3

New Advances in Voice Activity Detection using HOS and Optimization Strategies

J.M. Górriz, J. Ramírez, C.G. Puntonet
University of Granada
Spain

1. Introduction

Nowadays, the emerging wireless communication applications 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/non-speech detection is an unsolved problem in speech processing and affects numerous applications including robust speech recognition (Karray & Martin, 2003), (Ramírez et al., 2003), discontinuous transmission (ETSI, 1999), (ITU, 1996), estimation and detection of speech signals (Krasny, 2000), 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., 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 & Kondoz 2001) and have evaluated the influence of the VAD effectiveness on the performance of speech processing systems (Bouquin-Jeannes & Faucon, 1995) (see also the preceding chapter about VAD). 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), (Marzinzik & 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 (Marzinzik & Kollmeier, 2002), zero-crossing rate (ITU, 1996), periodicity measures (Tucker, 1992) or combinations of different features (ITU, 1996), (ETSI, 1999).

In this Chapter we show three methodologies for VAD: i) statistical likelihood ratio tests (LRTs) formulated in terms of the integrated bispectrum of the noisy signal. The integrated bispectrum is defined as a cross spectrum between the signal and its square, and therefore a function of a single frequency variable. It inherits the ability of higher order statistics to detect signals in noise with many other additional advantages (Górriz, 2006a), (Ramírez et al, 2006); ii) 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; and iii) an effective method employing support vector machines

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

(SVM) , a paradigm of learning from examples based in Vapkik-Chervonenkis theory (Vapnik, 1995). The use of kernels in SVM enables to map the data, via a nonlinear transformation, into some other dot product space (called feature space) in which the classification task is settled.

2. Relevant Feature Vectors for VAD

In this section we show some standard and novel Feature Vectors (FVs) for VAD. After the framing procedure, that is processing the input signal in short time frames, these FVs are computed in the feature extraction stage. Usually, the features are extracted using overlapping frames, which results in correlation between consecutive frames, and smoothes the spectral change from frame to frame. Feature extraction attempts to present the content of the speech signal compactly, such that the characteristic information of the signal is preserved. Then, the VAD decision is made using information provided by the features in the decision-making module.

2.1 Feature Vector based on Power Spectrum

Let x(n) be a discrete zero-mean time signal. In the framing stage the input signal x(n) sampled at 8 kHz is decomposed into 25-ms overlapped frames with a 10-ms window shift. The current frame consisting of 200 samples is zero padded to 256 samples and power spectral magnitude X(ω) is computed through the discrete Fourier transform (DFT). Finally, the filterbank reduces the dimensionality of the feature vector to a suitable representation for detection including broadband spectral information. Thus, the signal is passed through a K-band filterbank which is defined by:

$$E_B(k) = \sum_{\omega=\omega_k}^{\omega_{k+1}} X(\omega); \quad \omega_k = \frac{\pi}{K}k \quad k = 0,\ldots,K-1 \tag{1}$$

This feature in combination with long term information is usually adopted as feature vector, i.e. the one used in the clustering based VAD (Górriz et al., 2006b). In addition, it also provides the definition of other feature vector, the subband SNRs (that is used in the SVM approach), which includes the environmental level of noise and can be computed as:

$$SNR(k) = 20\log_{10}\left(\frac{E_B(k)}{N_B(k)}\right); \quad k = 0,\ldots,K-1 \tag{2}$$

where N_B denotes the subband power spectral magnitude of the residual noise that is extracted from the noisy channel using the approach presented in section 3.1.

2.2 Feature Vector based on HOS

The bispectrum of a discrete-time zero-mean signal x(t) is defined as the 2-D discrete time Fourier transform:

$$B_x(\omega_1,\omega_2) = \sum_{i=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} C_{3x}(i,k)\exp\{-j(\omega_1 i + \omega_2 k)\}; \tag{3}$$

where C_{3x} is the third-order cumulant of the input signal (third order moment of a zero-mean signal). Note that, from the above definition the third-order cumulant can be expressed as:

$$C_{3x}(i,k)=\left(\frac{1}{2\pi}\right)^2\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}B_x(\omega_1,\omega_2)\exp\{j(\omega_1 i+\omega_2 k)\}d\omega_1 d\omega_2 \tag{4}$$

Let denote $y(t)= x^2(t)-E\{x^2(t)\}$, the cross correlation between y(t) and x(t) is defined to be:

$$r_{yx}(i,k)=E\{y(t)x(t+k)\}=E\{x^2(t)x(t+k)\}=C_{3x}(0,k) \tag{5}$$

so that its cross spectrum is given by:

$$S_{yx}(\omega)=\sum_{k=-\infty}^{\infty}C_{3x}(0,k)\exp(-j\omega k) \tag{6}$$

and the reverse transformation is also satisfied:

$$C_{3x}(0,k)=\frac{1}{2\pi}\int_{-\pi}^{\pi}S_{yx}(\omega)\exp(j\omega k)d\omega \tag{7}$$

If we compare Eq. (4) with Eq. (7) we obtain:

$$S_{yx}(\omega)=\frac{1}{2\pi}\int_{-\pi}^{\pi}B_x(\omega_1,\omega_2)d\omega_2=\frac{1}{2\pi}\int_{-\pi}^{\pi}B_x(\omega_1,\omega_2)d\omega_1 \tag{8}$$

