1

Voice Activity Detection. Fundamentals and Speech Recognition System Robustness

J. Ramírez, J. M. Górriz and J. C. Segura
University of Granada
Spain

1. Introduction

An important drawback affecting most of the speech processing systems is the environmental noise and its harmful effect on the system performance. Examples of such systems are the new wireless communications voice services or digital hearing aid devices. In speech recognition, there are still technical barriers inhibiting such systems from meeting the demands of modern applications. Numerous noise reduction techniques have been developed to palliate the effect of the noise on the system performance and often require an estimate of the noise statistics obtained by means of a precise voice activity detector (VAD).

Speech/non-speech detection is an unsolved problem in speech processing and affects numerous applications including robust speech recognition (Karray and Marting, 2003; Ramirez et al. 2003), discontinuous transmission (ITU, 1996; ETSI, 1999), real-time speech transmission on the Internet (Sangwan et al., 2002) or combined noise reduction and echo cancellation schemes in the context of telephony (Basbug et al., 2004; Gustafsson et al., 2002). 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 the approaches have focussed on the development of robust algorithms with special attention being paid to the derivation and study of noise robust features and decision rules (Woo et al., 2000; Li et al., 2002; Marzinzik and Kollmeier, 2002). The different VAD methods include those based on energy thresholds (Woo et al., 2000), pitch detection (Chengalvarayan, 1999), spectrum analysis (Marzinzik and Kollmeier, 2002), zero-crossing rate (ITU, 1996), periodicity measure (Tucker, 1992), higher order statistics in the LPC residual domain (Nemer et al., 2001) or combinations of different features (ITU, 1993; ETSI, 1999; Tanyer and Özer, 2000).

This chapter shows a comprehensive approximation to the main challenges in voice activity detection, the different solutions that have been reported in a complete review of the state of the art and the evaluation frameworks that are normally used. The application of VADs for speech coding, speech enhancement and robust speech recognition systems is shown and discussed. Three different VAD methods are described and compared to standardized and

Open Access Database www.i-techonline.com

Source: Robust Speech Recognition and Understanding, Book edited by: Michael Grimm and Kristian Kroschel,
ISBN 987-3-90213-08-0, pp.460, I-Tech, Vienna, Austria, June 2007

recently reported strategies by assessing the speech/non-speech discrimination accuracy and the robustness of speech recognition systems.

2. Applications

VADs are employed in many areas of speech processing. Recently, VAD methods have been described in the literature for several applications including mobile communication services (Freeman et al. 1989), real-time speech transmission on the Internet (Sangwan et al., 2002) or noise reduction for digital hearing aid devices (Itoh and Mizushima, 1997). As an example, a VAD achieves silence compression in modern mobile telecommunication systems reducing the average bit rate by using the discontinuous transmission (DTX) mode. Many practical applications, such as the Global System for Mobile Communications (GSM) telephony, use silence detection and comfort noise injection for higher coding efficiency. This section shows a brief description of the most important VAD applications in speech processing: coding, enhancement and recognition.

2.1 Speech coding

VAD is widely used within the field of speech communication for achieving high speech coding efficiency and low-bit rate transmission. The concepts of silence detection and comfort noise generation lead to dual-mode speech coding techniques. The different modes of operation of a speech codec are: *i*) the active speech codec, and *ii*) the silence suppression and comfort noise generation modes. 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. Figure 1 shows a block diagram of a dual mode speech codec. The full rate speech coder is operational during active voice speech, but a different coding scheme is employed for the inactive voice signal, using fewer bits and resulting in a higher overall average compression ratio. As an example, the recommendation G.729 Annex B (ITU, 1996) uses a feature vector consisting of the linear prediction (LP) spectrum, the full-band energy, the low-band (0 to 1 KHz) energy and the zero-crossing rate (ZCR). The standard was developed with the collaboration of researchers from France Telecom, the University of Sherbrooke, NTT and AT&T Bell Labs and the effectiveness of the VAD was evaluated in terms of subjective speech quality and bit rate savings (Benyassine et al., 1997). Objective performance tests were also conducted by hand-labeling a large speech database and assessing the correct identification of voiced, unvoiced, silence and transition periods. Another standard for DTX is the ETSI (Adaptive Multi-Rate) AMR speech coder (ETSI, 1999) developed by the Special Mobile Group (SMG) for the GSM system. The standard 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 2 VAD, which uses parameters of the speech encoder, and is more robust against environmental noise than AMR1 and G.729. The dual mode speech transmission achieves a significant bit rate reduction in digital speech coding since about 60% of the time the transmitted signal contains just silence in a phone-based communication.

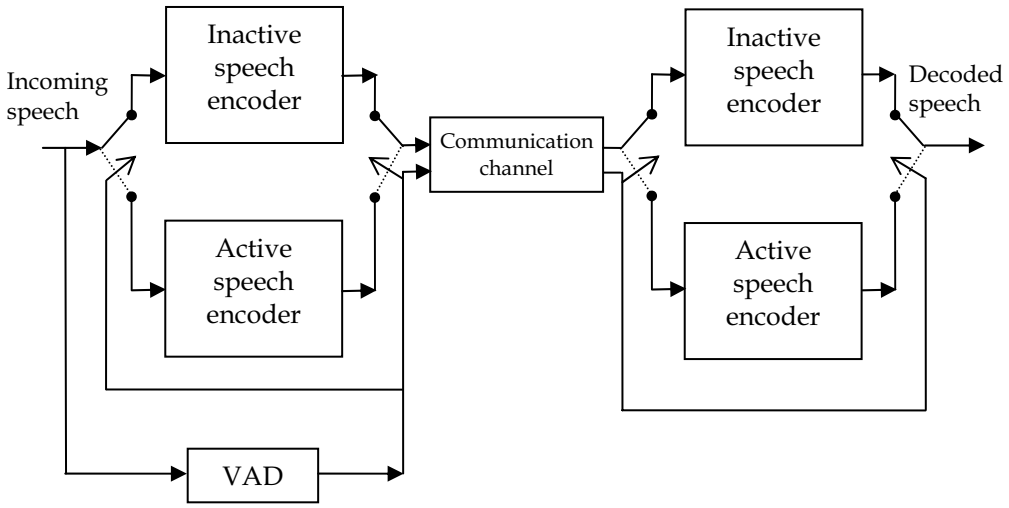


Figure 1. Speech coding with VAD for DTX.

