An Effective Subband OSF-Based VAD With Noise Reduction for Robust Speech Recognition

Javier Ramírez, José C. Segura, *Senior Member, IEEE*, Carmen Benítez, *Member, IEEE*, Ángel de la Torre, and Antonio Rubio, *Senior Member, IEEE*

***Abstract*—An effective voice activity detection (VAD) algorithm is proposed for improving speech recognition performance in noisy environments. The approach is based on the determination of the speech/nonspeech divergence by means of specialized order statistics filters (OSFs) working on the subband log-energies. This algorithm differs from many others in the way the decision rule is formulated. Instead of making the decision based on the current frame, it uses OSFs on the subband log-energies which significantly reduces the error probability when discriminating speech from nonspeech in a noisy signal. Clear improvements in speech/nonspeech discrimination accuracy demonstrate the effectiveness of the proposed VAD. It is shown that an increase of the OSF order leads to a better separation of the speech and noise distributions, thus allowing a more effective discrimination and a tradeoff between complexity and performance. The algorithm also incorporates a noise reduction block working in tandem with the VAD and showed to further improve its accuracy. A previous noise reduction block also improves the accuracy in detecting speech and nonspeech. The experimental analysis carried out on the AURORA databases and tasks provides an extensive performance evaluation together with an exhaustive comparison to the standard VADs such as ITU G.729, GSM AMR, and ETSI AFE for distributed speech recognition (DSR), and other recently reported VADs.**

***Index Terms*—Noise reduction, robust speech recognition, speech/nonspeech detection, subband order statistics filters.**

I. Introduction

CURRENTLY, there are technical barriers inhibiting speech recognition systems from meeting the requirements of modern applications. An important drawback affecting most of the applications is the environmental noise and its harmful effect on the system performance. Examples of such systems are the new wireless communications voice services or digital hearing aid devices.

Numerous techniques have been derived to palliate the effect of noise on the system performance. Most of the noise reduction algorithms often require an estimate of the noise statistics by means of a precise voice activity detector (VAD). Speech/nonspeech detection is an unsolved problem in speech processing and affects numerous applications including robust speech recognition [1], [2], discontinuous transmission [3], [4], real-time speech transmission on the Internet [5] or combined noise reduction and echo cancellation schemes in the context of telephony [6], [7]. The speech/nonspeech classification task is not as trivial as it appears, and most of the VAD algorithms fail when the level of background noise increases. During the last decade, numerous researchers have developed different strategies for detecting speech on a noisy signal [8]–[11] and have evaluated the influence of the VAD effectiveness on the performance of speech processing systems [12]. Most of them have focussed on the development of robust algorithms with special attention on the derivation and study of noise robust features and decision rules [13]–[15]. The different approaches include those based on energy thresholds [13], pitch detection [16], spectrum analysis [15], zero-crossing rate [4], periodicity measure [17], higher order statistics in the LPC residual domain [18] or combinations of different features [3], [4], [19].

A representative set of the reported VAD methods formulates the decision rule on a frame by frame basis using instantaneous measures of the divergence between speech and noise [8], [13]. Recently, it has been shown that VAD robustness can be improved by using long-term information about the speech signal to formulate the decision rule [2], [20], [21]. An interesting approach is the endpoint detection algorithm proposed by Li [14], which is based on the optimal edge detector first established by Canny [22], and uses optimal FIR filters for edge detection. However, alternative approaches such as nonlinear filters are still a non fully developed research topic for speech end-point detection. Order statistic filters (OSFs) have been proposed for many applications including edge detection in images [23] and the optimal design of a class of OSFs called L-filters has been studied [24]. The design of an optimal L-filter is not a trivial task and simplified design procedures are normally used. A first approach is the use of moving quasirange filters [25] that are defined as the difference between symmetric rank-order filters. Although these techniques were developed mainly for image processing, several authors have studied them for robust speech/nonspeech discrimination. Cox [26] studied a nonparametric rank-order statistical signal detection scheme. This method is based on a four-channel filter bank decomposition with the ranking operation being performed in the time domain over 15-ms speech frames and involving samples of the four channels. Although the algorithm performed well and showed to be robust against noise, it suffers several drawbacks that need to be addressed. First, it is computationally intensive since it requires ranking a 400-sample data set every 15 ms. Second, the decision procedure was only adequate for white noise and would fail when other noises with a low-pass spectral profile such as

Manuscript received December 23, 2003; revised August 30, 2004. This work was supported in part by the Spanish Government under CICYT Project TIC2001-3323. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Futoshi Asano.

The authors are with the Department of Signal Theory, Networking and Communications, University of Granada, 18071 Granada, Spain (e-mail: javierrp@ugr.es).

Digital Object Identifier 10.1109/TSA.2005.853212

car noise are considered. All these preliminary results motivated further studies in this field and the exploration of its application to related areas such as effective voice activity detection in noisy environments.

On the other hand, noneffective speech/nonspeech detection is an important source of performance degradation in automatic speech recognition (ASR) systems. There are two main motivations for that.

i) Most of the speech enhancement algorithms make use of the VAD module in order to estimate the statistics of noise. Therefore, the effectiveness of the noise compensation algorithms is strongly affected by the accuracy of the VAD.
ii) Frame-dropping (FD) is a frequently used technique in speech recognition to reduce the number of insertion errors. Since it is based on the VAD, speech frames incorrectly labeled as silence causes unrecoverable deletion errors, and silence frames incorrectly labeled as speech could increase the insertion errors.

This paper explores a new alternative toward improving speech detection robustness in adverse environments and the performance of speech recognition systems. The proposed VAD includes a noise reduction block that preceeds the VAD, and uses OSFs to formulate a robust decision rule. The rest of the paper is organized as follows. Section II reviews the theoretical background on OSFs and shows the proposed algorithm. Section III analyzes the motivations for the proposed algorithm by comparing the speech/nonspeech distributions for different filter lengths and when noise reduction is optionally applied. Section IV describes the experimental framework considered for the evaluation of the proposed endpoint detection algorithm. Finally, Section V summarizes the conclusions of this work.

II. Order Statistics Filters for Endpoint Detection

Nonlinear filters including OSFs [24], also known as L-filters, have been shown to be more effective and robust than linear filters in many applications [27]–[30]. As an example, filters based on order statistics have been successfully employed in restoration of signals and images corrupted by additive noise. The most common OSF is the median filter that is easy to implement and exhibits good performance in removing impulsive noise. The output of an L-order OSF is defined on the data set $\{x(l-N),\ldots,x(l),\ldots,x(l+N)\}$ by

$$y(l)=\sum_{i=1}^{L} a_i x_{(i)} \tag{1}$$

where $L=2N+1$ and $x_{(i)}$ represents the past data set rearranged in ascending order, that is

$$\{x_{(1)}\le x_{(2)}\le\cdots\le x_{(L)}\}. \tag{2}$$

Note that $x_{(1)}$ is the minimum, $x_{(L)}$ is the maximum, and $x_{(N+1)}$ is the median. The weights a_i define the OSF. As an example, the median filter is a special type of L-filter whose coefficients are $a_i=1$ if $i=N+1$ and $a_i=0$ otherwise.

