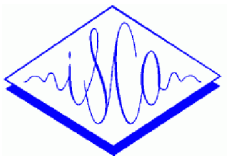
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Time Delay Estimation for Speech Signal Based on FOC-Spectrum

Hong Liu[1], *Xiaofei Li*[2]

[1]Key Laboratory of Machine Perception and Intelligence, Peking University, Shenzhen, CHINA
[2]Key Laboratory of Integrated Micro-system, Peking University, Shenzhen, CHINA

hongliu@pku.edu.cn, lixiaofei0111@sohu.com

Abstract

Higher-order statistics can be used for time delay estimation (TDE) to suppress spatially correlated Gaussian noise, since the higher-order cumulant of Gaussian signal is always zero. However, third-order statistics is invalid for those signals with zero skewness, speech signal as a typical one. In this paper, the fourth-order cumulant (FOC) spectrum is derived, based on which a TDE algorithm that is valid for speech signal and immune to spatially correlated Gaussian noise is proposed. This method can estimate the time delay between two sensor signals or simultaneously estimate the time delays between one sensor signal and other three. In addition, just like generalized cross correlation method, this spectrum domain algorithm is more robust than time domain FOC-based TDE algorithm, especially for speech signal due to its periodicity. Experiments verify the effectiveness of this TDE method for speech signal with spatially correlated Gaussian noise.

Index Terms: time delay estimation, FOC-spectrum, spatially correlated Gaussian noise, speech signal

10.21437/Interspeech.2012-473

1. Introduction

Time delay means the time difference of arrival of a signal which propagates from one source to two sensors. Time delay estimation (TDE) is widely used for radar, sonar, biomedicine, geophysics, etc.. Second-order statistics has been extensively studied for time delay estimation. Knapp proposed Generalized Cross Correlation (GCC) algorithm with many weighting functions for TDE [1]. There are many other methods of TDE based on second-order statistics, such as eigenvalue decomposition [2], acoustic transfer functions ratio [3], and so on.

Higher-order statistics has been widely used for many applications[4-6], which is proposed for TDE to suppress spatially correlated Gaussian noise, since the higher-order cumulant of Gaussian signal equals zero. Nikias estimated time delay using bispectrum between two sensor signals [7]. Using triple sensor signals, Zhang extracted two time delay simultaneously [8]. However, the third-order cumulant and bispectrum of speech signal equal zero because of the zero skewness of speech signal. Therefore, third-order cumulant and bispectrum are invalid for time delay estimation of speech data. Time delay is estimated through maximizing the estimator based on fourth-order cumulant (FOC) in time domain by Tugnait [9] and Liang [10]. However, just like the cross correlation method, for time delay estimation, the periodicity of speech signal and non-Gaussian noise often bring about confused peak of these time domain estimators, which worsen their performance.

In this paper, we derive the fourth-order cumulant (FOC) spectrum and cross spectrum. The multiplicative relationship of each frequency spectrum in fourth-order cumulant spectrum indicates the independence of multiple time delays from one signal to the others. FOC spectrum and cross spectrum are applied to estimate the time delay of speech sensor signals. This TDE method is immune to spatially correlated Gaussian noise and can estimate the time delay between two or multiple sensor signals. Moreover, just like SCOT weighting function in [1], a whitening function is proposed to suppress noise of each channel and weaken confused peaks that time domain estimators suffer.

The rest of this paper is organized as follows: In Section 2, FOC spectrum is derived. TDE algorithms of two and multiple signals are presented in Section 3 and 4. Experimental results and conclusion are presented in Section 5 and 6.

2. Fourth-order cumulant spectrum

The Fourier transform of fourth-order cumulant will be derived in this section. A non-Gaussian signal $x(t)$ is considered. Unlike third-order cumulant, the skewness of $x(t)$ is unrestricted, which can be zero or nonzero.