2.2 Speech enhancement

Speech enhancement aims at improving the performance of speech communication systems in noisy environments. It mainly deals with suppressing background noise from a noisy signal. A difficulty in designing efficient speech enhancement systems is the lack of explicit statistical models for the speech signal and noise process. In addition, the speech signal, and possibly also the noise process, are not strictly stationary processes. Speech enhancement normally assumes that the noise source is additive and not correlated with the clean speech signal. One of the most popular methods for reducing the effect of background (additive) noise is spectral subtraction (Boll, 1979). The popularity of spectral subtraction is largely due to its relative simplicity and ease of implementation. The spectrum of noise N(f) is estimated during speech inactive periods and subtracted from the spectrum of the current frame $X(f)$ resulting in an estimate of the spectrum $S(f)$ of the clean speech:

$$|S(f)| = |X(f)| - |N(f)| \tag{1}$$

There exist many refinements of the original method that improve the quality of the enhanced speech. As an example, the modified spectral subtraction enabling an over-subtraction factor α and maximum attenuation β for the noise is given by:

$$|S(f)| = \max\{|X(f)| - \alpha |N(f)|, \beta |X(f)|\} \quad (2)$$

Generally, spectral subtraction is suitable for stationary or very slow varying noises so that the statistics of noise could be updated during speech inactive periods. Another popular method for speech enhancement is the Wiener filter that obtains a least squares estimate of the clean signal $s(t)$ under stationary assumptions of speech and noise. The frequency response of the Wiener filter is defined to be:

$$W(f) = \frac{\Phi_{ss}(f)}{\Phi_{ss}(f) + \Phi_{nn}(f)} \quad (3)$$

and requires an estimate of the power spectrum $\Phi_{ss}(f)$ of the clean speech and the power spectrum $\Phi_{nn}(f)$ of the noise.

2.3 Speech recognition

Performance of speech recognition systems is strongly influenced by the quality of the speech signal. Most of these systems are based on complex hidden Markov models (HMM) that are trained using a training speech database. The mismatch between the training conditions and the testing conditions has a deep impact on the accuracy of these systems and represents a barrier for their operation in noisy environments. Fig. 2 shows an example of the degradation of the word accuracy for the AURORA2 database and speech recognition task when the ETSI recommendation (ETSI, 2000) not including noise compensation algorithm is used as feature extraction process. Note that, when the HMMs are trained using clean speech, the recognizer performance rapidly decreases when the level of background noise increases. Better results are obtained when the HMMs are trained using a collection of clean and noisy speech records.

VAD is a very useful technique for improving the performance of speech recognition systems working in these scenarios. A VAD module is used in most of the speech recognition systems within the feature extraction process for speech enhancement. The noise statistics such as its spectrum are estimated during non-speech periods in order to apply the speech enhancement algorithm (spectral subtraction or Wiener filter). On the other hand, non-speech frame-dropping (FD) is also a frequently used technique in speech recognition to reduce the number of insertion errors caused by the noise. It consists on dropping non-speech periods (based on the VAD decision) from the input of the speech recognizer. This reduces the number of insertion errors due to the noise that can be a serious error source under high mismatch training/testing conditions. Fig. 3 shows an example of a typical robust speech recognition system incorporating spectral noise reduction and non-speech frame-dropping. After the speech enhancement process is applied, the Mel frequency cepstral coefficients and its first- and second-order derivatives are computed in a frame by frame basis to form a feature vector suitable for recognition. Figure 4 shows the improvement provided by a speech recognition system incorporating the VAD presented in (Ramirez et al., 2005) within an enhanced feature extraction process based on a Wiener filter and non-speech frame dropping for the AURORA 2 database and tasks. The relative improvement over (ETSI, 2000) is about 27.17% in multicondition and 60.31% in clean condition training/testing.

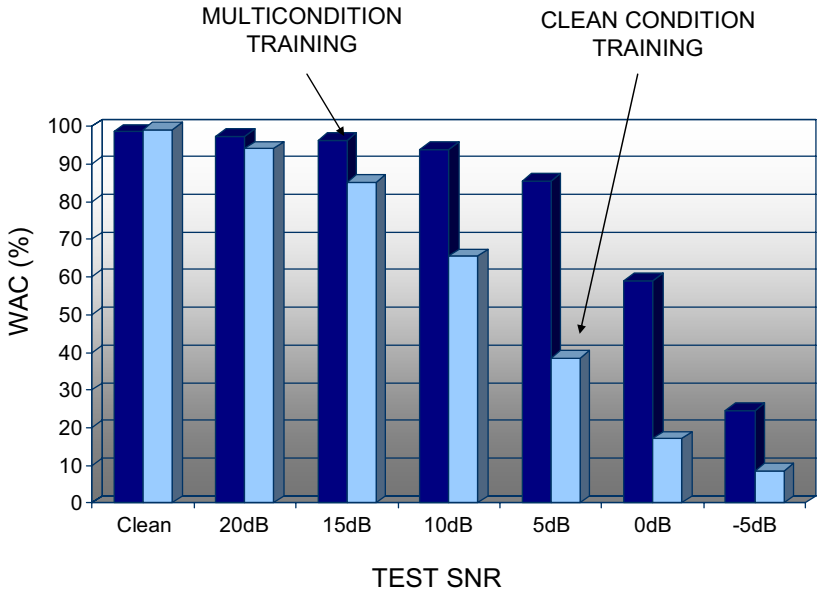


Figure 2. Speech recognition performance for the AURORA-2 database and tasks.

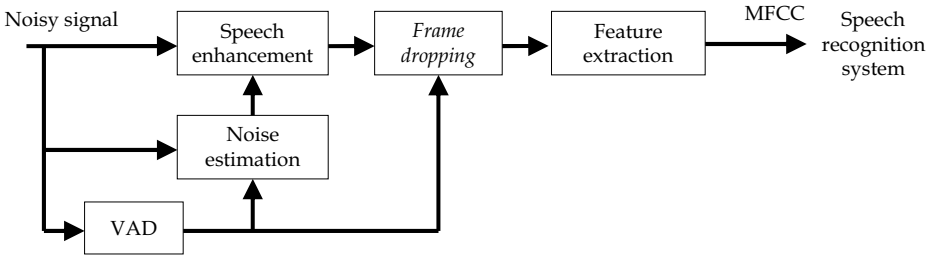


Figure 3. Feature extraction with spectral noise reduction and non-speech frame-dropping.

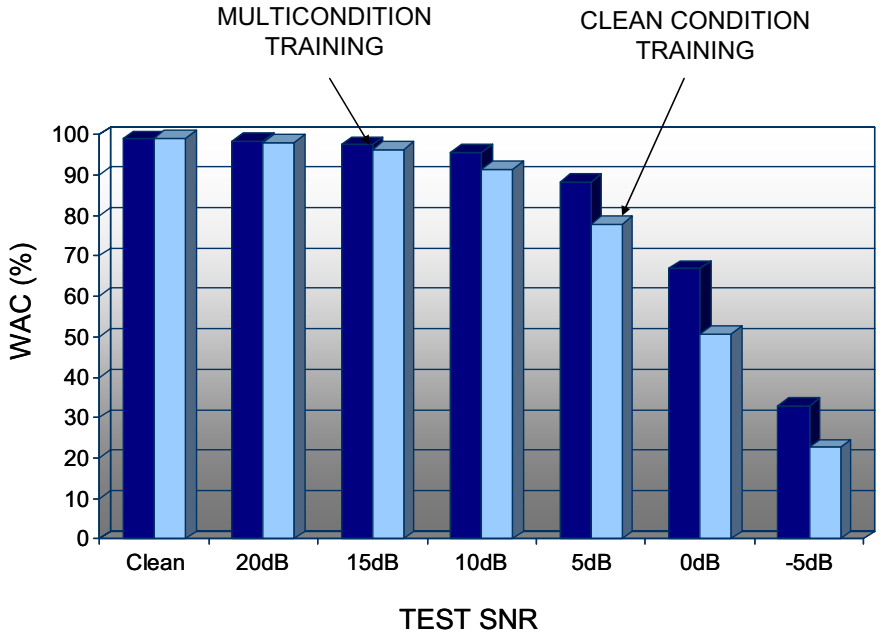


Figure 4. Results obtained for an enhanced feature extraction process incorporating VAD-based Wiener filtering and non-speech frame-dropping.

