Bispectrum Estimators for Voice Activity Detection and Speech Recognition

J.M. Górriz, C.G. Puntonet, J. Ramírez, and J.C. Segura

E.T.S.I.I., Universidad de Granada,
C/ Periodista Daniel Saucedo, 18071 Granada, Spain
gorriz@ugr.es

Abstract. A new Bispectra Analysis application is presented is this paper. A set of bispectrum estimators for robust and effective voice activity detection (VAD) algorithm are 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 auto-cumulants. 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. The experimental analysis carried out on the AURORA databases and tasks provides an extensive performance evaluation together with an exhaustive comparison to the standard VADs such as ITU G.729, GSM AMR and ETSI AFE for distributed speech recognition (DSR), and other recently reported VADs.

1 Introduction

Speech/non-speech detection is an unsolved 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, 7] and have evaluated the influence of the VAD effectiveness on the performance of speech processing systems [8]. 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 [9, 10, 11]. The different approaches include those based on energy

thresholds [9], pitch detection [12], spectrum analysis [11], zero-crossing rate [3], periodicity measure [13], higher order statistics in the LPC residual domain [14] 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. Section 3 analyzes the statistical tests used in this aproach and compare the speech/non-speech distributions for our decision function based on bispectra and when noise reduction is optionally applied (see section 4). Section 5 describes the experimental framework considered for the evaluation of the proposed end-point detection algorithm. Finally, section 6 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, $k = 0, \pm 1 \ldots \pm M$ 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 $n = 1 \ldots N$ samples of the input signal. It can be represented using the associated Toeplitz matrix:

$$T_{x(t_0)} = \begin{pmatrix} y_{-M}(t_0) & \ldots & y_{-m}(t_N) \\ y_{-M+1}(t_0) & \ldots & y_{-M+1}(t_N) \\ \ldots & \ldots & \ldots \\ y_M(t_0) & \ldots & y_M(t_N) \end{pmatrix} \tag{2}$$

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

$$H_o = \begin{pmatrix} \mathbf{y}_0 = n_0 \\ \mathbf{y}_{\pm 1} = n_{\pm 1} \\ \ldots \\ \mathbf{y}_{\pm M} = n_{\pm M} \end{pmatrix} \tag{3}$$

$$H_1 = \begin{pmatrix} \mathbf{y}_0 = s_0 + n_0 \\ \mathbf{y}_{\pm 1} = s_{\pm 1} + n_{\pm 1} \\ \ldots \\ \mathbf{y}_{\pm M} = s_{\pm M} + n_{\pm M} \end{pmatrix} \tag{4}$$

where s_k's/n_k's are the speech (see section /refsec: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:

$$C_{y_k y_l} \equiv E[y_0 y_k y_l] \qquad (5)$$

and the two-dimensional discrete Fourier transform (DFT) of $C_{y_k y_l}$, the bispectrum function:

$$\mathcal{C}_{y_k y_l}(\omega_1, \omega_2) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} C_{y_k y_l} \cdot \exp(-j(\omega_1 k + \omega_2 l))) \qquad (6)$$

2.1   Bispectrum Estimators

The set of estimators used in the statistical tests of section 3 are described in the following[1].

Indirect Methods: Sampling the equation 6 and assuming a finite number of samples, the "indirect" bispectrum estimator can be written as:

$$\hat{\mathcal{C}}_{\mathbf{y}_k \mathbf{y}_l}(n, m) = \sum_{k=-M}^{M} \sum_{l=-M}^{M} C_{\mathbf{y}_k \mathbf{y}_l} \cdot w(k, l) \cdot \exp(-j(\omega_n k + \omega_m l)) \qquad (7)$$

where $\omega_{n,m} = \frac{2\pi}{M}(n, m)$ with $n, m = -M, \ldots, M$ are the sampling frequencies, $w(k, l)$ is the window function (to get smooth estimates [15]) and $C_{\mathbf{y}_k \mathbf{y}_l} = \frac{1}{N} \sum_{i=0}^{N-1} y_0(t_i) y_k(t_i) y_l(t_i) = \frac{1}{N} \mathbf{y}_0 \mathbf{y}_k \mathbf{y}_l |_{t_0}$. Under the assumption that the bispectrum $C_{\mathbf{y}_k \mathbf{y}_l}$ is sufficiently smooth, the smoothed estimate is known to be consistent, with variance given by:

