

# Voice Activity Detection Using Higher Order Statistics

J.M. Górriz, J. Ramírez, J.C. Segura, and S. Hornillo

Dept. Teoría de la Señal, Telemática y comunicaciones,  
Facultad de Ciencias , Universidad de Granada,  
Fuentenueva s/n, 18071 Granada, Spain  
`gorriz@ugr.es`

**Abstract.** A robust and effective voice activity detection (VAD) algorithm is proposed for improving speech recognition performance in noisy environments. The approach is based on filtering the input channel to avoid high energy noisy components and then the determination of the speech/non-speech bispectra by means of third order autocumulants. This algorithm differs from many others in the way the decision rule is formulated (detection tests) and the domain used in this approach. Clear improvements in speech/non-speech discrimination accuracy demonstrate the effectiveness of the proposed VAD. It is shown that application of statistical detection test leads to a better separation of the speech and noise distributions, thus allowing a more effective discrimination and a tradeoff between complexity and performance. The algorithm also incorporates a previous noise reduction block improving the accuracy in detecting speech and non-speech.

## 1 Introduction

Nowadays speech/non-speech detection is a complex problem in speech processing and affects numerous applications including robust speech recognition [1], discontinuous transmission [2, 3], real-time speech transmission on the Internet [4] or combined noise reduction and echo cancellation schemes in the context of telephony [5]. 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 [6] and have evaluated the influence of the VAD effectiveness on the performance of speech processing systems [7]. 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 [8, 9, 10]. The different approaches include those based on energy thresholds [8], pitch detection [11], spectrum analysis [10], zero-crossing rate [3], periodicity measure [12], higher order statistics in the LPC residual domain [13] or combinations of different features [3, 2]. This paper explores a new alternative towards improving speech detection robustness in adverse environments and the performance of speech recognition systems. The proposed VAD proposes a noise

reduction block that precedes the VAD, and uses Bispectra of third order cumulants to formulate a robust decision rule. The rest of the paper is organized as follows. Section 2 reviews the theoretical background on Bispectra analysis and shows the proposed signal model, analyzing the motivations for the proposed algorithm by comparing the speech/non-speech distributions for our decision function based on bispectra and when noise reduction is optionally applied. Section 3 describes the experimental framework considered for the evaluation of the proposed statistical decision algorithm. Finally, section summarizes the conclusions of this work.

## 2 Model Assumptions

Let  $\{x(t)\}$  denote the discrete time measurements at the sensor. Consider the set of stochastic variables  $y_k$ ,  $k = 0, \pm 1 \dots \pm M$  obtained from the shift of the input signal  $\{x(t)\}$ :

$$\mathbf{y}_k(t) = \mathbf{x}(t + k \cdot \tau) \tag{1}$$

where  $k \cdot \tau$  is the differential delay (or advance) between the samples. This provides a new set of  $2 \cdot m + 1$  variables by selecting  $n = 1 \dots N$  samples of the input signal which can be represented using the associated Toeplitz matrix.

Using this model the speech-non speech detection can be described by using two essential hypothesis(re-ordering indexes):

$$H_o = \begin{pmatrix} \mathbf{y}_0 = n_0 \\ \mathbf{y}_{\pm 1} = n_{\pm 1} \\ \dots \\ \mathbf{y}_{\pm M} = n_{\pm M} \end{pmatrix}; \quad H_1 = \begin{pmatrix} \mathbf{y}_0 = s_0 + n_0 \\ \mathbf{y}_{\pm 1} = s_{\pm 1} + n_{\pm 1} \\ \dots \\ \mathbf{y}_{\pm M} = s_{\pm M} + n_{\pm M} \end{pmatrix} \tag{2}$$

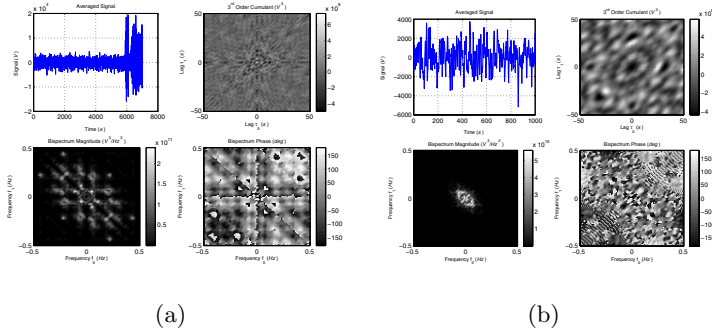
where  $s_k$ 's/ $n_k$ 's are the speech/non-speech (any kind of additive background noise i.e. gaussian) signals, related themselves with some differential parameter. All the process involved are assumed to be jointly stationary and zero-mean. Consider the third order cumulant function  $C_{y_k y_l}$  defined as:

$$C_{y_k y_l} \equiv E[y_0 y_k y_l]; \quad C_{y_k y_l}(\omega_1, \omega_2) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} C_{y_k y_l} \cdot \exp(-j(\omega_1 k + \omega_2 l)) \tag{3}$$

and the two-dimensional discrete Fourier transform (DFT) of  $C_{y_k y_l}$ , the bispectrum function. The sequence of cumulants of the voice speech is modelled as a sum of coherent sine waves:

$$C_{y_k y_l} = \sum_{n,m=1}^K a_{nm} \cos[kn\omega_0^1 + lm\omega_0^2] \tag{4}$$

where  $a_{nm}$  is amplitude,  $K \times K$  is the number of sinusoids and  $\omega$  is the fundamental frequency in each dimension. It follows from equation 4 that  $a_{mn}$  is related to the energy of the signal  $\mathcal{E}_s = E\{s^2\}$ . The VAD proposed in the later



**Fig. 1.** Different Features allowing voice activity detection. (a) Features of Voice Speech Signal. (b) Features of non Speech Signal

reference only works with the coefficients in the sequence of cumulants and is more restrictive in the model of voice speech. Thus the Bispectra associated to this sequence is the DTF of equation 4 which consist in a set of Dirac’s deltas in each excitation frequency  $n\omega_0^1, m\omega_0^2$ . Our algorithm will detect any high frequency peak on this domain matching with voice speech frames, that is under the above assumptions and hypotheses, it follows that on  $H_0$ ,

$$C_{y_k y_l}(\omega_1, \omega_2) \equiv C_{n_k n_l}(\omega_1, \omega_2) \simeq 0 \tag{5}$$

and on  $H_1$ :

$$C_{y_k y_l}(\omega_1, \omega_2) \equiv C_{s_k s_l}(\omega_1, \omega_2) \neq 0 \tag{6}$$

Since  $s_k(t) = s(t + k \cdot \tau)$  where  $k = 0, \pm 1 \dots \pm M$ , we get

$$C_{s_k s_l}(\omega_1, \omega_2) = \mathcal{F}\{E[s(t + k \cdot \tau)s(t + l \cdot \tau)s(t)]\} \tag{7}$$

The estimation of the bispectra (equation 3) is deep discussed in [14] and many others, where conditions for consistency are given. The estimate is said to be (asymptotically) consistent if the squared deviation goes to zero, as the number of samples tends to infinity.

**2.1 Detection Tests for Voice Activity**

The decision of our algorithm implementing the VAD is based on statistical tests from references [15] (Generalized likelihood ratio tests) and [16] (Central  $\chi^2$ -distributed test statistic under  $H_0$ ). We will call the tests GLRT and  $\chi^2$  tests. The tests are based on some asymptotic distributions and computer simulations in [17] show that the  $\chi^2$  tests require larger data sets to achieve a consistent theoretical asymptotic distribution. Then we decline to use it unlike the GLRT tests.

If we reorder the components of the set of  $L$  Bispectrum estimates  $\hat{C}(n_l, m_l)$  where  $l = 1, \dots, L$ , on the fine grid around the bifrequency pair into a  $L$  vector  $\beta_{ml}$  where  $m = 1, \dots, P$  indexes the coarse grid [15] and define  $P$ -vectors

$\phi_i(\beta_{1i}, \dots, \beta_{Pi})$ ,  $i = 1, \dots, L$ ; the generalized likelihood ratio test for the above discussed hypothesis testing problem:

$$H_0 : \mu = 0 \quad \text{against} \quad H_1 : \eta \equiv \mu^T \sigma^{-1} \mu > 0 \quad (8)$$

where  $\mu = 1/L \sum_{i=1}^L \phi_i$  and  $\sigma = 1/L \sum_{i=1}^L (\phi_i - \mu)(\phi_i - \mu)^T$ , leads to the activity voice speech detection if:

$$\eta > \eta_0 \quad (9)$$

where  $\eta_0$  is a constant i.e. the probability of false alarm.

## 2.2 Noise Reduction Block

Almost any VAD can be improved just placing a noise reduction block in the data channel before it. The noise reduction block for high energy noisy peaks, consists of four stages(1) Spectrum smoothing 2)Noise estimation 3)Wiener Filter (WF) design and 4)Frequency domain filtering) and was first developed in [18].