3. Voice activity detection in noisy environments

An important problem in many areas of speech processing is the determination of presence of speech periods in a given signal. This task can be identified as a statistical hypothesis problem and its purpose is the determination to which category or class a given signal belongs. The decision is made based on an observation vector, frequently called feature vector, which serves as the input to a decision rule that assigns a sample vector to one of the given classes. The classification task is often not as trivial as it appears since the increasing level of background noise degrades the classifier effectiveness, thus leading to numerous detection errors. Fig. 5 illustrates the challenge of detecting speech presence in a noisy signal when the level of background noise increases and the noise completely masks the speech signal. The selection of an adequate feature vector for signal detection and a robust decision rule is a challenging problem that affects the performance of VADs working under noise conditions. Most algorithms are effective in numerous applications but often cause detection errors mainly due to the loss of discriminating power of the decision rule at low SNR levels (ITU, 1996; ETSI, 1999). For example, a simple energy level detector can work satisfactorily in high signal-to-noise ratio (SNR) conditions, but would fail significantly when the SNR drops. VAD results more critical in non-stationary noise environments since it is needed to update the constantly varying noise statistics affecting a misclassification error strongly to the system performance.

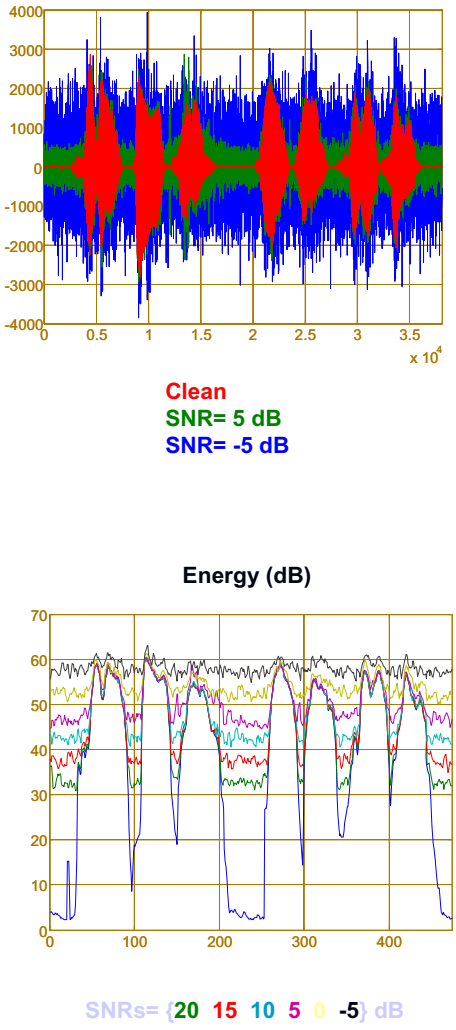


Figure 5. Energy profile of a speech utterance corrupted by additive backgorund noise at decreasing SNRs.

3.1 Description of the problem

The VAD problem considers detecting the presence of speech in a noisy signal. The VAD decision is normally based on a feature vector **x**. Assuming that the speech signals and the noise are additive, the VAD module has to decide in favour of the two hypotheses:

$$\begin{aligned} H_0 &: \mathbf{x} = \mathbf{n} \\ H_1 &: \mathbf{x} = \mathbf{n} + \mathbf{s} \end{aligned} \tag{4}$$

A block diagram of VAD is shown in figure 6. It consists of: *i*) the feature extraction process, *ii*) the decision module, and *iii*) the decision smoothing stage.

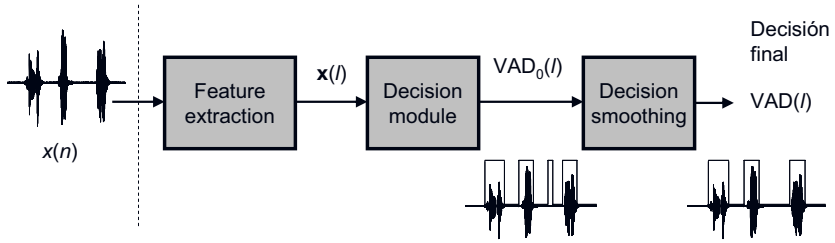


Figure 6. Block diagram of a VAD.

3.2 Feature extraction

The objective of feature extraction process is to compute discriminative speech features suitable for detection. A number of robust speech features have been studied in this context. The different approaches include: *i*) full-band and subband energies (Woo et al., 2000), *ii*) spectrum divergence measures between speech and background noise (Marzinzik and Kollmeier, 2002), *iii*) pitch estimation (Tucker, 1992), *iv*) zero crossing rate (Rabiner et al., 1975), and *v*) higher-order statistics (Nemer et al. 2001; Ramírez et al., 2006a; Górriz et al., 2006a; Ramírez et al., 2007). Most of the VAD methods are based on the current observation (frame) and do not consider contextual information. However, using long-term speech information (Ramírez et al., 2004a; Ramírez et al. 2005a) has shown significant benefits for detecting speech presence in high noise environments.

3.3 Formulation of the decision rule

The decision module defines the rule or method for assigning a class (speech or silence) to the feature vector **x**. Sohn et al. (Sohn et al., 1999) proposed a robust VAD algorithm based on a statistical likelihood ratio test (LRT) involving a single observation vector. (Sohn et al., 1999). The method considered a two-hypothesis test where the optimal decision rule that minimizes the error probability is the Bayes classifier. Given an observation vector to be classified, the problem is reduced to selecting the class (H_0 or H_1) with the largest posterior probability $P(H_i|\mathbf{x})$:

$$P(H_1|\mathbf{x}) \underset{H_0}{\overset{H_1}{\gtrless}} P(H_0|\mathbf{x}) \tag{5}$$

Using the Bayes rule leads to statistical likelihood ratio test:

$$\frac{P(\mathbf{x}\,|\,H_1)}{P(\mathbf{x}\,|\,H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P(H_0)}{P(H_1)} \tag{6}$$

In order to evaluate this test, the discrete Fourier transform (DFT) coefficients of the clean speech (S_j) and the noise (N_j) are assumed to be asymptotically independent Gaussian random variables:

$$\begin{aligned} p(x\,|\,H_0) &= \prod_{j=0}^{J-1} \frac{1}{\pi\lambda_N(j)} \exp\left\{-\frac{|X_j|^2}{\lambda_N(j)}\right\} \\ p(x\,|\,H_1) &= \prod_{j=0}^{J-1} \frac{1}{\pi[\lambda_S(j)+\lambda_N(j)]} \exp\left\{-\frac{|X_j|^2}{\lambda_S(j)+\lambda_N(j)}\right\} \end{aligned} \tag{7}$$

where X_j represents the noisy speech DFT coefficients, and $\lambda_N(j)$ and $\lambda_S(j)$ denote the variances of N_j and S_j for the j-th bin of the DFT, respectively. Thus, the decision rule is reduced to:

$$\frac{1}{J}\sum_{j=0}^{J-1}\left[\frac{\gamma_j\xi_j}{1+\xi_j} - \log(1+\xi_j)\right] \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{8}$$

and η defines the decision threshold and J is the DFT order. ξ_j and γ_j define the *a priori* and *a posteriori* SNRs:

$$\gamma_j = \frac{|X_j|^2}{\lambda_N(j)} \qquad \xi_j = \frac{\lambda_S(j)}{\lambda_N(j)} \tag{9}$$

that are normally estimated using the Ephraim and Malah minimum mean-square error (MMSE) estimator (Ephraim and Malah, 1984).