This paper addresses the use of OSFs for endpoint detection. The proposed approach is defined to operate on the subband log-energies. Noise reduction is performed first and the VAD decision is formulated on the de-noised signal. The noisy speech signal $x(n)$ is decomposed into 25-ms frames with a 10-ms window shift. Let $X(m,l)$ be the spectrum magnitude for the mth band $(m=0,1,\ldots,\mathrm{NFFT}-1)$ at frame l. The design of the noise reduction block is based on Wiener filter (WF) theory whereby the attenuation is function of the signal-to-noise ratio (SNR) of the input signal. A block diagram of the system is shown in Fig. 1. Note that the VAD decision is formulated in terms of the de-noised signal, being the subband log-energies processed by means of order statistics filters.

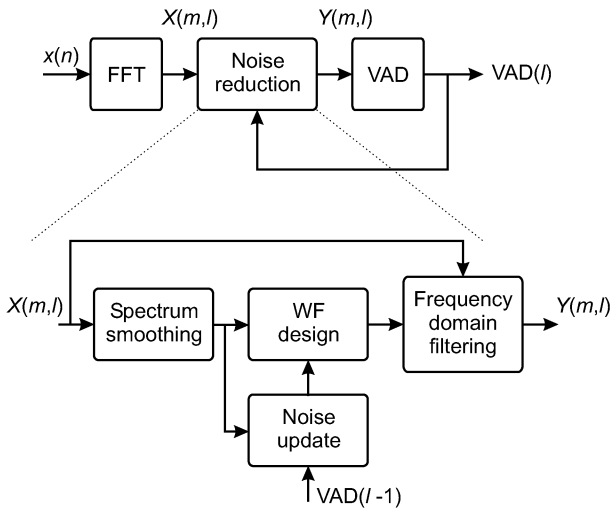

Fig. 1. Block diagram of the proposed OSF-based VAD.

A. Noise Reduction Block

The noise reduction block consists of four stages.

i) Spectrum smoothing. The power spectrum is averaged over two consecutive frames and two adjacent spectral bands.
ii) Noise estimation. The noise spectrum $N_e(m,l)$ is updated by means of a 1st order IIR filter on the smoothed spectrum $X_s(m,l)$, that is, $N_e(m,l)=\lambda N_e(m,l-1)+(1-\lambda)X_s(m,l)$ where $\lambda=0.99$ and $m=0,1,\ldots,\mathrm{NFFT}/2$.
iii) WF design. First, the clean signal $S(m,l)$ is estimated by combining smoothing and spectral subtraction:

$$S(m,l)=\gamma S'(m,l-1)+(1-\gamma)\max(X_s(m,l)-N_e(m,l),0) \tag{3}$$

where $\gamma=0.98$. Then, the WF $H(m,l)$ is designed as

$$H(m,l)=\frac{\eta(m,l)}{1+\eta(m,l)} \tag{4}$$

where

$$\eta(m,l)=\max\left[\frac{S(m,l)}{N_e(m,l)},\eta_{\min}\right] \tag{5}$$

and $\eta_{\min}$ is selected so that the filter H yield a 20 dB maximum attenuation. Note that $S'(m,l)$ is the spectrum of the cleaned speech signal, assumed to be zero at the beginning of the process and needed for designing the WF through (3) to (5). It is given by

$$S'(m,l)=H(m,l)X(m,l). \tag{6}$$

The filter $H(m,l)$ is smoothed in order to eliminate rapid changes between neighbor frequencies that may often

cause musical noise. Thus, the variance of the residual noise is reduced and consequently, the robustness when detecting nonspeech is enhanced. The smoothing is performed by truncating the impulse response of the corresponding causal FIR filter to 17 taps using a Hanning window. With this operation performed in the time domain, the frequency response of the Wiener filter is smoothed and the performance of the VAD is improved.

iv) Frequency domain filtering. The smoothed filter H_s is applied in the frequency domain to obtain the de-noised spectrum $Y(m,l) = H_s(m,l)X(m,l)$.

B. OSF-Based Endpoint Detection

Once the input speech has been de-noised, the log-energies for the lth frame, $E(k,l)$, in K subbands ($k = 0, 1, \ldots, K-1$), are computed by means of

$$E(k,l) = \log\left(\frac{K}{\mathrm{NFFT}}\sum_{m=m_k}^{m_{k+1}-1}|Y(m,l)|^2\right)$$
$$m_k = \left\lfloor \frac{\mathrm{NFFT}}{2K}k \right\rfloor \quad k = 0, 1, \ldots, K-1 \tag{7}$$

where an equally spaced subband assignment is used.

The noise suppression block performs noise reduction of the block $\{X(m,l-N), X(m,l-N+1), \ldots, X(m,l-1), X(m,l), X(m,l+1), \ldots, X(m,l+N)\}$ before the subband log-energies $E(k,l)$ are computed. This is carried out as follows. During the initialization process, the noise suppression algorithm is applied to the first $2N+1$ frames and, in each iteration, the $(l+N+1)$th frame is de-noised, so that $Y(m,l+N+1)$ become available for the next iteration. It is worthwhile clarifying that the noise spectrum estimated up to the lth frame, which does not depend on future frames, is used for denoising N frames forward and that this estimate is updated if the VAD decides the lth frame to be a noise-only frame. The only assumption that is made here is that the noise spectrum does not change significantly within an N-frame neighborhood of the lth frame.

The algorithm uses two OSFs for the multiband quantile (MBQ) SNR estimation. The implementation of both OSFs is based on a sequence of $2N+1$ log-energy values $E(k,l-N), \ldots, E(k,l), \ldots, E(k,l+N)$ around the frame to be analyzed. The rth order statistics of this sequence, $E_{(r)}(k,l)$, is defined as the rth largest number in algebraic order. A first OSF estimates the subband signal energy by means of

$$Q_p(k,l) = (1-f)E_{(s)}(k,l) + fE_{(s+1)}(k,l) \tag{8}$$

where $Q_p(k,l)$ is the p sampling quantile, $s = \lfloor 2pN \rfloor$ and $f = 2pN - s$.

Finally, the SNR in each subband is measured by

$$\mathrm{QSNR}(k,l) = Q_p(k,l) - E_N(k) \tag{9}$$

where $E_N(k)$ is the noise level in the kth band that needs to be estimated. For the initialization of the algorithm, the first N frames are assumed to be nonspeech frames and the noise level in the kth band, $E_N(k)$, is estimated as the median of the set $\{E(0,k), E(1,k), \ldots, E(N-1,k)\}$. In order to track nonstationary noisy environments, the noise references are updated during nonspeech periods by means of a second OSF (a median filter)

$$E_N(k) = \alpha E_N(k) + (1-\alpha)Q_{0.5}(k,l)$$
$$k = 0, 1, \ldots, K-1 \tag{10}$$

where $Q_{0.5}(k,l)$ is the output of the median filter and $\alpha = 0.97$ was experimentally selected. On the other hand, the sampling quantile $p = 0.9$ is selected as a good estimation of the subband spectral envelope.