$$var\left(\hat{\mathcal{C}}_{\mathbf{y}_k \mathbf{y}_l}(n, m)\right) = \frac{1}{N} S_{\mathbf{y}_0} S_{\mathbf{y}_k} S_{\mathbf{y}_l} \int \int w(t, s) dt ds \qquad (8)$$

where S is the power spectrum. That is, the data are segmented into possibly overlapping records; biased or unbiased sample estimates of third-order cumulants are computed for each record and then averaged across records; a lag window is applied to the estimated cumulants, and the bispectrum is obtained as the $2-D$ FFT (fast fourier transform) of the windowed cumulant function. This is the classical method for estimating the Bispectrum function which is known to be consistent. An alternative approach is to perform the smoothing $W(\omega_n, \omega_m)$ in the frequency domain.

Direct Methods: The "direct" class of methods for higher-order spctrum estimation are similar to the "averaged periodogram" or Welch method for power spectrum estimation [16]. In this approach the data are segmented into

---

[1] A deep discussion can be found in the "HOSA" software toolbox (Higher-Order Spectral Analysis Toolbox User's Guide) by Ananthram Swami, Jerry M. Mendel and Chrysostomos L. (Max) Nikias. http://www.mathworks.com/hosa.html

possibly overlapping records; the mean is removed from each record, and the FFT computed; the bispectrum of the Kth record is computed as:

$$\hat{\mathcal{C}}^K_{\mathbf{y}_k \mathbf{y}_l}(n, m) = \mathbf{Y}_K(m)\mathbf{Y}_K(n)\mathbf{Y}_K(m+n) \tag{9}$$

where $\mathbf{Y}_K$ denotes the FFT of the Kth record. The bispectral estimates are averaged across records, and an optional frequency-domain smoother is also applied. Brillinger [17] pointed out that the higher order periodogram is asymptotically "unbiased", and "consistent" if neighbor frequency smoothing is applied.

Integrated Bispectrum: Various VAD algorithms used to date use an averaged Bispectrum function to obtain the decision rule, i.e. in [18, 19]. The following estimation based on Tugnait's work [20] where the connection between the cross-spectrum of a given signal $x(t)$ and its square $y(t) = x^2(t) - E(x^2(t))$ and the integrated bispectrum of the signal is established:

$$S_{yx}(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathcal{C}_x(\omega, \widetilde{\omega}) d\widetilde{\omega} = \sum_{k=-\infty}^{+\infty} E(y(t)x(t+k)) \exp(j\omega) d\omega \tag{10}$$

This implementation improves VAD efficiency and reduces computational effort since just a single FFT has to be computed.

The estimation of the bispectrum is deep discussed in [21] 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.

3   Detection Tests for Voice Activity

The decision of our algorithm is based on statistical tests including the Generalized Likelihood ratio tests (GLRT) [22] and the Central χ^2-distributed test statistic under H_O [23]. We will call them GLRT and χ^2 tests. The tests are based on some asymptotic distributions and computer simulations in [24] show that the χ^2 tests require larger data sets to achieve a consistent theoretical asymptotic distribution.

GRLT: Consider the complete domain in bispectrum frequency for $0 \leq \omega_{n,m} \leq 2\pi$ and define P uniformly distributed points in this grid (m, n), called coarse grid. Define the fine grid of L points as the L nearest frequency pairs to coarse grid points. We have that $2M + 1 = P \cdot L$. If we reorder the components of the set of L Bispectrum estimates $\hat{\mathcal{C}}(n_l, m_l)$ where $l = 1, \ldots, L$, on the fine grid around the bifrequency pair into a L vector β_{ml} where $m = 1, \ldots P$ indexes the coarse grid [22] and define P-vectors $\phi_i(\beta_{1i}, \ldots, \beta_{Pi})$, $i = 1, \ldots L$; the generalized likelihood ratio test for the above discussed hypothesis testing problem:

$$H_0 : \mu = \mu_n \quad against \quad H_1 : \eta \equiv \mu^T \sigma^{-1} \mu > \mu_n^T \sigma_n^{-1} \mu_n \tag{11}$$