Several methods for VAD formulate the decision rule based on distance measures like the Euclidean distance (Gorriz et al., 2006b), Itakura-Saito and Kullback-Leibler divergence (Ramírez et al., 2004b). Other techniques include fuzzy logic (Beritelli et al., 2002), support vector machines (SVM) (Ramírez et al. 2006b) and genetic algorithms (Estevez et al., 2005).

3.4 Decision smoothing

Most of the VADs that formulate the decision rule on a frame by frame basis normally use decision smoothing algorithms in order to improve the robustness against the noise. The motivations for these approaches are found in the speech production process and the reduced signal energy of word beginnings and endings. The so called hang-over algorithms extends and smooth the VAD decision in order to recover speech periods that are masked by the acoustic noise.

4. Robust VAD algorithms

This section summarizes three VAD algorithms recently reported that yield high speech/non-speech discrimination in noisy environments.

4.1 Long-term spectral divergence

The speech/non-speech detection algorithm proposed in (Ramírez et al., 2004a) assumes that the most significant information for detecting voice activity on a noisy speech signal remains on the time-varying signal spectrum magnitude. It uses a long-term speech window instead of instantaneous values of the spectrum to track the spectral envelope and is based on the estimation of the so called Long-Term Spectral Envelope (LTSE). The decision rule is then formulated in terms of the Long-Term Spectral Divergence (LTSD) between speech and noise.

Let $x(n)$ be a noisy speech signal that is segmented into overlapped frames and, $X(k,l)$ its amplitude spectrum for the k-th band at frame l. The N-order Long-Term Spectral Envelope (LTSE) is defined as:

$$LTSE_N(k,l) = \max\left\{X(k,l+j)\right\}_{j=-N}^{j=+N} \tag{9}$$

The VAD decision rule is then formulated by means of the N-order Long-Term Spectral Divergence (LTSD) between speech and noise is defined as the deviation of the LTSE respect to the average noise spectrum magnitude $N(k)$ for the k band, k= 0, 1, …, $NFFT$-1, and is given by:

$$LTSD_N(l) = 10\log_{10}\left(\frac{1}{NFFT}\sum_{k=0}^{NFFT-1}\frac{LTSE^2(k,l)}{N^2(k)}\right)\underset{H_0}{\overset{H_1}{\gtrless}}\eta \tag{9}$$

4.2 Multiple observation likelihood ratio test

An improvement over the LRT proposed by Sohn (Sohn et al., 1999) is the multiple observation LRT (MO-LRT) proposed by Ramírez (Ramírez et al., 2005b). The performance of the decision rule was improved by incorporating more observations to the statistical test. The MO-LRT is defined over the observation vectors $\{\mathbf{x}_{l-m},...,\mathbf{x}_l,...\mathbf{x}_{l+m}\}$ as follows:

$$\ell_{l,m} = \sum_{k=l-m}^{l+m}\ln\left(\frac{p(\mathbf{x}_k\,|\,H_1)}{p(\mathbf{x}_k\,|\,H_0)}\right)\underset{H_0}{\overset{H_1}{\gtrless}}\eta \tag{10}$$

where l denotes the frame being classified as speech (H_1) or silence (H_0). Thus, the decision rule is formulated over a sliding window consisting of observation vectors around the current frame. The so-defined decision rule reported significant improvements in speech/non-speech discrimination accuracy over existing VAD methods that are defined on a single observation and need empirically tuned hangover mechanisms.

4.3 Order statistics filters

The MO-LRT VAD takes advantage of using contextual information for the formulation of the decision rule. The same idea can be found in other existing VADs like the Li et al. (Li et

al., 2002) that considers optimum edge detection linear filters on the full-band energy. Order statistics filters (OSFs) have been also evaluated for a low variance measure of the divergence between speech and silence (noise). The algorithm proposed in (Ramírez et al., 2005a) uses two OSFs for the multiband quantile (MBQ) SNR estimation. The algorithm is described as follows. Once the input speech has been de-noised by Wiener filtering, the log-energies for the l-th frame, $E(k,l)$, in K subbands (k= 0, 1, …, K-1), are computed by means of:

$$E(k,l) = \log\left(\frac{K}{NFFT}\sum_{m=m_k}^{m_{k-1}-1}|Y(m,l|^2\right) \qquad m_k = \left\lfloor \frac{NFFT}{2K}k \right\rfloor \qquad k = 0,1,...,K-1 \tag{11}$$

The implementation of both OSFs is based on a sequence of log-energy values {$E(k,l\text{-}N)$, …, $E(k,l)$, …, $E(k,l\text{+}N)$} around the frame to be analyzed. The r-th order statistics of this sequence, $E_{(r)}(k,l)$, is defined as the r-th largest number in algebraic order. A first OSF estimates the subband signal energy by means of

$$Q_p(k,l) = (1-f)E_{(s)}(k,l) + fE_{(s+1)}(k,l) \tag{12}$$

where $Q_p(k,l)$ is the sampling quantile, s= $\lfloor 2pN \rfloor$ and f= $2pN$-s. Finally, the SNR in each subband is measured by:

$$QSNR(k,l) = Q_p(k,l) - E_N(k) \tag{13}$$

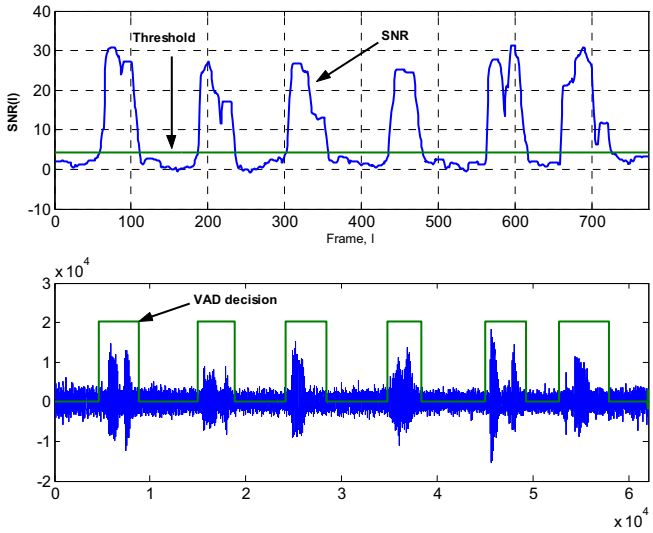
where $E_N(k)$ is the noise level in the k-th band that needs to be estimated. For the initialization of the algorithm, the first frames are assumed to be non-speech frames and the noise level $E_N(k)$ in the k-th band is estimated as the median of the set {$E(0,k)$, $E(1,k)$, …, $E(N\text{-}1,k)$}.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

