Speech/Non-Speech Discrimination Based on Contextual Information Integrated Bispectrum LRT

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

***Abstract*—This letter shows an effective statistical voice activity detection algorithm based on the integrated bispectrum, which is defined as a cross spectrum between the signal and its square and inherits the ability of higher order statistics to detect signals in noise with many other additional advantages: *1*) its computation as a cross spectrum leads to significant computational savings, and *2*) the variance of the estimator is of the same order as that of the power spectrum estimator. The decision rule is formulated in terms of an average likelihood ratio test (LRT) involving successive integrated bispectrum speech features. With these and other innovations, the proposed method reports significant improvements in speech/pause discrimination as well as in speech recognition over standardized techniques such as ITU-T G.729, ETSI AMR, and AFE VADs, and over recently published VADs.**

***Index Terms*—Contextual likelihood ratio test, higher order statistics, robust speech recognition, voice activity detection.**

I. Introduction

Detecting the presence of speech in a noisy signal is a problem affecting numerous applications, including robust speech recognition [1], [2], discontinuous transmission voice communications [3], [4], real-time speech transmission on the Internet [5], or combined noise reduction and echo cancellation schemes in the context of telephony [6]. These systems often benefit from voice activity detectors (VADs), which are frequently used in such application scenarios for different purposes. The classification task is not as trivial as it appears, and most of the VAD algorithms often fail in high noise conditions. During the last decade, numerous researchers have developed different strategies for detecting speech in a noisy signal [7]–[10] and have evaluated the influence of the VAD effectiveness on the performance of speech processing systems [11]. One of the most important disadvantages of these approaches is that no *a priori* information about the statistical properties of the signals is used. Higher order statistics methods rely on an *a priori* knowledge of the input processes and have been considered for VAD since they can distinguish between Gaussian signals (which has a vanishing bispectrum) from non-Gaussian signals. However, the main limitations of bispectrum-based techniques are that they are computationally expensive [12], and the variance of the bispectrum estimators is much higher than that of power spectral estimators for identical data record size. These problems were addressed by Tugnait [13], [14], who showed a computationally efficient and reduced variance statistical test based on the integrated polyspectra for detecting an unknown random, stationary, non-Gaussian signal in Gaussian noise. This letter advances in the field and shows an effective VAD based on a likelihood ratio test (LRT) that is defined on the integrated bispectrum of the noisy speech. The proposed approach also incorporates contextual information to the decision rule, a strategy first proposed in [2] that has reported significant benefits [15], particularly in robust speech recognition applications [16], [17].

II. Background

The bispectrum of a discrete-time signal $x(t)$ is defined as

$$B_x(\omega_1,\omega_2)=\sum_{i=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}C_{3x}(i,k)e^{-j(\omega_1 i+\omega_2 k)} \tag{1}$$

where $C_{3x}(i,k)=E\{x^*(t)x(t+i)x(t+k)\}$ is the third-order cumulant of the process $x(t)$. Note that, from the above definition, the third-order cumulant can be expressed as

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

Although the bispectra have all the advantages of cumulants/polyspectra, their direct use has two serious limitations: *1*) the computation of bispectra in the whole triangular region is huge, and *2*) the two-dimensional (2-D) template matching score in the classification is impractical. To use efficiently bispectra, integrated bispectrum methods [13], [14] were proposed for different applications [18], [19].

A. Definition

Let $x(t)$ be a zero mean stationary random process. If we define $\tilde{y}(t)=x^2(t)-E\{x^2(t)\}$, the cross correlation between $\tilde{y}(t)$ and $x(t)$ is defined to be

$$r_{\tilde{y}x}(k)=E\left\{\tilde{y}(t)x(t+k)\right\}=E\left\{x^2(t)x(t+k)\right\}=C_{3x}(0,k) \tag{3}$$

and its cross spectrum is given by

$$S_{\tilde{y}x}(\omega)=\sum_{k=-\infty}^{k=+\infty}C_{3x}(0,k)e^{-j\omega k} \tag{4}$$

with

$$C_{3x}(0,k)=\frac{1}{2\pi}\int_{-\pi}^{\pi}S_{\tilde{y}x}(\omega)\exp\left\{j(\omega_k)\right\}d\omega. \tag{5}$$

Manuscript received October 18, 2005; revised February 15, 2006. This work was supported in part by the Spanish Government under TEC2004-03829/FEDER MEC project and in part by the European Commission under the HIWIRE project (Contract no. 507943). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jitendra K. Tugnait.