The integrated bispectrum is defined as a cross spectrum between the signal and its square, and therefore a function of a single frequency variable. Hence, its computation as a cross spectrum leads to significant computational savings, but more important is that the variance of the estimator is of the same order as that of the power spectrum estimator. On the other hand, Gaussian processes have vanishing third order moments so that the bispectrum and integrated bispectrum functions are zero as well, preserving their detection ability. From figure 1 it can be clearly concluded that higher order statistics or polyspectra provide discriminative features for speech/non-speech classification (Górriz et. al, 2005).
Given a finite data set {x(1), x(2), ..., x(N)} the integrated bispectrum is normally estimated by splitting the data set into blocks. Thus, the data set is divided into K_B non-overlapping blocks of data each of size M_B samples so that $N= K_B M_B$. Then, the cross periodogram of the ith block of data is given by

$$\hat{S}_{yx}^{i}(\omega)=\frac{1}{M_B}X^i(\omega)\left[Y^i(\omega)\right]^* \tag{9}$$

where $X^i(\omega)$ and $Y^i(\omega)$) denote the discrete Fourier transforms of x(t) and y(t) for the ith block. Finally, the estimate is obtained by averaging K_B blocks:

$$\hat{S}_{yx}(\omega) = \frac{1}{K_B} \sum_{i=1}^{K_B} \hat{S}_{yx}^{i}(\omega) \tag{10}$$

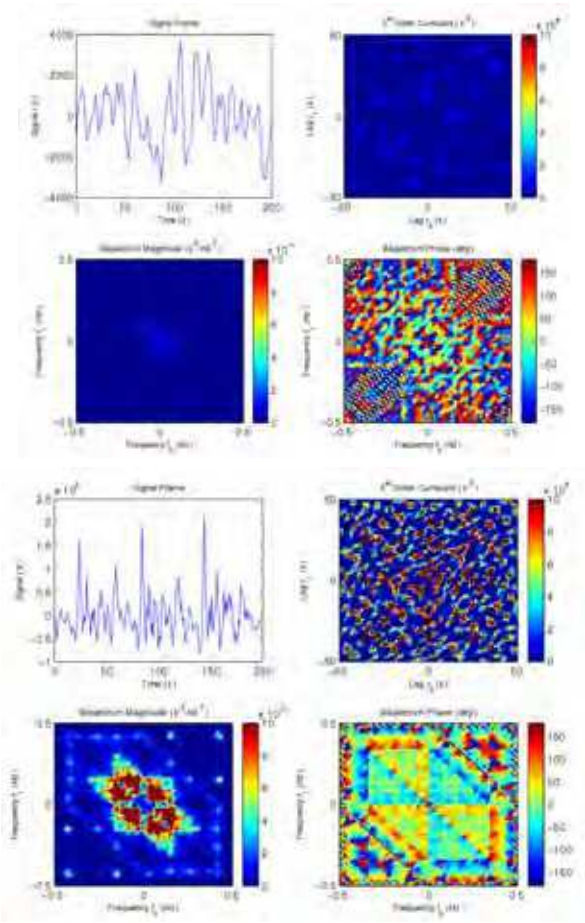


Figure 1. Third order statistics of a: left) noise only signal, right) speech signal corrupted by car noise.

3. Modelling the Noise subspace.

In the VAD problem, the background noise is usually assumed to be stationary over longer period of time than the speech signal. This enables to develop a smoothed noise model during a initialization period, and the estimation of noise statistics during non-speech periods for noise model update. Of course updating the noise parameters requires the information provided by the VAD decision in the previous frames (i.e. the non-speech periods). Then, the decision-making module can be divided to two separate parts: a primary decision, which makes the actual decision for the frame and a secondary decision, which

monitors the updating of the noise parameters. These parameters or noise model are related to the feature vectors used in the VAD algorithm, the decision-making module compares them with the actual FV to detect the current speech/non-speech period.

3.1 Spectral Noise Model

Assuming that the first N overlapped frames are nonspeech frames the noise level in the kth band, $N_B(k)$ can be estimated as the median of the set $\{N_B(0, k),..., N_B(N-1,k)\}$. In order to track non stationary noisy environments, the noise spectrum $N_B(k)$ is updated during non-speech periods by means of a 1st order IIR filter on the smoothed spectrum $X_s(k)$, that is:

$$N_B(l,k) = \lambda N_B(l-1,k) + (1-\lambda)X_s(l,k); \quad \lambda \cong 0.99 \tag{11}$$

where $X_s(k)$ is obtained averaging over consecutive frames and adjacent spectral bands, i.e. two consecutive frames and two adjacent bands, l is the frame index and λ is selected empirically. This update is applied to non-speech periods exclusively where $X_s(k)$ provides the instant noise characterization.

3.2 Clustering Based Model

Recently, cluster analysis has been applied to a set of noise spectral FVs to obtain a soft model for VAD (Górriz et al., 2006b). Given an initial set of noise FVs $\{N_B(j,k)\}$ with j=1,…,N, we apply hard decision-based clustering to assign them to a prespecified number of prototypes C<N, labelled by an integer i=1,…,C. This allocation is achieved in such a way that the *similarity measure* is minimized in terms of the squared Euclidean distance between noise energy vectors:

$$d\left(N_B(j), N_B(j')\right) = \sum_{k=1}^{K} \left|N_B(j,k) - N_B(j',k)\right|^2 = \left\|N_B(j) - N_B(j')\right\|_2; \tag{12}$$

and is defined as:

$$J(\Phi) = \frac{1}{2}\sum_{i=1}^{C} \sum_{\Psi(j)=i}^{C} \left\|N_B(j) - \tilde{N}_B(i)\right\|_2 = \tag{13}$$

where Ψ(j)=i denotes a many-to-one mapping, that assigns the jth observation to the ith prototype and

$$\tilde{N}_B(i) = mean\left(N_B(j)\right); \quad \forall j \,/\, \Phi(j) = i \in \{1,\ldots,C\} \tag{14}$$

is the mean vector associated with the ith prototype (the sample mean for the ith prototype. Thus, the loss function is minimized by assigning N noise spectral observations to C noise 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 modelled by C K-dimensional noise prototype vectors (see figure 2). We call this set of clusters C-partition or noise prototypes since, in this scenario, the word "cluster" is assigned to different classes of *labelled data*, that is **K** is fixed to 2, i.e. we define two clusters: "noise" and "speech" and the cluster "noise" consists of C prototypes (Górriz et al., 2006b).