$$E_N(k) = \alpha E_N(k) + (1-\alpha)Q_{0.5}(k,l) \tag{14}$$

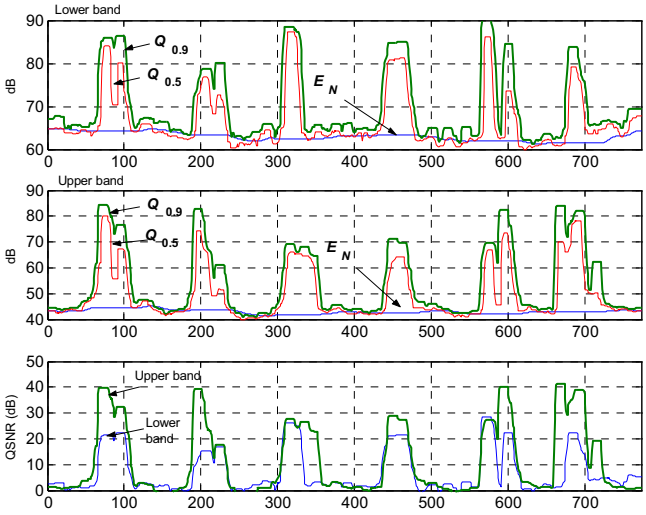
where $Q_{0.5}(k,l)$ is the output of the median filter and α= 0.97 was experimentally selected. On the other hand, the sampling quantile p= 0.9 is selected as a good estimation of the subband spectral envelope. The decision rule is then formulated in terms of the average subband SNR:

$$SNR(l) = \frac{1}{K}\sum_{k=0}^{K-1} QSNR(k,l) \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{15}$$

Figure 7 shows the operation of the MBQ VAD on an utterance of the Spanish SpeechDat-Car (SDC) database (Moreno et al., 2000). For this example, K= 2 subbands were used while N= 8 . The optimal selection of these parameters is studied in (Ramirez et al., 2005a). It is clearly shown how the SNR in the upper and lower band yields improved speech/non-speech discrimination of fricative sounds by giving complementary information. The VAD performs an advanced detection of beginnings and delayed detection of word endings which, in part, makes a hang-over unnecessary.



(a)



(b)

Figure 7. Operation of the VAD on an utterance of Spanish SDC database. (a) SNR and VAD decision. (b) Subband SNRs.

5. Experimental framework

Several experiments are commonly conducted to evaluate the performance of VAD algorithms. The analysis is mainly focussed on the determination of the error probabilities or classification errors at different SNR levels (Marzinzik and Kollmeier, 2002), and the influence of the VAD decision on the performance of speech processing systems (Bouquin-Jeannes and Faucon, 1995). Subjective performance tests have also been considered for the evaluation of VADs working in combination with speech coders (Benyassine et al., 1997). The experimental framework and the objective performance tests commonly conducted to evaluate VAD methods are described in this section.

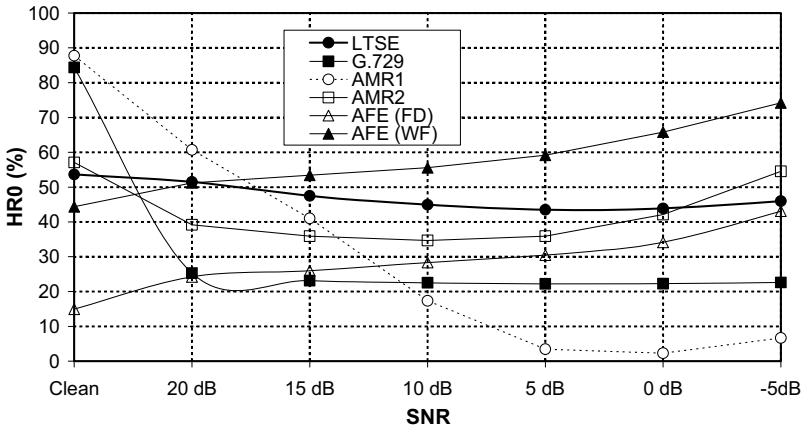
5.1 Speech/non-speech discrimination analysis

VADs are widely evaluated in terms of the ability to discriminate between speech and pause periods at different SNR levels. In order to illustrate the analysis, this subsection considers the evaluation of the LTSE VAD (Ramírez et al., 2004). The original AURORA-2 database (Hirsch and Pearce, 2000) was used in this analysis since it uses the clean TIdigits database consisting 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 20dB, 15dB, 10dB, 5dB, 0dB and -5dB. 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 TIdigits database was used to manually label each utterance as speech or non-speech frames for reference. Detection performance as a function of the SNR was assessed in terms of the non-speech 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:

$$\mathrm{HR0} = \frac{N_{0,0}}{N_0^{ref}} \qquad \mathrm{HR1} = \frac{N_{1,1}}{N_1^{ref}} \tag{15}$$

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

Figure 8 provides the results of this analysis and compares the proposed LTSE VAD algorithm to standard G.729, AMR and AFE (ETSI, 2002) VADs in terms of non-speech hit-rate (HR0, Fig. 8.a) and speech hit-rate (HR1, Fig. 8.b) 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 (ETSI, 2002) for estimating the noise spectrum in the Wiener filtering stage and non-speech frame-dropping are provided. It can be concluded that LTSE achieves the best compromise among the different VADs tested; it obtains a good behavior in detecting non-speech periods as well as exhibits a slow decay in performance at unfavorable noise conditions in speech detection.



(a)

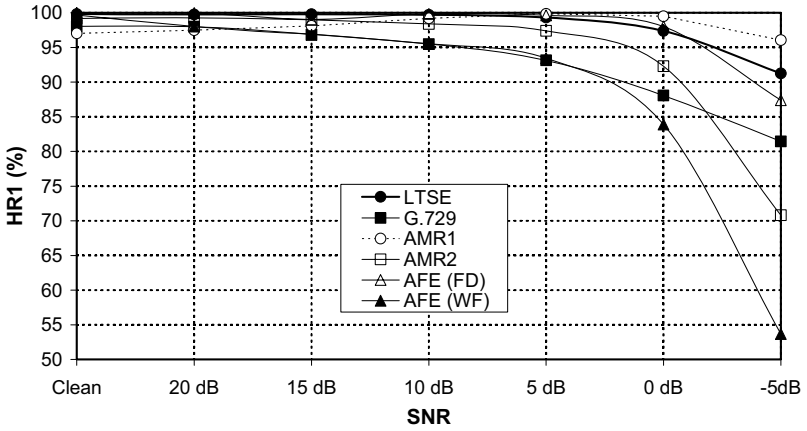


Figure 8. Speech/non-speech discrimination analysis. (a) Non-speech hit-rate (HR0). (b) Speech hit rate (HR1).

