INTEGRATED POLISPECTRUM ON SPEECH RECOGNITION

Asunción Moreno, Miquel Rutllán

Universidad Politécnica de Cataluña

Dep. Teoría de la Señal y Comunicaciones

Barcelona, Spain

e-mail: asuncion@gps.tsc.upc.es

ABSTRACT

This paper deals with the application of Higher Order Statistics (HOS) in a recognition system where the speech input is corrupted by noise. Cumulants, Biespectrum and Trispectrum are being used in noisy speech signal processing because of their immunity to noise. In this paper we introduce the use of integrated polispectrum and triespectrum in a recognition system based on a filter bank analysis. The results improves other HOS based methods without a substantial increase of the computational load.

1. INTRODUCTION

It is well known [1] that Higher Order Statistics (HOS) (greater than two) of a Gaussian (white or colored) process are zero, third order cumulant of a process with a symmetric p.d.f. is zero and the cumulant of the sum of two independent processes is the sum of the cumulants of both processes. Based on those properties, we apply HOS in the first stage of a recognition system where the speech is corrupted by noise. If the noise is gaussian, the cumulant of the speech input is the cumulant of the clean speech and the same assumption can be done if the noise has a symmetric p.d.f. which is true for most of real noises, and the analysis is performed by third order.

The advantages of using HOS in a recognition system are shown in [2], [3]. Those papers apply HOS to extract the all pole model and the recognition is done based on this model. From those papers we can deduce that HOS works better than second order in noisy conditions and a very simple method based on one slice of the third order cumulant achieves the best results.

Filter bank analysis is commonly applied instead of all pole modeling under noisy conditions, so, it is reasonable to assume that the results will improve using a bank filter. Biespectrum is a two frequencies function and some drawbacks come in its application to a recognition system:

a) The calculus of the bispectrum is computationally expensive

b) The variance of the bispectrum estimator is greater than the variance of the Spectrum estimator.

c) A two dimensional bank filter is computationally expensive and difficult to model physically

Those problems are addressed in this paper. From the experience of using only one slice cumulant and the good results obtained [3], we obtain the spectrum from one slice of the third (or four) order cumulant. The method is computationally not expensive, the main drawback of HOS, and can be related with the Integrated Polispectrum [4]. The integrated polispectrum is a one dimensional function of the frequency and can be directly applied to a mel scale filter bank in the recognition system.

This paper is organized as follows: Section II introduces some basic concepts of HOS and Integrated Spectrum. In Section III we discuss the estimation of the Integrated polispectrum in speech analysis. Section IV shows the recognition system used in this experiment and the results are compared against other methods.

2. INTEGRATED POLISPECTRUM

Let s(t) a stationary, discrete time, zero mean non Gaussian signal. Let n(t) a stationary, discrete time, zero mean, Gaussian signal statistical independent of s(t). The signal s(n) is observed in additive noise x(t) = s(t) +n(t).

The third order cumulant of x(t) is:

$$C_{3x}(i,k) = E\{x(t)\, x(t+i)\, x(t+k)\} \quad (1)$$

Assuming that s(t) and n(t) are independent processes:

$$C_{3x}(i,k) = C_{3s}(i,k) + C_{3n}(i,k) \quad (2)$$

And finally, as the third order cumulant of a Gaussian (white or colored) process is zero

$$C_{3x}(i,k) = C_{3s}(i,k) \quad (3)$$

This equation shows that the cumulant of a process with noise added is the cumulant of the clean process.

The bispectrum of an stationary process is defined as the Fourier Transform of the third order cumulant:

$$B_s(\omega_1,\omega_2) = \sum_{-\infty}^{\infty}\sum_{-\infty}^{\infty} C_{3s}(i,k)\exp(-j(\omega_1 i+\omega_2 k)) \quad (4)$$

From the definition we see that:

$$C_{3s}(i,k) = \frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} B_s(\omega_1,\omega_2)\exp(j(\omega_1 i+\omega_2 k))\, d\omega_1\, d\omega_2 \quad (5)$$

Define the Integrated Bispectrum

$$IB_s(\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} B_s(\omega_1,\omega)\, d\omega_1 \quad (6)$$

From (5) and (6) the one slice zero lag cumulant can be easily related with the Integrated Bispectrum

$$C_{3s}(0,k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} IB_s(\omega)\exp(j\omega k)d\omega \quad (7)$$

can be interpreted as the inverse Fourier transform of the integrated bispectrum.

