Robust Time Delay Estimation for Sound Source Localization in Noisy Environments

Panayiotis G. Georgiou, Chris Kyriakakis, Panagiotis Tsakalides *

Integrated Media Systems Center
University of Southern California
Los Angeles, CA 90089
georgiou@sipi.usc.edu

ABSTRACT

This paper addresses the problem of robust localization of a sound source in a wide range of operating environments. We use fractional lower order statistics in the frequency domain of two-sensor measurements to accurately locate the source in impulsive noise. We demonstrate a significant improvement in detection via simulation experiments of a sound source in α-Stable noise. Applications of this technique include the efficient steering of a microphone array in teleconference applications.

1. INTRODUCTION

Numerous applications can be envisioned in which microphone array steering is desired. For example, in teleconferencing and telepresence systems it is often required to redirect a video camera so that the person speaking is in the field-of-view. In multi-participant environments it is desirable to provide spatially-selective speech acquisition as well as noise and echo cancellation[1, 2]. Furthermore, the integration of microphone-based tracking with vision-based tracking and facial expression recognition can provide a significant increase in system functionality.

The localization of a source in audio applications has an added complexity not commonly found in other signal processing fields, which arises from the signal being wideband. Additionally, the statistics are not known *a-priori* and they vary with time.

Inter-sensor *Time Delay Estimation* (TDE) is a method commonly used [3, 4] to estimate the position of the source using bearing information ([5] and references therein). The majority of TDE methods proposed so far in audio applications use second or higher order statistics of the measurements to locate the signal of interest. A drawback of these methods is that in impulsive noise or severe interference environments, which are best described by the α-Stable family of distributions, second or higher order statistics are not theoretically defined.

In this paper we interoduce a new method for TDE based on *Fractional Lower Order Statistics* (FLOS) of the received signals. We also examine the behavior of the *Phase Transform* (PHAT) [6] algorithm, which uses second order statistics, under stable noise. We show that when the Gaussian noise assumption fails – and instead α-Stable distribution is a better approximation for the noise – then the FLOS-PHAT algorithm gives better detection than the PHAT.

*The research has been funded by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center with additional support from the Annenberg Center for Communication at the University of Southern California and the California Trade and Commerce Agency.

2. MATHEMATICAL FORMULATION

2.1. TDE in Gaussian Noise

Consider a two-element microphone array receiving

$$\begin{aligned} r_1(t) &= x(t) + n_1(t) \\ r_2(t) &= x(t-\tau) + n_2(t) \end{aligned} \tag{1}$$

in which the noise components $n_1(t)$ and $n_2(t)$ are assumed to be zero mean and uncorrelated with the audio signal $x(t)$, *i.e.*

$E[n_1(t_1)n_1^*(t_2)] = E[n_2(t_1)n_2^*(t_2)] = k\delta(t_1 - t_2)$,
$E[s(t_1)n_1^*(t_2)] = E[s(t_1)n_2^*(t_2)] = 0$, and
$E[n_1(t_1)n_2^*(t_2)] = 0 \ \forall\, t_1, t_2$.

The goal is to estimate the delay τ from measurements of r_1 and r_2, in order to be able to localize the sound source $x(t)$. We are interested in localizing wideband signals, hence we transform the measurements into the frequency domain:

$$\begin{aligned} R_1(k) &= [X(k) + N_1(k)] \\ R_2(k) &= [X(k) \cdot e^{-j\omega_k \tau} + N_2(k)] \end{aligned} \tag{2}$$

We formulate the second-order cross-correlation function of the measurements, assuming the noise has finite second moment statistics:

$$\begin{aligned} C_{R_1R_2}(k) &= E\left\{R_1(k) \cdot R_2(k)^*\right\} \\ &= E\left\{|X(k)|^2 e^{j\omega_k\tau}\right\} + \underbrace{E\left\{N_1(k)N_2^*(k)\right\}}_{0} \\ &+ \underbrace{E\left\{X(k)N_2^*(k)\right\}}_{0} + \underbrace{E\left\{X^*(k)N_1(k)e^{j\omega_k\tau}\right\}}_{0} \end{aligned} \tag{3}$$

A fast method to use for the estimation of the delay between two signals is the *Phase Transform* method [6]. According to PHAT the signal cross spectrum $C_{R_1R_2}(k)$ is smoothed by a window inversely proportional to the magnitude cross spectrum.

$$C^w_{R_1R_2}(k) = \frac{C_{R_1R_2}(k)}{|C_{R_1R_2}(k)|} = e^{j\omega_k\tau} \tag{4}$$

The inverse Fourier transform will result in a sharp peak in the time domain corresponding to the delay τ. Although this method was expected to be quite sensitive to noise, we found that it performed well even for low SNR's.