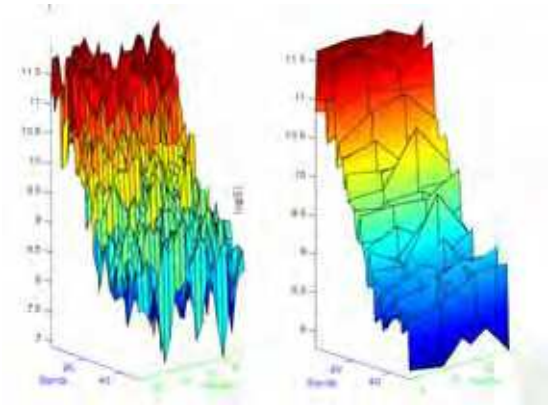


Figure 2. 20 noise log-energy frames, computed using N_{FFT}=256 and averaged over 50 subbands and clustering approach to the latter set of frames using hard decision C-means (C=4 prototypes).

4. Novel methodologies for VAD

The VAD decision is done in the decision-making module of the VAD algorithm, according to the decision rule. This section addresses the problem of voice activity detection formulated in terms of a classical binary hypothesis testing framework:

$$\begin{aligned} H_0 &: x(t) = n(t) \\ H_1 &: x(t) = s(t) + n(t) \end{aligned} \tag{15}$$

There are several ways of defining this decision rule which selects between both hypotheses, depending of the application where the VAD decision is employed. In the following, we show the most representative methodologies presented by the authors in several works over the past few years: i) multiple observation likelihood ratio tests (MO-LRT) for VAD over the integrated bispectrum (IBI) FV ii) the clustering distance based VAD using spectral FVs and iii) Support Vector Machines for solving a binary classification problem using subband SNRs as FVs.

4.1 Integrated Bispectrum based MO-LRT VAD

In a two-hypothesis test, the optimal decision rule that minimizes the error probability is the Bayes classifier. Given an observation vector $\mathbf{y}$ to be classified, the problem is reduced to selecting the class (H_0 or H_1) with the largest posterior probability $P(H_i \mid \mathbf{y})$. From the Bayes rule a statistical LRT can be defined by:

$$L(\mathbf{y}) = \frac{p_{\mathbf{y}|H_1}(\mathbf{y} \mid H_1)}{p_{\mathbf{y}|H_0}(\mathbf{y} \mid H_0)} \tag{16}$$

where the observation vector $\mathbf{y}$ is classified as H_1 if $L(\mathbf{y})$ is greater than $P(H_0)/P(H_1)$ otherwise it is classified as H_0. The LRT first proposed by Sohn (Sohn et al., 1999) for VAD,

which was defined on the power spectrum, is generalized and extended to the case where successive observations. The proposed multiple-observation likelihood ratio test (MO-LRT) formulates the decision for the central frame of a (2*m*+1)-observation buffer {y_{l-m}, ..., y_{l-1}, y_l, y_{l+1}, ..., y_{l+m}}:

$$L_{l,m}(\mathbf{y}) = \frac{p_{\mathbf{y}_{l-m},\ldots,\mathbf{y}_{l+m}|H_1}(\mathbf{y}_{l-m},\ldots,\mathbf{y}_{l+m} \mid H_1)}{p_{\mathbf{y}_{l-m},\ldots,\mathbf{y}_{l+m}|H_0}(\mathbf{y}_{l-m},\ldots,\mathbf{y}_{l+m} \mid H_0)} \tag{17}$$

where l denotes the frame being classified as speech (H_1) or non-speech (H_0). Note that, assuming statistical independence between the successive observation vectors, the corresponding log-LRT:

$$l_{l,m}(\mathbf{y}) = \sum_{k=l-m}^{l+m} \ln\left(\frac{p_{\mathbf{y}_k|H_1}(\mathbf{y}_k \mid H_1)}{p_{\mathbf{y}_k|H_0}(\mathbf{y}_k \mid H_0)}\right) \tag{18}$$

is recursive in nature and if each term of equation (18) is defined as Φ(k) the latter equation can be recursively calculated as:

$$l_{l+1,m}(\mathbf{y}) = l_{l,m}(\mathbf{y}) - \Phi(l-m) + \Phi(l+m+1) \tag{19}$$

The so called multiple observation LRT (MO-LRT) reports significant improvements in robustness as the number of observations increases. Assuming the integrated bispectrum $\{S_{yx}(\omega): \omega\}$ as the FV **y** and to be independent zero-mean Gaussian variables in presence and absence of speech with variances λ_1 and λ_0, resp. we obtain that the ratio of likelihoods can be expressed as:

$$\Phi(\mathbf{y}_k) = \sum_{\omega} \frac{\xi_k(\omega)\gamma_k(\omega),}{1+\xi_k(\omega)} - \log(1+\xi_k(\omega)) \tag{20}$$

where the *a priori* and *a posteriori* variance ratios are defined as:

$$\xi_k(\omega) \equiv \frac{\lambda_1^k(\omega)}{\lambda_0^k(\omega)} - 1; \quad \gamma_k(\omega) \equiv \frac{\left|S_{yx}^k(\omega)\right|}{\lambda_0^k(\omega)} \tag{21}$$

In order to evaluate the decision function the computation of variances under both hypotheses must be properly established. This is discussed further in (Ramírez et al., 2006) where a complete derivation of them is obtained in terms of the power spectrum of the noise (computed using the model presented in section 3.1) and the clean signal.