In practical applications, $x(1), \cdots, x(N)$ denote discrete samples of signal $x(t)$, and N is sample length. The biased estimation of fourth-order cumulant is:

$$\begin{aligned}\hat{\mathrm{c}}_{4x}(\tau_1,\tau_2,\tau_3) = \hat{\mathrm{m}}_{4x}(\tau_1,\tau_2,\tau_3) - \hat{\mathrm{R}}_x(\tau_1)\hat{\mathrm{R}}_x(\tau_3-\tau_2) \\ - \hat{\mathrm{R}}_x(\tau_2)\hat{\mathrm{R}}_x(\tau_3-\tau_1) - \hat{\mathrm{R}}_x(\tau_3)\hat{\mathrm{R}}_x(\tau_2-\tau_1)\end{aligned} \quad (1)$$

where $\hat{\mathrm{m}}$ and $\hat{\mathrm{R}}$ are the biased estimation of fourth-order moment and correlation function, respectively.

FOC spectrum is defined as the $3D$ Fourier transform

$$
\begin{aligned}
P_1 &= \sum_{\tau_1}\sum_{\tau_2}\sum_{\tau_3}\hat{\mathrm{R}}_x(\tau_1)\hat{\mathrm{R}}_x(\tau_3-\tau_2)e^{-j(\omega_1\tau_1+\omega_2\tau_2+\omega_3\tau_3)} \\
&= \sum_{\tau_1}\sum_{\tau_2}\sum_{\tau_3}\{\frac{1}{N}\sum_{n=1}^{N}x(n)x(n+\tau_1)\}\{\frac{1}{N}\sum_{n=1}^{N}x(n+\tau_2)x(n+\tau_3)\}e^{-j(\omega_1\tau_1+\omega_2\tau_2+\omega_3\tau_3)} \\
&= \frac{1}{N^2}\{\sum_{n=1}^{N}x(n)\sum_{\tau_1}x(n+\tau_1)e^{-j\omega_1\tau_1}\}\{\sum_{n=1}^{N}(\sum_{\tau_2}x(n+\tau_2)e^{-j\omega_2\tau_2})(\sum_{\tau_3}x(n+\tau_3)e^{-j\omega_3\tau_3})\} \\
&= \frac{1}{N^2}\{\sum_{n=1}^{N}x(n)e^{j\omega_1 n}X(\omega_1)\}\{\sum_{n=1}^{N}e^{j(\omega_2+\omega_3)n}X(\omega_2)X(\omega_3)\} \\
&= \frac{1}{N^2}X(-\omega_1)(e^{-j(-\omega_2-\omega_3)(N+1)/2}\frac{sin((-\omega_2-\omega_3)N/2)}{sin((-\omega_2-\omega_3)/2)})X(\omega_1)X(\omega_2)X(\omega_3) \\
&\approx \frac{1}{N}X(-\omega_1)\delta(-\omega_2-\omega_3)X(\omega_1)X(\omega_2)X(\omega_3)
\end{aligned}
\tag{3}
$$

of $\hat{\mathrm{c}}_{4x}$. It is well known that Fourier transform of $\hat{\mathrm{m}}_{4x}$ is:

$$
\begin{aligned}
&M_{4x}(\omega_1,\omega_2,\omega_3) \\
&= \frac{1}{N}X(-\omega_1-\omega_2-\omega_3)X(\omega_1)X(\omega_2)X(\omega_3)
\end{aligned}
\tag{2}
$$

where $X(\omega)$ is the Fourier transform of $x(n)$. In addition, Fourier transform of the second term of $\hat{\mathrm{c}}_{4x}$ can be derived as Formula (3). Where $\delta(\omega)$ is the unit sample signal and the relationship $\approx$ is valid when N is big enough. Similarly, the Fourier transform of the third and fourth term of $\hat{\mathrm{c}}_{4x}$ can be obtained as P_2 and P_3.