The decision rule is then formulated in terms of the average subband SNR

$$\mathrm{SNR}(l) = \frac{1}{K}\sum_{k=0}^{K-1}\mathrm{QSNR}(k,l). \tag{11}$$

If the SNR is greater than a threshold η, the current frame is classified as speech, otherwise it is classified as nonspeech. The algorithm for fixing the threshold is similar to that used in the AMR1 standard [3]. 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 (η_0) and noisiest conditions (η_1). Thus, the threshold is adaptive to the measured full-band noise energy E

$$\eta = \begin{cases} \eta_0 & E < E_0 \\ \frac{\eta_1-\eta_0}{E_1-E_0}(E-E_0)+\eta_0 & E_0 \le E \le E_1 \\ \eta_1 & E_1 < E \end{cases} \tag{12}$$

thus enabling the VAD selecting the optimum working point for different SNR conditions. Note that, the threshold is linearly decreased as the noise level is increased between (E_0, η_0) and (E_1, η_1) which represent optimum thresholds for the cleanest and noisiest conditions defined by the noise energies E_0 and E_1, respectively. The sensitivity of the proposed VAD to the adaptive threshold update is not an important issue for several reasons: i) the use of an adaptable threshold is only meant for improving the performance of the VAD in low noise conditions. Increasing the detection threshold when the noise level is low helps to better identify nonspeech periods without damaging the performance of the VAD under high-noise conditions, ii) in order to reduce the sensitivity of the VAD to the noise level, the threshold is bounded to be linear between 0 and 1 in a predetermined interval of noise energy values, and iii) its use has reported benefits but it is not the key for the high performance achieved by the proposed VAD. Moreover, the experiments conducted on up to five different databases lead to high levels of performance maintained for all the noises and SNR conditions.

The proposed algorithm takes advantage of the noise reduction block for improving its robustness against the background noise. It is worth clarifying here how the convergence of the feedback loop shown in Fig. 1 has been guaranteed. The solution adopted has been to assume that each utterance of the database contains a noise-only period at the beginning of the sentence for the initialization of the feedback. If the utterance does not start with a nonspeech period the algorithm could fail at the beginning to evaluate the noise spectrum and the detection afterwards could be totally erroneous. However, reported VAD algorithms

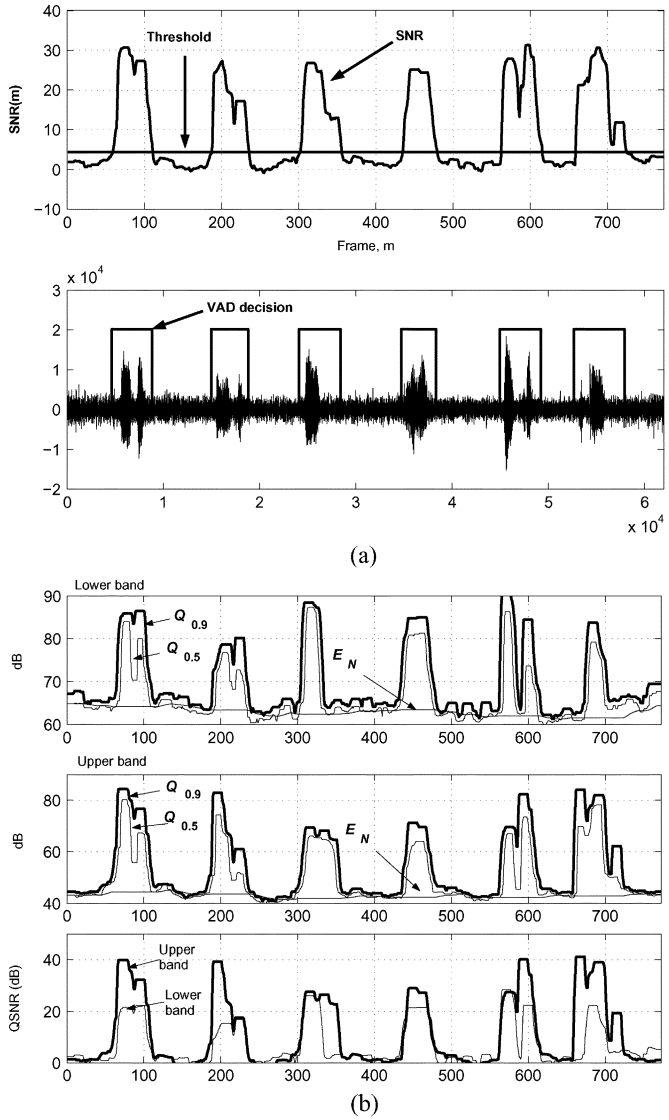

Fig. 2. Operation of the VAD on an utterance of Spanish SDC database. (a) SNR and VAD Decision. (b) Subband SNRs.

need an estimation of the noise parameters and normally make these assumptions for their initialization. After the initialization, an incorrect estimation of the noise statistics when the VAD fails is a common problem to most of the VADs and needs to be addressed. The results presented in the next sections will show the effectiveness of the proposed VAD that is free of convergence problems.

On the other hand, the performance improvements are only damaged by the N-frame delay that is required for the operation of the VAD. This fact can be an implementation obstacle for several applications, but for others, such as speech recognition, the benefits in robustness will justify its use as it will be shown in the next sections.

Fig. 2 shows the operation of the proposed VAD on an utterance of the Spanish SpeechDat-Car (SDC) database [31]. The phonetic transcription is ["siete", "θinko", "dos", "uno", "otSo", "seis"]. For this example, $K = 2$ subbands were used while $N = 8$. The optimal selection of these parameters will be studied later in this paper. It is clearly shown how the SNR in the upper and lower band yields improved speech/nonspeech discrimination of fricative sounds by giving complementary information. The VAD performs an advanced detection of beginnings and delayed detection of word endings which, in part, makes a hang-over unnecessary.

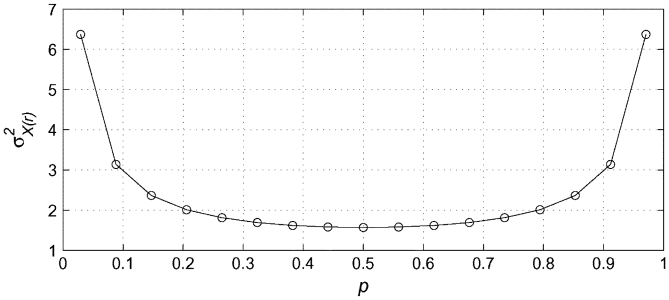

Fig. 3. Sampling quantile variance for $N = 8$.

C. OSF Selection

The selection of the OSFs used below can be justified as follows. Let $X_1, X_2, \ldots, X_{2N+1}$ be a set of uniformly distributed random variables with probability distribution function (pdf) $f(x)$. From the asymptotic theory for order statistics [32], the variance of the rth order statistic $X_{(r)}$ defined by

$$\sigma^2_{X_{(r)}} = \int_{-\infty}^{\infty} (x - \mu_r)^2 f_{X_{(r)}}(x) dx \tag{13}$$

can be approximated for N sufficiently large by means of

$$\sigma^2_{X_{(r)}} = \frac{p(1-p)}{f^2(F^{-1}(p))} \tag{14}$$

where μ_r is the mean of the rth order statistics, $f_{X_{(r)}}(x)$ is the pdf of the ordered variable $X_{(r)}$ and $F^{-1}(p)$ is the quantile function evaluated at p.