Figure 3 shows an example of the operation of the contextual MO-LRT VAD on an utterance of the Spanish SpeechDat-Car database (Moreno et al., 2000). The figure shows the decision variables for the tests defined by equation 20 and, alternatively, for the test with the second log-term in equation 20 suppressed (approximation) when compared to a fixed threshold η= 1.5. For this example, MB= 256 and m= 8. This approximation reduces the variance during non-speech periods. It can be shown that using an 8-frame window reduces the variability

of the decision variable yielding to a reduced noise variance and better speech/non-speech discrimination. On the other hand, the inherent anticipation of the VAD decision contributes to reduce the number of speech clipping errors.

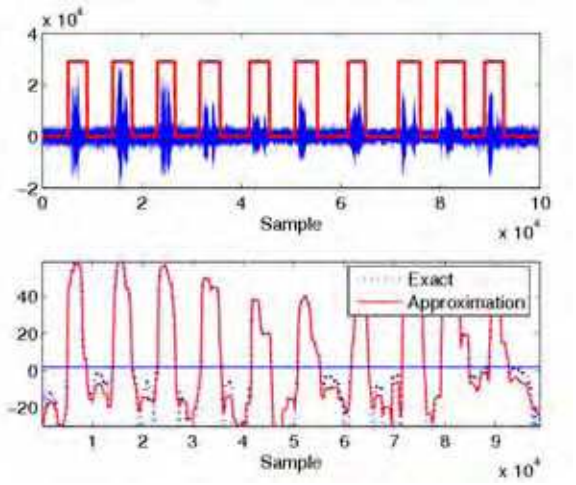


Figure 3. Operation of the MO-LRT VAD defined on the integrated bispectrum.

4.2 Long Term C-Means VAD

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 (Anderberg et al., 1973), image processing (Jain & Flynn, 1996) and information retrieval (Rasmussen, 1992). Clustering has a rich history in other disciplines 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. Consider noise model proposed in section 3.2 and let $\hat{E}(l)$ be the decision feature vector at frame l that is defined on the MO window as follows:

$$\hat{E}(l) \equiv \max\{E_B(j)\}; \quad j = l-m,\ldots,l+m \tag{22}$$

where $E_B(j)$ is defined in section 2.1 and the function *max* is applied in each frequency band k. The selection of this envelope feature vector, describing not only a single instantaneous frame but also a $(2m+1)$ entire neighbourhood, 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 & Kollmeier, 2002), (Li et al., 2002).

In the LTCM-VAD (Górriz et al., 2006b) the presence of the second "cluster" (speech frame) is detected if the following ratio holds:

$$\eta(l) \equiv \log\left(\frac{1}{K}\sum_{k=1}^{K}\frac{\hat{E}(k,l)}{\left\langle \widetilde{N}_B(i,k)\right\rangle}\right) > \gamma; \tag{23}$$

where $<\tilde{N}_B(i)>=(1/C)\Sigma^C_{i=1}\tilde{N}_B(i)=(1/C)\Sigma^C_{i=1}\Sigma^N_{j=1}\gamma_{ij}N_B(j)$ is the averaged noise prototype center defined in terms of the noise model presented in section 3.2 and γ is the decision threshold.

In order to adapt the operation of the proposed VAD to non-stationary and noise environments, the set of noise prototypes are updated according to the VAD decision during non-speech periods (not satisfying equation 23) in a competitive manner (only the closer noise prototype $\tilde{N}_B(l)$ at time l is moved towards the current feature vector $\hat{E}(l)$):

$$\begin{aligned} \widetilde{N}_B^C(l) &\equiv \arg_{\min}\left(\left\|\hat{E}(l)-\widetilde{N}_B(i)\right\|_2\right); \quad i=1,\ldots,C \\ \widetilde{N}_B^C(l+1) &= \alpha\widetilde{N}_B^C(l)+(1-\alpha)\hat{E}(l); \quad \alpha\approx 0.9 \end{aligned} \tag{24}$$

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

4.3 SVM enabled VAD

SVMs have recently been proposed for pattern recognition in a wide range of applications by its ability for learning from experimental data (Vapnik, 1995). The reason is that SVMs are much more effective than other conventional parametric classifiers. In SVM-based pattern recognition, the objective is to build a function f: $R^N \rightarrow \pm 1$ using training data that is, N-dimensional patterns $\mathbf{x}_i$ and class labels y_i:

$$(\mathbf{x}_1, y_1),(\mathbf{x}_2, y_2),\ldots,(\mathbf{x}_l, y_l) \in R^N \times \{\pm 1\} \tag{25}$$

so that f will correctly classify new examples ($\mathbf{x}$,y). The use of kernels in SVM enables to map the data into some other dot product space (called feature space) F via a nonlinear transformation Φ: $R^N \rightarrow F$ and perform the linear SVM algorithm (Vapnik, 1995) in F. The kernel is related to the Φ function by $k(\mathbf{x}_i,\mathbf{x}_j)= (\Phi(\mathbf{x}_i)\bullet\Phi(\mathbf{x}_j))$. In the input space, the hyperplane corresponds to a nonlinear decision function whose form is determined by the kernel. Thus, the decision function can be formulated in terms of a nonlinear function as:

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{\ell}\nu_i k(\mathbf{x}_i,\mathbf{x})+b\right) \tag{26}$$

where the parameters υ_i are the solution of a quadratic programming problem that are usually determined by the well known Sequential Minimal Optimization (SMO) algorithm (Platt, 1999). Many classification problems are always separable in the feature space and are

able to obtain better results by using RBF kernels instead of linear and polynomial kernel functions (Clarkson & Moreno, 1999).