5.2 Receiver operating characteristics curves

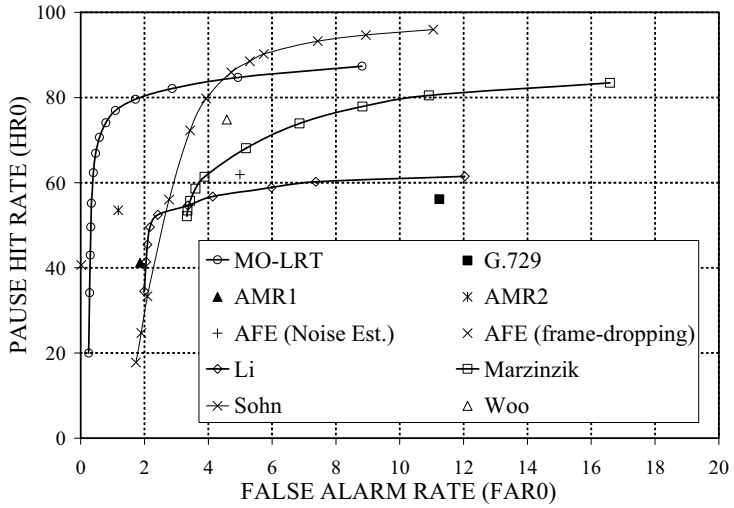
The ROC curves are frequently used to completely describe the VAD error rate. The AURORA subset of the original Spanish SpeechDat-Car (SDC) database (Moreno et al., 2000) was used in this analysis. This database contains 4914 recordings using close-talking and distant microphones from more than 160 speakers. The files are categorized into three noisy conditions: quiet, low noisy and highly noisy conditions, which represent different driving conditions with average SNR values between 25 dB, and 5 dB. The non-speech hit rate (HR0) and the false alarm rate (FAR0= 100-HR1) were determined in each noise condition being the actual speech frames and actual speech pauses determined by hand-labeling the database on the close-talking microphone.

Figure 9 shows the ROC curves of the MO-LRT VAD (Ramírez et al., 2005b) and other frequently referred algorithms for recordings from the distant microphone in quiet and high noisy conditions. The working points of the G.729, AMR, and AFE VADs are also included. The results show improvements in detection accuracy over standard VADs and over a representative set of VAD algorithms. Thus, among all the VAD examined, our VAD 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. The benefits are especially important over G.729, which is used along with a speech codec for discontinuous transmission, and over the Li's algorithm, that is based on an optimum linear filter for edge detection. The proposed VAD also improves Marzinzik's VAD that tracks the power spectral envelopes, and the Sohn's VAD, that formulates the decision rule by means of a statistical likelihood ratio test.

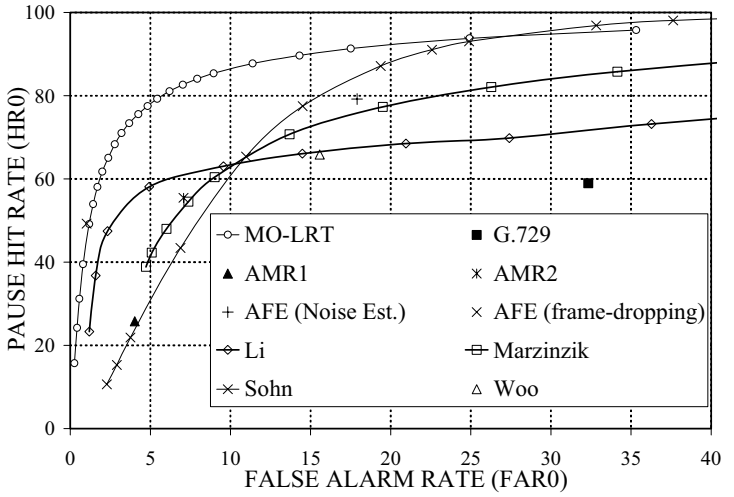
5.3 Improvement in speech recognition systems

Performance of ASR systems working over wireless networks and noisy environments normally decreases and non efficient speech/non-speech detection appears to be an important degradation source (Karray and Martin, 2003). Although the discrimination analysis or the ROC curves are effective to evaluate a given algorithm, this section evaluates the VAD according to the goal for which it was developed by assessing the influence of the VAD over the performance of a speech recognition system.

The reference framework considered for these experiments was the ETSI AURORA project for DSR (ETSI, 2000; ETSI, 2002). 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 modeled as whole word HMMs (Hidden Markov Models) with the following parameters: 16 states per word, simple left-to-right models, mixture of three Gaussians per state (diagonal covariance matrix) while speech pause models consist of three states with a mixture of six Gaussians per state. The 39-parameter feature vector consists of 12 cepstral coefficients (without the zero-order coefficient), the logarithmic frame energy plus the corresponding delta and acceleration coefficients.



(a)



(b)

Figure 9. ROC curves. (a) Stopped Car, Motor Running. (b) High Speed, Good Road.

	G.729	AMR1	AMR2	AFE	**MBQW**

WF	66.19	74.97	83.37	81.57	**84.12**
WF+FD	70.32	74.29	82.89	83.29	**86.09**
	Woo	Li	Marzinzik	Sohn	Hand-labeled
WF	83.64	77.43	84.02	83.89	84.69
WF+FD	81.09	82.11	85.23	83.80	86.86

Table 1. Average Word Accuracy (%) for the AURORA 2 database for clean and multicondition training experiments. Results are mean values for all the noises and SNRs ranging from 20 to 0 dB.

Two training modes are defined for the experiments conducted on the AURORA-2 database: *i*) training on clean data only (Clean Training), and *ii*) training on clean and noisy data (multicondition training). For the AURORA-3 SpeechDat-Car databases, the so called well-matched (WM), medium-mismatch (MM) and high-mismatch (HM) conditions are used. These databases contain recordings from the close-talking and distant microphones. In WM condition, both close-talking and hands-free microphones are used for training and testing. In MM condition, both training and testing are performed using the hands-free microphone recordings. In HM condition, training is done using close-talking microphone material from all driving conditions while testing is done using hands-free microphone material taken for low noise and high noise driving conditions. Finally, recognition performance is assessed in terms of the word accuracy (WAcc) that considers deletion, substitution and insertion errors. An enhanced feature extraction scheme incorporating a noise reduction algorithm and non-speech frame-dropping 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 achieved by the different VADs that were compared. These results are averaged over the three test sets of the AURORA-2 recognition experiments and SNRs between 20 and 0 dBs. Note that, for the recognition experiments based on the AFE VADs, the same configuration of the standard (ETSI, 2002), which considers different VADs for WF and FD, was used. The MBQW VAD outperforms G.729, AMR1, AMR2 and AFE standard VADs in both clean and multi condition training/testing experiments. When compared to recently reported VAD algorithms, it yields better results being the one that is closer to the "ideal" hand-labeled speech recognition performance.

