Hard C-means clustering for voice activity detection

J.M. Górriz a,*, J. Ramírez a, E.W. Lang b, C.G. Puntonet c

a Department of Signal Theory, Networking and Communications, University of Granada, 18071 Granada, Spain
b Institute of Biophysics, University of Regensburg, 93040 Regensburg, Germany
c Department of Computer Architecture and Technology, University of Granada, 18071 Granada, Spain

Received 27 February 2006; received in revised form 19 July 2006; accepted 20 July 2006

Abstract

An effective voice activity detection (VAD) algorithm is proposed for improving speech recognition performance in noisy environments. The proposed speech/pause discrimination method is based on a hard-decision clustering approach built on a set of subband log-energies and noise prototypes that define a cluster. Detecting the presence of speech (a new cluster) is achieved using a basic sequential algorithm scheme (BSAS) according to a given “distance” (in this case, geometrical distance) and a suitable threshold. The accuracy of the Cluster VAD (ClVAD) algorithm lies in the use of a decision function defined over a multiple-observation (MO) window of averaged subband log-energies and a suitable noise subspace model defined in terms of prototypes. In addition, the reduced computational cost of the clustering approach makes it adequate for real-time applications, i.e. speech recognition. An exhaustive analysis is conducted on the Spanish SpeechDat-Car databases in order to assess the performance of the proposed method and to compare it to existing standard VAD methods. The results show improvements in detection accuracy over standard VADs such as ITU-T G.729, ETSI GSM AMR and ETSI AFE and a representative set of recently reported VAD algorithms for noise robust speech processing.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Voice activity detection; Speech recognition; Clustering; C-means; Prototypes; Subband energy

1. Introduction

The emerging wireless communication systems are demanding increasing levels of performance of 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/non-speech detection is an unsolved problem in speech processing and affects numerous applications including robust speech recognition (Karray and Martin, 2003; Ramírez et al., 2003), discontinuous transmission (ETSI, 1999; ITU, 1996), real-time speech transmission on the Internet (Sangwan et al., 2002) or combined noise reduction and echo cancelation schemes in the context of telephony (Basbug et al., 2003). The speech/non-speech 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 (Sohn et al., 1999; Cho and Kondoz, 2001; Gazor and Zhang, 2003; Armani et al., 2003) and have evaluated the influence of the VAD effectiveness on the performance of speech processing systems (Bouquin-Jeannes and Faucon, 1995). 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 (Woo et al., 2000; Li et al., 2002; Marzinik and Kollmeier, 2002; Sohn et al., 1999). The different approaches include those based on energy thresholds (Woo et al., 2000), pitch detection (Chengalvarayan, 1999), spectrum analysis (Marzinik and Kollmeier, 2002), zero-crossing rate (ITU, 1996), periodicity measures (Tucke, 1992) or combinations of different features (Tanyer and Özer, 2000; ITU, 1996; ETSI, 1999).

The speech/pause discrimination can 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 (Anderberg, 1973), image processing (Jain and Flynn, 1996) and information retrieval (Rasmussen, 1992; Salton, 1991). Clustering has a rich history in other disciplines (Jain and Dubes, 1988; Fisher, 1987) such as machine learning, biology, psychiatry, psychology, archaeology, geology, geography, and marketing. Cluster analysis, also called data segmentation, has a variety of goals. All 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 consist of a set of distinct subgroups, each group representing objects with substantially different properties.

The essay is organized as follows: in Section 2 we describe a suitable signal model to detect the presence of speech in noisy environments. Section 3 considers cluster analysis to form “descriptive statistics” transforming the noise sample set into a soft-noise model consisting of low-dimensional features. A description of the algorithm is given in Section 4 and a complete experimental framework is shown in Section 5. Finally, we state some conclusions and acknowledgements in the last part of the paper.

