VOICE ACTIVITY DETECTION BASED ON HIGHER ORDER CUMULANTS AND CONVOLUTION

Miguel Enrique Iglesias Martínez[1*], Fidel Ernesto Hernández Montero[2**]
*Universidad de Pinar del Río Hermanos Saíz Montes de Oca, Pinar del Río, Cuba, Email: **Instituto Superior Politécnico José Antonio Echeverría (CUJAE), La Habana, Cuba, Email:

ABSTRACT
This paper refers to the application of higher-order statistical signal processing techniques (cumulant calculation) on noise reduction. The performed procedure, joined to a convolution process, results in the complete estimation (i.e., amplitude, frequency and phase recovery) of any corrupted periodic signal. The aim of this work lies in its application to the voice activity detection (VAD) for environments with high noise levels. The minimum signal to noise ratio for all experiments using the proposed algorithm was -5dB. Obtained results are highly satisfactory compared with existing models.

KEYWORDS: Higher-Order Statistics; Noise Reduction; Convolution; Voice.

MSC: 93E11 93E10 93C40

RESUMEN
Este trabajo esta encaminado a la aplicación de las técnicas de análisis estadístico de orden superior, básicamente basada en el cálculo de cumulantes, en tareas de educción de ruido. El procedimiento propuesto unido a un proceso de convolución, resulta en la estimación completa (amplitud, frecuencia y fase), de cualquier señal periódica contaminada por ruido. El objetivo fundamental de esta investigación es probar su aplicación a la detección de actividad de voz en ambientes con altos niveles de ruido. La relación señal/ruido mínima utilizada en todos los experimentos y aplicando el algoritmo propuesto fue de – 5dB. Los resultados obtenidos son altamente satisfactorios comparados con los modelos existentes.

PALABRAS CLAVE: Estadística de Orden Superior; Reducción de Ruido; Convolución; Voz.

1. INTRODUCTION

Although the voice is a non-stationary signal every 20 to 30 milliseconds, the signal emitted by the vocal tract can be considered stationary, which can be represented by assuming periodicity interval in Fourier series as a sum of harmonic signals. Thus experimentally the proposed algorithm was applied for voice activity detection (VAD) . Real samples of the speech and noise signals obtained from the "REVERB Challenge Speech Enhancement Task" conference of 2013, and NOISEX-92 data base were used.
There are several approaches and algorithms used to treat the application of voice activity detection using principal component analysis (PCA) techniques [22] based on subspaces model, artificial intelligence methods as support vector machines [17],[3] or using adaptive algorithms [5],[4],[12] among others [20],[6],[1],[23],[7],[8],[15]. The main drawback of adaptive methods is that everyone needs a reference signal for the adaptation process, also the method based on artificial intelligence, which in the case of supervised models, a training process is necessary and the computational complexity could be raised. On the other hand, methods based on higher order statistics applied to VAD as in [14],[9],[11],[21],[23], [2], y [13], use the bispectrum, kurtosis and skewness for detecting voice activity in combination with another method, which may complicate the algorithm computationally, because of the higher order spectrum involves the calculation of a multidimensional function.
In relation to the proposed voice activity detection algorithm, this is based on the use of one-dimensional component of the fourth order cumulant and a convolution process, for removing the incident noise on the voice signal. In addition to a linear gain setting and a correlation process between data segments before and

[1] mgi@upr.edu.cu
[2] fhernandez@electrica.cujae.edu.cu

after processing for checking the existence of voice areas.
Although the proposed algorithm uses statistics as higher order cumulants, the convolution-based procedure does not involve a high computational cost. Furthermore, only the one-dimensional component of fourth order cumulant is used.
The work is organized as follows: In section 1 are described the theoretical foundations of the proposed algorithm and a brief description of higher order statistical analysis modeling speech signal is commented. Also in section 2 are shown the experimental results, to subsequently present the conclusions and references consulted.

