Generalized LRT-Based Voice Activity Detector

Juan Manuel Górriz, Javier Ramírez, Carlos G. Puntonet, and José Carlos Segura, *Senior Member, IEEE*

***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 well-known statistical tests based on 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 being the statistical tests built on a multiple observation (MO) window consisting of averaged bispectrum coefficients of the speech signal. Clear improvements in speech/non-speech discrimination accuracy demonstrate the effectiveness of the proposed VAD. It is shown that application of a 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 experimental analysis carried out on the AURORA 3 databases 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.**

***Index Terms*—Bispectra analysis, higher order statistics, noise reduction, speech/non-speech detection.**

I. Introduction

SPEECH/NON-SPEECH detection is an unsolved problem in speech processing and affects numerous applications, including robust speech recognition, discontinuous transmission, real-time speech transmission on the Internet, or combined noise reduction and echo cancellation schemes in the context of telephony [1], [2]. The speech/non-speech classification task is not as trivial as it appears, and most of the voice activity detection (VAD) algorithms often 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 [3] and have evaluated the influence of the VAD effectiveness on the performance of speech processing systems. Most of them have focused on the development of robust algorithms with special attention on the derivation and study of noise robust features and decision rules [4]–[6]. The different approaches include those based on energy thresholds, pitch detection, spectrum analysis, zero-crossing rate, periodicity measure, higher order statistics in the LPC residual domain, or combinations of different features.

This letter explores a new alternative toward improving speech detection robustness in adverse environments and the performance of speech recognition systems. The proposed VAD incorporates a noise reduction block that precedes the VAD and uses bispectra of third-order cumulants to formulate a robust decision rule.

II. Model Assumptions

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

$$y_k(t) = x(t+k) \tag{1}$$

where k is the differential delay (or advance) between the samples. This provides a new set of $2M+1$ vector variables $\mathbf{y}_k = [y_k(1), \ldots, y_k(N)]$ by selecting N samples of the input signal. It can be represented using the associated Toeplitz matrix

$$T_{x(t)} = \begin{pmatrix} y_{-M}(1) & \ldots & y_{-M}(N) \\ y_{-M+1}(1) & \ldots & y_{-M+1}(N) \\ \ldots & \ldots & \ldots \\ y_M(1) & \ldots & y_M(N) \end{pmatrix}. \tag{2}$$

Using this model, the speech/non-speech detection can be described by using two essential hypotheses (reordering indexes)

$$\begin{aligned} &H_0 : y_k(t) = n_k(t); \quad t = 1, \ldots, N \\ &H_1 : y_k(t) = s_k(t) + n_k(t) \quad k = 0, \pm 1, \ldots, \pm M \end{aligned} \tag{3}$$

where $s_k(t)$ denotes the common non-Gaussian speech signal with delay k, and $n_k(t)$ are the additive non-speech noise sequences, respectively. All the processes involved are assumed to be jointly stationary and zero-mean. Consider the third-order cumulant function defined as $C_{\mathbf{y}_k\mathbf{y}_l} \equiv E[y_0(t)y_k(t)y_l(t)]$ and the two-dimensional discrete Fourier transform (DFT) of $C_{\mathbf{y}_k\mathbf{y}_l}$, the bispectrum function

$$\mathcal{C}_{\mathbf{y}_k\mathbf{y}_l}(\omega_1, \omega_2) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} C_{\mathbf{y}_k\mathbf{y}_l} \cdot \exp(-j(\omega_1 k + \omega_2 l)). \tag{4}$$

Sampling (4), the bispectrum estimate can be written as

$$\hat{\mathcal{C}}_{\mathbf{y}_k\mathbf{y}_l}(n,m) = \sum_{k=-M}^{M} \sum_{l=-M}^{M} C_{\mathbf{y}_k\mathbf{y}_l} \cdot w(k,l) \cdot \exp(-j(\omega_n k + \omega_m l)) \tag{5}$$

where $\omega_{n,m} = (2\pi/M)(n,m)$ with $n, m = -M, \ldots, M$ are the sampling frequencies, $w(k,l)$ is the window function (to reduce aliasing [7]), and $C_{\mathbf{y}_k\mathbf{y}_l} = (1/N)\sum_{t=1}^{N} y_0(t)y_k(t)y_l(t) = (1/N)\mathbf{y}_0 \cdot \mathbf{y}_k \cdot \mathbf{y}_l$. The estimation of the bispectrum is discussed in depth in [8] and many others, where conditions for consistency are given. The estimate is said to be asymptotically consistent if the

Manuscript received July 1, 2005; revised March 7, 2006. This work was supported in par by the Spanish Government under the CICYT projects SR3-VoIP (TEC2004-03829/FEDER) and SESIBONN (TEC2004-06096-C03-00) and in part by the European Commission (HIWIRE, IST No. 507943). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Frederic Bimbot.