2. Segmentation and feature extraction

Let \( x(n) \) be a discrete time signal. Consider a frame of signal containing the elements:

\[
X_j = \{ x(i + j \cdot D) \}; \quad i = 1, \ldots, L, \tag{1}
\]

where \( D \) is the window shift (~10 ms), \( L \) is the number of samples in each frame and \( j \) selects a certain data window. Consider the set of \( 2 \cdot m + 1 \) frames \( \{ x^{l-m}, \ldots, x^l, \ldots, x^{l+m} \} \) centered on frame \( x^l \), namely multiple observation (MO) window, and denote by \( X(s,j), j = l - m, \ldots, l, \ldots, l + m \) its Discrete Fourier Transform (DFT) resp.:

\[
X_j(\omega_s) = X(s,j) = \sum_{n=0}^{N_{FFT}-1} x(n + j \cdot D) \cdot \exp(-j \cdot n \cdot \omega_s), \tag{2}
\]

where \( \omega_s = \frac{2\pi s}{N_{FFT}}, 0 \leq s \leq N_{FFT} - 1 \), \( N_{FFT} \) is the FFT resolution (if \( N_{FFT} \geq L \) then the DFT is padded with zeros) and \( j \) denotes the imaginary unit. The log-energies for the \( j \)th frame in the MO window, \( E(k,j) \), in \( K \) subbands \( (k = 0, 1, \ldots, K - 1) \), are computed by means of

\[
E(k,j) = \log \left( \frac{2K}{N_{FFT}} \sum_{s=0}^{N_{FFT} - 1} |X(s,j)|^2 \right), \tag{3}
\]

\[
s_k = \left[ \frac{N_{FFT}}{2K} k \right], \quad k = 0, 1, \ldots, K - 1
\]

where an equally spaced subband assignment is used and \( [\cdot] \) denotes the “floor” function. Hence, the signal log-energy is averaged over \( K \) subbands obtaining a suitable representation of the input signal for VAD (Ramirez et al., in press), the observation vector at frame \( j \):

\[
E_j = (E(0,j), \ldots, E(K - 1, j))^T \tag{4}
\]

The VAD decision rule is formulated over a sliding window consisting of \( 2m + 1 \) observation vectors around the frame \( l \) for which the decision is being made, as we will show in the following sections. This strategy, known as “long term information”, provides very good results using several approaches for VAD such as Górriz et al. (2005).
3. Hard partitional C-means clustering applied to VAD

Hard C-means clustering is a method for finding clusters and cluster centers in a set of unlabeled data (MacQueen, 1967). 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 (Hastie et al., 2001):

- for each cluster we identify the subset of training points (its cluster) that is closer to it than any other center;
- 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.

Given a set of input vectors defined in Eq. (4),
\[
E = \{E_{1}, \ldots, E_{l}, \ldots, E_{l+m}\}
\]
where \( E_{j} = (E(0,j), \ldots, E(k,j), \ldots, E(K-1,j)) \in \mathbb{R}^{K} \) and each measurement \( E(k,j) \) is said to be a feature, hard partitional Clustering attempts to seek a \( C \)-partition of \( E \),
\[
P = \{P_{1}, \ldots, P_{C}\}, C \leq N, \text{such that:}
\]

- \( P_{i} \neq \emptyset, i = 1, \ldots, C; \)
- \( \bigcup_{i=1}^{C} P_{i} = E; \)
- \( P_{i} \cap P_{l} = \emptyset; i, i' = 1, \ldots, C \text{ and } i \neq i'. \)

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 of a criterion function. The sum of squares error of a criterion function is reached when the average dissimilarity becomes the new center for that cluster.

\[
J(\Gamma, M) = \sum_{j=1}^{l} \sum_{j=1}^{l+\ldots} \gamma_{ij} \|E_{j} - m_{i}\|^{2}, \quad (5)
\]

where \( \Gamma = \gamma_{ij} \) is a partition matrix, \( \gamma_{ij} = \begin{cases} 1 & \text{if } E_{j} \in P_{i} \\ 0 & \text{otherwise} \end{cases} \) with \( \sum_{i=1}^{C} \gamma_{ij} = 1, \forall j \), \( M = [m_{1}, \ldots, m_{C}] \) is the cluster prototype or centroid (means) matrix with \( m_{i} = 1/N_{i} \sum_{j=1}^{N} \gamma_{ij} E_{j} \), the sample mean for the \( i \)th cluster and \( N_{i} \) the number of objects in the \( i \)th 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 (MacQueen, 1967).