Finally, the FOC spectrum can be represented as:

$$
\begin{aligned}
P_{4x}(\omega_1,\omega_2,\omega_3) =& M_{4x}(\omega_1,\omega_2,\omega_3)+P_1+P_2+P_3 \\
=& X'(\omega_1,\omega_2,\omega_3)X(\omega_1)X(\omega_2)X(\omega_3)
\end{aligned}
\tag{4}
$$

where,

$$
\begin{aligned}
&X'(\omega_1,\omega_2,\omega_3) \\
=&\frac{1}{N}\{X(-\omega_1-\omega_2-\omega_3)+X(-\omega_1)\delta(-\omega_2-\omega_3) \\
&+X(-\omega_2)\delta(-\omega_1-\omega_3)+X(-\omega_3)\delta(-\omega_1-\omega_2)\}
\end{aligned}
\tag{5}
$$

Similarly, the cross FOC spectrum of four different signals $x_0(n)$, $x_1(n)$, $x_2(n)$ and $x_3(n)$ can be computed as:

$$
\begin{aligned}
&P_{0123}(\omega_1,\omega_2,\omega_3) \\
&= X_0'(\omega_1,\omega_2,\omega_3)X_1(\omega_1)X_2(\omega_2)X_3(\omega_3)
\end{aligned}
\tag{6}
$$

Formula (4) and (6) show that the relationship of each frequency spectrum is multiplicative in FOC spectrum, which causes the additive relationship of each phase spectrum. This property indicates the independence of multiple time delays from one signal to the others.

3. Time delay estimation

In the case of non-Gaussian sound source, higher-order cumulant can suppress spatial correlated Gaussian noise completely. For example, speech sound source and air-conditioning noise are common in practical applications. Four-order cumulant can deal with those signals that have zero skewness, such as speech signal, for which the third-order cumulant is valid.

Two discrete sensor signals are written as:

$$
\begin{aligned}
x_0(n) &= s(n)+v_0(n) \\
x_1(n) &= s(n-D_1)+v_1(n)
\end{aligned}
\tag{7}
$$

where $s(n)$ is the sound source signal, and $v(n)$ denotes spatially correlated Gaussian noise which has zero fourth-order cumulant. Signal $x_0(n)$ denotes reference signal that is considered to have zero time delay with source signal. Signal $s(n)$ is independent with $v(n)$. D_1 denotes the time difference between $x_0(n)$ and $x_1(n)$. Because of the semi-invariance of cumulant, the FOC spectrum of $x_0(n)$ are:

$$
P_{0000}(\omega_1,\omega_2,\omega_3) = P_{4s}(\omega_1,\omega_2,\omega_3) \tag{8}
$$

where the semi-invariance causes that the cumulant of the sum of two independent signals equals to the sum of two cumulants of these two signals. And the FOC of Gaussian signal c_{4v_0} equals to zero. Similarly, the cross FOC and spectrum of these two signals can be calculated by (1) and (6) as:

$$
c_{0100}(\tau_1,\tau_2,\tau_3) = c_{4s}(\tau_1-D_1,\tau_2,\tau_3) \tag{9a}
$$

$$
P_{0100}(\omega_1,\omega_2,\omega_3) = P_{4s}(\omega_1,\omega_2,\omega_3)e^{j\omega_1 D_1} \tag{9b}
$$

Define function:

$$
I(\omega_1,\omega_2,\omega_3) = \psi\frac{P_{0100}(\omega_1,\omega_2,\omega_3)}{P_{0000}(\omega_1,\omega_2,\omega_3)} = e^{j\omega_1 D_1} \tag{10}
$$

where ψ is a weighting function to whiten the spectrum and suppress noise of each channel, which is:

$$
\psi(\omega_1,\omega_2,\omega_3) = \frac{|P_{0000}(\omega_1,\omega_2,\omega_3)|}{|P_{0100}(\omega_1,\omega_2,\omega_3)|} \tag{11}
$$

