mathbf{x}_3|0)\end{bmatrix} \tag{56}$$

$$\mathbf{L}_1=\begin{bmatrix}\log p(\mathbf{x}_1|1)+\log p(\mathbf{x}_2|1)+\log p(\mathbf{x}_3|1)\\ \log p(\mathbf{x}_1|1)+\log p(\mathbf{x}_2|1)+\log p(\mathbf{x}_3|0)\\ \log p(\mathbf{x}_1|0)+\log p(\mathbf{x}_2|1)+\log p(\mathbf{x}_3|1)\end{bmatrix}. \tag{57}$$

The bias terms that appear in (50) due to $\mathbf{P}_0$ and $\mathbf{P}_1$ have a small influence on the decision rule and can be omitted. The maximum difference between the components of $\mathbf{P}_1$ is between the first and second row being its value $\log((\rho/\phi)-2N)$. This difference is 3.5 for typical values $\rho=0.5$, $\phi=0.01$, and $N=8$. On the other hand, for a white signal, the minimum expected value of the difference between the corresponding elements of $\mathbf{L}_1$ and $\mathbf{L}_0$ is $J((\xi/(1+\xi))-log(1+\xi))$ with absolute values of 49.4, 18.47, and 5.92 for $\xi=0$ dB, -3 and -6 dB, respectively. Therefore, in most practical situations, the selection of the maximum value of $(\mathbf{L}_1+\mathbf{P}_1)$ is fully conditioned by the values of $\mathbf{L}_1$. The same situation occurs for the selection of the maximum of $(\mathbf{L}_0+\mathbf{P}_0)$. As a conclusion, these bias terms do not have influence in the selection of the maximum in the statistical test for SNRs above -6 dB so that a simplified decision rule can be defined

$$\Lambda^{\mathrm{RMO}}=\frac{1}{J(N+1)}(\max\mathbf{L}_1-\max\mathbf{L}_0)\underset{G_0}{\overset{G_1}{\gtrless}}\eta. \tag{58}$$

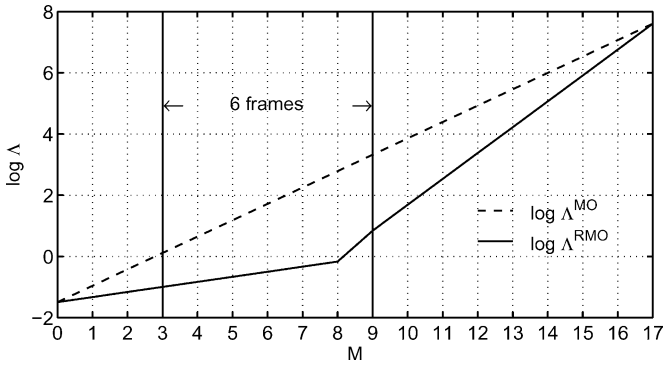

Fig. 4. Evolution of the expected values of Λ^{MO} and Λ^{RMO} as a function of the number M of voice frames in the buffer. For values of $N = 8$ and $\xi = 10$ dB.

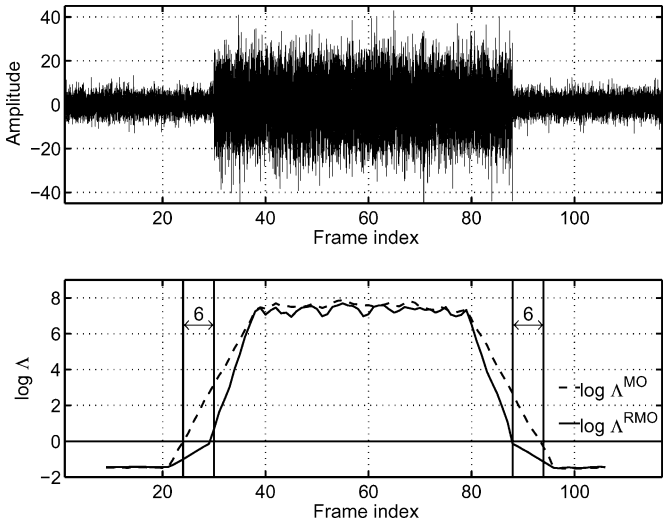

Fig. 5. Evolution of Λ^{MO} and Λ^{RMO} for a white signal in additive uncorrelated white noise with $\xi = 10$ dB and $N = 8$.

The selection of the scale factor $J(N+1)$ makes the rule to have a range comparable to the statistical tests Λ^{SO} and Λ^{MO}. The decision threshold η is used to tune the operating point of the VAD. In order to enable the VAD to work in optimum conditions for a wide range of SNRs, the threshold is made adaptive and linear-dependent on the measured noise energy as in previous works [6], [7].