J. M. Górriz, J. Ramírez, and J. C. Segura are with the Departmento Teoría de la Señal, Telemática y Comunicaciones, Periodista Daniel Saucedo Aranda, 18071 Granada, Spain (e-mail: gorriz@ugr.es; javierrp@ugr.es; segura@ugr.es).

C. G. Puntonet is with the Departmento Arquitectura y Tecnología de Computadores, Periodista Daniel Saucedo Aranda, 18071 Granada, Spain (e-mail: carlos@atc.ugr.es).

Digital Object Identifier 10.1109/LSP.2006.876340

squared deviation goes to zero as the number of samples tends to infinity.

III. Tests for Voice Activity Detection

The decision of our algorithm is based on statistical tests, including the generalized likelihood ratio tests (GLRT) [9] and the central χ^2-distributed test statistic under H_0 [10]. We will call them GLRT and χ^2 tests. The tests are based on some asymptotic distributions, and computer simulations in [11] show that the χ^2 tests require larger data sets to achieve a consistent theoretical asymptotic distribution.

A. GLRT

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, as shown in Fig. 1. Define the fine grid of L points as the L nearest frequency pairs to the 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 an L vector β_{ml}, where $m = 1, \ldots, P$ indexes the coarse grid [9], and define P-vectors $\phi_i(\beta_{1i}, \ldots, \beta_{Pi})$, $i = 1, \ldots, L$; the GLRT, for the above-discussed hypothesis testing problem

$$H_0 : \mu = \mu_n \text{ against } H_1 : \eta \equiv \mu^T \sigma^{-1} \mu > \mu_n^T \sigma_n^{-1} \mu_n \tag{6}$$

where μ and σ are the mean and covariance maximum likelihood Gaussian estimates of vector $\mathcal{C} = (\mathcal{C}_{\mathbf{y}_k\mathbf{y}_l}(m_1,n_1) \ldots \mathcal{C}_{\mathbf{y}_k\mathbf{y}_l}(m_P,n_P))$, that is

$$\mu = \frac{1}{L}\sum_{i=1}^{L} \phi_i$$
$$\sigma = \frac{1}{L}\sum_{i=1}^{L} (\phi_i - \mu)(\phi_i - \mu)^T. \tag{7}$$

Thus, presence of speech is detected if

$$\eta > \eta_n \tag{8}$$

where η_n is the threshold determined by a certain significance level, i.e., the probability of false alarm. Note the following.

1) We assume independence between the components of bispectrum of signal $s(t)$ and additive noise $n(t)$;[1] thus

$$\mu = \mu_n + \mu_s; \quad \sigma = \sigma_n + \sigma_s. \tag{9}$$

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

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.

B. χ^2 Tests

In this section, we consider the χ^2_{2L} distributed test statistic [10]

$$\eta = \sum_{m,n} 2KN_B^{-1} |\Gamma_{\mathbf{y}_k\mathbf{y}_l}(m,n)|^2 \tag{10}$$

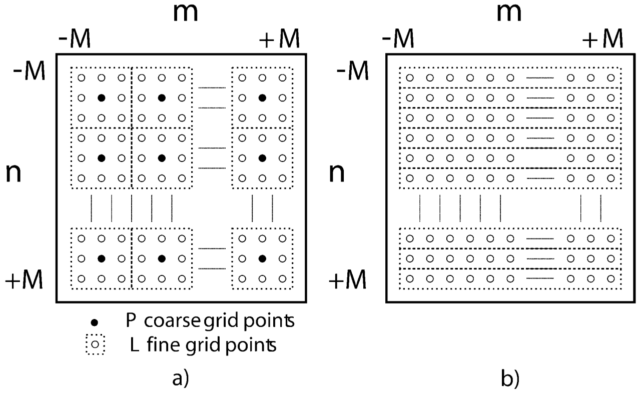


Fig. 1. a) Fine and coarse grids. P points are uniformly distributed with L boundary points. b) Row averaging for integrated bispectrum estimation.

where K is the number of nonoverlapping segments, each of size N_B, given the sample sequence, $\Gamma_{\mathbf{y}_k\mathbf{y}_l}(m,n) = (|\hat{\mathcal{C}}_{\mathbf{y}_k\mathbf{y}_l}(n,m)|/[S_{\mathbf{y}_0}(m)S_{\mathbf{y}_k}(n)S_{\mathbf{y}_l}(m+n)]^{0.5})$, which is asymptotically distributed as $\chi^2_{2L}(0)$, where $S_{\mathbf{y}_k}$ represents the power spectrum of $y_k(t)$, and L denotes the number of points in the principal domain. The Neyman–Pearson test for a significant level (false-alarm probability) α turns out to be

$$H_1 \; if \; \eta > \eta_\alpha \tag{11}$$

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 earlier in this section.