From (7) and the definition (1)

$$\begin{aligned} IB_s(\omega) &= \sum_{-\infty}^{\infty} C_{3x}(0,k)\exp(-j\omega k) \\ &= \sum_{-\infty}^{\infty} E\{s^2(t)\, s(t+k)\}\exp(-j\omega k) \\ &= S_{s^2 s}(\omega) \end{aligned} \quad (8)$$

The third order one slice zero lag cumulant can also be interpreted as the cross correlation function of the signal and its square function and the integrated bispectrum is the cross spectrum of $s^2(t)$ and $s(t)$.

The integrated bispectrum will be used in the recognition system. As far as the Gaussian noise vanishes in the Bispectrum, also is zero in the Integrated Bispectrum and the property of noise immunity holds.

The fourth order Cumulant of x(t) is defined as:

$$\begin{aligned} C_{4x}(i,k,l) &= E\{x(t)\, x(t+i)\, x(t+k)\, x(t+l)\} \\ &- E\{x(t)\, x(t+i)\}\, E\{x(t+k)\, x(t+l)\} \\ &- E\{x(t)\, x(t+k)\}\, E\{x(t+i)\, x(t+l)\} \\ &- E\{x(t)\, x(t+l)\}\, E\{x(t+k)\, x(t+i)\} \end{aligned} \quad (9)$$

Assuming s(t) and n(t) independent and n(t) Gaussian

$$C_{4x}(i,k,l) = C_{4s}(i,k,l) \quad (10)$$

The Trispectrum is the Fourier Transform of the fourth order cumulant:

$$C_{4s}(i,k,l) = \frac{1}{(2\pi)^3}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} T_s(\omega_1,\omega_2,\omega_3) \times \exp(j(\omega_1 i+\omega_2 k+\omega_3 l))\, d\omega_1\, d\omega_2\, d\omega_3 \quad (11)$$

Define the Integrated Trispectrum $IT(\omega)$

$$IT_s(\omega) = \frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} T_s(\omega_1,\omega_2,\omega)\, d\omega_1\, d\omega_2 \quad (12)$$

From (11) and (12) the one slice zero lag fourth order cumulant is the Fourier Transform of the integrated Trispectrum:

$$C_{4s}(0,0,l) = \frac{1}{2\pi}\int_{-\pi}^{\pi} IT(\omega)\ \exp.(jwl)\, d\omega \quad (13)$$

If we define the process:

$v(t) = s^3(t) - 3\, s(t)\, E\{s^2(t)\}$

and the zero mean process

$$r(t) = v(t) - E\{v(t)\} \quad (14)$$

The fourth order zero lag

$$C_{4s}(0,0,k) = E\{v(t)\, s(t+k)\} = E\{r(t)\, s(t+k)\} \quad (15)$$

and the Integrated Trispectrum

$$\begin{aligned} IT_s(\omega) &= \sum_{-\infty}^{\infty} C_{4x}(0,0,k)\exp(-j\omega k) \\ &= \sum_{-\infty}^{\infty} E\{r(t)\, s(t+k)\}\exp(-j\omega k) \\ &= S_{rs}(\omega) \end{aligned} \quad (16)$$

The Integrated Trispectrum can be seen either as the Fourier Transform of the one slice zero lag fourth order cumulant or as the Fourier transform of the crosscorrelation function of the signal s(t) and a cubic function of s(t).