where $\mu = 1/L \sum_{i=1}^{L} \phi_i$ and $\sigma = 1/L \sum_{i=1}^{L} (\phi_i - \mu)(\phi_i - \mu)^T$ are the maximum likelihood gaussian estimates of vector $\mathcal{C} = (\mathcal{C}_{\mathbf{y}_k \mathbf{y}_l}(m_1, n_1) \ldots \mathcal{C}_{\mathbf{y}_k \mathbf{y}_l}(m_P, n_P))$, leads to the activity voice detection if:

$$\eta > \eta_0 \tag{12}$$

where η_0 is a constant determined by a certain significance level, i.e. the probability of false alarm. Note that:

1. We suppose independence between signal s_k and additive noise n_k bispectrum coeffcients[2] thus:

$$\mu = \mu_n + \mu_s; \quad \sigma = \sigma_n + \sigma_s \tag{13}$$

2. The right hand side of H_1 hypothesis must be estimated in each frame (it's unknown a-priori). In our algorithm the approach is based on the information in the previous non-speech detected intervals.

These assumptions are very restrictive, indeed, somehow the results shown in the experimental section allow them. The statistic considered here η is distributed as a central $F_{2P,2(L-P)}$ under the null hypothesis. Therefore a Neyman-Pearson test can be designed for a significance level α.

χ^2 **Tests:** In this section we consider the χ^2_{2L} distributed test statistic[23]:

$$\eta = \sum_{m,n} 2M^{-1} |\Gamma_{\mathbf{y}_k \mathbf{y}_l}(m,n)|^2 \tag{14}$$

where $\Gamma_{\mathbf{y}_k \mathbf{y}_l}(m,n) = \frac{|\hat{\mathcal{C}}_{\mathbf{y}_k \mathbf{y}_l}(n,m)|}{[S_{\mathbf{y}_0}(m) S_{\mathbf{y}_k}(n) S_{\mathbf{y}_l}(m+n)]^{0.5}}$ which is asymptotically distributed as $\chi^2_{2L}(0)$ where L denotes the number of points in interior of the principal domain. The Neyman-Pearson test for a significant level (false-alarm probability) α turns out to be:

$$H_1 \quad if \quad \eta > \eta_\alpha \tag{15}$$

where η_α is determined from tables of the central χ^2 distribution. Note that the denominator of $\Gamma_{\mathbf{y}_k \mathbf{y}_l}(m,n)$ is unknown a priori so they must be estimated as the bispectrum function (that is calculate $\hat{\mathcal{C}}_{\mathbf{y}_k \mathbf{y}_l}(n,m)$). This requires a larger data set as we mentioned above in this section.

4   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 and was first developed in [25]:

i) Spectrum smoothing. The power spectrum is averaged over two consecutive frames and two adjacent spectral bands.

---

[2] Observe that now we do not assume that n_k $k = 0 \ldots \pm M$ are gaussian.

ii) Noise estimation. The noise spectrum $N_e(m,l)$ is updated by means of a 1st order IIR filter on the smoothed spectrum $X_s(m,l)$, that is, $N_e(m,l) = \lambda N_e(m,l-1) + (1-\lambda)X_s(m,l)$ where $\lambda = 0.99$ and $m= 0, 1, ..., NFFT/2$.

iii) Wiener Filter (WF) design. First, the clean signal $S(m,l)$ is estimated by combining smoothing and spectral subtraction and then, the WF $H(m,l)$ is designed. The filter $H(m,l)$ is smoothed in order to eliminate rapid changes between neighbor frequencies that may often cause musical noise. Thus, the variance of the residual noise is reduced and consequently, the robustness when detecting non-speech is enhanced. The smoothing is performed by truncating the impulse response of the corresponding causal FIR filter to 17 taps using a Hanning window. With this operation performed in the time domain, the frequency response of the Wiener filter is smoothed and the performance of the VAD is improved.

iv) Frequency domain filtering. The smoothed filter H_s is applied in the frequency domain to obtain the de-noised spectrum $Y(m,l) = H_s(m,l)X(m,l)$.