IV. Averaged Bispectrum Function and Long-Term Information

In order to observe the potential of the proposed method, we propose an approximated decision based on an average of the components of the bispectrum in one frequency dimension instead of averaging over coarse and fine grids [9]. In this way, we define η as

$$\eta = \frac{1}{L \cdot P}\sum_{i=1}^{P}\sum_{j=1}^{L} \hat{\mathcal{C}}(i,j) = \frac{1}{L}\sum_{j=1}^{L} \mu(j) \tag{12}$$

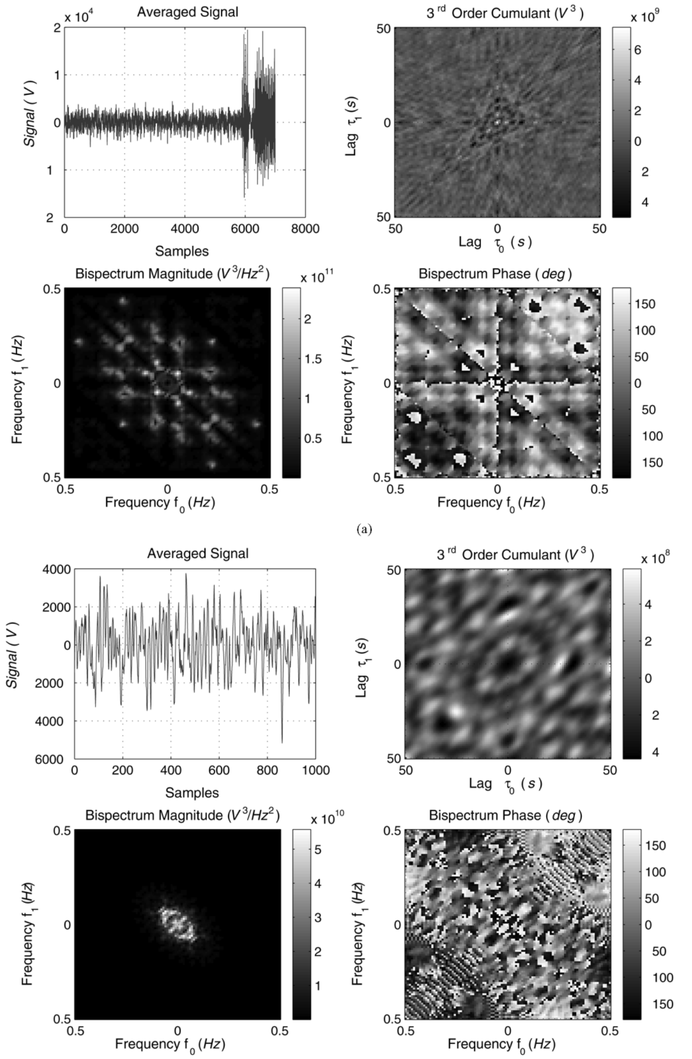
where L, P defines the selected grid (high frequencies with noteworthy variability). It can be easily deduced that

$$S_{sx}(\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{C}(\omega, \omega_2)d\omega_2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{C}(\omega_1, \omega)d\omega_1. \tag{13}$$

That is, the cross spectrum between the signal $x(t)$ and its square $s(t)$ can be interpreted as an integrated bispectrum of $x(t)$. This integrated bispectrum will form the basis for the test statistic used in this letter for detecting the presence of the non-Gaussian signal $x(t)$ in noisy data. The advantage of this implementation is the low computational cost, unlike the bispectrum-based test for voice activity detection presented in [12].

Fig. 2 shows the differences between the cumulants and bispectrum of speech and non-speech. It can be clearly concluded

[1] This is an acceptable assumption [12] since the results obtained from it are quiet significant. Here, we do not assume that $n_k(t)$ $k = 0 \cdots \pm M$ are Gaussian; they are modeled as an adaptive Bispectrum bias instead.