		Base	Base + WF					Base + WF + FD				
			G.729	AMR1	AMR2	AFE	**MBQW**	G.729	AMR1	AMR2	AFE	**MBQW**
Finnish	WM	92.74	93.27	93.66	95.52	94.28	95.36	88.62	94.57	95.52	94.25	94.70
	MM	80.51	75.99	78.93	75.51	78.52	74.49	67.99	81.60	79.55	82.42	80.08
	HM	40.53	50.81	40.95	55.41	55.05	56.40	65.80	77.14	80.21	56.89	83.67
	Average	**71.26**	73.36	71.18	75.48	75.95	**75.42**	74.14	84.44	85.09	77.85	**86.15**
Spanish	WM	92.94	89.83	85.48	91.24	89.71	91.66	88.62	94.65	95.67	95.28	96.79
	MM	83.31	79.62	79.31	81.44	76.12	83.95	72.84	80.59	90.91	90.23	91.85
	HM	51.55	66.59	56.39	70.14	68.84	70.47	65.50	62.41	85.77	77.53	87.25
	Average	**75.93**	78.68	73.73	80.94	78.22	**82.03**	75.65	74.33	90.78	87.68	**91.96**
German	WM	91.20	90.60	90.20	93.13	91.48	92.87	87.20	90.36	92.79	93.03	93.73
	MM	81.04	82.94	77.67	86.02	84.11	85.58	68.52	78.48	83.87	85.43	87.40
	HM	73.17	78.40	70.40	83.07	82.01	82.56	72.48	66.23	81.77	83.16	83.49
	Average	**81.80**	83.98	79.42	87.41	85.87	**87.00**	76.07	78.36	86.14	87.21	**88.21**
Average		**76.33**	78.67	74.78	81.28	80.01	**81.48**	75.29	79.04	87.34	84.25	**88.77**

(a)

		Base	Base + WF					Base + WF + FD				
			Woo	Li	Marzinzik	Sohn	**MBQW**	Woo	Li	Marzinzik	Sohn	**MBQW**
Finnish	WM	92.74	95.25	95.15	95.39	95.21	95.36	86.81	85.60	93.73	93.84	94.70
	MM	80.51	77.70	76.74	73.94	72.16	74.49	66.62	55.63	76.47	80.10	80.08
	HM	40.53	57.74	53.85	57.28	57.24	56.40	62.54	58.34	68.37	75.34	83.67
	Average	**71.26**	76.90	75.25	75.54	74.87	**75.42**	71.99	66.52	79.52	83.09	**86.15**
Spanish	WM	92.94	90.85	91.24	91.31	91.25	91.66	95.35	91.82	94.29	96.07	96.79
	MM	83.31	81.07	84.00	82.90	82.21	83.95	89.30	77.45	89.81	91.64	91.85
	HM	51.55	61.38	64.72	65.05	69.89	70.47	83.64	78.52	79.43	84.03	87.25
	Average	**75.93**	77.77	79.99	79.75	81.12	**82.03**	89.43	82.60	87.84	90.58	**91.96**
German	WM	91.20	92.83	92.25	93.17	93.17	92.87	91.59	89.62	91.58	93.23	93.73
	MM	81.04	85.58	85.21	85.29	86.09	85.58	80.28	70.87	83.67	83.97	87.40
	HM	73.17	83.02	82.98	83.02	83.53	82.56	78.68	78.55	81.27	82.19	83.49
	Average	**81.80**	87.14	86.81	87.16	87.60	**87.00**	83.52	79.68	85.51	86.46	**88.21**
Average		**76.33**	80.60	80.68	80.82	81.20	**81.48**	81.65	76.27	84.29	86.71	**88.77**

(b)

Table 2. Average Word Accuracy for the SpeechDat-Car databases. (a) Comparison to standardized VADs. (b) Comparison to other recently reported methods.

Table 2 shows the recognition performance for the Finnish, Spanish, and German SDC databases for the different training/test mismatch conditions (HM, high mismatch, MM: medium mismatch and WM: well matched) when WF and FD are performed on the base system (ETSI, 2000). Again, MBQW 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 MBQW VAD to Marzinzik's VAD. The word accuracy of both VADs is quite similar for the AURORA 2 task. However, MBQW yields a significant performance improvement over Marzinzik's VAD for the SDC databases.

6. Conclusions

This chapter has shown an overview of the main challenges in robust speech detection and a review of the state of the art and applications. VADs are frequently used in a number of applications including speech coding, speech enhancement and speech recognition. A precise VAD extracts a set of discriminative speech features from the noisy speech and formulates the decision in terms of well defined rule. The chapter has summarized three robust VAD methods that yield high speech/non-speech discrimination accuracy and improve the performance of speech recognition systems working in noisy environments. The evaluation of these methods showed the experiments most commonly conducted to compare VADs: *i*) speech/non-speech discrimination analysis, *ii*) the receiver operating characteristic curves, and iii) speech recognition system tests.

7. 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 SR3-VoIP project (TEC2004-03829/TCM) from the Spanish government. The views expressed here are those of the authors only. The Community is not liable for any use that may be made of the information contained therein.

8. References

Karray, L.; Martin. A. (2003). Toward improving speech detection robustness for speech recognition in adverse environments, *Speech Communication*, no. 3, pp. 261–276.

Ramírez, J.; Segura, J.C.; Benítez, C.; de la Torre, A.; Rubio, A. (2003). A new adaptive long-term spectral estimation voice activity detector, *Proc. EUROSPEECH 2003*, Geneva, Switzerland, pp. 3041–3044.

ITU-T Recommendation G.729-Annex B. (1996). A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70.

ETSI EN 301 708 Recommendation. (1999). Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels.

Sangwan, A.; Chiranth, M.C.; Jamadagni, H.S.; Sah, R.; Prasad, R.V.; Gaurav, V. (2002). VAD Techniques for Real-Time Speech Transmission on the Internet, *IEEE International Conference on High-Speed Networks and Multimedia Communications*, pp. 46-50.

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

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

Sohn, J.; Kim, N.S.; Sung, W. (1999). A statistical model-based voice activity detection, *IEEE Signal Processing Letters*, vol. 16, no. 1, pp. 1–3.

Cho, Y.D.; Kondoz, A. (2001). Analysis and improvement of a statistical model-based voice activity detector, *IEEE Signal Processing Letters*, vol. 8, no. 10, pp. 276–278.

Gazor, S.; Zhang, W. (2003). A soft voice activity detector based on a Laplacian-Gaussian model, *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 498–505.

Armani, L.; Matassoni, M.; Omologo, M.; Svaizer, P. (2003). Use of a CSP-based voice activity detector for distant-talking ASR, *Proc. EUROSPEECH 2003*, Geneva, Switzerland, pp. 501–504.

Bouquin-Jeannes, R.L.; Faucon, G. (1995). Study of a voice activity detector and its influence on a noise reduction system, *Speech Communication*, vol. 16, pp. 245–254.

Woo, K.; Yang, T.; Park, K.; Lee, C. (2000). Robust voice activity detection algorithm for estimating noise spectrum, *Electronics Letters*, vol. 36, no. 2, pp. 180–181.

Li, Q.; Zheng, J.; Tsai, A.; Zhou, Q. (2002). Robust endpoint detection and energy normalization for real-time speech and speaker recognition, *IEEE Trans. Speech Audio Processing*, vol. 10, no. 3, pp. 146–157.