2. REMOVING NOISE FROM SPEECH SIGNAL THROUGH HIGHER-ORDER STATISTICS

For real value signals, in the problem that concerns removing noise from periodic signal, observed data can be described in terms of a finite sum of harmonic signals. According to the sinusoidal mode1 of [20],[18], a short speech segment is modeled as a sum of sinusoids that are coherent (in-phase) during voiced speech and incoherent during unvoiced speech:. Then the speech signal over a short-term window may be expressed as:

$$y(t) = \sum_{k=1}^{N} A_k \cos(w_k t + \psi_k + \phi_k) + n(t) = x(t) + w(t) \quad \textbf{(1)}$$

where $x(t)$ is the voice signal segment (signal to be detected) and $n(t)$ is additive zero mean Gaussian noise. Besides, A_k, f_k and ϕ_k are the amplitude, frequency and phase, respectively, of the signal. The phase term ψ_k defined as the time when the pitch pulse occurred relative to the beginning of the frame. Since higher-order cumulants of a zero mean Gaussian noise is equal to zero, the estimation of cumulants for noise cancellation can be made starting from the third order, but from [10] all third-order cumulants of complex harmonic are always zero. Consequently this research continues with the use of fourth-order cumulant.

2.1. Fourth-Order Cumulant Calculation.

For a zero-mean stationary random process *z(t)*, and for *k*=3,4, the kth order cumulant of *z(t)* can be defined in term of its joint moments as [19][16]:

$$C_k^z(\tau_1, \tau_2, ..., \tau_{k-1}) = E\{z(\tau_1)...z(\tau_{k-1})\} - E\{g(\tau_1)...g(\tau_{k-1})\} \quad (2)$$

Using equation (3) and working with only the one-dimensional component of the fourth-order cumulant, $C_4^y(\tau_1,0,0)$, by setting $\tau_2 = \tau_3 = 0$, leads to a result similar to that obtained in [16] by setting $\tau_1 = \tau_2 = \tau_3 = \tau$. This one-dimensional component contains original amplitude and frequency of the signal to detect, $x(t)$, although the phase is missed; on the other hand, the noise is entirely removed:

$$C_4^y(\tau_1,0,0) = E\{x(t)^3 \cdot x(t+\tau_1)\} - 3 \cdot E\{x(t) \cdot x(t+\tau_1)\} \cdot E\{x^2(t)\} \quad (3)$$

Then, developing the left term of the equation (4) by substituting $x(t)$ declared in (1):

$$E\{x(t)^3 \cdot x(t+\tau_1)\} = E\left\{\sum_{k=1}^{N} A_k \cos(w_k t + \phi_k)^3 \cdot \sum_{k=1}^{N} A_k \cos(w_k t + w_k \tau_1 + \phi_k)\right\}$$

$$= \sum_{k=1}^{N} A_k^4 E\left\{\cos(w_k t + \phi_k)^3 \cdot \cos(w_k t + w_k \tau_1 + \phi_k)\right\}$$

$$= \sum_{k=1}^{N} \frac{3A_k^4}{8} \cos(w_k \tau_1) \quad (4)$$

Developing the right term of the equation (3) :

$$3 \cdot E\{x(t) \cdot x(t+\tau_1)\} \cdot E\{x^2(t)\} = 3 \cdot \sum_{k=1}^{N} \frac{A_k^{\,2}}{2} \cos(w_k \tau_1) \cdot \sum_{k=1}^{N} \frac{A_k^{\,2}}{2} \quad (5)$$
$$= \sum_{k=1}^{N} \frac{3A_k^{\,4}}{4} \cos(w_k \tau_1)$$

Substituting, in the equation (6), the result obtained in the expression (4) and (5) respectively:

$$C_4^y(\tau_1,0,0) = \sum_{k=1}^{N} \frac{3A_k^{\,4}}{8} \cos(w_k \tau_1) - \sum_{k=1}^{N} \frac{3A_k^{\,4}}{4} \cos(w_k \tau_1) \quad (6)$$

obtaining as result (similar to that obtained in [16]):

$$C_4^y(\tau_1,0,0) = \sum_{k=1}^{N} -\frac{3A_k^{\,4}}{8} \cos(w_k \tau_1) \quad (7)$$

It is clear from equation (7), that the waveform of the original signal is not preserved, which is due to the loss of the phase information of the original signal in the noise cancellation procedure. This is the problem to face in the following section.