In the following sections, we show the way we apply this procedure to modelling the noise subspace and to find a soft decision rule for VAD.

3.1. Noise subspace modelling

The above mentioned procedure is applied to a set of initial pause frames (log-energies) obtaining a set of clusters that characterizes the noise subspace. We call this set of clusters, individuals or “noise prototypes”.\(^1\) Let reorder indexes of each observation vector of the set \( E \) by the integer \( j \in \{1, \ldots, N\} \), and uniquely assign to a prespecified number of prototypes \( C < N \) labeled by an integer \( i \in \{1, \ldots, C\} \). The dissimilarity measure between observation vectors is the squared Euclidean distance:

\[
d(E_{j}, E_{j'}) = \sum_{k=0}^{K-1} (E(k,j) - E(k,j'))^{2} = \|E_{j} - E_{j'}\|^{2}; \quad j, j' \in \{1, \ldots, N\} \quad (6)
\]

and the loss function to be minimized is defined as in Eq. (5). 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, \( N \times K \)-dimensional pause feature vectors are efficiently modelled by \( C \times K \)-dimensional noise prototype centers \( (m_{i}) \) denoted by

\[
M = [m_{1}, \ldots, m_{C}]. \quad (7)
\]

In Fig. 1 we observed how the complex nature of noise can be simplified (smoothed) using a clustering approach.

In order to adapt the noise model to a non-stationary environment we propose a competitive rule which adjusts noise prototypes centers to the current detected pause frame, that is, only the closest noise prototype \( m_{c} \) is moved towards the current feature vector:

\[
m_{c} = \arg \min_{i} (\|m_{i} - E_{l}\|^{2}) \Rightarrow m_{c}^{\text{new}} = \|x \cdot m_{i}^{\text{old}} + (1 - x) \cdot E_{l}, \quad (8)
\]

where \( \alpha \) is a normalized constant. Its value is close to one for a soft decision function (i.e. we selected in simulation \( x = 0.99 \)), that is, uncorrected

\(^1\) The word cluster is usually assigned to different classes of labeled data, that is, in this case, the number of clusters \( K \) is fixed to 2, corresponding to noise and speech frames.
classified speech frames contributing to the false alarm rate, do not affect the noise space model significantly.

3.2. Multiple-observation based decision function for VAD

Once a suitable model for noise has been built, we need to describe how to detect speech. In order to classify the speech data class we use an on-line basic sequential algorithm scheme (BSAS)\(^2\) using a multiple-observation (MO) window centered at frame \(l\), which is defined in Section 2. For this purpose let consider the same dissimilarity measure, i.e. the Euclidean distance, a threshold of dissimilarity or detection \(\gamma\) and a maximum number of classes \(K = 2\) for labeled data (the class “noise”, which contains \(C\) noise prototypes, and the class “speech”). The condition for the new class (speech frame detection) is satisfied if the decision function is greater than the detection threshold:

\[
\eta(l) \equiv \|\hat{E}(l) - \langle M \rangle\|^2 > \gamma,
\]

where \(\langle M \rangle\) is the averaged noise prototype center, \(\gamma\) is the decision threshold and \(\hat{E}(l)\) is the energy envelope, that is defined on the MO window as follows:

\[
\hat{E}(l) = \max\{E_j\}, \quad j = l - m, \dots, l + m,
\]

where the function “\(\max\)” is applied component wise to the set of feature vectors in the MO window. The selection of this MO feature vector, describing the actual frame, 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 (Marzinzik and Kollmeier, 2002; Li et al., 2002)), as shown in Fig. 2. The use of a MO window imposes an \(m\)-frame delay on the algorithm that, for several applications including robust speech recognition, is not a serious implementation obstacle.

