An effective cluster-based model for robust speech detection and speech recognition in noisy environments

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

C. G. Puntonet
Department of Computer Architecture and Technology, University of Granada, Spain

(Received 29 December 2005; revised 3 May 2006; accepted 5 May 2006)

This paper shows an accurate speech detection algorithm for improving the performance of speech recognition systems working in noisy environments. The proposed method is based on a hard decision clustering approach where a set of prototypes is used to characterize the noisy channel. Detecting the presence of speech is enabled by a decision rule formulated in terms of an averaged distance between the observation vector and a cluster-based noise model. The algorithm benefits from using contextual information, a strategy that considers not only a single speech frame but also a neighborhood of data in order to smooth the decision function and improve speech detection robustness. The proposed scheme exhibits reduced computational cost making it adequate for real time applications, i.e., automated speech recognition systems. An exhaustive analysis is conducted on the AURORA 2 and AURORA 3 databases in order to assess the performance of the algorithm and to compare it to existing standard voice activity detection (VAD) methods. The results show significant improvements in detection accuracy and speech recognition rate over standard VADs such as ITU-T G.729, ETSI GSM AMR, and ETSI AFE for distributed speech recognition and a representative set of recently reported VAD algorithms. © *2006 Acoustical Society of America.*
[DOI: 10.1121/1.2208450]

PACS number(s): 43.72.Ne, 43.72.Dv [EJS] Pages: 470–481

I. INTRODUCTION

The emerging wireless communication systems require increasing levels of performance and speech processing systems working in noise adverse environments. These systems often benefit from using voice activity detectors (VADs) which are frequently used in such application scenarios for different purposes. Speech/nonspeech detection is an unsolved problem in speech processing and affects numerous applications including robust speech recognition,[1,2] discontinuous transmission,[3,4] estimation and detection of speech signals,[5,6] real-time speech transmission on the Internet[7] or combined noise reduction and echo cancelation schemes in the context of telephony.[8] 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[9–13] and have evaluated the influence of the VAD effectiveness on the performance of speech processing systems.[14] Most of them have focused on the development of robust algorithms with special attention on the derivation and study of noise robust features and decision rules.[12,15–17] The different approaches include those based on energy thresholds,[15] pitch detection,[18] spectrum analysis,[17] zero-crossing rate,[4] periodicity measures[19] or combinations of different features.[3,4,20]

The speech/pause discrimination may be described as an unsupervised learning problem. Clustering is an appropriate solution for this case where the data set is divided into groups which are related "in some sense." Despite the simplicity of clustering algorithms, there is an increasing interest in the use of clustering methods in pattern recognition,[21] image processing[22] and information retrieval.[23,24] Clustering has a rich history in other disciplines[25,26] such as machine learning, biology, psychiatry, psychology, archaeology, geology, geography, and marketing. Cluster analysis, also called data segmentation has a variety of goals. All of these are related to grouping or segmenting a collection of objects into subsets or "clusters" such that those within each cluster are more closely related to one another than objects assigned to different clusters. Cluster analysis is also used to form descriptive statistics to ascertain whether or not the data consists of a set of distinct subgroups, each group representing objects with substantially different properties.

The paper is organized as follows. Section II introduces the necessary background information on clustering analysis. Section III shows the feature extraction process and a description of the proposed long term information *C*-means (LTCM) VAD algorithm is given in Sec. IV. Section V discusses some remarks about the proposed method. A complete experimental evaluation is conducted in Sec. VI in order to compare the proposed method with a representative set of VAD methods and to assess its performance for robust speech recognition applications. Finally, we state some conclusions and acknowledgments in the last part of the paper.

[a]URL: http://www.ugr.es/~gorriz; Electronic mail: gorriz@ugr.es

 0001-4966/2006/120(1)/470/12/$22.50

TABLE I. Hard C-means pseudocode.

(1). Initialize a C-partition randomly or based on some prior knowledge. Calculate the cluster prototype matrix $\mathbf{M}=[\mathbf{m}_1,\ldots,\mathbf{m}_C]$
(2) Assign each object in the data set to the nearest cluster P_i[a].
(3). Recalculate the cluster prototype matrix based on the current partition
(4). Repeat steps (2)–(3) until there is no change for each cluster.

[a]That is, $\mathbf{x}_j \in P_i$ if $\|\mathbf{x}_j-\mathbf{m}_i\|<\|\mathbf{x}_j-\mathbf{m}_{i'}\|$ for $j=1,\ldots,N, i\neq i'$, and $i'=1,\ldots,C$

II. HARD PARTITIONAL CLUSTERING BASIS

Partitional clustering algorithms partition data into certain number of clusters, in such a way that, patterns in the same cluster should be "similar" to each other unlike patterns in different clusters. Given a set of input patterns $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_j,\ldots,\mathbf{x}_N\}$, where $\mathbf{x}_j=(x_{j1},\ldots,x_{ji},\ldots,x_{jK})\in\mathbb{R}^K$ and each measure x_{jk} is said to be a feature, hard partitional clustering attempts to seek a C-partition of $\mathbf{X}$, $P=\{P_1,\ldots,P_C\}, C\leq N$, such that

(i) $P_i\neq\emptyset, i=1,\ldots,C$;
(ii) $\cup_{i=1}^{C}P_i=\mathbf{X}$;
(iii) $P_i\cap P_{i'}=\emptyset; i,i'=1,\ldots,C$ and $i\neq i'$.

The "similarity" measure is established in terms of a criterion function. The sum of squares error function is one of the most widely used criteria and is defined as

$$J(\mathbf{\Gamma},\mathbf{M})=\sum_{i=1}^{C}\sum_{j=1}^{N}\gamma_{ij}\|\mathbf{x}_j-\mathbf{m}_i\|^2, \tag{1}$$

where $\mathbf{\Gamma}=\gamma_{ij}$ is a partition matrix,

$$\gamma_{ij}=\begin{pmatrix}1 & \text{if } \mathbf{x}_j\in P_i\\ 0 & \text{otherwise}\end{pmatrix}$$

with $\Sigma_{i=1}^{C}\gamma_{ij}=1, \forall j, \mathbf{M}=[\mathbf{m}_1,\ldots,\mathbf{m}_C]$ is the cluster prototype or centroid (means) matrix with $\mathbf{m}_i=1/N_i\Sigma_{j=1}^{N}\gamma_{ij}\mathbf{x}_j$, the sample mean for the ith cluster and N_i the number of objects in the ith cluster. The optimal partition resulting of the minimization of the latter criterion can be found by enumerating all possibilities. It is unfeasible due to costly computation and heuristic algorithms have been developed for this optimization instead.

Hard C-means clustering is the best-known heuristic squared error-based clustering algorithm.[27] The number of cluster centers (prototypes) C is *a priori* known and the C-means iteratively moves the centers to minimize the total cluster variance. Given an initial set of centers the hard C-means algorithm alternates two steps:[28]

(i) for each cluster we identify the subset of training points (its cluster) that is closer to it than any other center;
(ii) the means of each feature for the data points in each cluster are computed, and this mean vector becomes the new center for that cluster.

In Table I we show a more detailed description of the C-means algorithm.