Fig. 3 shows the variance of the p sampling quantiles for a set of uniformly distributed random variables with zero mean and unity variance Gaussian pdf. The variance reaches the minimum value for $p = 0.5$, that is, the median is the minimum variance order statistic. In this way, the median filter used for the estimation of the noise level in each subband provides a robust estimation of the noise statistics with the minimal variance. On the other hand, for the purpose of detecting speech in noise, a higher quantile is needed. The maximum $X_{(2N+1)}$ could be a candidate but its performance is far from being optimal because of its high variance as shown in Fig. 3, which normally leads to high false alarm rates. Then, a good compromise between detectability of speech in noise and low false alarm rates is to select a reduced variance quantile like the $p = 0.9$ sampling quantile. That is the reason for the selection of $Q_{0.5}$ and $Q_{0.9}$ in the proposed VAD.

III. Speech/Non-Speech Distributions

In order to clarify the motivations for the algorithm proposed, the distributions of the SNR defined by (11) were studied as a function of the OSF length. A hand-labeled version of the Spanish SDC database [31] was used in the analysis. This database contains recordings from close-talking and distant microphones at different driving conditions: a) stopped car, motor

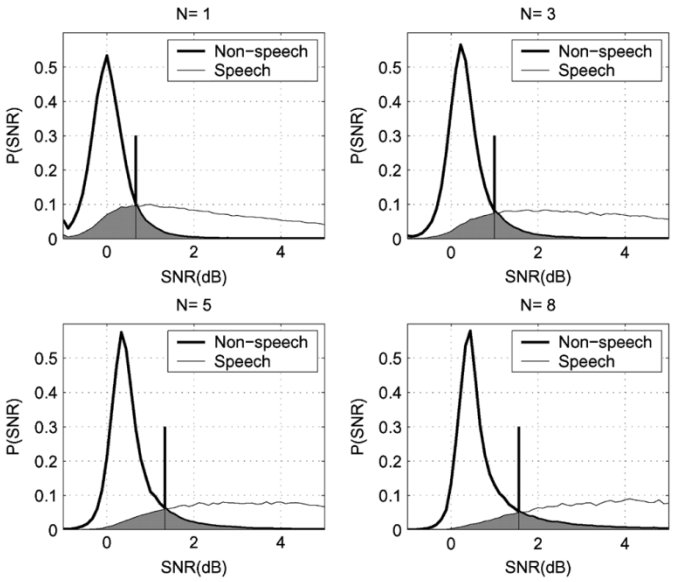

Fig. 4. Speech/nonspeech distributions and error probabilities of the optimum Bayes classifier for $N = 1, 3, 5$, and 8.

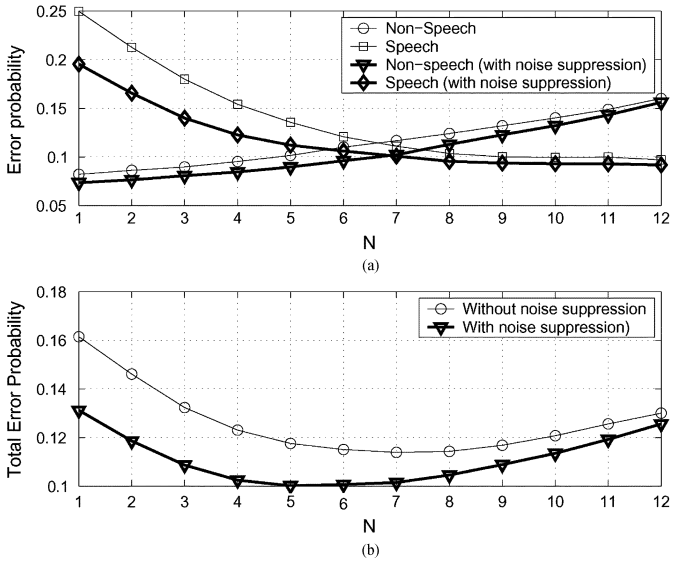

Fig. 5. Probability of error as a function of *N*. (a) Speech and nonspeech errors. (b) Total classification error of speech and nonspeech.

running, b) town traffic, low speed, rough road, and c) high speed, good road. The most unfavorable noise environment (high speed, good road and distant microphone) with an average SNR of about 5 dB was selected for the experiments. Thus, the N-order SNR was measured during speech and nonspeech periods, and the histogram and probability distributions were built. Fig. 4 shows the distributions of speech and noise for $N = 1, 3, 5$, and 8. When the length of the OSFs increases, the noise variance decreases and the speech distribution is shifted to the right being more separated from the nonspeech distribution. Thus, the distributions of speech and nonspeech are less overlapped and consequently, the error probabilities are reduced. As a result, it is clearly shown that the speech and noise distributions are better discriminated when more log-energy observations are considered, thus increasing the VAD robustness against environmental noises.

The reduction of the distribution overlap yields improvements in speech/pause discrimination. This fact can be shown by calculating the misclassification errors of speech and noise for an optimal Bayes classifier. Note that Fig. 4 also shows the areas representing the probabilities of incorrectly detecting speech and nonspeech and the optimal decision threshold. Fig. 5(a) shows the independent decision errors for speech and nonspeech when the noise reduction block is optionally applied. The speech detection error is clearly reduced when increasing the length of the window while the increased robustness is only damaged by a moderate increase in the nonspeech detection error. These improvements are achieved by reducing the overlap between the distributions when N is increased as shown in Fig. 4. Increasing the length of the window is beneficial in high noise environments since the VAD introduces an artificial "hang-over" period which reduces front and rear-end clipping errors. This saving period is the reason for the increase of the nonspeech detection error shown in Fig. 5(a). On the other hand, if no noise reduction is performed, the speech detection error is reduced from 25% to 10% when the order of the VAD is increased from 1 to 8 frames. Further reductions of the speech detection error ranging from 20% to 9% are achieved if denoising is considered prior to OSF-based endpoint detection. As as result, if the noise reduction block is considered, the speech classification errors are reduced, thus increasing the VAD robustness against noise.

Fig. 5(b) shows the total error defined as the average of speech and nonspeech errors weighted by the a priori speech and nonspeech probabilities. The total error is reduced with the increasing length of the OSFs and exhibits a minimum value for a fixed order. If no noise reduction algorithm is applied before endpoint detection, the minimum error is 11.39% for $N = 7$ while the minimum error is 10.01% for $N = 5$ if noise reduction is performed. Thus, the delay of the algorithm is reduced by incorporating a preceeding noise reduction stage as described in Section II.A. According to Fig. 5, the optimal value of the order of the VAD would be $N = 8$. Therefore, using a noise reduction block previous to endpoint detection together with a long-term measure of the SNR using OSFs reports important benefits for detecting speech in noise since misclassification errors are significantly reduced.