C. Hangover Analysis of the Revised RMO-LRT

This section analyzes the decision rule defined above and shows that the inconvenient hangover behavior of the MO-LRT does not appear in the revised method. As a starting point, the expected value of the decision function defined in (58) is derived as a function of the number of speech frames present in the buffer. If a similar situation to that shown in Fig. 1 is assumed, the expected value of the decision rule, for a white signal in white noise with SNR ξ, is given by (see the Appendix)

$$E[\Lambda^{\mathrm{RMO}}]=\begin{cases}\frac{N+1-M}{N+1}\\ \quad\left(\frac{\xi}{1+\xi}-\log(1+\xi)\right), & 0\le M<N+1\\ \frac{M-N}{N+1}\left(\xi-\log(1+\xi)\right), & N+1\le M\le 2N+1.\end{cases} \tag{59}$$

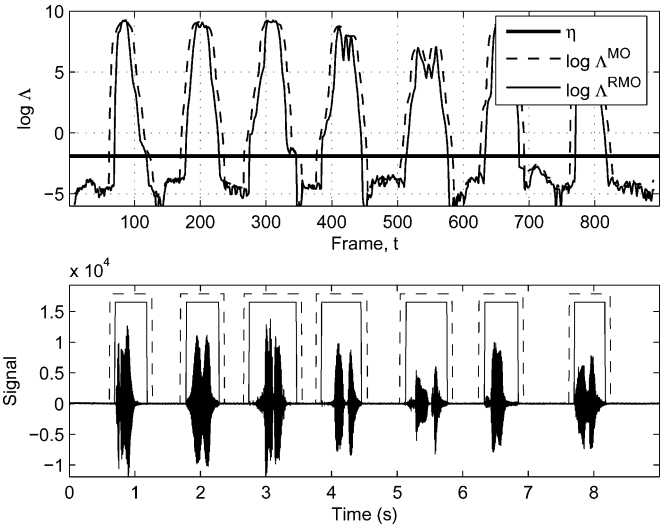

Fig. 6. Comparison between the original MO-LRT and the revised MO-LRT for VAD in clean conditions.

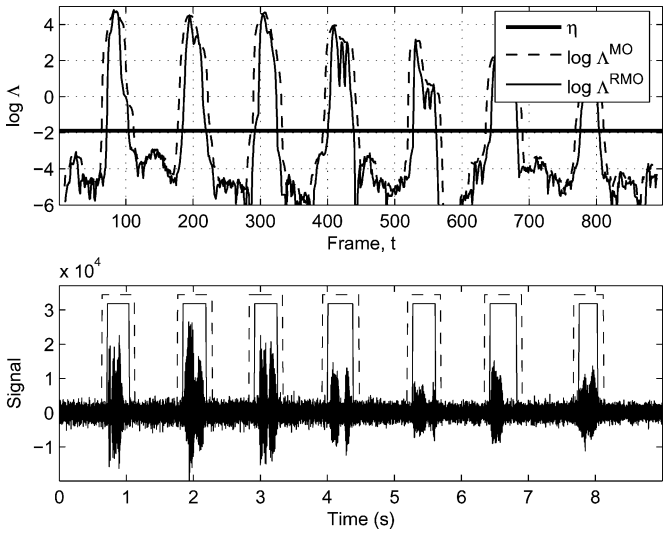

Fig. 7. Comparison between the original MO-LRT and the revised MO-LRT for VAD in high-noise car environment.

Fig. 4 shows the expected value of the decision functions Λ^{MO} and Λ^{RMO} as a function of the number M of speech observations in the buffer for a nonspeech-to-speech transition. Note that, $(\xi/(1+\xi)) - \log(1+\xi) < 0$ and $\xi - \log(1+\xi) > 0$ for $\xi > 0$, and that the threshold $\eta = 0$ is achieved when $M \geq N+1$ independently of the value of ξ. As a conclusion, the revised statistical test does not have an implicit hangover. This fact can be also corroborated in Fig. 5 where both Λ^{MO} and Λ^{RMO} decision rules are shown for the same conditions in Fig. 3. The hangover is eliminated since now the more probable hypothesis in the sets $\mathbf{H}_1^*$ and $\mathbf{H}_0^*$ are evaluated leading to an adaptive averaging during nonspeech-to-speech and speech-to-nonspeech transitions.