The feature vector used in the SVM approach is the subband SNR described in section 2.1 but also including long term information: the long-term spectral envelope (LTSE) (Ramírez et al., 2006b) including contextual information of the speech signal is computed by: = max{$X_j(\omega_m)$, where j=*l*-L,…,l-1, l, l+1,…,l+L} and its dimensionality is reduced to a wide K-band spectral representation as in section 2.1.

An optional denoising process is applied to the input signal which consist of: i) Spectrum smoothing, ii) Noise estimation., iii) Wiener Filtering (WF) design and iv) Frequency domain filtering (Ramirez et al, 2006b). Once the SVM model is trained (i.e using the LIBSVM software tool (Chang & Lin, 2001); a training set consisting of 12 utterances of the AURORA 3 Spanish SpeechDat-Car (SDC) was used), we obtain the relevant support vectors and parameters in the classification of noise and speech features which enable to build the non-linear decision function shown in equation 26. The evaluation of the latter function is computational expensive, however evaluation methods have been also proposed (Ramírez et al. 2006b) in order to speed up the VAD decision. The method is based on a off-line computation of the decision rule over an input space grid and storing it in an *N*-dimensional look-up table. Finally, given a feature vector **x** we look for the nearest point in the grid previously defined and then perform a table look-up to assign a class (speech or non-speech) to the former feature vector.

Fig. 4..left shows the training data set in the 3-band input space. It is shown that the two classes can not be separated without error in the input space. Fig. 4.right shows the SVM decision rule that is obtained after the training process. Note that, i) the non-speech and speech classes are clearly distinguished in the 3-D space, and ii) the SVM model learns how the signal is masked by the noise and automatically defines the decision rule in the input space. Fig. 4. right also suggests a fast algorithm for performing the decision rule defined by equation 26 that becomes computationally expensive when the number of support vectors and/or the dimension of the feature vector are high. Note that all the information needed for deciding the class a given feature vector **x** belongs resides in figure 4.right. Thus, the input space can be discretized over the different components of the feature vector **x**.

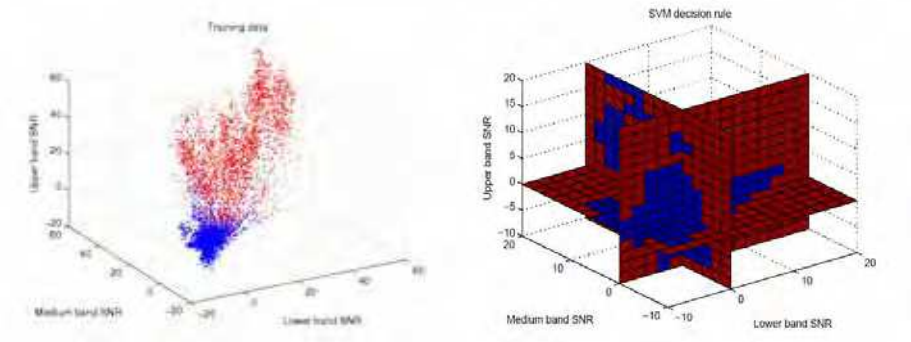


Figure 4. Classification rule in the input space after training a 3-band SVM model. left) Training data set, right) SVM classification rule.

5. Experimental Section.

Several experiments are commonly conducted in order 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 & Kollmeier, 2002) and the influence of the VAD decision on the performance of speech processing systems (Bouquin-Jeannes & Faucon, 1995). Subjective performance tests have also been considered for the evaluation of VADs working in combination with speech coders (Benyassine et al., 1997). This section describes the experimental framework and the objective performance tests conducted in this paper to evaluate the proposed algorithms.

5.1 Receiver operating characteristics (ROC) curves

The ROC curves are frequently used to completely describe the VAD error rate. They show the tradeoff between speech and non-speech detection accuracy as the decision threshold varies. The AURORA subset of the original Spanish SpeechDat-Car 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 25dB and 5dB. The non-speech hit rate (HR0) and the false alarm rate (FAR0= 100-HR1) were determined for each noise condition being the actual speech frames and actual speech pauses determined by hand-labelling the database on the close-talking microphone.

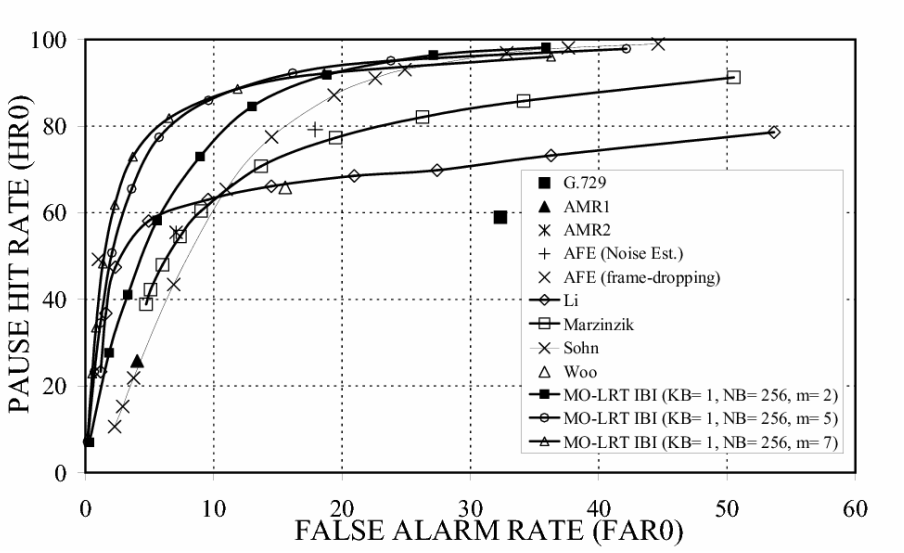


Figure 5. Classification rule in the input space after training a 3-band SVM model. left) Training data set, right) SVM classification rule.