IV. Experimental Framework

Several experiments are commonly conducted to evaluate the performance of VAD algorithms. The analysis is mainly focussed on the determination of the error probabilities or classification errors at different SNR levels [15], and the influence of the VAD decision on the performance of speech processing systems [12]. Subjective performance tests have also been considered for the evaluation of VADs working in combination with speech coders [33]. The experimental framework and the objective performance tests conducted to evaluate the proposed algorithm are described in this section.

A. Speech/Nonspeech Discrimination Analysis in Noise

First, the proposed VAD was evaluated in terms of the ability to discriminate speech from nonspeech at different

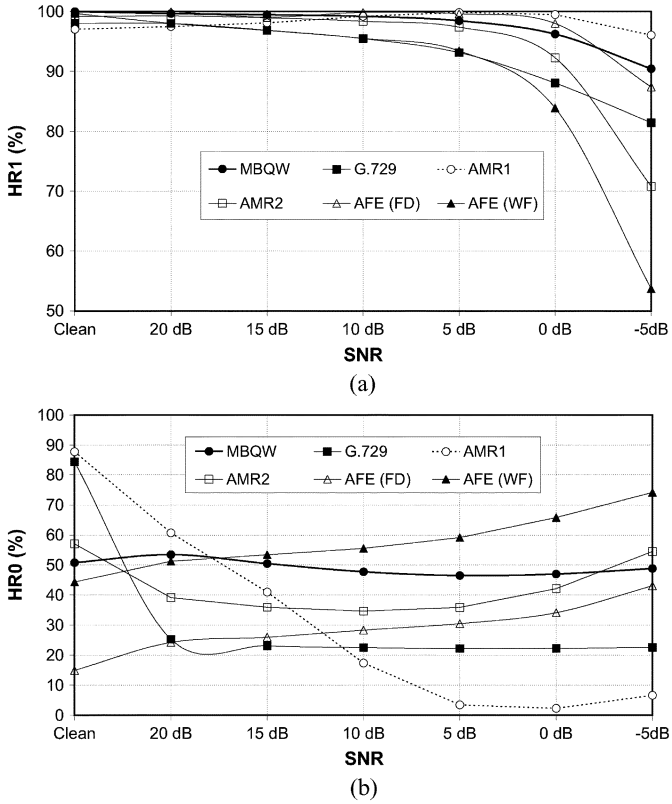

Fig. 6. Speech/nonspeech discrimination analysis as a function of the SNR. Results are averaged for all the noises considered in the AURORA 2 database. (a) Speech hit-rate (compared to standard VADs). (b) Nonspeech hit-rate (compared to standard VADs).

SNRs. The original AURORA-2 database [34] was used. 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. 6 compares the performance of the proposed VAD to standard G.729, AMR and AFE VADs for clean conditions and SNR levels ranging from 20 to -5 dB. These results are averaged over the entire set of noises. Note that results for the two VADs defined in the AFE standard for distributed speech recognition (DSR) [35] for noise spectrum estimation in Wiener filtering and nonspeech FD are also provided.

The proposed VAD scheme (MBQW: multiband quantile VAD with Wiener filtering) 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. G.729 VAD suffers poor speech detection accuracy

TABLE I
AVERAGE SPEECH/NON-SPEECH HIT RATES FOR SNRS BETWEEN CLEAN CONDITIONS AND -5 dB. COMPARISON OF THE PROPOSED MBQW VAD TO STANDARD AND RECENTLY REPORTED VADS

	G.729	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
	Woo	Li	Marzinzik	Sohn	**MBQW**
HR0 (%)	55.40	57.03	52.69	43.66	49.27
HR1 (%)	88.41	83.65	93.04	94.46	97.64

with the increasing noise level while nonspeech detection is good in clean conditions (85%) and poor (20%) in noisy conditions. AMR1 has an extreme conservative behavior with high speech detection accuracy for the whole range of SNR levels but very poor nonspeech detection results at increasing noise levels. Although AMR1 seems to be well suited for speech detection at unfavorable noise conditions, its extremely conservative behavior results in only 10% of the actual nonspeech frames getting correctly detected, making it of little use in practical speech processing system. AMR2 leads to considerable improvements over G.729 and AMR1 yielding better nonspeech detection accuracy, but still suffering fast degradation of the speech detection ability at unfavorable noisy conditions. The VAD used in the AFE standard for estimating the noise spectrum in the Wiener filtering stage is based on the full energy band and yields a poor speech detection performance with a fast decay of the speech hit-rate at low SNR values. On the other hand, the VAD used in the AFE for FD achieves a high accuracy in speech detection but moderate results in nonspeech detection.

Table I summarizes these results and the benefits reported by the proposed VAD in terms of the average speech/nonspeech hit-rates (for all the noises and SNR conditions). Note that, results for recently reported VAD methods [8], [13]–[15] are also included. The proposed VAD yields a 49.27% HR0 average value, while the G.729, AMR1, AMR2, WF, and FD AFE VADs yield 31.77%, 31.31%, 42.77%, 57.68%, and 28.74%, respectively. On the other hand, MBQW attains a 97.64% HR1 average value in speech detection while G.729, AMR1, AMR2, WF and FD AFE VADs provide 93.00%, 98.18%, 93.76%, 88.72%, and 97.70%, respectively. Marzinzik's VAD [15] tracking the power spectral envelope dynamics is the one of the four nonstandard VADs that yields the best compromise between speech and pause hit rates followed by the Sohn's [8], Woo's [13] and Li's [14] algorithms.

These results clearly demonstrate that there is no optimal VAD for all the applications. Each VAD is developed and optimized for specific purposes. Hence, the evaluation has to be conducted according to the specific goal of the VAD. Frequently, VADs avoid loosing speech periods leading to an extremely conservative behavior in detecting speech pauses (for instance, the

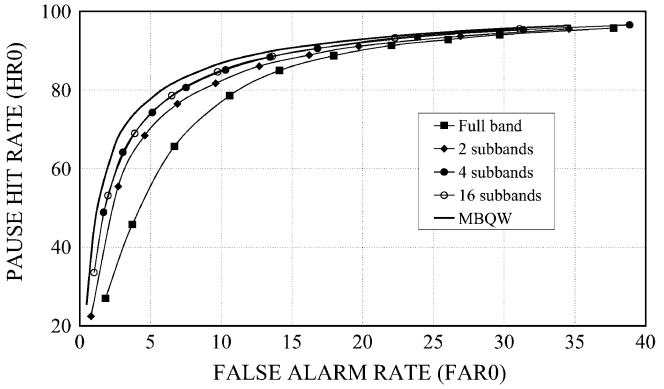

Fig. 7. Selection of the number of subbands (High: high speed, good road, 5 dB average SNR).

AMR1 VAD). Thus, in order to correctly describe the VAD performance, both parameters have to be considered. A more accurate analysis of this compromise is conducted in the following section with the receiver operating characteristic (ROC) curves.