III. FEATURE EXTRACTION INCLUDING CONTEXTUAL INFORMATION

Let $x(n)$ be a discrete time signal. Denote by $\mathbf{y}_{n'}$ a frame containing the samples

$$\mathbf{y}_{n'}=\{x(i+n'\cdot D)\},\quad i=0,\ldots,L-1,\quad n=i+n'\cdot D, \tag{2}$$

where D is the window shift, L is the number of samples in each frame and n' selects a certain data window. Consider the set of $2\cdot m+1$ frames $\{\mathbf{y}_{l-m},\ldots,\mathbf{y}_l,\ldots,\mathbf{y}_{l+m}\}$ centered on frame $\mathbf{y}_l$, and denote by $Y(s,n')$, $n'=l-m,\ldots,l,\ldots,l+m$ its discrete Fourier transform (DFT), respectively,

$$Y_{n'}(\omega_s)\equiv Y(s,n')=\sum_{i=0}^{N_{\mathrm{FFT}}-1}x(i+n'\cdot D)\cdot\exp(-\mathbf{j}\cdot i\cdot\omega_s), \tag{3}$$

where $\omega_s=2_{\pi\cdot s}/N_{\mathrm{FFT}}$, $0\leq s\leq N_{\mathrm{FFT}}-1$, N_{FFT} is DFT resolution (if $N_{\mathrm{FFT}}>L$ then the DFT is padded with zeros) and $\mathbf{j}$ denotes the imaginary unit. The averaged energies for each n'th frame, $E(k,n')$, in K subbands ($k=1,2,\ldots,K$), are computed by means of

$$E(k,n')=\left(\frac{2K}{N_{\mathrm{FFT}}}\sum_{s=s_k}^{s_{k+1}-1}|Y(s,n')|^2\right)$$

$$s_k=\left\lfloor\frac{N_{\mathrm{FFT}}}{2K}(k-1)\right\rfloor,\quad k=1,2,\ldots,K, \tag{4}$$

where an equally spaced subband assignment is used and $\lfloor\cdot\rfloor$ denotes the "floor" function. Hence, the signal energy is averaged over K subbands obtaining a suitable representation of the input signal for VAD,[29] the observation vector at each frame n', defined as

$$\mathbf{E}(n')=(E(1,n'),\ldots,E(K,n'))^T\in\mathbb{R}^K. \tag{5}$$

The VAD decision rule is formulated over a sliding window consisting of $2m+l$ observation (feature) vectors around the frame for which the decision is being made (l), as we will show in the following sections. This strategy, known as "long term information,"[30] provides very good results using several approaches for VAD, however it imposes an m-frame delay on the algorithm that, for several applications including robust speech recognition, is not a serious implementation obstacle.

In the following section we show the way we apply **C**-means to modeling the noise subspace and to find a soft decision rule for VAD.

IV. HARD C-MEANS FOR VAD

In the LTCM VAD algorithm, the clustering method described in Sec. II is applied to a set of initial pause frames in order to characterize the noise subspace, that is, the generic feature vector described in Sec. II is defined in terms of energy observation vectors as we show in the following: each observation vector in Eq. (5) is uniquely labeled, by the integer $j\in\{1,\ldots,N\}$, and uniquely assigned (hard decision-

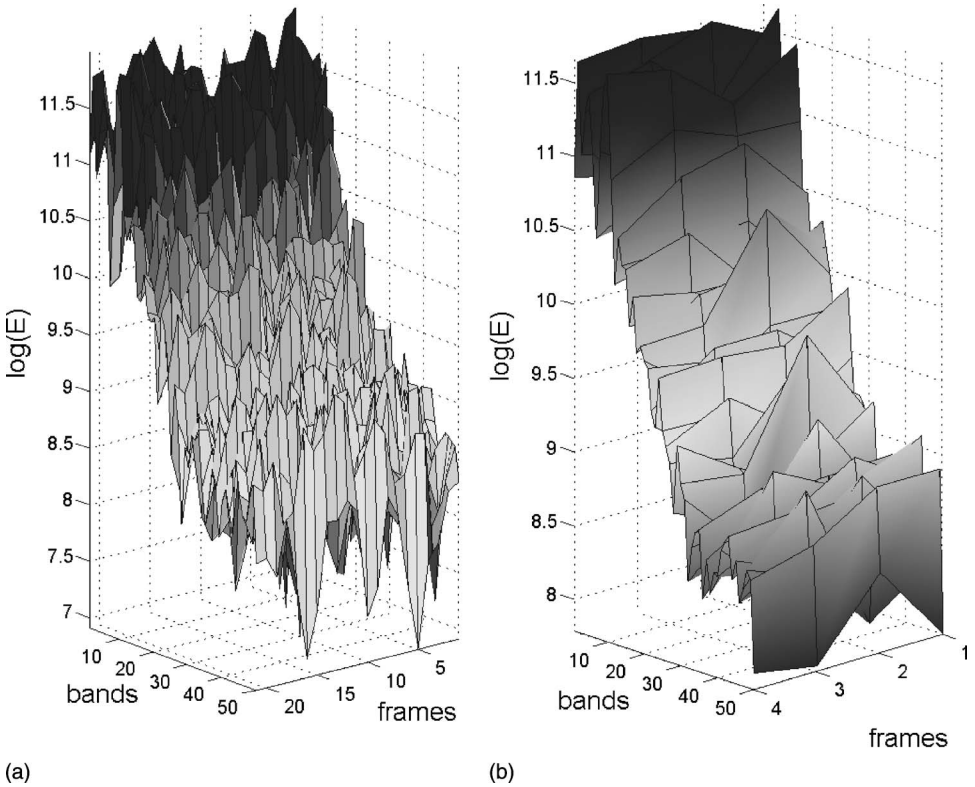

FIG. 1. (a) 20 noise log-energy frames, computed using $N_{\mathrm{FFT}}=256$ and averaged over 50 subbands. (b) Clustering approach to the latter set of frames using hard decision C-means ($C=4$ prototypes).

based clustering) to a prespecified number of prototypes $C<N$, labeled by an integer $i\in\{1,\ldots,C\}$. Thus, we are selecting the generic feature vector as $\mathbf{x}_j\equiv\mathbf{E}_j$.

The similarity measure to be minimized in terms of energy vectors is based on the squared Euclidean distance:

$$d(\mathbf{E}_j,\mathbf{E}_{j'})=\sum_{k=1}^{K}(E(k,j)-E(k,j'))^2=\|\mathbf{E}_j-\mathbf{E}_{j'}\|^2 \tag{6}$$

and can be equivalently defined as[28]

$$J(\mathcal{C})=\tfrac{1}{2}\sum_{i=1}^{C}\sum_{\mathcal{C}(j)=i}\sum_{\mathcal{C}(j')=i}\|\mathbf{E}_j-\mathbf{E}'_j\|^2=\tfrac{1}{2}\sum_{i=1}^{C}\sum_{\mathcal{C}(j)=i}\|\mathbf{E}_j-\mathbf{E}_i\|^2, \tag{7}$$

where $\mathcal{C}(j)=i$ denotes a many-to-one mapping, that assigns the jth observation to the ith prototype and

$$\bar{\mathbf{E}}_i=(\bar{E}(1,i),\ldots,\bar{E}(K,i))^T=\mathrm{mean}(\mathbf{E}_j),$$
$$\forall j,\quad \mathcal{C}(j)=i,\quad i=1,\ldots,C \tag{8}$$

is the mean vector associated with the ith prototype (the sample mean for the ith prototype $\mathbf{m}_i$ defined in Sec. II). Thus, the loss function is minimized by assigning N observations to C prototypes in such a way that within each prototype the average dissimilarity of the observations is minimized. Once convergence is reached, N K-dimensional pause frames are efficiently modeled by C K-dimensional noise prototype vectors denoted by $\bar{\mathbf{E}}_i^{\mathrm{opt}}$, $\quad i=1,\ldots,C$. We call this set of clusters C-partition or noise prototypes since, in this work, the word cluster is assigned to different classes of *labeled data*, that is $\mathbf{K}$ is fixed to 2, i.e., we define two clusters: "noise" and "speech" and the cluster "noise" consists of C prototypes. In Fig. 1 we observed how the complex nature of noise can be simplified (smoothed) using a this clustering approach. The clustering approach speeds the decision function in a significant way since the dimension of feature vectors is reduced substantially ($N\rightarrow C$).