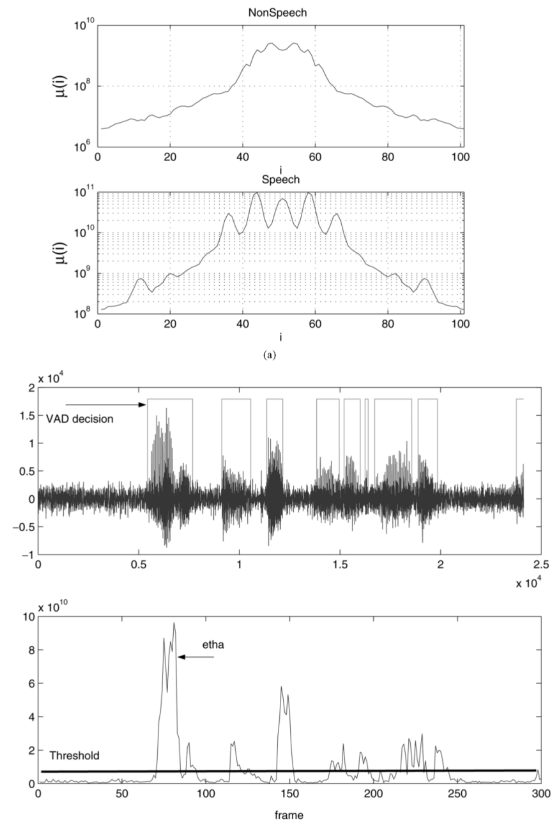
(a)

(b)

Fig. 2. Different features enabling voice activity detection (third-order cumulants, bispectrum magnitude, and phase $\mathbf{y}_k$). (a) Features of a speech signal. (b) Features of non-speech signal.

that the bispectrum of the signal exhibits discriminative characteristics for speech/non-speech classification.

Fig. 3(a) shows the fine-to-coarse grid transformation of (12) when it is defined as an average over rows (integrated bispectrum) of the 2-D bispectrum representation. The so-defined bispectrum estimation retains discriminative behavior and exhibits important differences between speech and non-speech signals. Fig. 3(b) shows the operation of the proposed VAD on an utterance of the Spanish SpeechDat-Car (SDC) database [13]. The phonetic transcription is: ["siete," "θinko," "dos," "uno," "otSo," "seis"]. Fig. 3(b) shows the value of η versus time. Observe how assuming a decision threshold η_0 slightly above the initial value of the magnitude η over the first frame (noise), we can achieve a good VAD decision. The figure also shows the good behavior of the proposed method for detecting fricative sounds (as derived from the phonetic transcription), even when a single frame is used in these preliminary experiments. Alternatively, we have also included long-term information (LTI) in the VAD decision, as proposed in [11], which essentially improves the efficiency of the proposed method. With this approach, the VAD decision is formulated not only over the current frame l, using $y_k(t)$, $k = 0, \ldots, \pm M$, but also over $2m + 1$ previous and future frames, that is, the frame set $\{l - m, \ldots, l, \ldots, l + m\}$ that includes multiple consecutive observations of the input signal $y_k(t + S \cdot j)$, $k = 0, \ldots, \pm M$, $j = 0, \ldots, \pm m$, where S defines the VAD frame-shift. In this way, the VAD performs an advanced detection of beginnings and delayed detection of word endings that, in part, makes a hangover unnecessary.



(a)

(b)

Fig. 3. Operation of the VAD on an utterance of Spanish SDC database. (a) Bispectrum averaged over rows for speech and non-speech. (b) Evaluation of η and VAD decision.

V. Experimental Framework

The ROC curves are frequently used to completely describe the VAD error rate. The AURORA 3 subset of the original Spanish SDC database [13] was used in this analysis. 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. Fig. 4 shows the ROC curves of the proposed bispectra GLRT based-VAD when defined on a single observation $(m = 1)$ and over multiple observations ($m = 8$ frame delay) under different noise conditions. The working points of the ITU-T G.729 [2], ETSI AMR [1], and ETSI AFE VADs [14] are also included as well as other frequently referred to algorithms [3]–[6] for recordings from the distant microphone in quiet and high noisy conditions.

The proposed VAD yields clear improvements in detection accuracy working closer to the upper left corner than any other