## 2.3 Some Remarks About the Algorithm

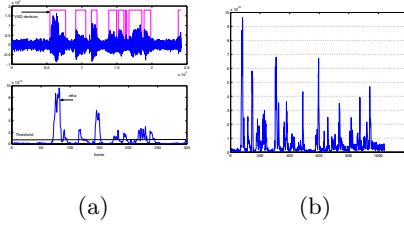
We propose a alternative decision based on an average of the components of the bispectrum (the absolute value of it). In this way we define  $\eta$  as:

$$\eta = \frac{1}{L \cdot N} \sum_{i=1}^L \sum_{j=1}^N \left| \hat{C}(i, j) \right| \quad (10)$$

where  $L, N$  defines the selected grid (high frequencies with noteworthy variability). We also include long term information (LTI) in the decision of the on-line VAD [19] which essentially improves the efficiency of the proposed method as is shown the following pseudocode:

- Initialize variables
- Determine  $\eta_0$  of noise in the first frame
- for  $i=1$  to end:
  1. Consider a new frame ( $i$ )
    - calculate  $\eta(i)$
  2. if  $H_1$  then
    - VAD( $i$ )=1
    - apply LTI to VAD( $i-\tau$ )
  - else
    - Slow Update of noise parameters:  $\eta_0(i+1) = \alpha\eta_0 + \beta\eta(i)$ ,  
 $\alpha + \beta = 1 \quad \alpha \rightarrow 1$
    - apply LTI to VAD( $i-\tau$ )

Fig. 2 shows the operation of the proposed VAD on an utterance of the Spanish SpeechDat-Car (SDC) database [20]. The phonetic transcription is: [“siete”, “thinko”, “dos”, “uno”, “otSo”, “seis”]. Fig 2(b) shows the value of  $\eta$  versus time. Observe how assuming  $\eta_0$  the initial value of the magnitude  $\eta$  over the

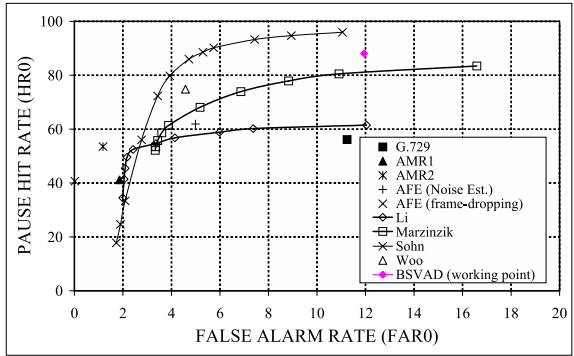


**Fig. 2.** Operation of the VAD on an utterance of Spanish SDC database. (a) Evaluation of  $\eta$  and VAD Decision. (b) Evaluation of the test hypothesis on an example utterance of the Spanish SpeechDat-Car (SDC) database [20]

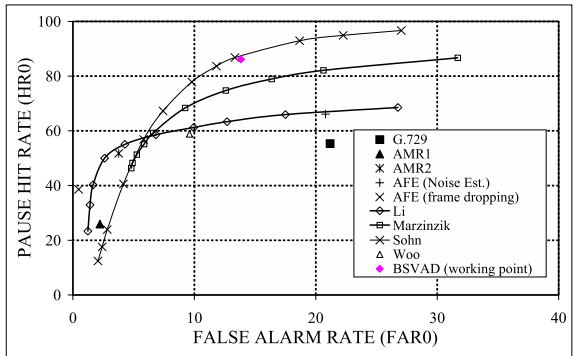
first frame (noise), we can achieve a good VAD decision. It is clearly shown how the detection tests yield improved speech/non-speech discrimination of fricative sounds by giving complementary information. The VAD performs an advanced detection of beginnings and delayed detection of word endings which, in part, makes a hang-over unnecessary. In Fig 1 we display the differences between noise and voice in general and in figure we settle these differences in the evaluation of  $\eta$  on speech and non-speech frames.

### 3 Experimental Framework

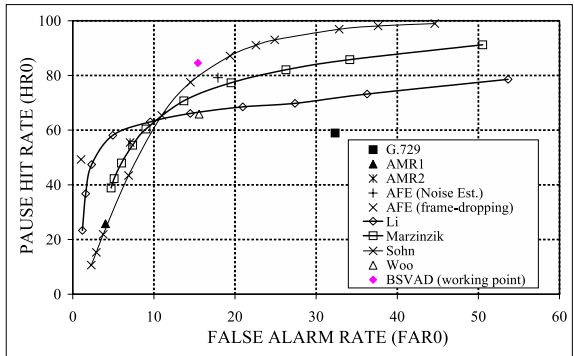
The ROC curves are frequently used to completely describe the VAD error rate. The AURORA subset of the original Spanish SpeechDat-Car (SDC) database [20] was used in this analysis. This database contains 4914 recordings using close-talking and distant microphones from more than 160 speakers. The files 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 in each noise condition being the actual speech frames and actual speech pauses determined by hand-labelling the database on the close-talking microphone. Fig. 3 shows the ROC curves of the proposed VAD (BiSpectra based-VAD) and other frequently referred algorithms [8, 9, 10, 6] for recordings from the distant microphone in quiet, low and high noisy conditions. The working points of the G.729, AMR and AFE VADs are also included. The results show improvements in detection accuracy over standard VADs and similarities over representative set VAD algorithms [8, 9, 10, 6]. 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. On average ( $\frac{HR0+HR1}{2}$ ), the proposed VAD is similar to Marzinik'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. These results clearly demonstrate that there is no optimal VAD for all the applications. Each VAD is developed and optimized for specific purposes. Hence, the evaluation has to be conducted according to the