Soft decision function for VAD

In order to classify the second data class (energy vectors of speech frames) we use a basic sequential algorithm scheme, related to Kohonen's leaning vector quantization (LVQ),[31] using a multiple observation (MO) window centered at frame l, as shown in Sec. II. For this purpose let us consider the same dissimilarity measure, a threshold of dissimilarity γ and the maximum clusters allowed $\mathbf{K}=2$.

Let $\hat{\mathbf{E}}(l)$ be the decision feature vector at frame l that is defined on the MO window as follows:

$$\hat{\mathbf{E}}(l)=\max\{\mathbf{E}(j)\},\quad j=l-m,\ldots,l+m. \tag{9}$$

The selection of this envelope feature vector, describing not only a single instantaneous frame but also a $(2m+1)$ entire neighborhood, is useful as it detects the presence of voice beforehand (pause-speech transition) and holds the detection flag, smoothing the VAD decision (as a hangover based algorithm in speech-pause transition[16,17]), as shown in Fig. 2.

Finally, the presence of the second "cluster" (speech frame) is detected if the following ratio holds:

$$\eta(l)=\log\left(1/K\sum_{k=1}^{K}\frac{\hat{E}(k,l)}{<\bar{\mathbf{E}}_i>}\right)>\gamma, \tag{10}$$

where $<\bar{\mathbf{E}}_i>=1/C\Sigma_{i=1}^{C}\bar{\mathbf{E}}_i=1/C\Sigma_{i=1}^{C}\Sigma_{j=1}^{N}\gamma_{ij}\mathbf{E}_j$ is the averaged noise prototype center and γ is the decision threshold.

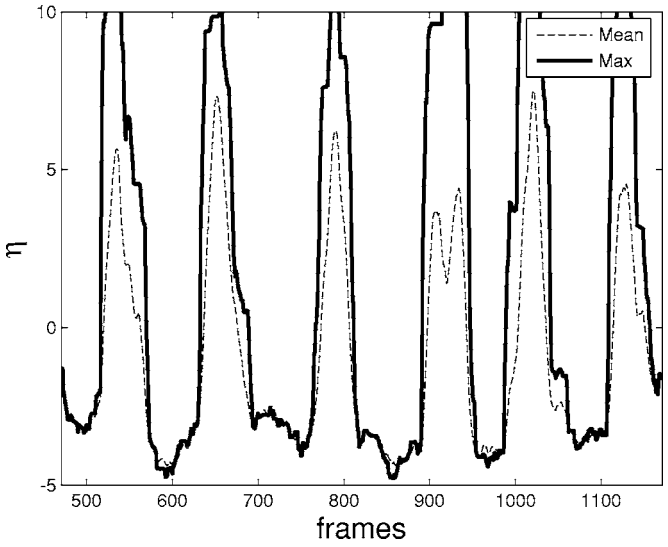

FIG. 2. Decision function in Eq. (10) for two different criteria: energy envelope [Eq. (9)] and energy average.

In order to adapt the operation of the proposed VAD to nonstationary and noise environments, the set of noise prototypes are updated according to the VAD decision during nonspeech periods [not satisfying Eq. (10)] in a competitive manner (only the closer noise prototype is moved towards the current feature vector):

$$\bar{\mathbf{E}}_{i'} = \arg_{\min}(\|\bar{\mathbf{E}}_i - \hat{\mathbf{E}}(l)\|^2) \quad i = 1, \ldots, C$$

$$\Rightarrow \bar{\mathbf{E}}_{i'}^{\text{new}} = \alpha \cdot \bar{\mathbf{E}}_{i'}^{\text{old}} + (1-\alpha) \cdot \hat{\mathbf{E}}(l), \tag{11}$$

where α is a normalized constant. Its value is close to one for a soft decision function (i.e., we selected in simulation $\alpha = 0.99$), that is, uncorrected classified speech frames contributing to the false alarm rate will not affect the noise space model significantly.

V. SOME REMARKS ON THE LTCM VAD ALGORITHM

The main advantage of the proposed algorithm is its ability to deal with on line applications such as DSR systems. The above-mentioned scheme is optimum in computational cost. First, we apply a batch hard C-means to a set of initial pause frames once, obtaining a fair description of the noise subspace and then, using Eq. (11), we move the nearest prototype to the previously detected as silence current frame. Any other on-line approach would be possible but it would be necessary to update the entire set of prototypes for each detected pause frame. In addition, the proposed VAD algorithm belongs to the class of VADs which model noise and apply a distance criterion to detect the presence of speech, i.e., Ref. 17.

A. Selection of an adaptive threshold

In speech recognition experiments (Sec. VI), the selection of the threshold is based on the results obtained in detection experiments [working points in receiving operating curves (ROC) for all conditions]. The working point (selected threshold) should correspond with the best tradeoff between the hit rate and false alarm rate, then the threshold is adaptively chosen depending on the noisy condition.

The VAD makes the speech/nonspeech detection by comparing the unbiased LTCM VAD decision to an adaptive threshold,[32] that is the detection threshold is adapted to the observed noise energy E. It is assumed that the system will work under different noisy conditions characterized by the energy of the background noise. Optimal thresholds (working points) γ_0 and γ_1 can be determined for the system working in the cleanest and noisiest conditions. These thresholds define a linear VAD calibration curve that is used during the initialization period for selecting an adequate threshold as a function of the noise energy E:

$$\gamma = \begin{cases} \gamma_0, & E \le E_0, \\ \dfrac{\gamma_0 - \gamma_1}{E_0 - E_1} + \gamma_0 - \dfrac{\gamma_0 - \gamma_1}{1 - E_1/E_2}, & E_0 < E < E_1, \\ \gamma_1, & E \ge E_1, \end{cases} \tag{12}$$

where E_0 and E_1 are the energies of the background noise for the cleanest and noisiest conditions that can be determined examining the speech databases being used. A high speech/nonspeech discrimination is ensured with this model since silence detection is improved at high and medium SNR levels while maintaining high precision detecting speech periods under high noise conditions.

The algorithm described so far is presented as pseudocode in the following:

(1) Initialize noise model:
  (a) Select N feature vectors $\{\mathbf{E}_j\}$, $j = 1, \ldots, N$.
  (b) Compute threshold γ.

(2) Apply C-means clustering to feature vectors, extracting C noise prototype centers

$$\{\bar{\mathbf{E}}_i\}, \quad i = 1, \ldots, C$$

(3) for l=init to end
  (a) Compute $\hat{\mathbf{E}}(l)$ over the MO window
  (b) if $\eta(l) > \gamma$ [Eq. (10)] than VAD = 1 else VAD = 0 and update noise prototype centers $\{\bar{\mathbf{E}}_i\}$, $i = 1, \ldots, C$ [Eq. (11)].