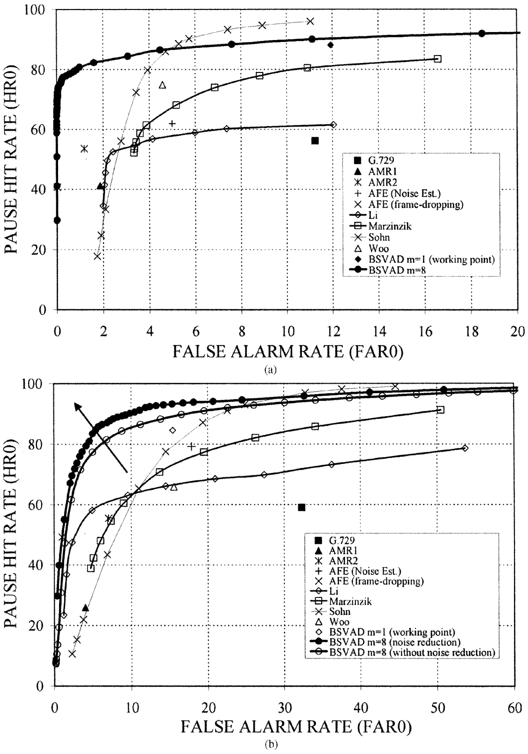


Fig. 4. ROC curves obtained for different subsets of the Spanish SDC database at different driving conditions. (a) Quiet (stopped car, motor running, 12 dB average SNR). (b) High (high speed, good road, 5 dB average SNR).

algorithm used as a reference. The benefits are especially important over G.729 and over the Li's algorithm [5]. The statistical test is more effective when multiple observation are considered. It improves Marzinzik's VAD [6], the Sohn's VAD [3], and all recently reported VADs to date for varying significance level. Fig. 4 also assesses the influence of the noise reduction block on the ROC curves in order to state clearly if the benefits reported by the proposed method are due to the use of a GLRT defined on the bispectrum of the input signal or to the noise reduction block. If noise reduction is not applied, the algorithm yields clear improvements over the competing algorithm. The purpose of the noise reduction is to achieve a more robust detection algorithm for use in high noise acoustic environments. Thus, denoising leads to an additional shift up and to the left of the ROC curve in the ROC space.

The results are conclusive over Aurora 3 on average (see Table I). The comparison with [12] is not included. Although the model is essentially different (in the latter reference, a Gaussian model is assumed in the vector of observations), the accuracy of both approaches is rather similar.

VI. Conclusion

This letter presented a new technique for improving speech detection robustness in noisy environments. The approach is based on higher order spectra analysis, i.e., the integrated bispectrum function. The VAD performs an advanced detection using the estimated components of the bispectrum function and robust statistical tests (i.e., GLRT) over the set of vector variables $\mathbf{y}_k$ (multiple-single observation). As a result, it leads to clear improvements in speech/non-speech discrimination, especially when the SNR drops. The proposed algorithm outperformed G.729, AMR and AFE standard VADs, as well as recently reported approaches for endpoint detection.

TABLE I
Average AURORA-3 Speech/Non-Speech Hit Rates

(%)	G.729	AMR1	AMR2	AFE (WF)	AFE (FD)
HR0	55.80	51.57	57.63	69.07	33.99
HR1	88.07	98.26	97.62	85.48	99.75
(%)	Woo	Li	Marzinzik	Sohn	GLRT/χ^2
HR0	62.17	57.03	51.21	66.20	75.11/63.16
HR1	94.53	88.32	94.27	88.61	96.28/88.74

References

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

[2] A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70, ITU, 1996, ITU-T Recommendation G.729-Annex B.

[3] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based VAD," *IEEE Signal Process. Lett.*, vol. 16, no. 1, pp. 1–3, Jan. 1999.

[4] K. Woo, T. Yang, K. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electron. Lett.*, vol. 36, no. 2, pp. 180–181, 2000.

[5] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 3, pp. 146–157, Mar. 2002.

[6] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 341–351, 2002.

[7] C. Nikias and A. Petropulu, *Higher Order Spectra Analysis: A Nonlinear Signal Processing Framework*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[8] D. Brillinger and M. Rossenblatt, *Spectral Analysis of Time Series*. New York: Wiley, 1975, ch. Asymptotic theory of estimates of kth order spectra.

[9] T. Subba-Rao, "A test for linearity of stationary time series," *J. Time Series Anal.*, vol. 1, pp. 145–158, 1982.

[10] J. Hinich, "Testing for Gaussianity and linearity of a stationary time series," *J. Time Series Anal.*, vol. 3, pp. 169–176, 1982.

[11] J. Tugnait, "Two channel tests for common non-Gaussian signal detection," *Proc. Inst. Elect. Eng. F, Radar Signal Process.*, vol. 140, no. 6, pp. 343–349, Dec. 1993.

[12] J. M. Górriz, J. Ramirez, J. C. Segura, and C. G. Puntonet, "An improved MO-LRT VAD based on a bispectra Gaussian model," *Electron. Lett.*, vol. 41, no. 15, pp. 877–879, 2005.

[13] A. Moreno, L. Borge, D. Christoph, R. Gael, C. Khalid, E. Stephan, and A. Jeffrey, "Speechdat-car: A large speech database for automotive environments," in *Proc. II LREC Conf.*, 2000.

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