2.2. Phase Recovery Method

In order to preserve the phase information of the original voice signal (deterministic) in $C_4^y(\tau_1,0,0)$, a method based on the convolution between corrupted signal, y(t), and $C_4^y(\tau_1,0,0)$ is proposed. In order to theoretically prove the proposed method, let $a(t)$ be a short term voice segment represented as a harmonic signal corrupted by Gaussian noise, $a(t) = \sum_{k=1}^{N} A_k \cos(w_k t + \psi_k + \phi_k) + n(t)$, and $b(t)$, the one-dimensional 4th-order cumulant of the corrupted signal, $C_4^y(\tau,0,0) = \sum_{k=1}^{N} -\frac{3A_k^{\,4}}{8} \cos(w_k \tau)$, (i.e., an equivalent of the free-noise periodic signal, the phase of which is equal to 0). The convolution procedure is developed as follows:

$$a(\tau) * b(\tau) = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} a(t) \cdot b(-t+\tau) dt \quad (8)$$

$$= \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} [\sum_{k=1}^{N} A_k \cos(w_k t + \psi_k + \phi_k) + n(t)][\sum_{k=1}^{N} -\frac{3A_k^{\,4}}{8} \cos(w_k \tau)] dt$$

$$a(\tau) * b(\tau) = \sum_{k=1}^{N} -\frac{3A_k^{\,5}}{16} \cos(w_k \tau + \psi_k + \phi_k) \quad (9)$$

Equation (9) reveals that an equivalent of the original voice signal segment, preserving phase information, is achieved. The method diagram can be observed in figure 1.

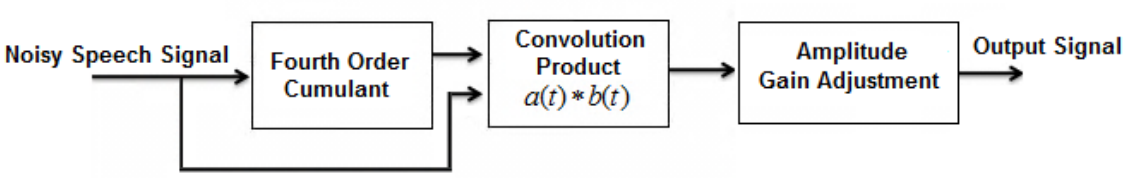


Figure. 1. Block diagram of the proposed algorithm for voice activity detection

3. EXPERIMENTAL RESULTS

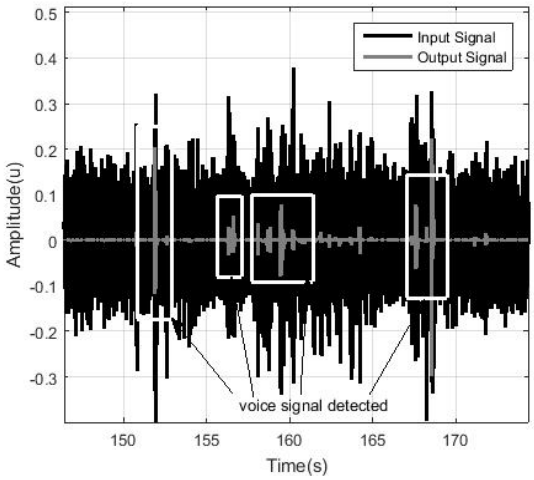
The purpose for the application of voice activity detection is to quantify or receive the voice activity in environments where the signal to noise ratio (SNR) can be low, medium or high. In case of the experiments, a low SNR of -5dB were used. This value is the minimum that has been used in accordance with the existing previous work in the literature consulted.
For the application of voice activity detection the proposed algorithm processes the data vector containing the voice signal of N samples contaminated, and divide it into segments of 25 milliseconds (ms) without overlap, until the length of the data window. The quiet zones present in the data vector are not a problem in processing, because in these areas the correlations between the useful voice signal and the noise it is low, there is only noise activity. Each time a 25ms segment is processed (equivalent to 200 samples, according to the sampling frequency used in the experiments), the correlation with respect to the input signal into this segment is evaluated, to verify if this value is greater than the initial correlation value in the full vector (average correlation sample to sample). At the end of the procedure, from the levels of correlation obtained you can identify where there is no voice activity and where there is not. As a first experiment the proposed algorithm was applied to a voice signal contaminated with pink noise of -5dB of signal to noise ratio. Table 1 shows the correlation levels obtained only in the presence of voice activity during the processing time interval, using a window of 200 samples for this case, equivalent to 25ms with a sampling frequency of 8 kHz, without overlap. The column "Output Correlation in the presence of voice activity" corresponds to the average value of the correlation data obtained during processing, but only in the presence of voice, quiet zones are omitted. This analysis was done in this way because the proposed algorithms in the works consulted in the literature for (VAD) application, omit the quiet zones. The audio signal was contaminated with real samples of pink noise and factory noise respectively.