where $|\cdot|$ denotes the amplitude spectrum. Let ω_2 and ω_3 take arbitrary constants, such as $\omega_2 = \omega_3 = 0$. Then function:

$$T(\tau_1) = \sum_{\omega_1=1}^{N} I(\omega_1, \omega, \omega) e^{-j\omega_1\tau_1} = \delta(\tau_1 - D_1) \quad (12)$$

takes the peak at $\tau_1 = D_1$, which is the time delay sample of two sensor signals. Through whitening the spectrum, ψ improves the time-delay resolution and suppresses the energy of sidelobe. In theory, the values of ω_2 and ω_3 do nothing about the estimation of time delay.

4. Simultaneous time delay estimation for four sensor signals

The cross FOC spectrum of two sensor signals mentioned above indicates the time delay of these two signals. Similarly, the cross FOC spectrum of four sensor signals indicates three time delays between the first signal and the other three, which can be extracted simultaneously.

Two discrete sensor signals are presented in (7), and the other two are written as:

$$\begin{aligned} x_2(n) &= s(n - D_2) + v_2(n) \\ x_3(n) &= s(n - D_3) + v_3(n) \end{aligned} \quad (13)$$

Similar to (9), the FOC cross spectrum of four sensor signals can be calculated by (6) as:

$$\begin{aligned} &P_{0123}(\omega_1, \omega_2, \omega_3) \\ &= P_{4s}(\omega_1, \omega_2, \omega_3) e^{j(\omega_1 D_1 + \omega_2 D_2 + \omega_3 D_3)} \end{aligned} \quad (14)$$

Define function:

$$\begin{aligned} I(\omega_1, \omega_2, \omega_3) &= \frac{|P_{0000}(\omega_1, \omega_2, \omega_3)|}{|P_{0123}(\omega_1, \omega_2, \omega_3)|} \frac{P_{0123}(\omega_1, \omega_2, \omega_3)}{P_{0000}(\omega_1, \omega_2, \omega_3)} \\ &= e^{j(\omega_1 D_1 + \omega_2 D_2 + \omega_3 D_3)} \end{aligned} \quad (15)$$

Data size of spectrum I is N^3, here N is the length of speech signal. Therefore the $3D$ Fourier transform of I has a high computational complexity. In order to reduce the complexity, radial slices of spectrum I are used in this paper, such as set $\omega_1 = k_1\omega$, $\omega_2 = k_2\omega$ and $\omega_3 = k_3\omega$, where $\omega = 1, \cdots, N$. So function I can be written as:

$$I(\omega) = I(\omega_1, \omega_2, \omega_3) = e^{j(k_1 D_1 + k_2 D_2 + k_3 D_3)\omega} \quad (16)$$

the Fourier transform of $I(\omega)$ is:

$$T(\tau) = \delta(\tau - (k_1 D_1 + k_2 D_2 + k_3 D_3)) \quad (17)$$

taking its peak at $\tau = k_1 D_1 + k_2 D_2 + k_3 D_3$.

Three slices are used to obtain three time delays respectively, namely $[k_1, k_2, k_3]$=$[1, 0, 0]$, $[0, 1, 0]$ and $[0, 0, 1]$. Obviously, estimating time delay of two signals mentioned in Section 3 is a specific case of multiple signals through using $x_0(n)$ instead of $x_2(n)$ and $x_3(n)$, and they have identical estimation result theoretically. In addition, two time delays between one signal and other two can also be estimated simultaneously by using this method. In the case of estimating time delays between one signal and other three, the estimator (12) computes cross spectrum and function I three times respectively, otherwise, estimator (17) only compute them once, which has smaller computational complexity.