Figure 5 shows the ROC curves of the proposed IBI-MO-LRT VAD and other frequently referred algorithms (Woo et al., 2000), (Li et al., 2002), (Marzinzik & Kollmeier, 2002), (Sohn

et al., 1999) for recordings from the distant microphone in high noisy conditions. The working points of the G.729, AMR and AFE standard VADs are also included. The efficient MO-LRT IBI VAD exhibits a shift of the ROC curve when the number of observations (m) increases as shown in figure 5. The method shows clear improvements in detection accuracy over standardized VADs and over a representative set of recently published VAD algorithms such us (Woo et al., 2000), (Li et al., 2002), (Marzinzik & Kollmeier, 2002), (Sohn et al, 1999).

The AURORA-2 database (Hirsch & Pearce, 2000) is also an adequate database for this analysis since it is built on the clean TIdigits 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 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 on a frame by frame basis for reference. Figure 6 provide comparative results of this analysis and compare the proposed LTCM VAD to standardized algorithms including the ITU-T G.729 (ITU, 1996), ETSI AMR (ETSI, 1999) and ETSI AFE (ETSI, 2002) and recently reported (Sohn et al., 1999), (Woo et al., 2000), (Li et al., 2002), (Marzinzik & Kollmeier, 2002) VADs in terms of speech hit-rate (HR1) 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 (WF) stage and non-speech frame-dropping (FD) are provided. The results shown in these figures are averaged values for the entire set of noises. Table 1 summarizes the advantages provided by LTCM VAD over the different VAD methods in terms of the average speech/non-speech hit-rates (over the entire range of SNR values). Thus, the proposed method with a mean of 97.57% HR1 and a mean of 47.81% HR0 yields the best trade-off in speech/non-speech detection.

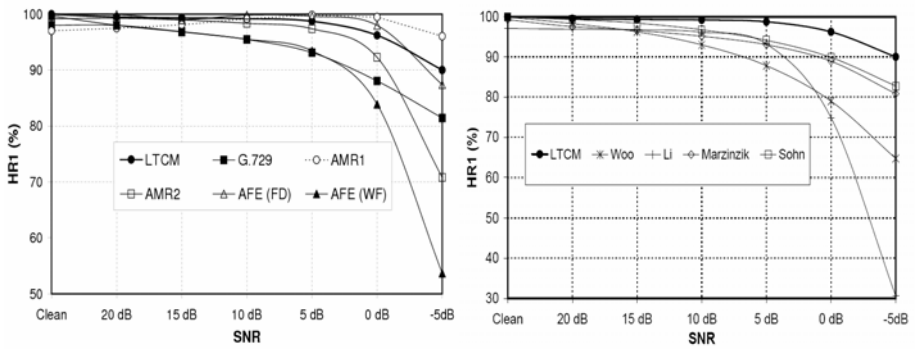


Figure 6. right) Speech hit rates (HR1) of standard VADs as a function of the SNR for the AURORA-2 database. Left) Speech hit rates (HR1) of other VADs as a function of the SNR for the AURORA-2 database.

	G729	AMR 1	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**
	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**	

Table 1. Average speech/non-speech hit rates for SNRs between clean conditions and -5 dB. Comparison to standardized VADs, and other VAD methods.

Figure 7 shows the ROC curves of the proposed SVM VAD, after a training process that consists of 12 utterances of the AURORA 3 Spanish SpeechDat-Car (SDC), using for varying L and K= 4 (number of subbands). It is shown that increasing the time span up to 8 frames also leads to a shift-up and to the left of the ROC curve. The optimal parameters for the proposed VAD are then K= 4 subbands and L= 8 frames. The results show significant improvements in speech/non-speech detection accuracy over standard VADs and over a representative set of VAD algorithms. These improvements are obtained by including contextual information in the feature vector and defining a non-linear decision rule over a wide band spectral representation of the data which enables a SVM-based classifier to learn how the speech signal is masked by the acoustic noise present in the environment.

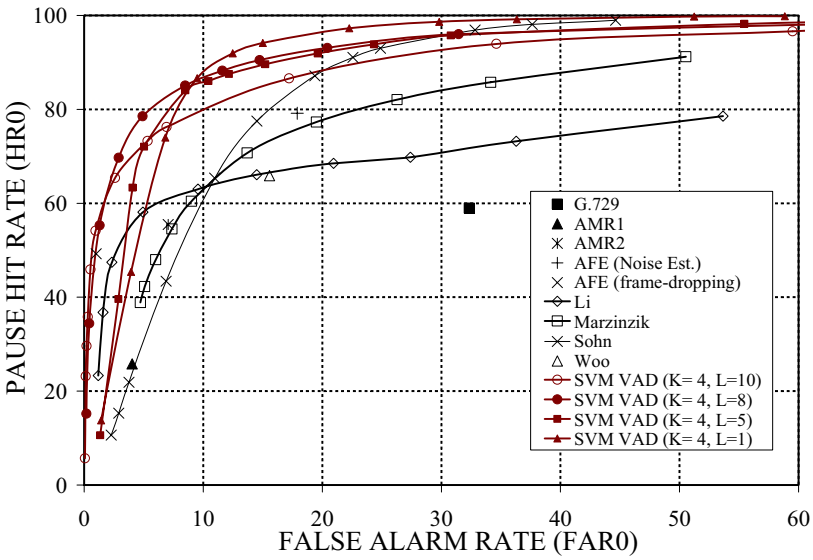


Figure 7. Influence of the time span L on the ROC curves. Comparison to standard and recently published VAD methods (High: high speed, good road, 5 dB average SNR).