Table 1. Nist_Clean Audio signal contaminated with pink noise and factory noise. Input SNR - 5dB.

Voice Signal	**Input correlation with Pink Noise**	**Input correlation with Factory Noise**	**Output correlation in the presence of voice activity with Pink Noise**	**Output correlation in the presence of voice activity with Pink Noise**
taavg_A.wav	0.4901	0.4907	0.7177	0.7152
tabqw_B.wav	0.4901	0.4915	0.7444	0.7532
tabst_B.wav	0.4908	0.4903	0.7334	0.7298
tadis_B.wav	0.4901	0.4910	0.6733	0.6729
tadti_A.wav	0.4901	0.4900	0.7034	0.6901
tadtr_B.wav	0.4905	0.4907	0.6103	0.6180
tafan_B.wav	0.4898	0.4903	0.6986	0.7026
tahag_B.wav	0.4912	0.4908	0.7059	0.7174
tahak_B.wav	0.4898	0.4899	0.7045	0.7128
taham_B.wav	0.4908	0.4902	0.6154	0.6195

As an example, in figure 2a it is shown a comparison between the contaminated signal "tabst_B.wav" (in black), and the output obtained after evaluating the algorithm (in gray scale). Encased in a white rectangle there are the voice areas detected, which at the beginning, before processing, was not noticeable. At the same time in figure 2b it is shown a comparison between the contaminated signal "tahag_B.wav" (in black) , and the obtained output after evaluating the algorithm(in gray scale).

a)



b)

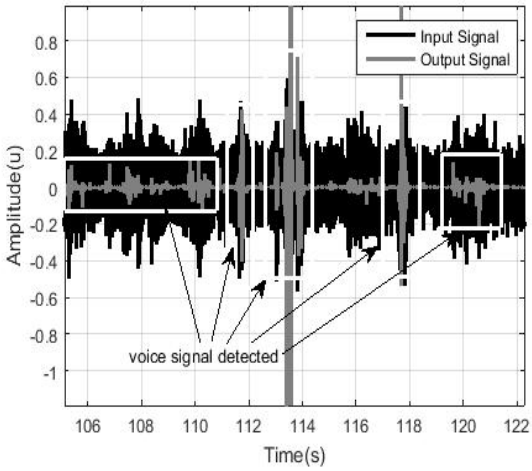


Figure 2 Voice detection areas in the "tabst_B.wav" , signal contaminated with pink noise. Input SNR - 5dB.

4. CONCLUSIONS

This research confirmed the advisability of the application of higher-order statistics combined to a convolution process, for voice activity detection in low signal to noise ratio environments. In this work, the joint use of 4th-order cumulant and convolution, was proposed and argued. Experimental results performed in Matlab were presented using real signals. Finally, an amplitude spectral manipulation was performed in order to restore the original amplitude of each spectral component. Results revealed a high effectiveness, given by the significant signal-to-noise rate and correlations levels enhancement achieved, preserving the amplitude, frequency and phase information of the voice signals to be detected.

RECEIVED: DECEMBER, 2016
REVISED: OCTOBER, 2017

REFERENCES

[1]. ARAKI S., SAWADA H., and MAKINO S. (2007): Blind speech separation in a meeting situation with maximum snr beamformers, **ICASSP**, 1, 41-44.

[2]. CHONG F. and CHUNHUI Z.(2014): Voice activity detection based on ensemble empirical mode decomposition and teager kurtosis, **ICSP** ,455-460.

[3]. DONGWEN YING, YU SHI, SOONG FRANK, DANG JIANWU, and XUGANG LU. (2006): A Robust Voice Activity Detection based on Noise Eigen space Projection. **ISCSLP, LNAI**, No.4274, 76-86.

[4]. ELIAS R., STERGIOU A., BOUKIS CH., SOURETIS G. PNEVMATIKAKIS A., and POLYMENAKOS L.(2006): An Adaptive Speech Activity Detector Based on signal Energy and LDA, **Joint Workshop on Multi-Modal Interaction and Related Machine Learning Algorithms**.

[5]. FUJIMOTO M., ISHIZUKA K., and NAKATANI T. (2008): A Voice Activity Detection based on the Adaptive Integration of Multiple Speech features and a Signal decision Scheme, **ICASSP**, 2, 4441 – 4444.