D. Examples

Fig. 6 shows the operation of the original MO-LRT VAD and the revised one over an utterance of the Spanish SpeechDat-Car database [9] in clean conditions (25-dB SNR). Note that the

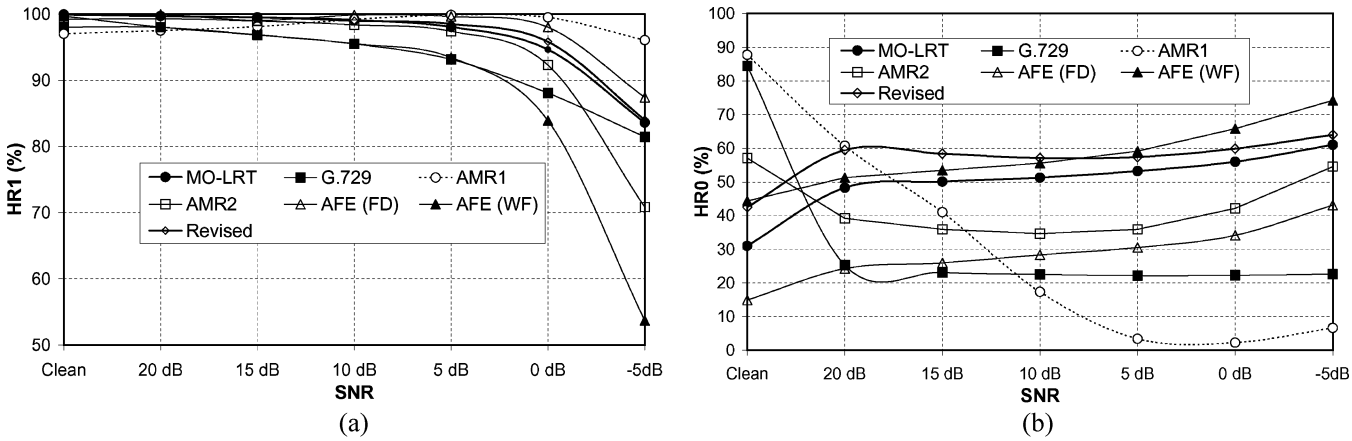

Fig. 8. Speech/nonspeech discrimination analysis as a function of the SNR. Results are averaged for all the noises considered in the AURORA 2 database. (a) Speech hit-rate. (b) Nonspeech hit-rate.

new algorithm removes the saving period at the word beginnings and endings being more accurate in such a low noise condition. It is interesting to point out that the hangover of the original MO-LRT was a result of extending the decision over a neighborhood of the current frame. However, the new statistical test exhibits the same smoothing process and reduced variance of the decision variable with the benefit of being suitable for a more effective hangover mechanism development. Under the noisiest conditions (5-dB SNR), the new algorithm has a similar behavior to the previous VAD as shown in Fig. 7.

V. Experimental Results

First, the proposed VAD was evaluated in terms of the ability to discriminate between speech and nonspeech periods at different noise scenarios and SNR levels. All the methods including SO, MO, as well as the proposed RMO were evaluated under the same conditions, that is, the same noise reduction method based on the Ephraim and Malah estimator [8] was used for estimating the *a priori* SNR and an adaptive threshold update similar to that previously used in [6] and [7] enables the effective tuning of the operating point for the wide range of SNR conditions.

A. Speech/Nonspeech Discrimination in Noise

The first experiments focus on the comparison of the revised method to the original MO-LRT VAD proposed in [6]. The AURORA-2 database [10] was considered in the analysis where the clean TIdigits database, consisting of sequences of up to seven connected digits spoken by American English talkers, is used as source speech, and a selection of eight different real-world noises are artificially added at SNRs from 20 dB 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). The clean TIdigits database was manually labeled for reference and detection performance was assessed as a function of the SNR in terms of the nonspeech hit-rate (HR0) and the speech hit-rate (HR1) which are defined as the fraction of all actual nonspeech or speech frames that are correctly detected as nonspeech or speech frames, respectively. Fig. 8 compares the performance of the revised VAD to MO-LRT [6], G.729, ETSI AMR and Advanced Front-End (AFE) standardized VADs for clean conditions and SNR levels ranging from 20 to −5 dB. These results are averaged hit-rates for the entire set of noises. Note that, results for the two VADs defined in the AFE standard for DSR [3] for noise spectrum estimation in Wiener filtering (WF) and nonspeech frame-dropping (FD) are also provided. It is clearly shown that, while the revised method yields similar speech detection accuracy when compared to MO-LRT [6], it exhibits an improved accuracy in detecting nonspeech periods. The improvements are especially important for SNRs ranging from clean conditions to about 5 dB due to the implicit hangover removal. Finally, Table II summarizes the average hit-rates for all the noises and SNR conditions of the previously analyzed methods, and other VADs recently reported. It is shown that the revised MO-LRT method yields a significant improvement in HR0, and simular results in HR1 when compared to the original