B. Receiver Operating Characteristics Curves

The ROC curves are frequently used to completely describe the VAD error rate. The AURORA subset of the original Spanish SpeechDat-Car (SDC) database [31] was used in this analysis. This database contains 4914 recordings using close-talking and distant microphones from more than 160 speakers. The files are categorized into three noisy conditions: quiet, low noisy and highly noisy conditions, which represent different driving conditions with average SNR values between 25 dB, and 5 dB. The nonspeech hit rate (HR0) and the false alarm rate $(\mathrm{FAR0} = 100 - \mathrm{HR1})$ were determined in each noise condition being the actual speech frames and actual speech pauses determined by hand-labeling the database on the close-talking microphone.

1) Selection of the Optimum Number of Subbands: Before showing comparative results, the selection of the optimal number of subbands is considered. Fig. 7 shows the influence of the noise reduction block and the number of subbands on the ROC curves. First, noise reduction is not performed to better show the influence of the number of subbands. Increasing the number of subbands improves the performance of the proposed VAD by shifting the ROC curves in the ROC space. For more than four subbands, the VAD reports no additional improvements. This value yields the best trade-off between computational cost and performance. On the other hand, the noise reduction block included in the proposed MBQW VAD reports an additional shift of the ROC curve as shown in Fig. 7.

2) Comparative Results: Fig. 8 shows the ROC curves of the proposed VAD and other frequently referred algorithms [8], [13]–[15] for recordings from the distant microphone in quiet, low and high noisy conditions. The working points of the G.729, AMR, and AFE VADs are also included. The results show improvements in detection accuracy over standard VADs and over a representative set VAD algorithms [8], [13]–[15]. The following can be concluded from these results.

i) The working point of the G.729 VAD shifts to the right in the ROC space with decreasing SNR.

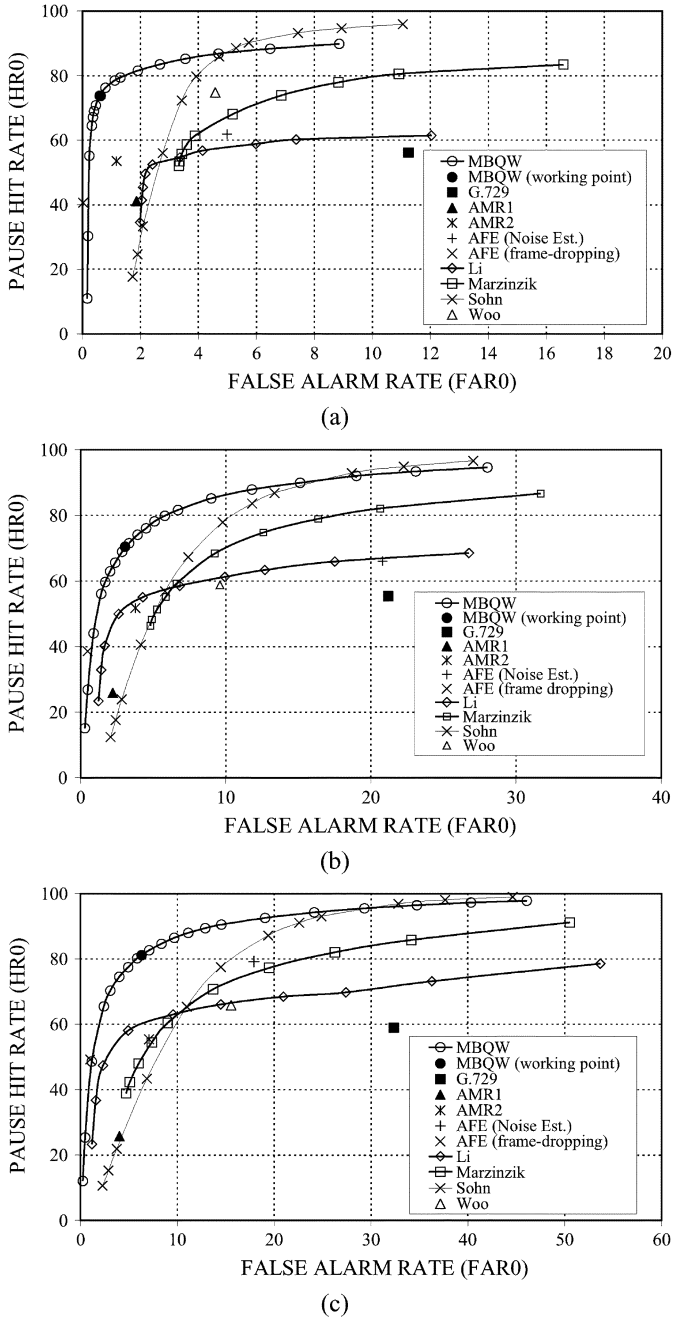

Fig. 8. ROC curves obtained for different subsets of the Spanish SDC database at different driving conditions: (a) Quiet (stopped car, motor running, 12 dB average SNR). (b) Low (town traffic, low speed, rough road, 9 dB average SNR). (c) High (high speed, good road, 5 dB average SNR).

ii) AMR1 works on a low false alarm rate point of the ROC space but exhibits poor nonspeech hit rate.

iii) AMR2 yields clear advantages over G.729 and AMR1 exhibiting important reduction of the false alarm rate when compared to G.729 and increased nonspeech hit rate over AMR1.

iv) The VAD used in the AFE for noise estimation yields good nonspeech detection accuracy but works on a high false alarm rate point on the ROC space. It suffers from rapid performance degradation when the driving conditions get noisier. On the other hand, the VAD used in the

AFE for FD has been planned to be conservative since it is only used in the DSR standard for that purpose. Thus, it exhibits poor nonspeech detection accuracy working on a low false alarm rate point of the ROC space.

v) The proposed VAD also works with lower false alarm rate and higher nonspeech hit rate when compared to the Sohn's [8], Woo's [13], Li's [14], and Marzinzik's [15] algorithms.

Thus, among all the VAD 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, that is based on an optimum linear filter for edge detection. The proposed VAD also improves Marzinzik's VAD that tracks the power spectral envelopes, and the Sohn's VAD, that formulates the decision rule by means of a statistical likelihood ratio test.

Fig. 8 shows the ability of this VAD to tune the decision threshold by means of the algorithm described by (12). The adaptive MBQW VAD defined by thresholds $\eta_0 = 2$ dB for $E_0 = 30$ dB and $\eta_1 = 1.4$ dB for $E_1 = 50$ dB enables working near the optimal point of the ROC curve for different SNR conditions ranging from 25 to 5 dB. On the other hand, it was found experimentally that using $K = 4$ subbands significantly increases the effectiveness of the proposed VAD. This fact is motivated by a shift up and to the left of the ROC curve when the number of subbands is increased.

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 [33]. 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. Influence of the VAD on an ASR System

Performance of ASR systems working over wireless networks and noisy environments normally decreases and nonefficient speech/nonspeech detection appears to be an important degradation source [1]. Although the discrimination analysis or the ROC curves are effective to evaluate a given algorithm, this section evaluates the VAD according to the goal for which it was developed by assessing the influence of the VAD over the performance of a speech recognition system.