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

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

Digital Object Identifier 10.1109/LSP.2006.873147

If (2) and (5) are compared, we obtain

$$S_{\tilde{y}x}(\omega)=\frac{1}{2\pi}\int_{-\pi}^{\pi} B_x(\omega,\omega_2)d\omega_2=\frac{1}{2\pi}\int_{-\pi}^{\pi} B_x(\omega_1,\omega)d\omega_1. \quad (6)$$

Thus, the integrated bispectrum is defined as a cross spectrum between the signal and its square, and therefore, it is a function of a single frequency variable. It is easy to see that the bispectrum of a Gaussian process is identically zero, and its integrated bispectrum is as well. Hence, its computation as a cross spectrum leads to significant computational savings. However, more important is that the variance of the estimator is of the same order as that of the power spectrum estimator [13].

B. Estimation

Let $\hat{S}_{yx}(\omega)$ denote a consistent estimator of $S_{yx}(\omega)$, where $y(t) = x^2(t) - E\{x^2(t)\}$. Given a finite data set $x(1), x(2), \ldots, x(N)$, the integrated bispectrum is normally estimated by dividing the sample sequence into segments or blocks [20]. Thus, the data set is divided into K_B nonoverlapping segments, each of size N_B samples, so that $N = K_B N_B$. Then, the cross periodogram of the kth block of data is given by

$$\hat{S}_{yx}^{(k)}(\omega) = \frac{1}{N_B} X^{(k)}(\omega)\left[Y^{(k)}(\omega)\right]^* \quad (7)$$

where $X^{(k)}(\omega)$ and $Y^{(k)}(\omega)$ denote the discrete Fourier transform (DFT) of the kth block. Finally, the estimate is obtained by averaging K_B blocks

$$\hat{S}_{yx}(\omega) = \frac{1}{K_B}\sum_{k=1}^{K_B} \hat{S}_{yx}^{(k)}(\omega). \quad (8)$$

III. Integrated Bispectrum Likelihood Ratio Test

This section addresses the VAD problem formulated in terms of a classical binary hypothesis testing framework

$$\begin{aligned} H_0 &: x(t) = n(t) \\ H_1 &: x(t) = s(t) + n(t). \end{aligned} \quad (9)$$

In a two-hypothesis test, the optimal decision rule that minimizes the error probability is the Bayes classifier. Given an observation vector $\hat{\mathbf{y}}$ to be classified, the problem is reduced to selecting the class (H_0 or H_1) with the largest posterior probability $P(H_i|\hat{\mathbf{y}})$. In [17], the LRT, first proposed by Sohn [7] for VAD, which was defined on the power spectrum, is generalized and extended to a multiple observation LRT (MO-LRT) when successive observations $\hat{\mathbf{y}}_{l-m}, \ldots, \hat{\mathbf{y}}_{l-1}, \hat{\mathbf{y}}_l, \hat{\mathbf{y}}_{l+1}, \ldots, \hat{\mathbf{y}}_{l+m}$ of the noisy signal are available, where l is the frame being classified as speech or non-speech. This test involves evaluating and comparing to a fixed threshold η, the LRT of the joint conditional distributions of the observations under H_0 and H_1

$$L_{l,m}(\hat{\mathbf{y}}_{l-m}, \ldots, \hat{\mathbf{y}}_{l+m}) = \frac{p_{\mathbf{y}_{l-m},\ldots,\mathbf{y}_{l+m}|H_1}(\hat{\mathbf{y}}_{l-m}, \ldots, \hat{\mathbf{y}}_{l+m}|H_1)}{p_{\mathbf{y}_{l-m},\ldots,\mathbf{y}_{l+m}|H_0}(\hat{\mathbf{y}}_{l-m}, \ldots, \hat{\mathbf{y}}_{l+m}|H_0)} \quad (10)$$