B. Decision variable distributions

In this section we study the distributions of the decision variable as a function of the long-term window length (m) in order to clarify the motivations for the algorithm proposed. A hand-labeled version of the Spanish SpeechDat-Car (SDC) (Ref. 33) database was used in the analysis. This database contains recordings from close-talking and distant microphones at different driving conditions: (a) stopped car, motor running, (b) town traffic, low speed, rough road, and (c) high speed, good road. The most unfavorable noise environment (i.e., high speed, good road) was selected and recordings from the distant microphone were considered. Thus, the m-order divergence measure between speech and silences was measured during speech and nonspeech periods, and the histogram and probability distributions were built. The 8 kHz input signal was decomposed into overlapping frames

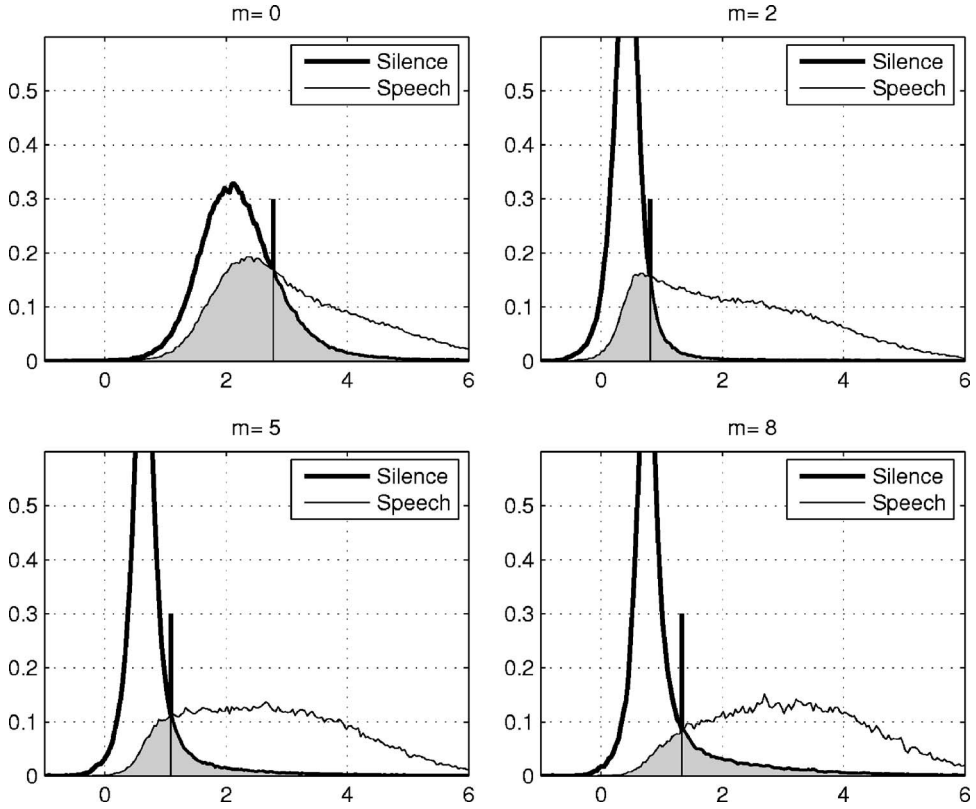

FIG. 3. Speech/nonSpeech distributions and error probabilities of the optimum Bayes classifier for $m=0,2,5$, and 8.

with a 10 ms window shift. Figure 3 shows the distributions of speech and noise for $m=0,2,5$, and 8. It is derived from this that speech and noise distributions are better separated when increasing the order of the long-term window. The noise is highly confined and exhibits a reduced variance, thus leading to high nonspeech hit rates. This fact can be corroborated by calculating the classification error of speech and noise for an optimal Bayes classifier. Figure 4 shows the misclassification errors as a function of the window length m. The speech classification error is approximately divided by three from 32% to 10% when the order of the VAD is increased from 0 to 8 frames. This is motivated by the separation of the distributions that takes place when m is increased as shown in Fig. 3. On the other hand, the increased speech detection robustness is only prejudiced by a moderate increase in the speech detection error. According to Fig. 4, the optimal value of the order of the VAD would be $m=8$. This analysis corroborates the fact that using long-term speech features[32] results beneficial for VAD since they reduce misclassification errors substantially.

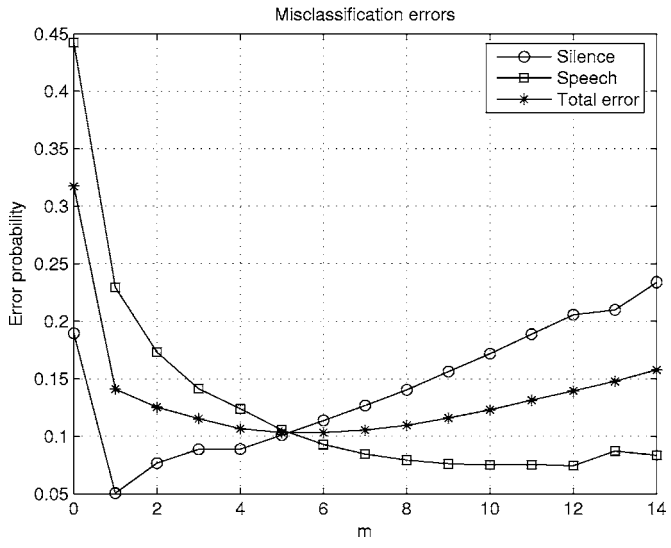

FIG. 4. Probability of error as a function of m.

VI. EXPERIMENTAL RESULTS

Several experiments are commonly carried out in order to assess the performance of VAD algorithms. The analysis is normally focused on the determination of the error probabilities in different noise scenarios and SNR values,[17,34] and the influence of the VAD decision on speech processing systems.[1,14] The experimental framework and the objective performance tests conducted to evaluate the proposed algorithm are described in this section.

A VAD achieves silence compression in modern mobile telecommunication systems reducing the average bit rate by using the discontinuous transmission (DTX) mode. The International Telecommunication Union (ITU) adopted a toll-quality speech coding algorithm known as G.729 to work in combination with a VAD module in DTX mode.[4] The ETSI AMR (Adaptive Multi-Rate) speech coder[3] developed by the Special Mobile Group (SMG) for the GSM system specifies two options for the VAD to be used within the digital cellular telecommunications system. In option 1, the signal is passed through a filterbank and the level of signal in each band is calculated. A measure of the SNR is used to make the VAD decision together with the output of a pitch detector, a tone detector and the correlated complex signal analysis module. An enhanced version of the original VAD is the AMR option

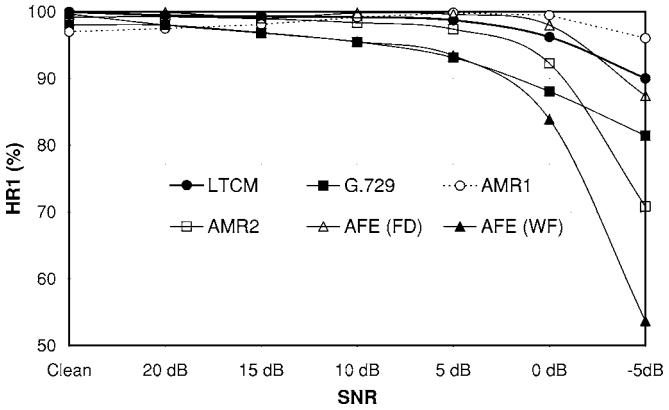

FIG. 5. Speech hit rates (HR1) of standard VADs as a function of the SNR for the AURORA 2 database.

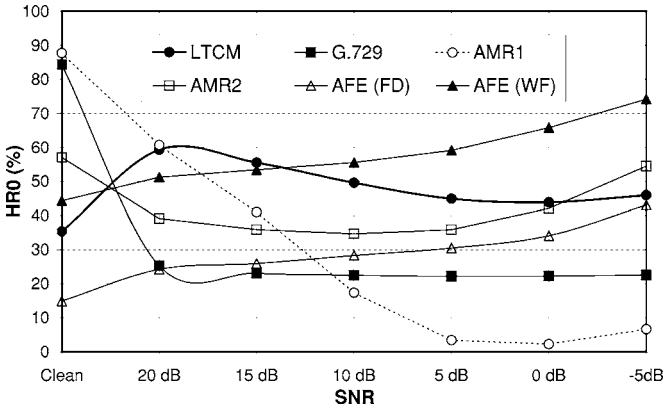

FIG. 7. Nonspeech hit rates (HR0) of standard VADs as a function of the SNR for the AURORA 2 database.