The motivations for the proposed model are the following: (i) the clustering approach speeds the decision function in a significant way since the size of the noise model is reduced substantially \((N \rightarrow C)\); (ii) it also filters the noise feature vectors obtaining a smoothed representation of the log-energies (see Fig. 1) as in Li’s algorithm (Li et al., 2002), which uses optimal FIR filters applied to averaged energies for edge detection. The latter

---

\(^2\) In a way, this process is related to Kohonen’s Learning Vector Quantization (LVQ) (Kohonen, 1989).
approach requires more computational effort in designing a high-pass filter and suffers of performance degradation when it works in high noise conditions (only a single energy average is used). The filter design and filtering stages are replaced by a clustering approach, that is performed in the initialization period exclusively, and an adaptation rule for this model; and (iii) the decision function of the proposed method includes long-term information which increases the accuracy of speech detection systems (Gorriz et al., 2005). With these and other innovations, our clustering approach is more robust, in the working area of speech recognition, than other recently reported VADs as we will see in Section 5.

4. Some remarks on the algorithm

The main feature of the proposed algorithm is its ability to deal with on-line applications such as Distributed Speech Recognition (DSR) systems. The Hard C-means described in the previous sections is applied to the noise space once and then, it is updated using Eq. (8) during pause frames, if Eq. (9) is not satisfied. In speech recognition experiments (Section 5) 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 noise condition, i.e. quiet, low and high condition. This way, in order to normalize the decision function, the left-hand side in Eq. (9) should be divided by the noise level.

4.1. Selection of threshold

The VAD makes the speech/non-speech detection by comparing the unbiased ClVAD decision to an adaptive threshold (Ramirez et al., 2004). The detection threshold is adapted to the observed noise log-energy $E$. It is assumed that the system will work at different noisy conditions characterized by the log-energy of the background noise. Optimal thresholds (working points) $\gamma_0$ and $\gamma_1$ 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 log-energy $E$:

$$\gamma = \begin{cases} \gamma_0; & E \leq E_0 \\ \frac{\gamma_0 + \gamma_1}{2} - \frac{\gamma_0 - \gamma_1}{E - E_0}; & E_0 < E < E_1 \\ \gamma_1; & E \geq E_1 \end{cases},$$

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

The algorithm described so far is presented as pseudo-code in the following:

1. Initialize Noise Model:
   - Select $N$ feature vectors $\{E_j\}, j = 1, \ldots, N$.
   - Compute threshold $\gamma$.
2. Apply C-means clustering to feature vectors extracting $C$ noise prototype centers $\{m_i\}, i = 1, \ldots, C$.
3. for $l = \text{init}$ to end
   - (a) compute $\hat{E}(l)$ over the MO window
   - (b) if $d(\hat{E}(l), \text{mean}(M)) > \gamma$ then $\text{VAD} = 1$
   - else Update noise prototype centers $m(i)$ with Eq. (8).

4.2. Example

Fig. 3 shows the operation of the proposed algorithm on an utterance of the Spanish SpeechDat-Car database (Moreno et al., 2000). It includes: (i) the noise log-energy model (top-left), (ii) the clustering C-means approach (top-right), (iii) the current...
Fig. 3. Step of the algorithm. The frame selected is classified as speech (VAD = 1) as it is shown in the decision function bottom-right. Top-left: Noise log-energy subbands. Top-right: prototypes centers obtained at this step using Hard C-means. Bottom-left: comparison between noise prototypes and the MO feature vector of the current frame plotted in the middle. Bottom-right: decision function and threshold versus frames.

Fig. 4. VAD operation. Top: decision function and constant threshold versus frames. Bottom: input signal and VAD decision versus time. frame log-energy (frame = 3) plotted together with the the noise prototypes (C = 4) (bottom-left) and (iv) the decision rule versus time (bottom-right). In Fig. 4 we show the evaluation of the decision
function of the proposed VAD on the same utterance of the Spanish SpeechDat-Car (SDC) database (Moreno et al., 2000). The phonetic transcription is: ‘tres, ‘nweﷺ, ‘heiro, ‘ʃejo, ‘dos, ‘uno, ‘oʃjo, ‘sejṣ, ‘kwatro. We also show the binary VAD decision and the selected threshold in the CIVAD operation for the same phrase.