TABLE II
Average Speech/Nonspeech Hit Rates for SNRs From Clean Conditions to −5 dB. Comparison of the Proposed VAD to (a) Standardized and (b) Recently Reported VADs

(a)

	G729	AMR1	AMR2	AFE WF	AFE FD
HR0	31.77	31.31	42.77	57.68	28.74
HR1	93.00	98.18	93.76	88.72	97.70

(b)

	Woo	Li	Sohn	Marzinzik	MO-LRT	Proposed
HR0	55.40	57.03	43.66	52.69	50.12	56.95
HR1	88.41	83.65	94.46	93.04	96.36	96.62

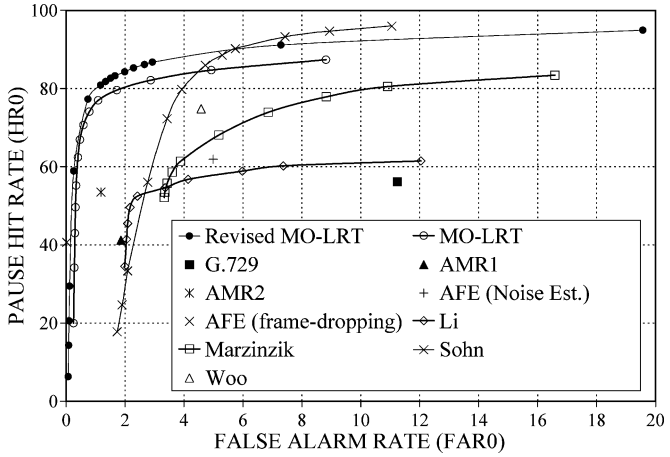

Fig. 9. ROC curves in quiet noise conditions (stopped car and engine running) and close-talking microphone.

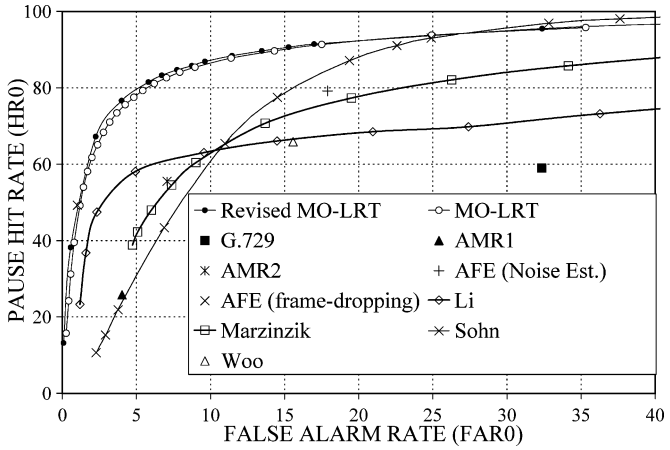

Fig. 10. ROC curves in high-noise conditions (high speed over a good road) and distant-talking microphone.

MO-LRT VAD. Moreover, the proposed VAD scheme achieves the best compromise among the different VADs tested. It yields good results in detecting nonspeech periods and exhibits a very slow performance degradation at unfavorable noise conditions in speech detection.

B. Receiver Operating Characteristics Curves

The receiving operating characteristic (ROC) curves have shown to be very effective for the evaluation of voice activity detectors [7], [11]. These plots, which show the tradeoff between the error probabilities of speech and nonspeech detection as the threshold η varies, completely describe the VAD error rate. In this analysis, the Spanish SpeechDat-Car (SDC)[9] database was used. This database consists of recordings from distant and close-talking microphones in car environments at different driving conditions. For the computation of the speech and nonspeech hit-rates, a semiautomatic "speech/nonspeech" labeling process was previously conducted on the close-talking microphone. Figs. 9 and 10 show the nonspeech hit rate (HR0) versus the false alarm rate [11] (FAR0 $=$ 1 $-$ HR1, where HR1 denotes the speech hit rate) for recordings from the distant microphone and under quiet and high noise conditions, respectively. The revised method yields better results than the previous method. These improvements are obtained by the robustness of the decision rule and by removing the implicit hangover found in the previous method and developing a more suitable design. The proposed algorithm also outperforms a number of standardized VAD methods including the ITU-T G.729 [1], ETSI AMR (opts. 1 and 2) [2] and the ETSI AFE [3] for DSR, as well as other recently published VAD methods [4], [11]–[13]. The best results are obtained for $N = 8$ while increasing the number of observations over this value reports no additional improvements. In particular, the proposed VAD outperforms Sohn's VAD [4], that assumes a single observation in the decision rule and a HMM-based hangover.