Fig. 1 shows the operation of the proposed VAD on an utterance of the Spanish SpeechDat-Car (SDC) database [26]. The phonetic transcription is: ["siete", "θinko", "dos", "uno", "otSo", "seis"]. Fig 1(b) shows the value of η versus time. Observe how assuming η_0 the initial value of the magnitude η over the 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 2 we display the differences between noise and voice in general and in figure we settle these differences in the evaluation of η on speech and non-speech frames.

According to [25], using a noise reduction block previous to endpoint detection together with a long-term measure of the noise parameters, reports important benefits for detecting speech in noise since misclassification errors are significantly reduced.

(a)   (b)
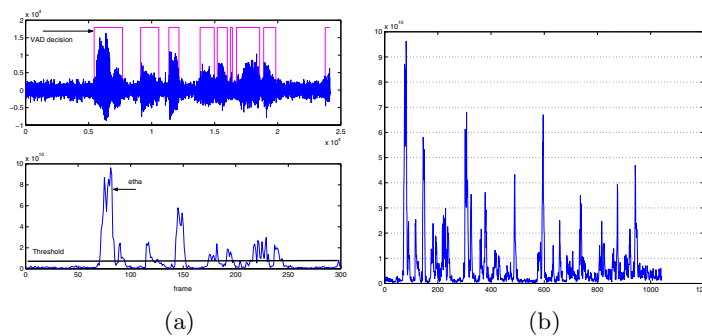
Fig. 1. Operation of the VAD on an utterance of Spanish SDC database. (a) Evaluation of η and VAD Decision. (b) Evaluation of the test hypothesis on an example utterance of the Spanish SpeechDat-Car (SDC) database [26].

(a)   (b)
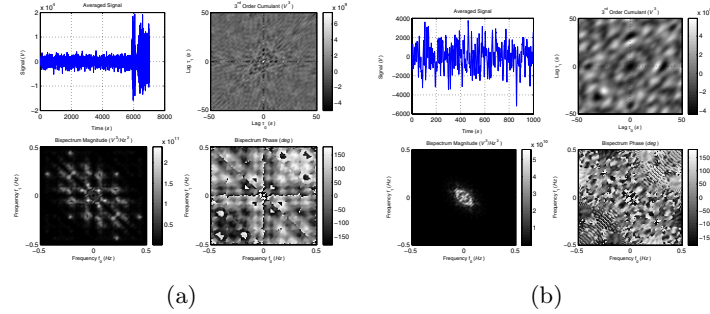
Fig. 2. Different Features allowing voice activity detection. (a) Features of Voice Speech Signal. (b) Features of non Speech Signal.

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 [11] vs. our VAD operation point, The work about the influence of the VAD decision on the performance of speech processing systems [8] is on the way. Subjective performance tests have also been considered for the evaluation of VADs working in combination with speech coders [27]. The experimental framework and the objective performance tests conducted to evaluate the proposed algorithm are partially showed for space reasons (we only show the results on AURORA-3 database)in this section.

First of all, let's compare the results we obtain using GLRT over the different Bispectrum estimators. The results over the Spanish database shows similar accuracy in voice activity detection depending on the parameters used in each estimator (resolution or number of FFT points "NFTT", different smoothing windows, number of records, etc.) as is shown in 4. Of course the computational effort of the cross-spectrum estimate, essential in on-line applications, is lower than the other approaches.
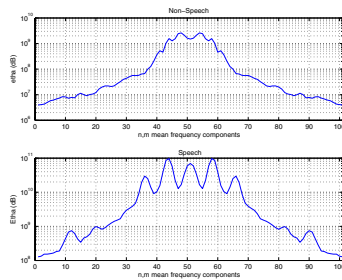
Fig. 3. Speech/non-Speech η values for Speech-Non Speech Frames
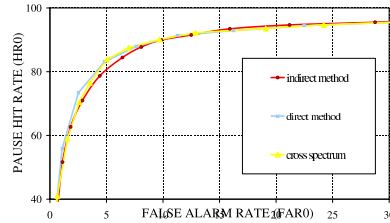
Fig. 4. Receiving Operating Curve in high noisy condition of Aurora 3 the Spanish-Car Database using three bispectrum estimators

The ROC curves are frequently used to completely describe the VAD error rate. The AURORA subset of the original Spanish SpeechDat-Car (SDC) database [26] 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. These noisy signals represent the most probable application scenarios for telecommunication terminals (suburban train, babble, car, exhibition hall, restaurant, street, airport and train station).