5.2 Speech recognition experiments

Although the ROC curves are effective for VAD evaluation, 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 (Woo et al., 2000) since non-efficient speech/non-speech classification is an important source of the degradation of recognition performance in noisy environments (Karray & Martin, 2003). There are two clear motivations for that: i) noise parameters such as its spectrum are updated during non-speech periods being the speech enhancement system strongly influenced by the quality of the noise estimation, and ii) frame-dropping (FD), a frequently used technique in speech recognition to reduce the number of insertion errors caused by the noise, is based on the VAD decision and speech misclassification errors lead to loss of speech, thus causing irrecoverable deletion errors. This section evaluates the VAD according to the objective it was developed for, that is, by assessing the influence of the VAD in a speech recognition system.

The reference framework (base) considered for these experiments is the ETSI AURORA project for distributed speech recognition (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 modeled as whole word HMMs (Hidden Markov Models) with 16 states per word, simple left-to-right models and 3-gaussian mixtures per state (diagonal covariance matrix). Speech pause models consist of 3 states with a mixture of 6 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 recordings from all the driving conditions while testing is done using the hands-free microphone at low and high noise driving conditions. Finally, recognition performance is assessed in terms of the word accuracy (WAcc) that considers deletion, substitution and insertion errors.

Table 2 shows the recognition performance for the Spanish SpeechDat-Car database when WF and FD are performed on the base system (ETSI, 2000) using the IBI-MO-LRT and LTCM VADs. Our VADs outperform all the algorithms used for reference yielding relevant improvements in speech recognition. Note that, this particular database 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 (Ramírez et al., 2006), (Górriz et al., 2006b). This fact can be clearly shown when comparing the performance of the proposed VADs to Marzinzik VAD (Marzinzik & Kollmeier, 2002). The word accuracy of the VADs is quite similar for the AURORA 2 task. However, the proposed VADs yield a significant performance improvement over Marzinzik VAD (Marzinzik & Kollmeier, 2002) for the AURORA 3 database (Górriz et al., 2006b), (Ramírez et al., 2006).

	Base	Woo	Li	Marzinzik	Sohn	G729
WM	**92.94**	95.35	91.82	94.29	96.07	88.62
MM	**83.31**	89.30	77.45	89.81	91.64	72.84
HM	**51.55**	83.64	78.52	79.43	84.03	65.50
Ave.	**75.93**	89.43	82.60	87.84	90.58	75.65
	AMR1	AMR2	AFE	**LTCM**	**IBI-LRT**	
WM	94.65	95.67	95.28	**96.41**	**96.39**	
MM	80.59	90.91	90.23	**91.61**	**91.75**	
HM	62.41	85.77	77.53	**86.20**	**86.65**	
Ave.	74.33	90.78	87.68	**91.41**	**91.60**	

Table 1. Average speech/non-speech hit rates for SNRs between clean conditions and -5 dB. Comparison to standardized VADs, and other VAD methods.

6. Conclusion

This paper showed three different schemes for improving speech detection robustness and the performance of speech recognition systems working in noisy environments. These methods are based on: i) statistical likelihood ratio tests (LRTs) formulated in terms of the integrated bispectrum of the noisy signal. The integrated bispectrum is defined as a cross spectrum between the signal and its square, and therefore a function of a single frequency variable. It inherits the ability of higher order statistics to detect signals in noise with many other additional advantages; ii) 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; and iii) an effective method employing support vector machines (SVM) , a paradigm of learning from examples based in Vapkik-Chervonenkis theory. The use of kernels in SVM enables to map the data, via a nonlinear transformation, into some other dot product space (called feature space) in which the classification task is settled.

The proposed methods incorporate contextual information to the decision rule, a strategy that has reported significant improvements in speech detection accuracy and robust speech recognition applications. The optimal window size was determined by analyzing the overlap between the distributions of the decision variable and the error rate. The experimental analysis conducted on the well-known AURORA databases has reported significant improvements over standardized techniques such as ITU G.729, AMR1, AMR2 and ESTI AFE VADs, as well as over recently published VADs. The analysis assessed: i) the speech/non-speech detection accuracy by means of the ROC curves, with the proposed VADs yielding improved hit-rates and reduced false alarms when compared to all the reference algorithms, and ii) the recognition rate when the VADs are considered as part of a complete speech recognition system, showing a sustained advantage in speech recognition performance.

7. References

Karray, L. & Martin, A. (2003). Towards improving speech detection robustness for speech recognition in adverse environments, *Speech Communitation*, No. 3, (2003) (261-276), 0167-6393

Ramírez, J. ; Segura, J. C. ; Benítez, M.C.; de la Torre, A. & Rubio, A. (2003). A New Adaptive Long-Term Spectral Estimation Voice Activity Detector, *Proceedings of EUROSPEECH*, pp. 3041-3044, 1018-4074, Geneva, Switzerland, September 2003, ISCA.

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

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

Krasny, L, (2000). Soft-decision speech signal estimation, *The Journal of the Acoustical Society of America* Vol.108 (2000), (25-75), 0001-4966.

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, *Proceedings of the IEEE International Conference on High-Speed Networks and Multimedia Communications*, pp 46-50, July 2002.