3. ESTIMATION OF INTEGRATED POLISPECTRUM

Speech signals can be considered locally stationary and the Integrated Polispectrum (IP) is calculated in frames. The method used to estimate the Integrated Polispectrum is

important in the recognition results. The above section shows two interpretations that let us to calculate the IP either by the Fourier Transform of the Cumulant or the Crossperiodogram of $x^2(t)$ (or r(t)) and x(t).

Given a noisy speech frame of N samples x(t) = x(0)...x(N-1), The two interpretations give the following estimation methods:

Define

$$y(t) = x^2(t) - E\{x^2(t)\} \qquad t=0\ldots N-1 \qquad (17)$$

$$v(t) = x^3(t) - 3\, x(t)\, E\{x^2(t)\} \qquad t=0\ldots N-1 \qquad (18)$$

and the zero mean process

$$r(t) = v(t) - E\{v(t)\} \qquad t=0\ldots N-1 \qquad (19)$$

a) Bicorrelogram based method:

$$C_{3x}(0,m) = \sum_{n=0}^{N-|m|-1} y(n)\, x(n+m) \qquad |m| \le N-1 \qquad (20)$$

$$IB_x(k) = \sum_{m} C_{3x}(0,m)\, w_L(m) \exp(-j2\pi km/L) \qquad k=0\ldots L-1 \qquad (21)$$

Where $w_L(m)$ is a window of length L

b) the method based in the cross periodogram is obtained from the following estimations

$$X(k) = \sum_{n=0}^{N-1} x(n) \exp(-j2\pi kn/N) \qquad (22)$$

$$Y(k) = \sum_{n=0}^{N-1} y(n) \exp(-j2\pi kn/N) \qquad (23)$$

$$IB_x(k) = \frac{1}{N} X(k)\, Y^*(k) \qquad (24)$$

To get a better estimator, in some applications is useful to divide the sample sequence in non overlapping blocks of length M<N (N=KM) and average the results.

The Integrated Trispectrum can be calculated in a similar manner:

a) Tricorrelogram based method:

$$C_{4x}(0,0,m) = \sum_{n=0}^{N-|m|-1} r(n)\, x(n+m) \qquad |m| \le N-1 \qquad (25)$$

$$IT_x(k) = \sum_{m} C_{4x}(0,0,m)\, w_L(m) \exp(-j2\pi km/L) \qquad k=0\ldots L-1 \qquad (26)$$

b) The method based in the cross periodogram is obtained from the following estimations

$$X(k) = \sum_{n=0}^{N-1} x(n) \exp(-j2\pi km/N) \qquad (27)$$

$$R(k) = \sum_{n=0}^{N-1} r(n) \exp(-j2\pi km/N) \qquad (28)$$

$$TB_x(k) = \frac{1}{N} X(k)\, R^*(k) \qquad (29)$$

4. RECOGNITION EXPERIMENT

We test the exposed methods in a isolated word recognition problem. For this purpose we use the TIMIT database composed by 11 isolated words (10 digits + "oh"). The database contains 225 speakers, 111 males and 114 females. and each speaker pronounces each word twice. The half of the database is used for training and the other half for testing.

Speech is downsampled to 8Khz. Speech frames are 30 ms long and delayed 15 ms.

The recognition experiment is based on a HTK system. The Integrated Bispectrum or Trispectrum are applied to a mel bank filter composed by 15 filters. 10 cepstrum and Δ cepstrum coefficients and Δ log Energy are used in the recognition.

Each word is modeled by a continuos density HMM of 15 states, and the silence models have 5 states; only transitions between consecutive states are allowed. Each state is modeled by a Gaussian function with diagonal covariance matrix by information vector.

Word models are trained with clean speech signals. Noisy signals are obtained adding Gaussian noise to form SNR of 20, 10, 5 and 0 dB. Noise samples are used to train the silence models.

The recognition rate is defined:

$$\text{Recognition rate: } \frac{\text{Num of correct recognitions}}{\text{Num of test signals}}\, 100\ \%$$

5. RESULTS

A comparison among the different estimation methods give the following conclusions:

1. Correlogram based methods give better results than cross-periodogram based methods.

2. An average of the estimations of M blocks doesn't improve the results obtained with the estimation based in only one block per frame.