[6]. GÓRRIZ J. M., RAMÍREZ J., and SEGURA J. C. (2006): Noise Subspace Fuzzy C-means Clustering for Robust Speech Recognition, **LNCS,** 3984, 772-779.

[7]. GÓRRIZ J. M., RAMÍREZ J., and SEGURA J. C.(2005): Bispectrum Estimators for Voice Activity Detection and Speech Recognition, **LNAI,** 3817,174–185.

[8]. GÓRRIZ J. M., RAMÍREZ J., and SEGURA J. C.(2006) : An effective cluster-based model for robust speech detection and speech recognition in noisy environments, **Journal of Acoustical Society of America**, 120, 470-481.

[9]. GORRIZ J. M., RAMÍREZ J., SEGURA J. C., and PUNTONET C.G.(2005): Improved MO-LRT

VAD based on bispectra Gaussian model, **ELECTRONICS LETTERS 21st**,. 41,
[10]. GORRIZ J. M., RAMÍREZ J., SEGURA J. C., PUNTONET C.G., and GARCÍA L.(2006):Effective speech/pause discrimination using an integrated bispectrum likelihood ratio test, **IEEE Xplore, ICASSP**, 1, 801-804.
[11]. GORRIZ J.M., RAMIREZ J., LANG E.W., and PUNTONET C.G. (2006): Hard C-means clustering for voice activity detection, **Elsevier Speech Communications** No. 48, 1638-1649.
[12]. HENNING P., and SOFFKE,O.(2002): An Approach to An Optimized Voice-Activity Detector for Noisy Speech Signals", **11th European Signal Processing Conference**.
[13]. J. NEMER ELIAS(1999): Speech Analysis and Quality Enhancement Using Higher Order Cumulants, **Ph.D Thesis, Ottawa-Carleton Institute for Electrical and Computer Engineering**, disponible en: http://www.collectionscanada.gc.ca/obj/s4/f2/dsk1/tape8/PQDD_0020/NQ48333.pdf, Consulted 15-10,2015.
[14]. KE LI, SWAMY M. N. S., and OMAIR AHMAD M.(2005): An Improved Voice Activity Detection Using Higher Order Statistics, **IEEE Transactions on Speech and Audio Processing**, 13, 965 - 974.
[15]. LEHMANN E.A., and JOHANSSON A.M.(2007): Particle Filter with Integrated Voice Activity Detection for Acoustic Source Tracking, **EURASIP Journal on Advances in Signal Processing**, Article ID 50870.
[16]. NIKIAS, C.L., and MENDEL, J.M. (1993): Signal Processsing with Higher-Order Spectra. IEEE **Signal Processing Magazine**, 10, 10-37.
[17]. Q. HAING JO, PARK, Y.S, LEE K.H, and HYUK CH. J. (2008): A Support Vector Machine-Based Voice Activity Detection Employing Effective Features Vectors, **IEICE Trans. Communication**, E91-B, 2090-2093.
[18]. R- MCAULAY and T. QUATIERI.(1986): Speech Analys id synthesis Based on a Sinusoidal Representation, **IEEE Trans. on Acoustics, Speech, and Signal Processing**, ASSP-34, 744 – 754.
[19]. SALAVEDRA M., and JOSEP M.(1995): Técnicas de Speech Enhancement Considerando Estadísticas de Orden Superior, **Tesis Doctoral, Universidad de Barcelona**, disponible en: http://hdl.handle.net/10803/6943 , Consulted 15-10, 2015.
[20]. SODOYER D., BERTRAND R., GIRIN L., SCHWARTZ JEAN-LUC, and JUTTEN CHRISTIAN. (2006): An Analysis of Visual Speech Information Applied to Voice Activity Detection, **ICASSP**, 1, 601-604.
[21]. SUNHO P., and SEUNGJIN CH.(2008):Gaussian Process Regression for Voice Activity Detection and Speech Enhancement", **IJCNN (IEEE World Congress on Computational Intelligence)**, 2879 – 2882.
[22]. TEMKO, A, MACHO D., and NADEU C. (2007): Enhanced Svm Training For Robust Speech Activity Detection, **ICASSP**, 4, 1025 - 1028.
[23]. XIAOKUN L., and YUNBIN D. (2008): Combining Speech Energy and Edge Information for Fast and Efficient Voice Activity Detection in Noisy Environments, **ICPR**, 1-4.