The reference framework considered for these experiments was the ETSI AURORA project for DSR [36]. The recognizer is based on the HTK (Hidden Markov Model Toolkit) software package [37]. The task consists of recognizing connected digits which are modeled as whole word HMMs (Hidden Markov Models) with the following parameters: 16 states per word, simple left-to-right models, mixture of three Gaussians per state (diagonal covariance matrix) while speech pause models consist of three states with a mixture of six Gaussians per state. The 39-parameter feature vector consists of 12 cepstral coefficients (without the zero-order coefficient), the logarithmic frame energy plus the corresponding delta and acceleration coefficients. Two training modes are defined for the experiments conducted on the AURORA-2 database: *i*) training on clean data only (Clean Training), and *ii*) training on clean and noisy data (multicondition training). For the AURORA-3 SpeechDat-Car databases, the so called well-matched (WM), medium-mismatch (MM) and high-mismatch (HM) conditions are used. These databases contain recordings from the close-talking and distant microphones. In WM condition, both close-talking and hands-free microphones are used for training and testing. In MM condition, both training and testing are performed using the hands-free microphone recordings. In HM condition, training is done using close-talking microphone 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) that considers deletion, substitution and insertion errors.

TABLE II
AVERAGE WORD ACCURACY (%) FOR THE AURORA 2 FOR CLEAN AND MULTICONDITION TRAINING EXPERIMENTS RESULTS ARE AVERGED FOR ALL THE NOISES AND SNRS RANGING FROM 20 TO 0 db

	G.729	AMR1	AMR2	AFE	**MBQW**
WF	66.19	74.97	83.37	81.57	**84.12**
WF+FD	70.32	74.29	82.89	83.29	**86.09**
	Woo	Li	Marzinzik	Sohn	**Hand-labelled**
WF	83.64	77.43	84.02	83.89	84.69
WF+FD	81.09	82.11	85.23	83.80	86.86

An enhanced feature extraction scheme incorporating a noise reduction algorithm and nonspeech FD was built on the base system [36]. The noise reduction algorithm has been implemented as a single Wiener filtering stage as described in the AFE standard [35] but without mel-scale warping. No other mismatch reduction techniques already present in the AFE standard have been considered since they are not affected by the VAD decision and can mask the impact of the VAD precision on the overall system performance.

Table II 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 [35], which considers different VADs for WF and FD, was used. The proposed VAD outperforms the standard G.729, AMR1, AMR2 and AFE VADs in both clean and

TABLE III
AVERAGE WORD ACCURACY (%) FOR THE SDC DATABASES AND TASKS

		Base	Woo	Li	Marzinzik	Sohn	G.729	AMR1	AMR2	AFE	**MBQW**
Finnish	WM	92.74	86.81	85.60	93.73	93.84	88.62	94.57	95.52	94.25	94.70
	MM	80.51	66.62	55.63	76.47	80.10	67.99	81.60	79.55	82.42	80.08
	HM	40.53	62.54	58.34	68.37	75.34	65.80	77.14	80.21	56.89	83.67
	Average	**71.26**	71.99	66.52	79.52	83.09	74.14	84.44	85.09	77.85	**86.15**
Spanish	WM	92.94	95.35	91.82	94.29	96.07	88.62	94.65	95.67	95.28	96.79
	MM	83.31	89.30	77.45	89.81	91.64	72.84	80.59	90.91	90.23	91.85
	HM	51.55	83.64	78.52	79.43	84.03	65.50	62.41	85.77	77.53	87.25
	Average	**75.93**	89.43	82.60	87.84	90.58	75.65	74.33	90.78	87.68	**91.96**
German	WM	91.20	91.59	89.62	91.58	93.23	87.20	90.36	92.79	93.03	93.73
	MM	81.04	80.28	70.87	83.67	83.97	68.52	78.48	83.87	85.43	87.40
	HM	73.17	78.68	78.55	81.27	82.19	72.48	66.23	81.77	83.16	83.49
	Average	**81.80**	83.52	79.68	85.51	86.46	76.07	78.36	86.14	87.21	**88.21**
Average		**76.33**	81.65	76.27	84.29	86.71	75.29	79.04	87.34	84.25	**88.77**

multi condition training/testing experiments. When compared to recently reported VAD algorithms, the proposed one yields better results being the one that is closer to the "ideal" hand-labeled speech recognition performance.

Table III shows the recognition performance for the Finnish, Spanish, and German SDC databases for the different training/test mismatch conditions (HM, high mismatch, MM: medium mismatch and WM: well matched) when WF and FD are performed on the base system [36]. Again, the VAD outperforms all the algorithms used for reference, yielding relevant improvements in speech recognition. Note that the SDC databases used in the AURORA 3 experiments have longer nonspeech periods than the AURORA 2 database and then, the effectiveness of the VAD results more important for the speech recognition system. This fact can be clearly shown when comparing the performance of the proposed VAD to Marzinzik's VAD. The word accuracy of both VADs is quite similar for the AURORA 2 task. However, the proposed VAD yields a significant performance improvement over Marzinzik's VAD for the SDC databases.

Finally, in order to compare the proposed method to the best available results, the VADs of the full AFE standard [35] (including both the noise estimation and FD VADs) were replaced by the proposed MBQW VAD and the AURORA 3 experiments were conducted again. Table IV shows the recognition results in terms of the word error rates. A significant performance improvement which is consistently maintained for all the databases is observed. The improvements were particularly important in high mismatch experiments. Furthermore, the average word error rate is significantly reduced and yields an

TABLE IV
WORD ERROR RATES (%) FOR THE AURORA 3 DATABASES. RESULTS FOR THE FULL AFE AND THE MODIFIED AFE WITH THE PROPOSED VAD BEING USED FOR NOISE ESTIMATION AND FRAME DROPPING

	Finnish	Spanish	German	Danish	Average
AFE					
WM(x0.40)	3.96	3.39	4.87	6.02	4.56
MM(x0.35)	19.49	6.21	10.40	22.49	14.65
HM(x0.25)	14.77	9.23	8.70	20.39	13.27
Overall	12.10	5.84	7.76	15.38	**10.27**
AFE + MBQW					
WM(x0.40)	4.00	3.09	4.75	6.15	4.50
MM(x0.35)	16.97	6.94	10.92	20.43	13.82
HM(x0.25)	11.03	7.22	8.46	17.03	10.94
Overall	10.40	5.51	7.84	13.88	**9.38**

overall relative improvement of about 8.67%, when using the proposed MBQW VAD instead of the original AFE. These improvements are achieved by replacing the VADs of the AFE by the proposed one, without altering any of the remaining signal processing functions. Moreover, the significance level of these improvements makes the probability that the proposed MBQW VAD improve over AFE is over 99.999%.

As a conclusion, the performance of the VAD has a strong impact in an ASR system. If speech pauses are very long and dominant over speech periods, insertion errors are an important error source. On the other hand, if pauses are short, maintaining a high speech hit rate can be beneficial to reduce the number of deletion errors since the insertion errors are not significant in this context. The mismatch between training and test conditions also affects the influence of the VAD on the overall system performance and when the system suffers a high mismatch between training and test, an effective VAD can be more important for increasing the performance of speech recognizers. This fact is mainly motivated by the efficiency of the nonspeech FD stage and the efficient application of the noise reduction algorithms.