It is worthwhile mentioning that the experiments described above yield a first measure of the performance of the VAD. Other measures of VAD performance that have been reported are the clipping errors. These measures provide valuable information about the performance of the VAD and can be used for optimizing its operation. Our analysis does not distinguish between the frames that are being classified 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 evaluated indirectly 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.

C. Assessment of the VAD on an Automatic Speech Recognition (ASR) System

Although the ROC analysis presented in the preceding section is effective for the evaluation of a given speech/nonspeech discrimination algorithm, 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 [12] since nonefficient speech/nonspeech discrimination is an important performance degradation source for speech recognition systems working in noisy environments [14]. There are two clear motivations for that: 1) 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 2) nonspeech frame-dropping, a frequently used technique in speech recognition to reduce the number of insertion errors caused by the acoustic noise, is based on the VAD decision and speech misclassification errors lead to loss of speech and cause unrecoverable deletion errors.

1) Recognition System Setup: The reference framework (Base) used in the experiments is the DSR front-end [15] proposed by the ETSI STQ working group for the evaluation of noise robust DSR feature extraction algorithms. The recognition system is based on the Hidden Markov Model Toolkit (HTK) software package [16]. The task consists on recognizing connected digits which are modeled as whole word hidden Markov models (HMMs) with the following parameters: 16 states per word, simple left-to-right models, mixture of three Gaussians per state and diagonal covariance matrix, while speech pause models consist of three states with a mixture

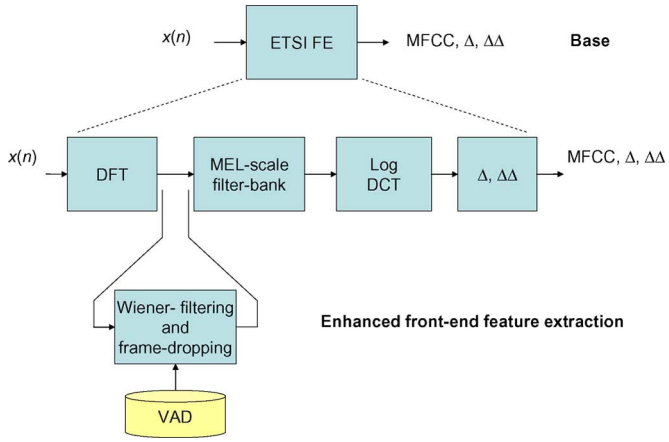

Fig. 11. Speech recognition experiments. Front-end feature extraction.

of six Gaussians per state. The 39-parameter feature vector consists of 12 cepstral coefficients (without the zero-order cepstral coefficient), the logarithmic frame energy plus the corresponding Δ and $\Delta\Delta$ coefficients.

Two training modes are defined for the experiments conducted on the AURORA-2 database: 2) training on clean data only (clean training), and 2) 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 consist of recordings from the close-talking and distant microphones. In the WM condition, both close-talking and hands-free microphones are used for training and testing. In the 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 material from all driving conditions, while testing is done using hands-free microphone material taken for low-noise and high-noise driving conditions. 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):

$$\text{WAcc}(\%) = \frac{N - D - S - I}{N} \times 100\% \qquad (60)$$

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

2) Comparative Results: The influence of the VAD decision on the performance of different feature extraction schemes was studied. The first approach (shown in Fig. 11) incorporates Wiener filtering to the Base system [15] as the noise suppression method. The second feature extraction algorithm that was evaluated uses Wiener filtering and nonspeech frame dropping. The noise reduction algorithm was implemented as described at the first stage of the Wiener filtering noise reduction system found in the advanced front-end AFE DSR standard [3]. The same feature extraction scheme was used for training and testing and no other mismatch reduction techniques already present in the AFE standard (waveform processing or blind equalization) were considered since they are not affected by the VAD decision and can mask the impact of the VAD on the overall system performance.