Marzinzik, M.; Kollmeier, B. (2002). Speech pause detection for noise spectrum estimation by tracking power envelope dynamics, *IEEE Trans. Speech Audio Processing*, vol. 10, no. 6, pp. 341–351.

Chengalvarayan, R. (1999). Robust energy normalization using speech/non-speech discriminator for German connected digit recognition, *Proc. EUROSPEECH 1999*, Budapest, Hungary, pp. 61–64.

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

Nemer, E.; Goubran, R.; Mahmoud, S. (2001). Robust voice activity detection using higher-order statistics in the lpc residual domain, *IEEE Trans. Speech Audio Processing*, vol. 9, no. 3, pp. 217–231.

Tanyer, S.G.; Özer, H. (2000). Voice activity detection in nonstationary noise, *IEEE Trans. Speech Audio Processing*, vol. 8, no. 4, pp. 478–482.

Freeman, D.K.; Cosier, G.; Southcott, C.B.; Boyd, I. (1989). The Voice Activity Detector for the PAN-European Digital Cellular Mobile Telephone Service, *International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 369-372.

Itoh, K.; Mizushima, M. (1997). Environmental noise reduction based on speech/non-speech identification for hearing aids, *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 419-422.

Benyassine, A.; Shlomot, E.; Su, H.; Massaloux, D.; Lamblin, C.; Petit, J. (1997). 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 Communications Magazine*, Vol. 35, No. 9, pp. 64-73.

Boll, S., F. Suppression of Acoustic Noise in Speech Using Spectral Subtraction, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, April 1979.

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

Ramírez, J.; Segura, J.C.; Benítez, C.; de la Torre, A.; Rubio, A. (2005a). An Effective Subband OSF-based VAD with Noise Reduction for Robust Speech Recognition, *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 6, pp. 1119-1129.

Ramírez, J.; Górriz, J.M; Segura, J.C.; Puntonet, C.G; Rubio, A. (2006a). Speech/Non-speech Discrimination based on Contextual Information Integrated Bispectrum LRT, *IEEE Signal Processing Letters*, vol. 13, No. 8, pp. 497-500.

Górriz, J.M.; Ramírez, J.; Puntonet, C.G.; Segura, J.C. (2006a). Generalized LRT-based voice activity detector, *IEEE Signal Processing Letters*, Vol. 13, No. 10, pp. 636-639.

Ramírez, J.; Górriz, J.M.; Segura, J.C. (2007). Statistical Voice Activity Detection Based on Integrated Bispectrum Likelihood Ratio Tests, to appear in *Journal of the Acoustical Society of America*.

Ramírez, J.; Segura, J.C.; Benítez, C.; de la Torre, Á.; Rubio, A. (2004a). Efficient Voice Activity Detection Algorithms Using Long-Term Speech Information, *Speech Communication*, vol. 42, No. 3-4, pp. 271-287.

Ramírez, J.; Segura, J.C.; Benítez, C.; de la Torre, Á.; Rubio, A. (2005). An effective OSF-based VAD with Noise Suppression for Robust Speech Recognition, *IEEE Transactions on Speech and Audio Processing*, vol. 13, No. 6, pp. 1119-1129.

Ephraim Y.; Malah, D. (1984). Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, IEEE Trans. Acoustics, Speech and Signal Processing, vol. ASSP-32, no. 6, pp. 1109–1121.

Górriz, J.M.; Ramírez, J.; Segura, J.C.; Puntonet, C.G. (2006b). An effective cluster-based model for robust speech detection and speech recognition in noisy environments, *Journal of the Acoustical Society of America*, vol. 120, No. 1, pp. 470-481.

Ramírez, J.; Segura, J.C.; Benítez, C.; de la Torre, Á.; Rubio. A. (2004b). A New Kullback-Leibler VAD for Robust Speech Recognition, *IEEE Signal Processing Letters*, vol. 11, No. 2, pp. 266-269.

Beritelli, F.; Casale, S.; Rugeri, G.; Serrano, S. (2002). Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors, *IEEE Signal Processing Letters*, Vol. 9, No. 3, pp. 85–88.

Ramírez, J.; Yélamos, P.; Górriz, J.M; Segura, J.C. (2006b). SVM-based Speech Endpoint Detection Using Contextual Speech Features, *IEE Electronics Letters*, vol. 42, No. 7.

Estevez, P.A.; Becerra-Yoma, N.; Boric, N.; Ramirez, J.A. (2005). Genetic programming-based voice activity detection, *Electronics Letters*, Vol. 41, No. 20, pp. 1141- 1142.

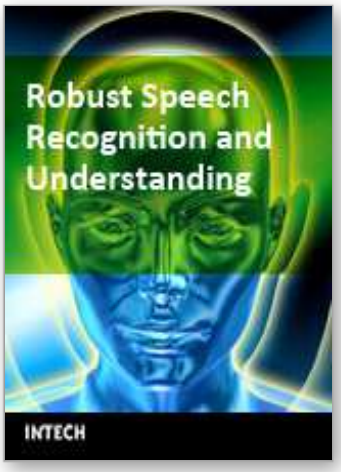
Ramírez, J.; Segura, J.C.; Benítez, C.; García, L.; Rubio, A. (2005b). Statistical Voice Activity Detection using a Multiple Observation Likelihood Ratio Test, *IEEE Signal Processing Letters*, vol. 12, No. 10, pp. 689-692.

Moreno, A.; Borge, L.; Christoph, D.; Gael, R.; Khalid, C.; Stephan, E.; Jeffrey, A. (2000). SpeechDat-Car: A large speech database for automotive environments, *Proc. II LREC Conf*.

Hirsch, H.G.; Pearce, D. (2000). 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*.

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

Young, S.; Odell, J.; Ollason, D.; Valtchev, V.; Woodland, P. (1997). The HTK Book. Cambridge, U.K.: Cambridge Univ. Press.



Robust Speech Recognition and Understanding

Edited by Michael Grimm and Kristian Kroschel

ISBN 978-3-902613-08-0
Hard cover, 460 pages
Publisher I-Tech Education and Publishing
Published online 01, June, 2007
Published in print edition June, 2007

This book on Robust Speech Recognition and Understanding brings together many different aspects of the current research on automatic speech recognition and language understanding. The first four chapters address the task of voice activity detection which is considered an important issue for all speech recognition systems. The next chapters give several extensions to state-of-the-art HMM methods. Furthermore, a number of chapters particularly address the task of robust ASR under noisy conditions. Two chapters on the automatic recognition of a speaker's emotional state highlight the importance of natural speech understanding and interpretation in voice-driven systems. The last chapters of the book address the application of conversational systems on robots, as well as the autonomous acquisition of vocalization skills.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

J. Ramirez, J. M. Gorriz and J. C. Segura (2007). Voice Activity Detection. Fundamentals and Speech Recognition System Robustness, Robust Speech Recognition and Understanding, Michael Grimm and Kristian Kroschel (Ed.), ISBN: 978-3-902613-08-0, InTech, Available from: http://www.intechopen.com/books/robust_speech_recognition_and_understanding/voice_activity_detection__fundamentals_and_speech_recognition_system_robustness

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821