which is easily performed if the observations are independent. The so-defined log-LRT

$$\ell_{l,m} = \sum_{k=l-m}^{l+m} \ln \frac{p_{\mathbf{y}_k|H_1}(\hat{\mathbf{y}}_k|H_1)}{p_{\mathbf{y}_k|H_0}(\hat{\mathbf{y}}_k|H_0)} \quad (11)$$

is recursive in nature, and if the Φ function is defined as

$$\Phi(k) = \ln \frac{p_{\mathbf{y}_k|H_1}(\hat{\mathbf{y}}_k|H_1)}{p_{\mathbf{y}_k|H_0}(\hat{\mathbf{y}}_k|H_0)} \quad (12)$$

equation (11) can be calculated in recursive fashion as

$$\ell_{l+1,m} = \ell_{l,m} - \Phi(l-m) + \Phi(l+m+1). \quad (13)$$

Assuming the integrated bispectrum $\{S_{yx}^{(k)} : \omega\}$ as the feature vector $\hat{\mathbf{y}}_k$ and to be independent zero-mean Gaussian variables

$$\begin{aligned} p\left(S_{yx}^{(k)}(\omega)|H_0\right) &= \frac{1}{\pi\lambda_0^{(k)}(\omega)} \exp\left[-\frac{\left|S_{yx}^{(k)}(\omega)\right|^2}{\lambda_0^{(k)}(\omega)}\right] \\ p\left(S_{yx}^{(k)}(\omega)|H_1\right) &= \frac{1}{\pi\lambda_1^{(k)}(\omega)} \exp\left[-\frac{\left|S_{yx}^{(k)}(\omega)\right|^2}{\lambda_1^{(k)}(\omega)}\right] \end{aligned} \quad (14)$$

$\Phi(k)$ is reduced to

$$\Phi(k) = \sum_{\omega}\left[\frac{\xi_k(\omega)\gamma_k(\omega)}{1+\xi_k(\omega)} - \log\left(1+\xi_k(\omega)\right)\right] \quad (15)$$

where

$$\xi^{(k)}(\omega) = \frac{\lambda_1^{(k)}(\omega)}{\lambda_0^{(k)}(\omega)} - 1 \quad \gamma^{(k)}(\omega) = \frac{\left|S_{yx}^{(k)}(\omega)\right|^2}{\lambda_0^{(k)}}. \quad (16)$$

Note that the decision rule is formulated over a sliding window consisting of $(2m+1)$ observation vectors around the frame for which the decision is being made. This fact imposes an m-frame delay to the algorithm that, for several applications, including robust speech recognition, is not a serious implementation obstacle. The two key issues to evaluate the proposed LRT are *1*) the estimation of the integrated bispectrum by means of a finite data set and *2*) the computation of the variances $\lambda_0^{(k)}(\omega)$ and $\lambda_1^{(k)}(\omega)$ of the integrated bispectrum under H_0 and H_1 hypothesis.

IV. Variance of the Integrated Bispectrum

The properties of the bispectrum estimators have been discussed in [20] and [21]. The test proposed in the previous section and the model assumed in (14) are justified since for large N_B, the estimate $\hat{S}_{yx}^{(i)}(\omega_m)$ is complex Gaussian and independent of $\hat{S}_{yx}^{(i)}(\omega_n)$ for $m \neq n$ $(m, n = 1, 2, \ldots, N_B/2 - 1)$. Moreover, its mean and variance for large values of N_B and K_B can be approximated [13] by

$$\begin{aligned} E\left\{\hat{S}_{yx}(\omega)\right\} &\approx S_{yx}(\omega) \\ \mathrm{var}\left\{\Re\left[\hat{S}_{yx}^{(i)}(\omega)\right]\right\} &\approx \frac{1}{2K_B}\left[S_{yy}(\omega)S_{xx}(\omega) + \Re\left\{S_{yx}^2(\omega)\right\}\right] \\ \mathrm{var}\left\{\Im\left[\hat{S}_{yx}^{(i)}(\omega)\right]\right\} &\approx \frac{1}{2K_B}\left[S_{yy}(\omega)S_{xx}(\omega) \right. \\ &\qquad \left. -\Re\left\{S_{yx}^2(\omega)\right\}\right]. \end{aligned} \quad (17)$$