2 VAD which uses parameters of the speech encoder being more robust against environmental noise than AMR1 and G.729. Recently, a new standard incorporating noise suppression methods has been approved by the ETSI for feature extraction and distributed speech recognition (DSR). The so-called advanced front-end (AFE) (Ref. 36) incorporates an energy-based VAD (WF AFE VAD) for estimating the noise spectrum in Wiener filtering speech enhancement, and a different VAD for nonspeech frame dropping (FD AFE VAD).

Recently reported VADs are based on the selection of discriminative speech features, noise estimation and classification methods. Sohn *et al.* showed a decision rule derived from the generalized likelihood ratio test by assuming that the noise statistics are known *a priori*.[12] An interesting approach is the endpoint detection algorithm proposed by Li,[16] which uses optimal FIR filters for edge detection. Other methods track the power spectrum envelope of the signal[17] or use energy thresholds for discriminating between speech and noise.[15]

A. Evaluation under different noise environments

First, the proposed VAD was evaluated in terms of the ability to discriminate between speech and nonspeech in different noise scenarios and at different SNR levels. The AURORA 2 database[35] is an adequate database for this analysis since it is built on the clean Tldigits database that consists of sequences of up to seven connected digits spoken by American English talkers as source speech, and a selection of eight different real-world noises that have been artificially added to the speech at SNRs of 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, and −5 dB. These noisy signals have been recorded at different places (suburban train, crowd of people (babble), car, exhibition hall, restaurant, street, airport, and train station), and were selected to represent the most probable application scenarios for telecommunication terminals. In the discrimination analysis, the clean Tldigits database was used to manually label each utterance as speech or nonspeech on a frame by frame basis for reference. Detection performance is then assessed in terms of the speech pause hit-rate (HR0) and the speech hit-rate (HR1) defined as the fraction of all actual pause or speech frames that are correctly detected as pause or speech frames, respectively,

$$HR1 = \frac{N_{1,1}}{N_1^{\text{ref}}}, \quad HR0 = \frac{N_{0,0}}{N_0^{\text{ref}}}, \tag{13}$$

where N_1^{ref} and N_0^{ref} are the number of real nonspeech and speech frames in the whole database and $N_{1,1}$ and $N_{0,0}$ are the number of real speech and nonspeech frames correctly classified, respectively.

Figures 5–8 provide comparative results of this analysis and compare the proposed VAD to standardized algorithms including the ITU-T G.729,[4] ETSI AMR,[3] and ETSI AFE

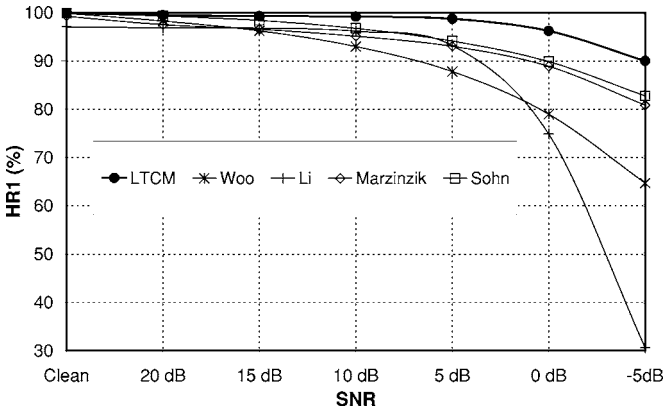

FIG. 6. Speech hit rates (HR1) of other VADs as a function of the SNR for the AURORA 2 database.

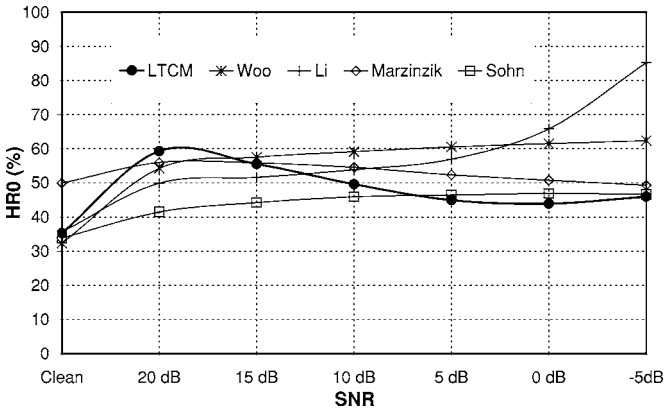

FIG. 8. Nonspeech hit rates (HR0) of other VADs as a function of the SNR for the AURORA 2 database.

TABLE II. Average speech/nonpeech hit rates for SNRs between clean conditions and −5 dB. Comparison to (a) standardized VADs and (b) other VAD methods.

			(a)			
	G.729	AMR1	AMR2	AFE (WF)	AFE (FD)	**LTCM**
HR0 (%)	31.77	31.31	42.77	57.68	28.74	**47.81**
HR1(%)	93.00	98.18	93.76	88.72	97.70	**97.57**
			(b)			
		Sohn	Woo	Li	Marzinzik	**LTCM**
HR0 (%)		43.66	55.40	57.03	52.69	**47.81**
HR1 (%)		94.46	88.41	83.65	93.04	**97.57**

(Ref. 36) in terms of the nonspeech hit-rate (HR0, Fig. 7) and speech hit-rate (HR1, Fig. 5) for clean conditions and SNR levels ranging from 20 to −5 dB. Note that results for the two VADs defined in the AFE DSR standard[36] for estimating the noise spectrum in the Wiener filtering (WF) stage and nonspeech frame dropping (FD) are provided. The results shown in these figures are averaged values for the entire set of noises.

It can be derived from Figures 7 and 5 that (i) ITU-T G.729 VAD suffers poor speech detection accuracy with the increasing noise level while nonspeech detection is good in clean conditions (85%) and poor (20%) in noisy conditions, (ii) ETSI AMR1 yields 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 degrades its nonspeech detection accuracy being HR0 less than 10% below 10 dB, making it less useful in a practical speech processing system, (iii) ETSI AMR2 leads to considerable improvements over G.729 and AMR1 yielding better nonspeech detection accuracy while still suffering fast degradation of the speech detection ability at unfavorable noisy conditions, (iv) The VAD used in the AFE standard for estimating the noise spectrum in the Wiener filtering stage is based in 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 frame dropping achieves a high accuracy in speech detection but moderate results in nonspeech detection, and (v) LTCM yields the best compromise among the different VADs tested. It obtains a good behavior in detecting nonspeech periods as well as exhibiting a slow decay in performance at unfavorable noise conditions in speech detection (90% at −5 dB).

Figures 6 and 8 compare the proposed VAD to a representative set of recently published VAD method.[12,15–17] It is worthwhile clarifying that the AURORA 2 database consists of recordings with very short nonspeech periods between digits and, consequently, it is more important to classify speech correctly than nonspeech in a speech recognition system. This is the reason to define a VAD method with a high speech hit rate even in very noisy conditions. Table II summarizes the advantages provided by LTCM VAD over the different VAD methods in terms of the average speech/nonspeech hit rates (over the entire range of SNR values). Thus, the proposed method with a 97.57% mean HR1 and a 47.81% mean HR0 yields the best trade-off in speech/nonspeech detection.