5. Experiments and analysis

To verify the performance of our method, four sensor speech signals collected in office environment are tested in each experiment. Speech signals are collected by four microphones with the sampling rate $44.1kHz$. Time delays between the first signal and other three are all 50 sampling points. Spatially correlated Gaussian noise are added into speech signals with certain SNR. Time delay of Gaussian noise between the first signal and other three are all 30 sampling points. In addition, another kind of environment noise, namely non-Gaussian noise or spatially uncorrelated noise, exist in office environment, such as computer fans noise. The average signal to environment noise ratio is detected as $12dB$. For convenience, SNR denotes the signal to spatially correlated Gaussian noise ratio below. In total of 200 sets of different speech signals ($200 \times 3 = 600$ time delays) are tested, and Gaussian noise are also different in each set. The duration of each speech signal is about $1s$, which are segmented with length $93ms$ and 50% overlap. For comparison, GCC method with PHAT weighting function in [1] and time domain FOC method in [10] are also tested, which are respectively named as *GCC* and *TFOC* here. The spectrum domain FOC method for two signals and multiple signals presented as estimator (12) and (17) are named as *SFOC1* and *SFOC2* here, respectively.

Wrong estimations brought by Gaussian noise are always around the peak of noise time delay, thus estimation correct rate is tested: If estimation result is around the time delay of speech signal, it is correct, otherwise, it is wrong. Table 1 shows the estimation correct rate versus SNR based on 200 sets of signals. It can be seen that the performance of *GCC* dramatically declines as the intensity of Gaussian noise increases. *TFOC* and our method have similar correct rate. Because of the limited sample number, the probability distribution of noise deviates from Gaussian distribution more or less [11], which leads FOC of noise unequal to zero. Therefore, the performance of *TFOC*, *SFOC1* and *SFOC2* will also slightly decline when SNR is less than $0dB$. The good performance proves that FOC-based method can suppress spatially correlated Gaussian noise effectively. As mentioned in Section 4, estimation results of estimator (12) and (17) are identical. As a result, *SFOC1* and *SFOC2* have identical performance.

Table 1: Time delay estimation correct rate

TDE METHOD	Correct rate (%)			
	$10dB$	$5dB$	$0dB$	$-5dB$
GCC	83.83	44.33	0	0
$TFOC$	100.0	100.0	91.32	83.24
$SFOC1$	100.0	100.0	91.00	83.60
$SFOC2$	100.0	100.0	91.00	83.60

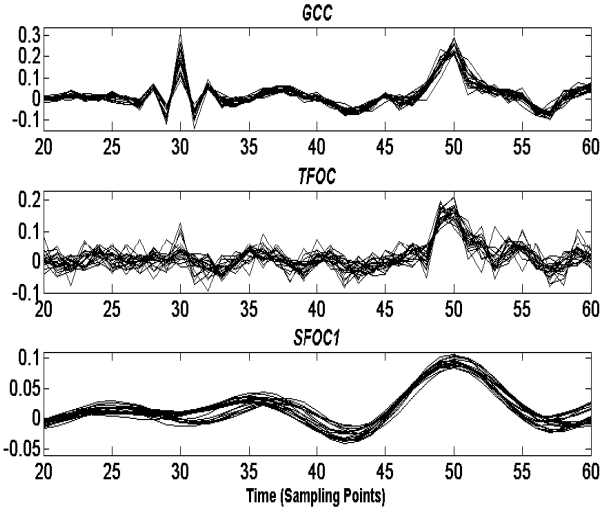


Figure 1: 20 overlapped time delay plots of *GCC*, *TFOC* and *SFOC1* versus time (sampling point) with SNR $5dB$.

Fig. 1 shows the 20 overlapped plots of *GCC*, *TFOC* and *SFOC1* versus time (sampling point) with SNR $5dB$. These 20 plots is randomly selected from 600 time delays. Obviously, *GCC* has a wrong peak at the time delay of Gaussian noise with sampling point 30. *TFOC* and *SFOC1* have correct peak at the 50th sampling point. However, *TFOC* has higher variance and confused peaks at neighboring sampling point of correct peak, which bring about bigger estimation error.