In this way, it is needed to estimate $S_{xx}(\omega)$ and $S_{yy}(\omega)$ under H_0 and H_1 hypothesis in order to compute $\lambda_0(\omega)$ and $\lambda_1(\omega)$. It can be shown [13], [14] that

$$\begin{aligned} \lambda_0(\omega) = \frac{1}{K_B}&\left[2S_{nn}(\omega) * S_{nn}(\omega) + 2\pi\sigma_n^4\delta(\omega)\right] \\ &\times S_{nn}(\omega) \end{aligned} \quad (18)$$

$$\lambda_1(\omega) = \frac{1}{K_B}\left[S_{ss}(\omega) + S_{nn}(\omega)\right] \times \left[2S_{ss}(\omega) * S_{ss}(\omega) + 2S_{nn}(\omega) * S_{nn}(\omega) + 4S_{ss}(\omega) * S_{nn}(\omega)\right]. \quad (19)$$

Finally, a way to estimate the power spectrum of the clean signal, $S_{ss}(\omega)$, is needed. In this letter, a method combining Wiener filtering and spectral subtraction is used to estimate $S_{ss}(\omega)$ in terms of the power spectrum of the noisy signal $S_{xx}(\omega)$. During a short initialization period, the power spectrum of the residual noise $S_{nn}(\omega)$ is estimated assuming a short non-speech period at the beginning of the utterance. Note that $S_{nn}(\omega)$ can be computed in terms of the DFT of the noisy signal $x(t) = n(t)$. After the initialization period, the integrated bispectrum of the noisy signal $S_{xx}(\omega)$ is computed for each frame using (7), and $S_{ss}(\omega)$ is then obtained by applying a denoising process. Denoising consists of a previous smoothed spectral subtraction followed by Wiener filtering. It is worthwhile clarifying that $S_{nn}(\omega)$ is not only estimated during the initialization period but also updated during non-speech frames based on the VAD decision. Thus, the denoising process consists of the following stages.

1) Spectral subtraction

$$S_1(\omega) = L_s S_{ss}(\omega) + (1 - L_s) \times \max\left(S_{xx}(\omega) - \alpha S_{nn}(\omega), \beta S_{xx}(\omega)\right). \quad (20)$$

2) First WF design and filtering

$$\mu_1(\omega) = \frac{S_1(\omega)}{S_{nn}(\omega)}$$
$$W_1(\omega) = \frac{\mu_1(\omega)}{(1 + \mu_1(\omega))}$$
$$S_2(\omega) = W_1(\omega) S_{xx}(\omega). \quad (21)$$

3) Second WF design and filtering

$$\mu_2(\omega) = \frac{S_2(\omega)}{S_{nn}(\omega)}$$
$$W_2(\omega) = \max\left(\frac{\mu_2(\omega)}{(1 + \mu_2(\omega))}, \beta\right)$$
$$S_{ss}(\omega) = W_2(\omega) S_{xx}(\omega) \quad (22)$$

where $L_s = 0.99$, $\alpha = 1$, and $\beta = 10^{(-22/10)}$ is selected to ensure a -22-dB maximum attenuation for the filter in order to reduce the high variance musical noise that normally appears due to rapid changes across adjacent frequency bins.

V. Receiver Operating Characteristics

This section analyzes the proposed VAD and compares its performance to other algorithms used as a reference. The analysis is based on the receiver operating characteristics (ROC) curves, a frequently used methodology to describe the VAD error rate. The Spanish SDC database [22] was used in this analysis. The non-speech hit rate (HR0) and the false alarm rate ($\mathrm{FAR0} = 100 - \mathrm{HR1}$) were determined as the threshold varies being the actual speech frames and actual speech pauses determined by hand-labeling the database on the close-talking microphone. Fig. 1 shows the ROC curves of the proposed VAD and other frequently referred to algorithms [7], [23]–[25] for recordings from the distant microphone in high noisy conditions (i.e., high speed, good road) with an average SNR of about 5 dB. The working points of ITU and ETSI VADs are also included just for reference since they are specifically designed for speech communications and not tunable. The proposed VAD exhibits a shift of the ROC curve when the number of observations (m) increases. The best results are obtained for m close to eight frames, which yields clear improvements in detection accuracy over standardized VADs and over a representative set of recently published VAD algorithms [7], [23]–[25]. Moreover, when contextual information is not used ($m = 0$), the proposed VAD using bispectrum also yields improvements over Sohn's VAD.