3. A rectangular window and a Hamming window have been tested to calculate the Integrated Bispectrum in equation (21). A causal Hamming window of length L= 128 samples has been selected. The L parameter doesn't seems very critical and this value let the use of the FFT algorithm.

4. A rectangular window and a Hamming window have been tested to calculate the Integrated Trispectrum in equation (26). A rectangular window of length L= 128 samples centered in the origin has been selected..

Method\SNR	Clean	20	10	5	0
YW2	98	97	77	49	24
1D3	88	85	82	75	56
1D4	97	93	85	72	49
FB Spec.	99	97	82	61	36
FB Bispec	94	91	88	80	58
FB Trispec	95	93	91	88	73

Table I. % recognition rate with different parameterizations

Table I shows the recognition rates obtained with the exposed methods and are compared against some previously published methods [3]. YW2, 1D3, 1D4 [3] are methods based on all pole modeling and are obtained from second, third and four order cumulants respectively. They are included in the table for comparison purposes.

YW2: Based on the classical estimation of the LPC parameters.

1D3: Obtains the all pole parameters from the unidimensional slice C_{3x} (0,k). This method [3] outperforms the results obtained from the third order Yule-Walker equations solved by LS or TLS.

1D4: Obtains the all pole parameters from the unidimensional slice C_{4x} (0,0,k) solving the equations by the correlation method. This method [3] outperforms the results obtained from the fourth order Yule-Walker equations solved by LS or TLS.

FB Esp, FB Biesp and FB Triesp are methods based on filter bank and are obtained from the Spectrum (classical), Integrated Bispectrum and Integrated Trispectrum respectively.

From the table can be appreciated that methods based on order two statistics are better in high SNR (clean and 20 dB).

HOS based methods achieve better results when SNR decreases. The new proposed methods have a clear advantage when the SNR is lower than 20 dB. Moreover, can be apreciated that the reduction in % recognition rate is small in a great range of SNR (from 5 to 20 dB). For example comparing SNR=5dB and SNR=20 dB, the reduction in %error rate is 37% for FB Spec while the reduction with FB Trispec is only 5%. This results confirm that Bispectrum and Trispectrum keep important information of the speech signal, and the noise reduction in the estimation is significant.

6. CONCLUSIONS

This paper shows the application of bispectrum or trispectrum in a standard recognition system based on a filter bank for recognition of noisy signals.

Integrated bispectrum or trispectrum are calculated from a unique unidimensional slice, reducing significantly the computational load of the calculus of the polispectrum. The chosen onedimensional slice has the additional advantage that can be better estimated from a sample frame of length N than other slices.

The results show a good recognition rate in the SNR range from 5 to 20dB.

The properties of the bispectrum or trispectrum against Gaussian colored noise or real noise, make them very suitable to work in real noise conditions. Any noise with a symmetric probability function has a third order cumulant equal to zero and the Integrated Bispectrum of the noisy signal is the same as the Integrated Bispectrum of the clean signal

The proposed methods can be easily improved applying general techniques used in robust speech recognition systems based on order two statistics.

Acknowledges

This work was supported by grant TIC95-1022-C05-03

7. REFERENCES

1. J. M. Mendel, "Tutorial on Higher-Order Statistics (Spectra) in Signal Processing and System Theory: Theoretical Results and Some Applications", *Proc. IEEE*, vol. 79, no. 3, pp. 278-305, March 1991.

2. K. K. Paliwal and M. M. Sondhi, "Recognition of Noisy Speech Using Cumulant-Based Linear Prediction Analysis", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 429-432, May 1991.

3. Asunción Moreno, Sergio Tortola, Josep Vidal, José A. R. Fonollosa, “New Hos-Based Parameter Estimation Methods For Speech Recognition In Noisy Environments” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1995

4. J. Tugnait. “Detection of non-Gaussian Signals using integrated polyspectrum. IEEE Transaction on Signal Processing, Vol 42, n11 Nov. 1994.