For comparing our method with *TFOC*, since their ability of suppressing Gaussian noise is comparable, estimation mean error (ME) is tested: the average error between estimation value and real value, which can be computed as: $ME = (\sum^{N} |D' - D|)/N$, here D', D and N denote the estimation value, real value and estimation times, respectively. In this experiment, $D = 50$ and $N = 600$. Fig. 2 shows the estimation mean error of *TFOC* and *SFOC1* versus SNR. It is obvious that *SFOC1* has smaller estimation error than *TFOC* when they are less affected by Gaussian noise, such as $5dB$, $10dB$ and "no noise".

6. Conclusions

In this paper the fourth-order cumulant spectrum and cross spectrum are derived. The multiplicative relationship of each frequency spectrum in fourth-order cumulant spectrum guarantees the additive relationship of each phase spectrum, which indicates the independence of multiple time delays from one signal to the others. Then, A time delay estimation method for speech signal is proposed based on FOC spectrum, which can suppress spatially correlated Gaussian noise. Moreover, It is more robust than time domain FOC-based method by weakening confused peaks. Experiments verify that our FOC spectrum method can effectively suppress spatially correlated Gaussian noise and has smaller estimation error than time domain estimator.

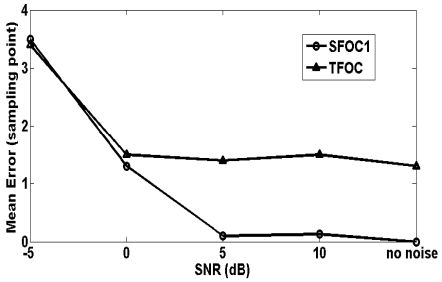


Figure 2: Estimation mean error of *TFOC* and *SFOC1*

7. Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC, No.60875050).

8. References

[1] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay", *in IEEE Trans. ASSP*, vol.24(4), 1976, pp.320-327.

[2] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization", *Journal of Acoustical Society of America*, vol.107(1), 2000, pp. 384-391.

[3] T.G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment", *IEEE Trans. Signal Processing*, vol.85(1), 2005, pp. 177-204.

[4] P. Chevalier, A. Ferrol and L. Albera, "High-Resolution Direction Finding From Higher Order Statistics: The 2q-MUSIC Algorithm", *IEEE Transactions on Signal Processing*, vol.54(8), 2006, pp. 2986-2997.

[5] LD. Lathauwer, J. Castaing and J. Cardoso, "Fourth-Order Cumulant-Based Blind Identification of Underdetermined Mixtures", *IEEE Trans. on Signal Processing*, vol.55(6), 2007, pp. 2965-2973.

[6] H.D. Han and Z. Ding, "A Blind Channel Shortening Criterion Based on High-order Cumulants", *ICASSP2010*, Dallas, Texas, USA, 2010, pp. 3210 - 3213.

[7] C.L. Nikias and R. Pan, "Time Delay Estimation in Unknown Gaussian Spatially Correlated Noise", *IEEE Trans. ASSP*, vol.36(11), 1988, pp. 1706-1714.

[8] W. Zhang and M. Raghuveer, "Nonparametric Bispectrum-based Time-Delay Estimators for Multiple Sensor Data", *IEEE Trans. on Signal Processing*, vol.39(3), 1991, pp. 770-774.

[9] J.K. Tugnait, "Time Delay Estimation with Unknown Spatially Correlated Gaussian Noise", *IEEE Trans. on Signal Processing*, vol.41(2), 1993, pp. 549-558.

[10] Y.C. Liang, A.R. Leyman and B.H. Soong, "Criteria and Algorithms for Time Delay Estimation Based on Cumulants", *IEEE International Symposium on Circuits and Systems*, HongKong, 1997, pp. 2493-2496.

[11] M. Feng and K.D. Kammeyer, "Suppression of Gaussian Noise Using Cumulants: A Quantitative Analysis", ICASSP1997, Stefan Moss, 1997, pp. 3813-3816.