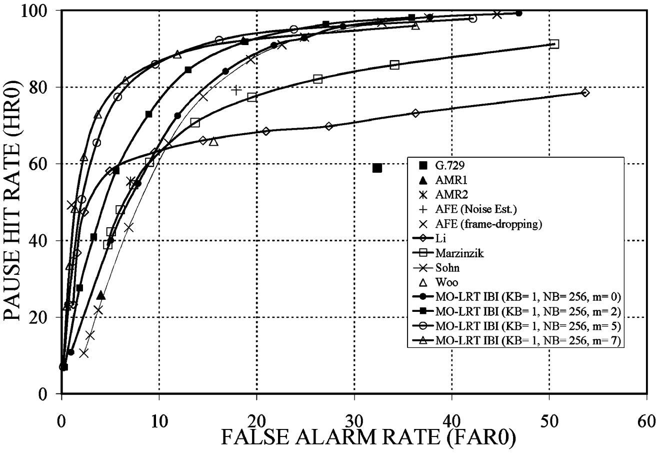

Fig. 1. ROC curves obtained in high noise conditions.

TABLE I
Average Word Accuracy (%) for the AURORA 2 Experiments

	G.729	AMR1	AMR2	AFE	**Proposed**
WF	66.19	74.97	83.37	81.57	**84.15**
WF+FD	70.32	74.29	82.89	83.29	**85.71**
	Woo	Li	Marzinzik	Sohn	**Hand-labelled**
WF	83.64	77.43	84.02	83.89	84.69
WF+FD	81.09	82.11	85.23	83.80	86.86

VI. Speech Recognition Experiments

Performance of speech recognition systems rapidly degrades in noisy environments due to the mismatch between training and testing conditions. In order to compensate for this effect, a previous noise reduction scheme working in combination with a precise VAD is normally used. The accuracy of the VAD has a strong influence on the system performance. There are two clear motivations for that: *1*) the noise parameters such as its spectrum are estimated during non-speech periods being the speech enhancement system strongly influenced by the accuracy of the noise estimation, and *2*) *frame-dropping*, a frequently used technique in robust 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 and irrecoverable errors. The reference (base) framework considered for these experiments was the ETSI AURORA project for distributed speech recognition (DSR) [26], while an enhanced feature extraction scheme incorporating a Wiener filter (WF) noise reduction system and non-speech frame-dropping (FD) was built on the base system.

Table I shows the word accuracies that yielded the different VADs compared. These results are averaged over the three test sets (A, B, and C) of the AURORA-2 recognition experiments [27] and SNRs between 20 and 0 dBs. The proposed integrated

TABLE II
AVERAGE WORD ACCURACY (%) FOR THE SPANISH SDC DATABASES

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

bispectrum MO-LRT VAD outperforms the standard G.729, AMR1, AMR2, and AFE VADs in both clean and multicondition training/testing experiments. When compared to recently reported VAD algorithms, the proposed one yields better results, being the one that is closer to the "ideal" (hand-labeling) speech recognition performance. Similar results were obtained for the experiments conducted on the AURORA 3 Spanish SpeechDat-Car database shown in Table II. Note that these particular databases have longer non-speech periods than the AURORA 2 database, and then, the effectiveness of the VAD results are more important for the speech recognition system. This fact can be clearly shown when comparing the performance of the proposed VAD to Marzinzik VAD [25]. The word accuracy of both VADs is quite similar for the AURORA 2 task. However, the proposed VAD yields significant performance improvements for the AURORA 3 database.