B. Receiver operating characteristic (ROC) curves

An additional test was conducted to compare speech detection performance by means of the ROC curves, a frequently used methodology in communications based on the hit and error detection probabilities,[17,29,37] that completely describes the VAD error rate. The AURORA subset of the Spanish SDC database[33] was used in this analysis. This database contains 4914 recordings using close-talking and distant microphones from more than 160 speakers. As in the whole SDC database, the files are categorized into three noisy conditions: quiet, low noise, and high noise conditions, which represent different driving conditions and average SNR values of 12 dB, 9 dB, and 5 dB. Thus, recordings from the close-talking microphone are used in the analysis to label speech/pause frames for reference, while recordings from the distant microphone are used for the evaluation of different VADs in terms of their ROC curves. The speech pause hit rate (HR0) and the false alarm rate (FAR0=100 -HR1) were determined in each noise condition for the proposed VAD and the G.729, AMR1, AMR2, and AFE VADs, which were used as a reference. For the calculation of the false-alarm rate as well as the hit rate, the "real" speech frames and "real" speech pauses were determined using the hand-labeled database on the close-talking microphone.

The sensitivity of the proposed method to the number of clusters used to model the noise space was studied. It was found experimentally that the behavior of the algorithm is almost independent of C, using a number of subbands $K=10$. Figure 9 shows that the accuracy of the algorithm (noise detection rate versus false alarm rate) in speech-pause discrimination is not affected by the number of prototypes selected as long as $C \geq 2$, thus the benefits of the clustering approach are evident. Note that the objective of the VAD is to work as close as possible to the upper left corner in this figure where speech and silence is classified with no errors. The effect of the number of subbands used in the algorithm is plotted in Fig. 10. The use of a complete energy average

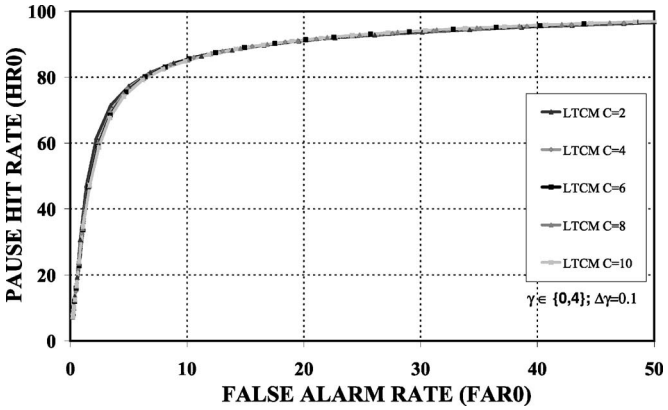

FIG. 9. ROC curves in high noisy conditions for different number of noise prototypes. The DFT was computed with $N_{\mathrm{FFT}}=256$, $K=10$ log-energy subbands were used to build features vectors and the MO-window contained $2\cdot m+1$ frames ($m=10$).

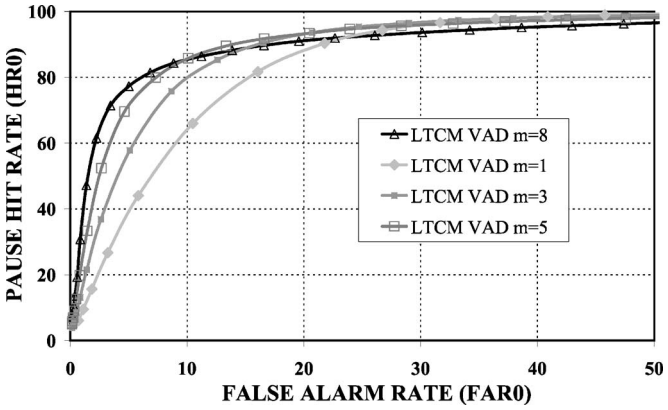

FIG. 11. Selection of the number of m (high, high speed, good road, 5 dB average SNR, $K=32$, $C=2$).

($K=1$) or raw data ($K=100$) reduces the effectiveness of the clustering procedure making its accuracy equivalent to other proposed VADs.

Figure 11 shows the speech pause hit rate (HR0) as a function of the false alarm rate (FAR0=100-HR1) of the proposed LTCM VAD for different values of the decision threshold and different values of the number of observations m. It is shown how increasing the number of observations (m) leads to better speech/nonspeech discrimination with a shift-up and to the left of the ROC curve in the ROC space. This enables the VAD to work closer to the "ideal" working point (HR0=100%, FAR0=0%) where both speech and nonspeech are classified ideally with no errors. These results are consistent with our preliminary experiments and the results shown in Figs. 3 and 4 that expected a minimum error rate for m close to eight frames.

Figure 12 shows the ROC curves of the proposed VAD and other reference VAD algorithms[12,15–17] for recordings from the distant microphone in high noisy conditions. The working points of the ITU-T G.729, ETSI AMR, and ETSI AFE VADs are also included. The results show improvements in detection accuracy over standardized VADs and over a representative set of VAD algorithms.[12,15–17] 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 ITU-T G.729,[4] which is used along with a speech codec for discontinuous transmission, and over the[16] algorithm, that is based on an optimum linear filter for edge detection. The proposed VAD also improves Marzinzik[17] VAD that tracks the power spectral envelopes, and the Sohn[12] VAD, that formulates the decision rule by means of a statistical likelihood ratio test (LRT) defined on the power spectrum of the noisy signal.

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.[38] 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 occurring when part of the word is lost after frame dropping.

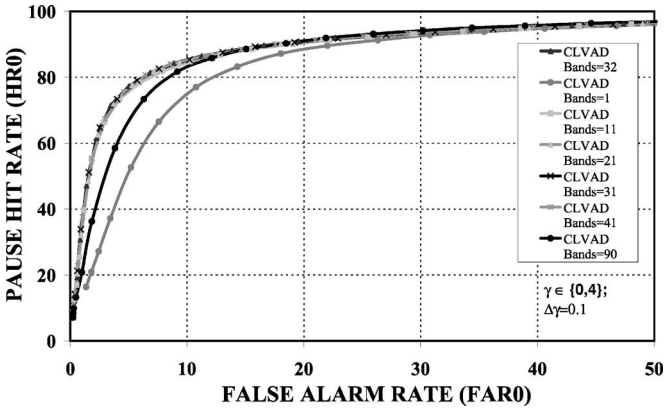

FIG. 10. ROC curves in high noisy conditions for different number of subbands. $N_{\mathrm{FFT}}=256$; $C=10$ and $m=10$.

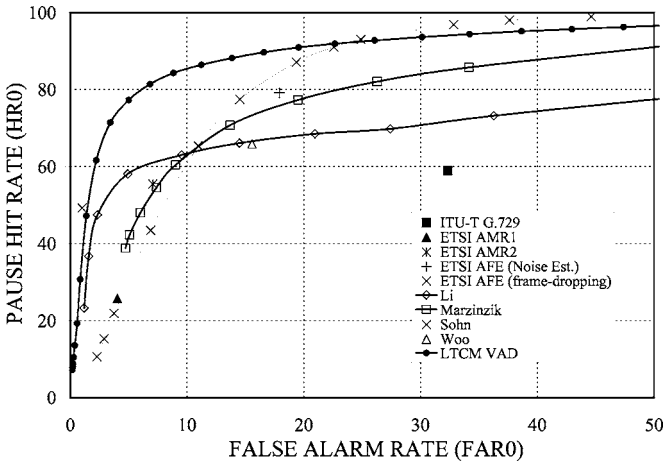

FIG. 12. ROC curves for comparison to standardized and other VAD methods (high, high speed, good road, 5 dB average SNR, $K=32$, $C=2$).