Basbug, F. ; Swaminathan, K. & Nandkumar S. (2003). Noise reduction and echo cancellation front-end for speech codecs, *IEEE Transactions on Speech and Audio Processing* Vol. 11, (2003), (1-13) , 1063-6676.

Sohn, J.; Kim, N.S. & Sung, W. (1999) A statistical model-based voice activity detection, *IEEE Signal Processing Letters*, Vol. 16, No. 1, (1-3), 1070-9908.

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, (276-278), 1070-9908.

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, (1995), (245-254), 0167-6393.

Woo, K. ; Yang, T. ; Park, K. & Lee, C. (2000). Robust voice activity detection algorithm for estimating noise spectrum, Electronics Letters, Vol 36, No. 2, (2000), (180-181) 0013-5194.

Li, Q. ; Zheng, J. ; Tsai, A. &Zhou, Q. (2002). Robust endpoint detection and energy normalization for real-time speech and speaker recognition, *IEEE Transactions on Speech and Audio Processing*, Vol 10, No.3, (2002) (146-157), 1063-6676.

Marzinzik, M. & Kollmeier, B. (2002). Speech pause detection for noise spectrum estimation by tracking power envelope dynamics, *IEEE Transactions on Speech and Audio Processing*, Vol 10, No. 6, (2002), (341-351), 1063-6676.

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

Tucker, R. (1992). Voice activity detection using a periodicity measure, IEE Proceedings, Communications, Speech and Vision, Vol. 139, No. 4 (1992), (377-380). 1350-2425.

Górriz, J.M. (2006a). New Advances in Voice Activiy Detection. Ph-D., 84-338-3863-6, University of Granada, July 2006, Granada.

Ramírez, J.; Górriz J.M. & Segura J.C. (2006). Statistical Voice Activity Detection Based on Integrated Bispectrum Likelihood Ratio Tests for Robust Speech Recognition. In press *The Journal of the Acoustical Society of America,* Vol.X, No.X, (2006) (XXX-XXX), 0001-4966.

Vapnik, V.N. (1995). The Nature of Statistical Learning Theory. Springer-Verlag, 9780387945590, Berlin, (1995).

Górriz, J.M. ; Ramírez, J. ; Puntonet, C.G. & Segura, J.C (2006b). An effective cluster-based model for robust speech detection and speech recognition in noisy environments. *The Journal of Acoustical Society of America*. No. 120, Vol. 470 (2006). (470-481). 0001-4966.

Górriz, J.M. ; Ramírez, J. ; Segura, J.C. & Puntonet, C.G. (2005). Improved MO-LRT VAD based on bispectra Gaussian model, *Electronics Letters*, Vol. 41, No. 15, (2005), (877-879).

Moreno, A. ; Borge, L. ; Christoph, D. ; Gael, R. ; Khalid, C. ; Stephan, E. & Jeffrey, A. (2000). SpeechDat-Car: A Large Speech Database for Automotive Environments. *Proceedings of the II Second International Conference on Language Resources and Evaluation Conference*, May 2000, Athens.

Jain, A. & Flynn, P. (1996) Image segmentation using clustering. In Advances in Image Understanding. A Festschrift for Azriel Rosenfeld, N. Ahuja and K. Bowyer, (Ed.), (65—83) , IEEE Press, Piscataway, NJ.

Anderberg, M.R.; Odell, J. ; Ollason, D. ; Valtchev, V. & Woodland, P. (1973). Cluster Analysis for Applications. Academic Press, Inc., 0120576503. New York, NY.

Rasmussen, E. (1992). Clustering algorithms, In Information Retrieval: Data Structures and

Algorithms, W. B. Frakes and R. Baeza-Yates, (Ed.). (419-442) Prentice-Hall, Inc., Upper Saddle River, NJ.

Platt, J.C. (1999). Fast Training of Support Vector Machines using Sequential Minimal Optimization In Advances in Kernel Methods - Support Vector Learning, (185-208). MIT Press.

Clarkson, P, & Moreno P.J. (1999). On the use of support vector machines for phonetic classification, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol 2, March 1999, pp 585-588. Phoenix.

Ramírez, J.; Yélamos, P.; Górriz, J.M. & Segura J.C. (2006b). SVM-based Speech Endpoint Detection Using Contextual Speech Features. *Electronic Letters* Vol 42 No.7 (65-66) 0013-5194.

Chang, C.C. & Lin, C.J. (2001), LIBSVM: a library for support vector machines, Dept. of Computer Science and Information Engineering, National Taiwan University, (2001, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

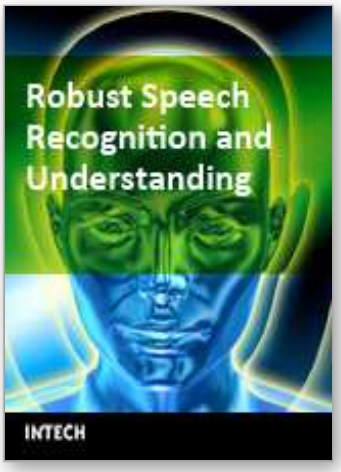
Benyassine, A.; Shlomot, E.; Su, H.; Massaloux, D.; Lamblin, C. & Petit, J (1997). ITU-T Recommendation G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications. *IEEE Communications Magazine*. Vol 35, (64-73) .

Hirsch H. & Pearce, D. (2000). The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions, In ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium. Paris, France.

ETSI, (2002). Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms, ETSI ES 202 050 Recommendation).

ETSI, (2000). Speech processing, transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms, ETSI ES 201 108 Recommendation.

Young, S; Odell, J.; Ollason, D.; Valtchev, V. & Woodland, P. (1997) The HTK Book. Cambridge University.



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.

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