Table III shows the AURORA-2 recognition results as a function of the SNR for speech recognition experiments based on the G.729, AMR, AFE, MO-LRT, and the proposed VAD for clean and multicondition train/test modes. The working points of the VADs for the AURORA 2 database are shown in Fig. 8. The results shown in the table are averaged WAcc values over the three test sets of the AURORA-2 recognition experiments. Notice that, particularly, for the recognition experiments based on the AFE VADs, we have used the same configuration of the standard [3] with different VADs for WF and FD. Only exact speech periods are kept in the FD stage and, consequently, all the frames classified by the VAD as nonspeech are discarded. The revised VAD outperforms the original MO-LRT method in both clean and multicondition training. The results show that the proposed VAD also outperforms the standard G.729, AMR1, AMR2, and AFE VADs when used for WF and also when the VAD is used for removing nonspeech frames. Note that the VAD decision is used in the WF stage for estimating the noise spectrum during nonspeech periods, and a good estimation of the SNR is critical for an efficient application of the noise reduction algorithm. In this way, the energy-based WF AFE VAD suffers fast performance degradation in speech detection as shown in Fig. 8, thus leading to numerous recognition errors and the corresponding increase of the word error rate. On the other hand, FD is strongly influenced by the performance of the VAD, and an efficient VAD for robust speech recognition needs a compromise between speech and nonspeech detection accuracy. When the VAD suffers a rapid performance degradation under severe noise conditions, it loses too many speech frames and leads to numerous deletion errors; if the VAD does not correctly identify nonspeech periods, it causes numerous insertion errors and the corresponding FD performance degradation. The best recognition performance is obtained when the revised MO-LRT VAD is used for WF and FD. Note that FD yields better results for the speech recognition system trained on clean speech. This is motivated by the fact that models trained using clean speech do not adequately model noise processes and normally cause insertion errors during nonspeech periods. Thus, efficiently removing speech pauses will lead to a significant reduction of this error source. On the other hand, noise is well modeled when models are trained using noisy speech, and the speech recognition system tends itself to reduce the number of insertion errors in multicondition training.

Table IV compares the word accuracies averaged for clean and multicondition training modes to the upper bound that could be achieved when the recognition system benefits from using the hand-labeled database for noise estimation and frame-dropping. The comparison is extended to other recently published VAD methods [4], [11]–[13]. These results show that the performance of the proposed algorithm is above the original MO-LRT VAD being the one that is closer to the ideal results obtained with the reference database.

Table V shows the recognition performance for the Spanish and Finnish SpeechDat-Car database when WF and FD are performed on the base system [15]. Again, the VAD outperforms the original MO-LRT and all the algorithms used for reference yielding relevant improvements in speech recognition. Note that these particular databases used in the AURORA 3 experiments

TABLE III
AVERAGE WORD ACCURACY FOR THE AURORA-2 DATABASE. (a) CLEAN TRAINING. (b) MULTICONDITION TRAINING

(a)

	Base	**Base + WF**						**Base + WF + FD**					
		G.729	AMR1	AMR2	AFE	MO-LRT	**Proposed**	G.729	AMR1	AMR2	AFE	MO-LRT	**Proposed**
Clean	99.03	98.81	98.80	98.81	98.77	98.87	98.92	98.41	97.87	98.63	98.78	99.11	99.18
20dB	94.19	87.70	97.09	97.23	97.68	97.40	97.45	83.46	96.83	96.72	97.82	98.12	98.15
15dB	85.41	75.23	92.05	94.61	95.19	95.14	95.21	71.76	92.03	93.76	95.28	96.32	96.43
10dB	66.19	59.01	74.24	87.50	87.29	89.01	89.23	59.05	71.65	86.36	88.67	91.18	91.45
5dB	39.28	40.30	44.29	71.01	66.05	73.31	73.62	43.52	40.66	70.97	71.55	77.17	78.24
0dB	17.38	23.43	23.82	41.28	30.31	44.11	44.38	27.63	23.88	44.58	41.78	49.97	51.35
-5dB	8.65	13.05	12.09	13.65	4.97	15.93	16.65	14.94	14.05	18.87	16.23	23.62	25.02
Average	60.49	57.13	66.30	78.33	75.30	79.79	**79.98**	57.08	65.01	78.48	79.02	82.55	**83.12**

(b)

	Base	**Base + WF**						**Base + WF + FD**					
		G.729	AMR1	AMR2	AFE	MO-LRT	**Proposed**	G.729	AMR1	AMR2	AFE	MO-LRT	**Proposed**
Clean	98.48	98.16	98.30	98.51	97.86	98.48	98.52	97.50	96.67	98.12	98.39	98.84	98.88
20dB	97.39	93.96	97.04	97.86	97.60	97.95	98.02	96.05	96.90	97.57	97.98	98.51	98.61
15dB	96.34	89.51	95.18	96.97	96.56	97.09	97.26	94.82	95.52	96.58	96.94	97.69	97.78
10dB	93.88	81.69	91.90	94.43	93.98	94.74	95.01	91.23	91.76	93.80	93.63	95.63	95.98
5dB	85.70	68.44	80.77	87.27	86.41	87.77	88.12	81.14	80.24	85.72	85.32	88.95	89.12
0dB	59.02	42.58	53.29	65.45	64.63	66.79	67.51	54.50	53.36	62.81	63.89	67.90	68.56
-5dB	24.47	18.54	23.47	30.31	28.78	31.89	33.25	23.73	23.29	27.92	30.80	33.82	35.18
Average	86.47	75.24	83.64	88.40	87.84	88.87	**89.18**	83.55	83.56	87.29	87.55	89.73	**90.01**

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

(a)

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

(b)

	Woo	Li	Marzinzik	Sohn	MO-LRT	**Proposed**
Base + WF	83.64	77.43	84.02	83.89	84.33	**84.58**
Base + WF+ FD	81.09	82.11	85.23	83.80	86.14	**86.56**

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.