VII. CONCLUSION

This letter showed an effective VAD for improving the performance of speech recognition systems working in noisy environments. The proposed approach is based on a statistical LRT defined on the integrated bispectrum, which is defined as a cross spectrum between the signal and its square, and inherits the ability of higher order statistics to detect signals in noise with many other additional advantages: *1*) its computation as a cross spectrum leads to significant computational savings, and *2*) the variance of the estimator is of the same order as that of the power spectrum estimator. The decision rule incorporates contextual information, a strategy that has reported significant improvements in speech detection and robust speech recognition. With these and other innovations, the proposed method reported significant improvements over standardized techniques such as ITU G.729, ETSI AMR, and AFE VADs, as well as over recently published VADs in speech/pause detection and recognition rate.

REFERENCES

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

[2] J. Ramírez, J. C. Segura, M. C. Benítez, A. de la Torre, and A. Rubio, "A new adaptive long-term spectral estimation voice activity detector," in *Proc. EUROSPEECH*, Geneva, Switzerland, Sep. 2003, pp. 3041–3044.

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

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

[5] A. Sangwan, M. C. Chiranth, H. S. Jamadagni, R. Sah, R. V. Prasad, and V. Gaurav, "VAD techniques for real-time speech transmission on the Internet," in *Proc. IEEE Int. Conf. High-Speed Networks Multimedia Communications*, 2002, pp. 46–50.

[6] F. Basbug, K. Swaminathan, and S. Nandkumar, "Noise reduction and echo cancellation front-end for speech codecs," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 1, pp. 1–13, Jan. 2003.

[7] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[8] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Process. Lett.*, vol. 8, no. 10, pp. 276–278, Oct. 2001.

[9] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 498–505, Sep. 2003.

[10] L. Armani, M. Matassoni, M. Omologo, and P. Svaizer, "Use of a CSP-based voice activity detector for distant-talking ASR," in *Proc. EUROSPEECH*, 2003, pp. 501–504.

[11] R. L. Bouquin-Jeannes and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech Commun.*, vol. 16, pp. 245–254, 1995.

[12] J. M. Górriz, J. Ramírez, 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] J. K. Tugnait, "Detection of nongaussian signals using integrated polyspectrum," *IEEE Trans. Signal Process.*, vol. 42, no. 11, pp. 3137–3149, Nov. 1994.

[14] ——, "Corrections to detection of non-Gaussian signals using integrated polyspectrum," *IEEE Trans. Signal Process.*, vol. 43, no. 11, pp. 2792–2793, Nov. 1995.

[15] A. Sangwan, W. Zhu, and M. Ahmad, "Improved voice activity detection via contextual information and noise suppression," in *Proc. IEEE Int. Symp Circuits Systems (ISCAS)*, May 2005, pp. 868–871.

[16] J. Ramírez, J. C. Segura, C. Benítez, A. de la Torre, and A. Rubio, "An effective subband OSF-based VAD with noise reduction for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 6, pp. 1119–1129, Nov. 2005.

[17] J. Ramírez, J. C. Segura, C. Benítez, L. García, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 689–692, Oct. 2005.

[18] X. Zhang, Y. Shi, and Z. Bao, "A new feature vector using selected bispectra for signal classification with application in radar target recognition," *IEEE Trans. Signal Process.*, vol. 49, no. 9, pp. 1875–1885, Sep. 2001.

[19] X. Liao and Z. Bao, "Circularly integrated bispectra: Novel shift invariant features for high-resolution radar target recognition," *Electron. Lett.*, vol. 34, no. 19, pp. 1879–1880, 1998.

[20] D. R. Brillinger and M. Rosenblatt, *Spectral Analysis of Time Series*. New York: Wiley, 1968, ch. Computation and interpretation of k-th order spectra.

[21] D. Brillinger, *Time Series Data Analysis and Theory*. New York: Holt, Rinehart and Winston, 1975.

[22] 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 *Proce. II LREC Conf.*, 2000.

[23] 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.

[24] 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, May 2002.

[25] 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, Nov. 2002.

[26] *Speech Processing, Transmission and Quality Aspects (stq); Distributed Speech Recognition; Front-End Feature Extraction Algorithm; Compression Algorithms*, ETSI ES 201 108 Recommendation, ETSI, 2000.

[27] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions," in *Proc. ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France, Sep. 2000.