5. Experimental framework

Several experiments are commonly conducted to evaluate the performance of CIVAD algorithm. The analysis is normally focused on the determination of misclassification errors at different SNR levels (Marzinzik and Kollmeier, 2002), and the influence of the VAD decision on speech processing systems (Bouquin-Jeannes and Faucon, 1995; Karray and Martin, 2003). The experimental framework and the objective performance tests conducted to evaluate the proposed algorithm are described in this section.

5.1. Speech/non-speech discrimination analysis

In this section, the ROC curves are used for the evaluation of the proposed VAD. These plots describe completely the VAD error rate and show the trade-off between the speech and non-speech error probabilities as the threshold γ varies. The Spanish SpeechDat-Car database was used in the analysis. This database contains recordings in a car environment from close-talking and hands-free microphones. Utterances from the close-talking device with an average SNR of about 25 dB were labeled as speech or non-speech for reference while the VAD was evaluated on the hands-free microphone. As in the whole SDC database, the files are categorized into three noisy conditions: quiet, low and highly noisy conditions, which represent different driving conditions and average SNR values of 12 dB, 9 dB and 5 dB. The speech and non-speech hit rates (HR1, HR0) were determined in each noise condition as a function of the decision threshold γ for each of the VAD tested. They are defined as the fraction of all actual pause or speech frames that are correctly detected as pause or speech frames, respectively:

\[
HR_1 = \frac{N_{1|1}}{N_{1^{\text{ref}}}}, \quad HR_0 = \frac{N_{0|0}}{N_{0^{\text{ref}}}}.
\]

where \(N_{1|1}^{\text{ref}}\) and \(N_{0|0}^{\text{ref}}\) are the number of real non-speech and speech frames in the whole database and \(N_{1|1}\) and \(N_{0|0}\) are the number of real speech and non-speech frames correctly classified, respectively. The false alarm rate (FAR) in each state is defined as \(\text{FAR}\{0,1\} = 1 - HR\{1,0\}\).

5.1.1. Parameter selection

The sensitivity of the proposed method to the number of prototypes using Aurora 3 database in the most unfavorable condition (high-speed, good road) with a 5 dB average SNR was studied. It was found experimentally that the behavior of the algorithm is almost independent of C. Fig. 5 shows

---

**Fig. 5.** ROC curves in high noisy conditions for different number of noise prototypes. The DFT was computed with \(N_{\text{FFT}} = 256\), \(K = 10\) log-energy subbands were used to build features vectors and the MO-window contained \(2 \cdot m + 1\) frames \((m = 10)\).
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 computational savings 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.

Fig. 6 shows the effect of incorporating “long term speech information” into the decision rule of the VAD operation. A high value of \( m \) increases the accuracy and robustness of the algorithm (that is, small variations in \( \gamma \) do not affect the performance of the VAD). This is motivated by a shift up and to the left of the ROC curve which enables working with improved speech and non-speech hit rates. The effect of number of subbands used in

![ROC curves in high noisy conditions for different number of frames](image1)

![ROC curves in high noisy conditions for different number of subbands.](image2)

Fig. 6. ROC curves in high noisy conditions for different number of frames in MO window (m). The DFT was computed with \( N_{\text{FFT}} = 256; K = 10 \) log-energy subbands were used to build features vectors and the number of noise prototypes was \( C = 8 \).

Fig. 7. ROC curves in high noisy conditions for different number of subbands. The DFT was computed with \( N_{\text{FFT}} = 256; C = 10 \) prototypes and a MO window of \( m = 10 \).
the algorithm is plotted in Fig. 7. The use of the full band energy average ($K = 1$) or raw data ($K = 100$) reduces the effectiveness of the clustering procedure making its accuracy equivalently to other recently proposed VADs, i.e. (Sohn et al., 1999).