In table 1 shows the averaged ROC curves of the proposed VAD (BiSpectra based-VAD) and other frequently referred algorithms [9, 10, 11, 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 over a representative set VAD algorithms [9, 10, 11, 6]. It can be concluded from these results that:

i) The working point of the G.729 VAD shifts to the right in the ROC space with decreasing SNR.

ii) AMR1 works on a low false alarm rate point of the ROC space but exhibits poor non-speech hit rate.

iii) AMR2 yields clear advantages over G.729 and AMR1 exhibiting important reduction of the false alarm rate when compared to G.729 and increased non-speech hit rate over AMR1.

iv) The VAD used in the AFE for noise estimation yields good non-speech detection accuracy but works on a high false alarm rate point on the ROC space. It suffers from rapid performance degradation when the driving conditions get noisier. On the other hand, the VAD used in the AFE for FD has been planned to be conservative since it is only used in the DSR standard for that purpose. Thus, it exhibits poor non-speech detection accuracy working on a low false alarm rate point of the ROC space.

v) The proposed VAD also works with lower false alarm rate and higher non-speech hit rate when compared to the Sohn's [6], Woo's [9], Li's [10] and

Marzinzik's [11] algorithms in poor SNR scenarios. The BSVAD works robustly as noise level increases.

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.

It is worthwhile mentioning that the experiments described above yields a first measure of the performance of the VAD. Other measures of VAD performance that have been reported are the clipping errors [27]. These measures provide valuable information about the performance of the VAD and can be used for optimizing its operation. Our analysis does not distinguish between the frames that are being classified and assesses the hit-rates and false alarm rates for a first performance evaluation of the proposed VAD. On the other hand, the speech recognition experiments conducted later on the AURORA databases will be a direct measure of the quality of the VAD and the application it was designed for. Clipping errors are evaluated indirectly by the speech recognition system since there is a high probability of a deletion error to occur when part of the word is lost after frame-dropping.

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 [1]. 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

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	χ^2/GLRT
HR0	62.17	57.03	51.21	66.200	66.520/68.048
HR1	94.53	88.323	94.273	88.614	85.192/90.536

Table 2. Average Word Accuracy (%) for the Spanish SDC databases and tasks

		Base	Woo	Li	Marzinzik	Sohn	G.729	AMR1	AMR2	AFE	**GLRT**
Sp.	WM	92.94	95.35	91.82	94.29	96.07	88.62	94.65	95.67	95.28	96.28
	MM	83.31	89.30	77.45	89.81	91.64	72.84	80.59	90.91	90.23	92.41
	HM	51.55	83.64	78.52	79.43	84.03	65.50	62.41	85.77	77.53	86.70
	Ave.	**75.93**	89.43	82.60	87.84	90.58	75.65	74.33	90.78	87.68	**91.80**

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 [28]. The recognizer is based on the HTK (Hidden Markov Model Toolkit) software package [29].

Table 2 shows the recognition performance for the Spanish 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 [28]. Again, the VAD outperforms all the algorithms used for reference, yielding relevant improvements in speech recognition. Note that the SDC databases used in the AURORA 3 experiments have longer non-speech periods than the AURORA 2 database and then, the effectiveness of the VAD results more important for the speech recognition system. This fact can be clearly shown when comparing the performance of the proposed VAD to Marzinzik's VAD. The word accuracy of both VADs is quite similar for the AURORA 2 task. However, the proposed VAD yields a significant performance improvement over Marzinzik's VAD for the SDC databases.

6   Conclusions

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.