C. Assessment of the VAD on an ASR system

Although the discrimination analysis or the ROC analysis presented in the preceding section are effective to evaluate 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[15] since nonefficient speech/nonspeech discrimination is an important performance degradation source for speech recognition systems working in noisy environments.[1] There are two clear motivations for that: (i) noise parameters such as its spectrum are updated during nonspeech periods being the speech enhancement system strongly influenced by the quality of the noise estimation, and (ii) 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, thus causing irrecoverable deletion errors.

The reference framework (Base) is the distributed speech recognition (DSR) front-end[39] proposed by the ETSI STQ working group for the evaluation of noise robust DSR feature extraction algorithms. The recognition system is based on the HTK (Hidden Markov Model Toolkit) software package.[40] The task consists in 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 3 Gaussians per state and only the variances of all acoustic coefficients (no full 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 cepstral coefficient), the logarithmic frame energy plus the corresponding derivatives (Δ) and acceleration ($\Delta\Delta$) coefficients.

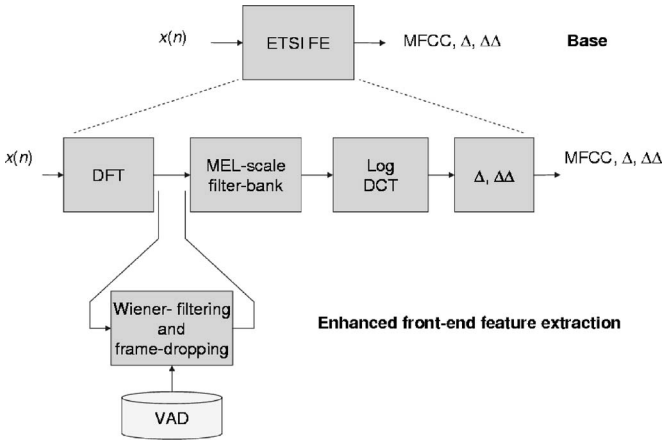

FIG. 13. Speech recognition experiments. Front-end feature extraction.

Two training modes are defined for the experiments conducted on the AURORA 2 database: (i) training on clean data only (Clean Training), and (ii) training on clean and noisy data (Multi-Condition Training). For the AURORA 3 SpeechDat-Car databases, the so-called well-matched (WM), medium-mismatch (MM) and high-mismatch (HM) conditions are used. AURORA 3 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-

TABLE III. Average word accuracy for the AUR0RA 2 database. (a) Clean training. (b) Multicondition training.

(a)

		Base + WF					Base + WF + FD				
	Base	G.729	AMR1	AMR2	AFE	**LTCM**	G.729	AMR1	AMR2	AFE	**LTCM**
Clean	99.03	98.81	98.80	98.81	98.77	98.88	98.41	97.87	98.63	98.78	99.18
20 dB	94.19	87.70	97.09	97.23	97.68	97.46	83.46	96.83	96.72	97.82	98.05
15 dB	85.41	75.23	92.05	94.61	95.19	95.14	71.76	92.03	93.76	95.28	96.10
10 dB	66.19	59.01	74.24	87.50	87.29	88.71	59.05	71.65	86.36	88.67	90.71
5 dB	39.28	40.30	44.29	71.01	66.05	72.48	43.52	40.66	70.97	71.55	75.82
0 dB	17.38	23.43	23.82	41.28	30.31	42.91	27.63	23.88	44.58	41.78	47.01
−5 dB	8.65	13.05	12.09	13.65	4.97	15.34	14.94	14.05	18.87	16.23	19.88
Average	60.49	57.13	66.30	78.33	75.30	79.34	57.08	65.01	78.48	79.02	**81.54**

(b)

		Base + WF					Base + WF + FD				
	Base	G.729	AMRl	AMR2	AFE	**LTCM**	G.729	AMRl	AMR2	AFE	**LTCM**
Clean	98.48	98.16	98.30	98.51	97.86	98.45	97.50	96.67	98.12	98.39	98.78
20 dB	97.39	93.96	97.04	97.86	97.60	97.93	96.05	96.90	97.57	97.98	98.41
15 dB	96.34	89.51	95.18	96.97	96.56	97.06	94.82	95.52	96.58	96.94	97.61
10 dB	93.88	81.69	91.90	94.43	93.98	94.64	91.23	91.76	93.80	93.63	95.39
5 dB	85.70	68.44	80.77	87.27	86.41	87.54	81.14	80.24	85.72	85.32	88.40
0 dB	59.02	42.58	53.29	65.45	64.63	66.23	54.50	53.36	62.81	63.89	66.92
−5 dB	24.47	18.54	23.47	30.31	28.78	31.21	23.73	23.29	27.92	30.80	32.91
Average	86.47	75.24	83.64	88.40	87.84	**88.68**	83.55	83.56	87.29	87.55	**89.35**

TABLE IV. Average word acccuracy for clean and multicondition AURORA 2 training/testing experiments. Comparison to (a) standard VADs and (b) recently presented VAD methods.

	(a)					
	G.729	AMR1	AMR2	AFE	**LTCM**	Hand-labeling
Base + WF	66.19	74.97	83.37	81.57	**84.01**	84.69
Base + WF+FD	70.32	74.29	82.89	83.29	**85.44**	86.86
	(b)					
	Woo	Li	Marzinzik	Sohn	**LTCM**	Hand-labeling
Base + WF	83.64	77.43	84.02	83.89	**84.01**	84.69
Base + WF+ FD	81.09	82.11	85.23	83.80	**85.44**	86.86

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),

$$\mathrm{WAcc}(\%) = \frac{N - D - S - I}{N} \times 100\,\%\,, \tag{14}$$

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

The influence of the VAD decision on the performance of different feature extraction schemes was studied. The first approach (shown in Fig. 13) incorporates Wiener filtering (WF) to the Base system as noise suppression method. The second feature extraction algorithm that was evaluated uses Wiener filtering and nonspeech frame dropping. The algorithm has been implemented as described for the first stage of the Wiener filtering noise reduction system present in the advanced front-end AFE DSR standard.[36] The same feature extraction scheme was used for training and testing and no other mismatch reduction techniques already present in the AFE standard (wave form processing or blind equalization) have been 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, and LTCM VAD algorithms. These results were averaged 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 used in the standard[36] 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. FD has impact on the training of silence models since less nonspeech frames are available for training. However, if FD is effective enough, few nonspeech periods will be handled by the recognizer in testing and consequently, the silence models will have little influence on the speech recognition performance. As a conclusion, the proposed VAD 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. 5, thus leading to numerous recognition errors and the corresponding increase of the word error rate, as shown in Table III. 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 proposed LTCM 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, removing efficiently 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 as shown in Table III, part (a).

Table IV, part (a), compares the word accuracies aver-

TABLE V. Average word accuracy (%) for the Spanish, SDC database.

	Base	Woo	Li	Marzinzik	Sohn	G729	AMRl	AMR2	AFE	**LTCM**
WM	92.94	95.35	91.82	94.29	96.07	88.62	94.65	95.67	95.28	96.41
MM	83.31	89.30	77.45	89.81	91.64	72.84	80.59	90.91	90.23	91.61
HM	51.55	83.64	78.52	79.43	84.03	65.50	62.41	85.77	77.53	86.20
Avg.	**75.93**	89.43	82.60	87.84	90.58	75.65	74.33	90.78	87.68	**91.41**

aged 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. These results show that the performance of the proposed algorithm is very close to that of the reference database. In all the test sets, the proposed VAD algorithm outperforms standard VADs obtaining the best results followed by AFE, AMR2, AMR1, and G.729. Table IV, part (b), extends this comparison to other recently presented VAD methods.[12,15–17]

Table V shows the recognition performance for the Spanish SpeechDat-Car database when WF and FD are performed on the base system.[39] Again, the VAD outperforms all the algorithms used for reference yielding relevant improvements in speech recognition. Note that, these particular databases used in the AURORA 3 experiments have longer nonspeech periods than the AURORA 2 database and then, the effectiveness of the VAD results more important for the speech recognition system. This fact can be clearly shown when comparing the performance of the proposed VAD to Marzinzik[17] VAD. The word accuracies of both VADs are quite similar for the AURORA 2 task. However, the proposed VAD yields a significant performance improvement over Marzinzik[17] VAD for the AURORA 3 database.