5.1.2. Comparison with referenced VADs

Fig. 8 shows the ROC curves in the most unfavorable condition (high-speed, good road) with a 5 dB average SNR. It was shown that increasing the number of observation vectors $m$ improves the performance of the proposed ClVAD. The best results are obtained for $m = 10$ while increasing the number of observations over this value reports no additional improvements. In the working area, the proposed VAD outperforms the Sohn’s VAD (Sohn et al., 1999), which assumes a single observation likelihood ratio test (LRT) in the decision rule together with an Hidden Markov Model (HMM)-based hangover mechanism, as well as standardized VADs such as G.729 and AMR (ITU, 1996; ETSI, 1999). It also improves recently reported methods (Sohn et al., 1999; Li et al., 2002; Woo et al., 2000 Marzinzik and Kollmeier, 2002). It is interesting mentioning the connection of the proposed method with Li’s algorithm which uses FIR filters for edge detection. The latter method uses a pre-filtering stage to the decision function which is based on a single averaged energy. In our method we replace this single feature by a set of noise prototypes which tracks the noise non-stationarity, making pre-filtering or hangover mechanisms unnecessary and speeding the evaluation of the decision function. Of course they could be included into our method but the computational efficiency would be damaged. It can be also derived from the complete analysis (quiet, low and high noise conditions) that:

- The working point of the G.729 VAD shifts to the right in the ROC space with decreasing SNR, while the proposed algorithm is less affected by the increasing level of background noise.
- AMR1 VAD works on a low false alarm rate point of the ROC space but it exhibits poor non-speech hit rate.
- AMR2 VAD yields clear advantages over G.729 and AMR1 exhibiting important reduction in the false alarm rate when compared to G.729 and increase in the non-speech hit rate over AMR1.
- Wiener Filtering (WF) AFE VAD yields good non-speech detection accuracy but works on a high false alarm rate point on the ROC space. It suffers rapid performance degradation when the driving conditions get noisier. On the other hand, frame dropping (FD) AFE VAD has been planned to be conservative since it is only used in the DSR standard for frame-dropping. Thus, it exhibits poor non-speech detection accuracy working on a low false alarm rate point of the ROC space.
- ClVAD yields the lowest false alarm rate for a fixed non-speech hit rate and also, the highest non-speech hit rate for a given false alarm rate in the working area displayed in Fig. 8.

![Fig. 8. ROC curves of proposed ClVAD in high noisy conditions for $m = 10$, $K = 10$ and $C = 8$ and comparison to standard and recently reported VADs.](image-url)
of the adaptive ClVAD to tune the detection threshold by means the algorithm described in Section 4 enables working on the optimal point of the ROC curve for different noisy conditions. Thus, the algorithm automatically selects the appropriate decision value for a given noisy condition in a similar way as it is carried out in the AMR (option 1) standard.

Thus, the proposed VAD works with improved speech/non-speech hit rates when compared to the most relevant algorithms to date. Although the discrimination analysis or the ROC curves are effective to evaluate a given algorithm, the influence of the VAD in a speech recognition system was also studied.

5.2. Influence of the VAD on a speech recognition system

The reference framework considered for these experiments was the ETSI AURORA project for DSR (ETSI, 2000). The recognizer is based on the HTK (Hidden Markov Model Toolkit) software package (Young et al., 1997). The task consists of recognizing connected digits which are modelled 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. 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.

An enhanced feature extraction scheme incorporating a noise reduction algorithm and non-speech FD was built on the base system (ETSI, 2000). The noise reduction algorithm has been implemented as a single Wiener filtering stage as described in the AFE standard (ETSI, 2002) 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 1 shows the recognition performance for the Spanish SDC database 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 (ETSI, 2000). 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 non-speech 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 Marzinik’s VAD (Marzinik and Kollmeier, 2002). The word accuracy of both VADs should be quite similar for the AURORA 2 task. However, the proposed VAD yields a significant performance improvement over Marzinik’s VAD for the SDC databases. Together with Aurora 3’s real recordings, this fact makes Aurora 3 a suitable database.