References

1. Karray, L., Martin, A.: Towards improving speech detection robustness for speech recognition in adverse environments. Speech Communition (2003) 261–276
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. Sangwan, A., Chiranth, M.C., Jamadagni, H.S., Sah, R., Prasad, R.V., Gaurav, V.: VAD techniques for real-time speech transmission on the Internet. In: IEEE International Conference on High-Speed Networks and Multimedia Communications. (2002) 46–50
5. Gustafsson, S., Martin, R., Jax, P., Vary, P.: A psychoacoustic approach to combined acoustic echo cancellation and noise reduction. IEEE Transactions on Speech and Audio Processing **10** (2002) 245–256

6. Sohn, J., Kim, N.S., Sung, W.: A statistical model-based voice activity detection. IEEE Signal Processing Letters **16** (1999) 1–3
7. Cho, Y.D., Kondoz, A.: Analysis and improvement of a statistical model-based voice activity detector. IEEE Signal Processing Letters **8** (2001) 276–278
8. Bouquin-Jeannes, R.L., Faucon, G.: Study of a voice activity detector and its influence on a noise reduction system. Speech Communication **16** (1995) 245–254
9. Woo, K., Yang, T., Park, K., Lee, C.: Robust voice activity detection algorithm for estimating noise spectrum. Electronics Letters **36** (2000) 180–181
10. Li, Q., Zheng, J., Tsai, A., Zhou, Q.: Robust endpoint detection and energy normalization for real-time speech and speaker recognition. IEEE Transactions on Speech and Audio Processing **10** (2002) 146–157
11. Marzinzik, M., Kollmeier, B.: Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. IEEE Transactions on Speech and Audio Processing **10** (2002) 341–351
12. Chengalvarayan, R.: Robust energy normalization using speech/non-speech discriminator for German connected digit recognition. In: Proc. of EUROSPEECH 1999, Budapest, Hungary (1999) 61–64
13. Tucker, R.: Voice activity detection using a periodicity measure. IEE Proceedings, Communications, Speech and Vision **139** (1992) 377–380
14. Nemer, E., Goubran, R., Mahmoud, S.: Robust voice activity detection using higher-order statistics in the lpc residual domain. IEEE Trans. Speech and Audio Processing **9** (2001) 217–231
15. Nikias, C., Petropulu, A.: Higher Order Spectra Analysis: a Nonlinear Signal Processing Framework. Prentice Hall (1993)
16. Kay, S.: Modern Spectral Esimation. Englewood Cliffs, NJ: Prentice Hall (1988)
17. Brillinger, D.: Analysis of a linear time invariant relation between a stochastic series and several deterministic series. In: Time Series Data Analysis and Theory. Siam in applied Maths (2001)
18. Górriz, J.M., Puntonet, C.G., Ramírez, J., Segura, J.C.: Bispectra analysis-based vad for robust speech recognition. In: IWINAC (2). (2005) 567–576
19. Górriz, J.M., Ramírez, J., Segura, J.C., Puntonet, C.G.: An improved mo-lrt vad based on a bispectra gaussian model. In press IEE Electronic Letters **X** (Jul 2005) XXX–XXX
20. Tugnait, J.: Detection of non-gaussian signals using integrated polyspectrum. IEEE Transaction on signal Processing **42** (1994) 3137–3149
21. Brillinger, D., Rossenblatt, M.: Asymptotic theory of estimates of kth order spectra. In: Spectral Analysis of Time Series. Wiley (1975)
22. Subba-Rao, T.: A test for linearity of stationary time series. Journal of Time Series Analisys **1** (1982) 145–158
23. Hinich, J.: Testing for gaussianity and linearity of a stationary time series. Journal of Time Series Analisys **3** (1982) 169–176
24. Tugnait, J.: Two channel tests fro common non-gaussian signal detection. IEE Proceedings-F **140** (1993) 343–349
25. Ramírez, J., Segura, J., Benítez, C., delaTorre, A., Rubio, A.: An effective subband osf-based vad with noise reduction for robust speech recognition. In press IEEE Transactions on Speech and Audio Processing **X** (2004) X–X
26. Moreno, A., Borge, L., Christoph, D., Gael, R., Khalid, C., Stephan, E., Jeffrey, A.: SpeechDat-Car: A Large Speech Database for Automotive Environments. In: Proceedings of the II LREC Conference. (2000)

27. Benyassine, A., Shlomot, E., Su, H., Massaloux, D., Lamblin, C., Petit, J.: 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 **35** (1997) 64–73
28. ETSI: Speech processing, transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms. ETSI ES 201 108 Recommendation (2000)
29. Young, S., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book. Cambridge University (1997)