VII. CONCLUSION

A new algorithm for improving speech detection and speech recognition robustness in noisy environments is shown. The proposed LTCM VAD is based on noise modeling using hard *C*-means clustering and employs long-term speech information for the formulation of a soft decision rule based on an averaged energy ratio. The VAD performs an advanced detection of beginnings and delayed detection of word endings which, in part, avoids having to include additional hangover schemes or noise reduction blocks. It was found that increasing the length of the long-term window yields to a reduction of the class distributions and leads to a significant reduction of the classification error. An exhaustive analysis conducted on the AURORA database showed the effectiveness of this approach. The proposed LTCM VAD outperformed recently reported VAD methods including Sohn's VAD, that defines a likelihood ratio test on a single observation, and the standardized ITU-T G.729, ETSI AMR for the GSM system and ETSI AFE VADs for distributed speech recognition. On the other hand, it also improved the recognition rate when the VAD is used for noise spectrum estimation, noise reduction and frame dropping in a noise robust ASR system.

ACKNOWLEDGMENTS

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

[1] L. Karray and A. Martin, "Towards improving speech detection robustness for speech recognition in adverse environments," Speech Commun. **43**, 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," Proceedings of EUROSPEECH 2003, Geneva, Switzerland, 2003, pp. 3041–3044.
[3] ETSI, "Voice activity detector (VAD) tor Adaptive Multi-Rate (AMR) speech traffic channels," ETSI EN 301 708 Recommendation, 1999.
[4] ITU, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," ITU-T/Recommendation G.729-Annex B, 1996.
[5] L. Krasny, "Soft-decision speech signal estimation," J. Acoust. Soc. Am. **108**, 2575 (2000).
[6] P. S. Veneklassen and J. P. Christoff, "Speech detection in noise," J. Acoust. Soc. Am. **32**, 1502 (1960).
[7] 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," IEEE International Conference on High-Speed Networks and Multimedia Communications, 2002, pp. 46–50.
[8] F. Basbug, K. Swaminathan, and S. Nandkumar, "Noise reduction and echo cancellation front-end for speech codecs," IEEE Trans. Speech Audio Process. **11**, 1–13 (2003).
[9] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," IEEE Signal Process. Lett. **8**, 276–278 (2001).
[10] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," IEEE Trans. Speech Audio Process. **11**, 498–505 (2003).
[11] L. Armani, M. Matassoni, M. Omologo, and P. Svaizer, "Use of a CSP-based voice activity detector for distant-talking ASR," Proceedings of EUROSPEECH 2003, Geneva, Switzerland, 2003, pp. 501–504.
[12] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," IEEE Signal Process. Lett. **16**, 1–3 (1999).
[13] I. Potamitis and E. Fishier, "Speech activity detection and enhancement of a moving speaker based on the wideband generalized likelihood ratio and microphone arrays," J. Acoust. Soc. Am. **116**, 2406–2415 (2004).
[14] R. L. Bouquin-Jeannes and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," Speech Commun. **16**, 245–254 (1995).
[15] K. Woo, T. Yang, K. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," Electron. Lett. **36**, 180–181 (2000).
[16] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," IEEE Trans. Speech Audio Process. **10**, 146–157 (2002).
[17] M. Marzinzik, and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," IEEE Trans. Speech Audio Process. **10**, 341–351 (2002).
[18] R. Chengalvarayan, "Robust energy normalization using speech/nonspeech discriminator for German connected digit recognition," Proceedings of EUROSPEECH 1999, Budapest, Hungary, 1999, pp. 61–64.
[19] R. Tucker, "Voice activity detection using a periodicity measure," IEE Proc.-Commun. **139**, 377–380 (1992).
[20] S. G. Tanyer and H. Özer, "Voice activity detection in nonstationary noise," IEEE Trans. Speech Audio Process. **8**, 478–482 (2000).
[21] M. R. Anderberg, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *Cluster Analysis for Applications* (Academic, New York, 1973).
[22] A. Jain and P. Flynn, "Image segmentation using clustering," in *In Advances in Image Understanding. A Festschrift for Azriel Rosenfeld*, edited by N. Ahuja and K. Bowyer, (IEEE, 1996), pp. 65–83.
[23] E. Rasmussen, "Clustering algorithms," in *Information Retrieval: Data Structures and Algorithms*, edited by W. B. Frakes and R. Baeza-Yates (Prentice-Hall, Upper Saddle River, NJ, 1992), pp. 419–442.
[24] G. Salton, "Developments in automatic text retrieval," Science **109**, 974–980 (1991).
[25] A. Jain and R. Dubes, *Algorithms for Clustering Data*, Prentice-Hall advanced reference series (Prentice-Hall, Upper Saddle River, NJ, 1988).
[26] D. Fisher, "Knowledge acquisition via incremental conceptual clustering," Mach. Learn. **2**, 139–172 (1987).
[27] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability (University of California Press, Berkeley, 1967).

[28]T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction Series*, Springer Series in Statistics, 1st ed. (Springer, New York, 2001).
[29]J. Ramírez, J. C. Segura, C. Benítez, A. de la Torre, and A. Rubio, "An effective subband OSF-based VAD with noise reduction for robust speech recognition," IEEE Trans. Speech Audio Process. **13**, 1119–1129 (2005).
[30]J. M. Górriz, J. Ramírez, J. C. Segura, and C. G. Puntonet, "Improved MO-LRT VAD based on bispectra Gaussian model," Electron. Lett. **41**, 877–879 (2005).
[31]T. Kohonen, *Self Organizing and Associative Memory*, 3rd ed. (Springer-Verlag, Berlin, 1989).
[32]J. Ramírez, J. C. Segura, M. C. Benítez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," Speech Commun. **42**, 271–287 (2004).
[33]A. Moreno, L. Borge, D. Christoph, R. Gael, C. Khalid, E. Stephan, and A. Jeffrey, "SpeechDat-Car: A Large Speech Database for Automotive Environments," Proceedings of the II LRCE Conference, 2000.
[34]F. Beritelli, S. Casale, G. Rugeri, and S. Serrano, "Performance evaluation and comparison of G.729/AMR/Fuzzy voice activity detectors," IEEE Signal Process. Lett. **9**, 85–88 (2002).
[35]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, Paris, France, 2000.
[36]ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," ETSI ES 202 050 Recommendation, 2002.
[37]J. M. Górriz, J. Ramírez, C. G. Puntonet, and J. C. Segura, "Generalized LRT-based Voice Activity Detector," IEEE Signal Process. Lett. (to be published).
[38]A. Benyassine, E. Shlomot, H. Su, D. Massaloux, C. Lamblin, and J. Petit, "ITU-T Recommendation G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," IEEE Commun. Mag. **35**, 64–73 (1997).
[39]ETSI, "Speech processing, transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms," ETSI ES 201 108 Recommendation, 2000.
[40]S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book* (Cambridge University Press, Cambridge, 1997).