Using Aurora 2 database two training modes are defined for the experiments: (i) training on clean data only (Clean Training, CT), and (ii) training on clean and noisy data (Multi-Condition Training, Table 1

<table>
<thead>
<tr>
<th></th>
<th>Base G.729</th>
<th>AMR1</th>
<th>AMR2</th>
<th>AFE</th>
<th>Woo Li</th>
<th>Marzinik</th>
<th>Sohn ClVAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>WM</td>
<td>92.94</td>
<td>88.62</td>
<td>94.65</td>
<td>95.67</td>
<td>95.28</td>
<td>97.53</td>
<td></td>
</tr>
<tr>
<td>MM</td>
<td>83.31</td>
<td>72.84</td>
<td>80.59</td>
<td>90.91</td>
<td>90.23</td>
<td>92.94</td>
<td></td>
</tr>
<tr>
<td>HM</td>
<td>51.55</td>
<td>65.50</td>
<td>62.41</td>
<td>85.77</td>
<td>77.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>75.93</td>
<td>75.65</td>
<td>74.33</td>
<td>90.78</td>
<td>87.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Woo Li</td>
<td>Marzinik</td>
<td>Sohn ClVAD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WM</td>
<td>95.35</td>
<td>91.82</td>
<td>94.29</td>
<td>96.07</td>
<td>97.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MM</td>
<td>89.30</td>
<td>77.45</td>
<td>89.81</td>
<td>91.64</td>
<td>92.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HM</td>
<td>83.64</td>
<td>78.52</td>
<td>79.43</td>
<td>84.03</td>
<td>85.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>89.43</td>
<td>82.60</td>
<td>87.84</td>
<td>90.58</td>
<td>91.89</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2 shows the AURORA-2 recognition results as a function of the SNR for speech recognition experiments based on the G.729, AMR, AFE, Long Term Integrated Bispectrum (LTIBI)-VAD (Górrez et al., 2006) and ClVAD algorithms. The LTIBI VAD is based on multiple observation likelihood ratio tests over the integrated bispectrum coefficients, thus its computation (robust speech/noise power spectrum estimation is also needed over the sliding window) is harder for real time applications compared to the cluster-based approach. As it is clearly shown in Table 2 our proposed method obtains almost the same averaged word accuracy in recognition experiments as the LTIBI VAD when they are applied to FD and WF stages. The higher discrepancy between both VADs is found for poor SNRs where the ability of higher order statistics prevails. The main benefit of our approach is reduced computational cost when compared to our previous methods and the high WACC obtained when compared to the standards VADs (ETSI, 2002; ITU, 1996; ETSI, 1999) and other recently proposed VADs (Sohn et al., 1999; Li et al., 2002; Marzinzik and Kollmeier, 2002). This fact makes it suitable for embedded speech recognition systems working in real-time applications.

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 non-speech FD stage and the efficient application of the noise reduction algorithms.

6. Conclusions

An effective VAD for improving voice activity detection robustness and speech recognition in noisy environments is shown. The proposed ClVAD is based on a partitional hard C-means clustering technique for noise subspace modelling and considers long term speech information for the formulation of a soft decision rule. The proposed ClVAD shows the best trade-off between detection accuracy and computational cost, resulting eminently suitable for real-time applications demanding high speech detection rates. Several experiments were conducted on the Aurora databases showing that the proposed ClVAD outperforms recently reported VAD methods for speech discrimination, including Sohn’s VAD (Sohn et al., 1999), that defines a likelihood ratio test on a single observation; Li’s VAD (Li et al., 2002), which is based on the optimal edge detector pre-filtering stage first established by Canny; and the standardized ITU-T G.729, ETSI AMR for the GSM system and ETSI AFE VADs for distributed speech recognition. The VAD performs an advanced detection of beginnings and delayed detection of word endings which, in part, avoids having to include additional hangover schemes. On the other hand, it also improved the word recognition rate when it was considered as part of a complete speech recognition system.

Acknowledgements

This work has received research funding from the EU 6th Framework Programme, under contract number 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.

References


ETSI, 2000. Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms, ETSI ES 201 108 Recommendation.

ETSI, 2002. Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms, ETSI ES 201 108 Recommendation.