V. Conclusion

This paper presented a new VAD for improving speech detection robustness in noisy environments. The approach is based on an effective endpoint detection algorithm employing noise reduction techniques and order statistic filters for the formulation of the decision rule. The VAD performs an advanced detection of beginnings and delayed detection of word endings which, in part, avoids having to include additional hangover schemes. As a result, it leads to clear improvements in speech/nonspeech discrimination especially when the SNR drops. With this and other innovations, the proposed algorithm outperformed G.729, AMR and AFE standard VADs as well as recently reported approaches for endpoint detection. It also improved the recognition rate when it was considered as part of a complete speech recognition system. Moreover, when the proposed VAD replaced the AFE VADs, a significant reduction of the word error rate was obtained.

References

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

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

[3] *Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels*, 1999. ETSI, ETS1 EN 301 708 Recommend..

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

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

[6] F. Basbug, K. Swaminathan, and S. Nandkumar, "Noise reduction and echo cancellation front-end for speech codecs," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 1, pp. 1–13, Jan. 2004.

[7] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 5, pp. 245–256, 2002.

[8] 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.

[9] 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, Aug. 2001.

[10] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 498–505, Sep. 2003.

[11] L. Armani, M. Matassoni, M. Omologo, and P. Svaizer, "Use of a CSP-based voice activity detector for distant-talking ASR," in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, Sep. 2003, pp. 501–504.

[12] 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.

[13] 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.

[14] 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, May 2002.

[15] 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. 6, pp. 341–351, Nov. 2002.

[16] R. Chengalvarayan, "Robust energy normalization using speech/nonspeech discriminator for German connected digit recognition," in *Proc. EUROSPEECH 1999*, Budapest, Hungary, Sep. 1999, pp. 61–64.

[17] R. Tucker, "Voice activity detection using a periodicity measure," *Proc. Inst. Elect. Eng.*, vol. 139, no. 4, pp. 377–380, 1992.

[18] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the lpc residual domain," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 217–231, May 2001.

[19] 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.

[20] J. Ramírez, J. C. Segura, M. C. Benítez, A. de la Torre, and A. Rubio, "A new Kullback-Leibler VAD for speech recognition in noise," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 666–669, Feb. 2004.

[21] ——, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, vol. 42, no. 3–4, pp. 271–287, 2004.

[22] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAM1-8, pp. 679–698, 1986.

[23] H. Hwang and R. Haddad, "Multilevel nonlinear filters for edge detection and noise suppression," *IEEE Trans. Signal Process.*, vol. 42, no. 2, pp. 249–258, Feb. 1994.

[24] R. Öten and R. J. P. de Figueiredo, "An efficient method for L-filter design," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 193–203, Jan. 2003.

[25] A. Restrepo, G. Hincapié, and A. Parra, "On the detection of edges using order statistic filters," in *Proc. IEEE Int. Image Processing Conf.*, vol. 1, 1994, pp. 308–312.

[26] B. V. Cox and L. K. Tinothy, "Nonparametric rank-order statistics applied to robust voiced-unvoiced-silence classification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 5, pp. 550–561, May 1980.

[27] J. Arce, N. Gallagher, and T. Nodes, *Advances in Computer Vision and Image Processing*. New York: JA1, 1986, vol. 2, ch. Median filters: Theory and applications.

[28] S. Ko and Y. Lee, "Center weighted median filters and their applications to image enhancement," *IEEE Trans. Circuits Syst.*, vol. 38, no. 9, pp. 984–993, Sep. 1991.

[29] I. Pitas and A. V. Pitas, "Application of adaptive order statistic filters in digital image/image sequence filtering," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS)*, vol. 2, 1993, pp. 327–330.

[30] J. Astola and P. Kuosmanen, *Fundamentals of Nonlinear Digital Filtering*. Boca Raton, FL: CRC, 1997.

[31] 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.

[32] H. David and H. Nagaraja, *Order Statistics*. New York: Wiley, 2003.

[33] 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, 1997.

[34] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions," *ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium*, Sep. 2000.

[35] *Speech Processing, Transmission, and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms*, 2002. ETSI, ETSI ES 201 108 Recommend..

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

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

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

Since 2001, he has been an Assistant Professor with the Department of Electronics and Computer Technology, University of Granada. His research interests includes robust speech recognition, speech enhancement, voice activity detection, and design and implementation of high-performance digital signal processing systems. He is author of more than 60 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.

Since 1986, he has been with the Research Group on Signals, Networking and Communications (GSTC), Department of Electronics and Computer Technology of the University of Granada. Since January 2004, he has been the Coordinator of this research group. He developed his Ph.D. thesis on a variant of HMM modeling. He has been the director of three Ph.D. dissertations on topics related to speech recognition. From 1987 to 1993, he was an Assistant Professor, and since 1993, Associate Professor in the same department and has also taught several national and international courses. His research interests are in robust speech recognition, distributed speech recognition, and signal processing.

Dr. Segura is member of ISCA and AERFAI.

Ma Carmen Benítez (M'00) received the M.Sc and Ph.D. degrees in physics from the University of Granada, Granada, Spain, in 1988 and 1998, respectively.

Since 1990, she has been with the Department of Electrónica y Tecnología de Computadores, Faculty of Sciences, University of Granada, first as a Research and Assistant Professor and since 2003 as Associate Professor. From 2001 to 2002, she was a Visiting Researcher at the International Computer Science Institute, Berkeley CA. Her research interests include speech processing, with a specific goal of speech recognition, confidence measures, and robust parameterization for speech recognition.

Dr. Benítez is a member of ISCA.

Ángel de la Torre received the M.Sc and Ph.D. degrees in physics from the University of Granada, Granada, Spain, in 1994 and 1999, respectively.

Since 1994, he has been with the Research Group on Signals, Networking and Communications of the Department of Electronics and Computer Technology, University of Granada. In 2000, he joined the PAROLE Group at the Laboratoire RFIA du LORIA, Nancy, France, for a postdoctoral stay in the field of robust speech recognition, and the Institut für Angewandte Physik, Innsbruck, Austria, for a postdoctoral stay in the field of cochlear implants. Since 2003, he has been an Associate Professor with the University of Granada. His research interests are in the field of signal processing, and particularly robust speech recognition, speech processing in noise conditions, and signal processing for cochlear implants. He is reviewer for several scientific journals.

Antonio J. Rubio (SM'00) received the Master degree from the University of Sevilla, Spain, in physic sciences in 1972 and the Ph.D. degree from the Universidad de Granada, Spain, in 1978.

He has been the Director of the Research Group on Signals, Networking, and Communications, University of Granada, since its creation. He is Full Professor in the Department of Electrónica y Tecnologia de Computadores, Universidad de Granada, in the area of signal theory and communications. His investigation is centered in the field of signal processing and, in particular, in the field of automatic speech recognition, in which he has directed several research projects. He has been the director of ten Ph.D. dissertations on topics related to speech recognition. He is a reviewer for several international journals.