When comparing the revised MO-LRT to the original method, the improvements shown in Tables III–V are mainly due to: 1) a reduction of the number of substitution errors when the VAD is only used for WF-based speech enhancement and 2) 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 nonspeech 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.

VI. CONCLUSION

This paper revises a multiple observation likelihood ratio test for voice activity detection in noisy environments. The new approach not only evaluates the two hypotheses consisting of all the observations to be speech or nonspeech but all the possible hypotheses defined over the individual observations. The method exhibits the same smoothing process and reduced variance of the MO-LRT decision rule with the benefit of being suitable for a more effective hangover mechanism development. The experimental results showed a high speech/nonspeech discrimination accuracy over a wide range of SNR conditions and significant improvements over standardized VADs such as ITU-T G.729, ETSI AMR, and ETSI AFE, as well as other publicly available approaches.

TABLE V
AVERAGE WORD ACCURACY (%) FOR THE AURORA-3 SPEECH DAT-CAR DATABASES: (a) SPANISH. (b) FINNISH

(a)

	Base	Woo	Li	Marzinzik	Sohn	G729	AMR1	AMR2	AFE	MO-LRT	**Proposed**
WM	92.94	95.35	91.82	94.29	96.07	88.62	94.65	95.67	95.28	96.33	96.63
MM	83.31	89.30	77.45	89.81	91.64	72.84	80.59	90.91	90.23	91.61	91.78
HM	51.55	83.64	78.52	79.43	84.03	65.50	62.41	85.77	77.53	87.43	87.68
Average	**75.93**	89.43	82.60	87.84	90.58	75.65	74.33	90.78	87.68	91.79	**92.03**

(b)

	Base	Woo	Li	Marzinzik	Sohn	G729	AMR1	AMR2	AFE	MO-LRT	**Proposed**
WM	92.74	86.81	85.60	93.73	93.84	88.62	94.57	95.52	94.25	94.50	94.72
MM	80.51	66.62	55.63	76.47	80.10	67.99	81.60	79.55	82.42	79.01	80.38
HM	40.53	62.54	58.34	68.37	75.34	65.80	77.14	80.21	56.89	83.12	84.34
Average	**71.26**	71.99	66.52	79.52	83.09	74.14	84.44	85.09	77.85	85.54	**86.48**

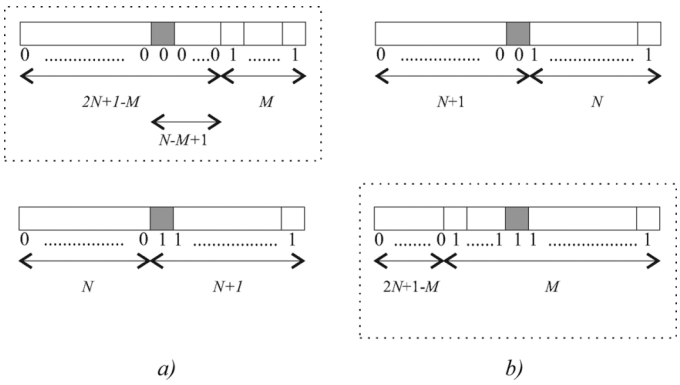

Fig. 12. Hypotheses considered for the derivation of the expected value of the decision criterion. (a) $M \leq N$. (b) $M > N$.

APPENDIX
DERIVATION OF $E\{\log \Lambda^{\mathrm{RMO}}\}$

This appendix shows the derivation of the expected value of the decision criterion shown in (58). The expected value is dependent on the *a priori* SNR and the position of the analysis window. Thus, assuming the same scenario shown in Fig. 1, where a nonspeech-to-speech transition is shown, the expected value can be derived by evaluating the probability of the most probable hypotheses in $\mathbf{H}_0^*$ and $\mathbf{H}_1^*$ for the detection problem in Fig. 1. Fig. 12 shows the most probable hypotheses for the two cases: (a) $M \leq N$ and (b) $M > N$. According to (58), The computation of the decision criterion is then reduced to the evaluation of the likelihood ratio for the frame locations with different partial hypothesis assumptions within the analysis window.