(a)



(b)



(c)

**Fig. 3.** ROC curves obtained for different subsets of the Spanish SDC database at different driving conditions: (a) Quiet (stopped car, motor running, 12 dB average SNR). (b) Low (town traffic, low speed, rough road, 9 dB average SNR). (c) High (high speed, good road, 5 dB average SNR)

**Table 1.** Average speech/non-speech hit rates for SNRs between 25dB and 5dB. Comparison of the proposed BSVAD to standard and recently reported VADs

	G.729	AMR1	AMR2	AFE (WF)	AFE (FD)
HR0 (%)	55.798	51.565	57.627	69.07	33.987
HR1 (%)	88.065	98.257	97.618	85.437	99.750
	Woo	Li	Marzinzik	Sohn	<b>BSVAD</b>
HR0 (%)	62.17	57.03	51.21	66.200	85.150
HR1 (%)	94.53	88.323	94.273	88.614	86.260

specific goal of the VAD. Frequently, VADs avoid losing speech periods leading to an extremely conservative behavior in detecting speech pauses (for instance, the AMR1 VAD). Thus, in order to correctly describe the VAD performance, both parameters have to be considered. On average the results are conclusive (see table 1).

## 4 Conclusion

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

## Acknowledgements

This work has received research funding from the EU 6<sup>th</sup> Framework Programme, under contract number IST-2002-507943 (HIWIRE, Human Input that Works in Real Environments) and SESIBONN project (TEC2004-06096-C03-00) from the Spanish government. The views expressed here are those of the authors only. The Community is not liable for any use that may be made of the information contained therein.

## References

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

2. ETSI, "Voice activity detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels," *ETSI EN 301 708 Recommendation*, 1999.
3. ITU, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," *ITU-T Recommendation G.729-Annex B*, 1996.
4. A. Sangwan, M. C. Chiranth, H. S. Jamadagni, R. Sah, R. V. Prasad, and V. Gau-rav, "VAD techniques for real-time speech transmission on the Internet," in *IEEE International Conference on High-Speed Networks and Multimedia Communica-tions*, 2002, pp. 46–50.
5. S. Gustafsson and et al., "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Trans. on S.&A. Proc.*, vol. 10, no. 5, pp. 245–256, 2002.
6. J. Sohn and et al., "A statistical model-based vad," *IEEE S.Proc.L.*, vol. 16, no. 1, pp. 1–3, 1999.
7. R. L. Bouquin-Jeannes and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech Communication*, vol. 16, pp. 245–254, 1995.
8. K. Woo and et al., "Robust vad algorithm for estimating noise spectrum," *Elec-tronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.
9. Q. Li and et al., "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Trans. on S.&A. Proc.*, vol. 10, no. 3, pp. 146–157, 2002.
10. M. Marzinzik and et al., "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. on S.&A. Proc.*, vol. 10, no. 6, pp. 341–351, 2002.
11. R. Chengalvarayan, "Robust energy normalization using speech/non-speech dis-criminator for German connected digit recognition," in *Proc. of EUROSPEECH 1999*, Budapest, Hungary, Sept. 1999, pp. 61–64.
12. R. Tucker, "Vad using a periodicity measure," *IEE Proceedings, Communications, Speech and Vision*, vol. 139, no. 4, pp. 377–380, 1992.
13. E. Nemer and et al., "Robust vad using hos in the lpc residual domain," *IEEE Trans. S.&A. Proc.*, vol. 9, no. 3, pp. 217–231, 2001.
14. D. Brillinger and et al., *Spectral Analysis of Time Series*. Wiley, 1975, ch. Asymp-totic theory of estimates of kth order spectra.
15. T. S. Rao, "A test for linearity of stationary time series," *Journal of Time Series Analysis*, vol. 1, pp. 145–158, 1982.
16. J. Hinich, "Testing for gaussianity and linearity of a stationary time series," *Journal of Time Series Analysis*, vol. 3, pp. 169–176, 1982.
17. J. Tugnait, "Two channel tests for common non-gaussian signal detection," *IEE Proceedings-F*, vol. 140, pp. 343–349, 1993.
18. J. Ramírez and et. al., "An effective subband osf-based vad with noise reduction for robust speech recognition," *In press IEEE Trans. on S.&A. Proc.*
19. J. Ramírez, J. C. Segura, M. C. Benítez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3-4, pp. 271–287, 2004.
20. A. Moreno and et al., "SpeechDat-Car: A Large Speech Database for Automotive Environments," in *II LREC Conference*, 2000.