a) When $M \leq N$, the two most probable hypotheses in $\mathbf{H}_0$ and $\mathbf{H}_1$ shown in Fig. 12 differ in $N+1-M$ frame locations so that the expected value can be expressed as:

$$E[\log \Lambda^{\mathrm{RMO}}] = \frac{N+1-M}{(N+1)J} E\left[\log\left(\frac{p(\mathbf{x}_k|1)}{p(\mathbf{x}_k|0)}\right)\right]. \quad (61)$$

By using (18)

$$E\left[\log\left(\frac{p(\mathbf{x}_k|1)}{p(\mathbf{x}_k|0)}\right)\right] = \sum_{j=0}^{J-1}\left(\frac{\xi_j E[\gamma_{j,k}]}{1+\xi_j} - \log(1+\xi_j)\right) \quad (62)$$

and assuming white signal an noise models leads to

$$E\left[\log\left(\frac{p(\mathbf{x}_k|1)}{p(\mathbf{x}_k|0)}\right)\right] = J\left(\frac{\xi}{1+\xi} - \log(1+\xi)\right). \quad (63)$$

Note that, in this case, $E[\gamma_{j,k}] = \xi$ since $\mathbf{x}_k$ corresponds to nonspeech observations. Finally, substituting (63) in (65)

$$E[\log \Lambda^{\mathrm{RMO}}] = \frac{N+1-M}{N+1}\left(\frac{\xi}{1+\xi} - \log(1+\xi)\right) \quad (64)$$

b) When $M > N$, in a similar way, the expected value of the decision criterion is given by

$$E[\log \Lambda^{\mathrm{RMO}}] = \frac{M-N}{(N+1)J} E\left[\log\left(\frac{p(\mathbf{x}_k|1)}{p(\mathbf{x}_k|0)}\right)\right] \quad (65)$$

that, under the same assumptions above, can be expressed as

$$E[\log \Lambda^{\mathrm{RMO}}] = \frac{M-N}{N+1}\left(\xi - \log(1+\xi)\right). \quad (66)$$

Note that, now $E[\gamma_{j,k}] = 1 + \xi$ since $\mathbf{x}_k$ corresponds to speech observations.

REFERENCES

[1] *A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70*, ITU-T Rec. G.729-Annex B, ITU, 1996.

[2] *Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels*, ETSI EN 301 708 Rec., ETSI, 1999.

[3] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms*, ETSI ES 202 050 Rec., ETSI, 2002.

[4] 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.

[5] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Process. Lett.*, vol. 8, no. 10, pp. 276–278, Oct.. 2001.

[6] 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. 689–692, Oct. 2005.

[7] 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.*, vol. 42, no. 3–4, pp. 271–287, 2004.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[9] 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.

[10] 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 Autom. Speech Recognition: Challenges for the Next Millennium*, Paris, France, Sep. 2000.

[11] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 2, pp. 109–118, Feb. 2002.

[12] 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.

[13] 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.

[14] L. Karray and A. Martin, "Towards improving speech detection robustness for speech recognition in adverse environments," *Speech Commun.*, vol. 40, no. 3, pp. 261–276, 2003.

[15] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front-End Feature Extraction Algorithm; Compression Algorithms*, ETSI ES 201 108 Rec., ETSI, 2000.

[16] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge, U.K.: Cambridge Univ., 1997.

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 Assistant Professor in the Department of Signal Theory Networking, and Communications (GSTC), 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.

José C. Segura (M'93–SM'03) was born in Alicante, Spain, in 1961. He received the M.S. and Ph.D. degrees in physics from the University of Granada, Granada, Spain, in 1984 and 1991, respectively. He developed his Ph.D. dissertation on a variant of HMM modeling.

Since 1986, he has been working with the Research Group on Signals, Networking and Communications (GSTC), Department of Electronics and Computer Technology, University of Granada. Since January 2004, he has been the Coordinator of this research group. He has been the director of three Ph.D. dissertations on topics related to speech recognition.

Juan M. 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 Assistant Professor with the Department of Signal Theory, Networking, and Communications, University of Granada. His present interests lie in the field of statistical signal processing and its application to speech and image processing.

Luz García received the M.Sc degree In telecommunication engineering from the Universidad Politécnica of Madrid, Madrid, Spain, in 2000. She is currently pursuing the PhD. degree in nonlinear feature transformations for automatic speech recognition.

She worked as a support and supply Telecom Engineer from 2000 to 2005 at Ericsson Spain S.A. She joined the Department of Signal Theory, Networking, and Communications (GTSC), University of Granada, Granada, Spain, in 2005. Her actual research field is speech processing.