However, when the process deviates from the ideal Gaussian assumption, and is better characterized by the α-Stable family of distributions, performance degrades.

2.2. TDE in Heavy-Tailed Noise

α-Stable Distributions

The α-Stable distribution which is of a more impulsive nature, is a generalization of the Gaussian distribution, and is appealing because of two main reasons.

- First, it satisfies the *stability property*, which states that if X, X_1 and X_2 are α-Stable independent random variables of the same distribution, then there exist μ_1 and μ_2 satisfying:

$$\nu_1 X_1 + \nu_2 X_2 \stackrel{d}{=} \mu_1 X + \mu_2 \tag{5}$$

  where ν_1, ν_2, μ_1 and μ_2 are constants and $\stackrel{d}{=}$ denotes equality in distribution.

- Second it satisfies, the *Generalized Central Limit Theorem* stating: X is α-Stable , if and only if X is the limit in distributions of the sum:

$$S_n = \frac{X_1 + X_2 + ...X_n}{a_n} - b_n \tag{6}$$

  where X_1, X_2..., are i.i.d. r.v.'s and $a_n \to \infty$

There is no closed form solution for the probability density function of α-Stable distributions, but the characteristic function, is given by:

$$\varphi(t) = \exp\left(j\lambda t - \gamma|t|^{\alpha}\left[1 + j\beta \text{sign}(t)\omega(t,\alpha)\right]\right) \tag{7}$$

where λ is called the *location parameter* ($-\infty < \lambda < \infty$), γ the *dispersion* ($\gamma > 0$) , α is the *characteristic exponent* satisfying $0 < \alpha \leq 2$ and β is the index of symmetry of the distribution ($-1 \leq \beta \leq 1$). For the cases when $\alpha > 1$, λ is the mean of the distribution whereas if $0 < \alpha < 1$, λ is the median, and no mean is defined.

The case of $\alpha = 2$, corresponds to the Gaussian distribution,while $\alpha = 1$, $\beta = 0$ corresponds to the Cauchy distribution. More generally, the smaller the characteristic exponent, the heavier the tails of the density function, and therefore, the more impulsive the noise. For the purposes of this paper, we will deal with the class of *Symmetric α-Stable* ($S\alpha S$) distributions, ($\beta = 0$) with finite mean, *i.e.* $1 < \alpha \leq 2$. References [7, 8, 9] treat the α-Stable theory further.

The class of α-Stable distributions, does not possess finite second (or higher) moment statistics. In fact, α-Stable distributions with $\alpha \neq 2$ have finite statistics only for order p lower than α:

$$\begin{aligned} &\alpha < 2, \quad E|X_\alpha|^p > \infty \quad \forall\, p \geq \alpha \\ &\alpha < 2, \quad E|X_\alpha|^p < \infty \quad \forall\, 0 \leq p < \alpha \\ \text{Gaussian: } &\alpha = 2, \quad E|X_\alpha|^p < \infty \quad \forall\, p \geq 0 \end{aligned} \tag{8}$$

Second-order based methods such as PHAT, operating in environments where the noise components follow a heavy-tailed density belonging to the $S\alpha S$family, suffer from severe performance degradation. In the following we propose a new method that uses the *Fractional Lower Order Statistics* (FLOS) of the received signals.

FLOS-PHAT

The *fractional lower order covariation* of two signals, x and y is defined as:

$$[X,Y]_\alpha = \int_S xy^{\alpha-1}\mu(ds) = \frac{\mathbf{E(XY^{<p-1>})}}{\mathbf{E(|Y|^p)}}\gamma_y \tag{9}$$

where S is the unit circle, $\mu(.)$ is the spectral measure of the $S\alpha S$ random vector (X,Y), $1 \leq p < \alpha$ and $y^{<k>} = |y|^{k-1}\, y^*$

For α-Stable distributions the frequency domain representation of a signal does not converge as $T \to \infty$, but under finite T (*i.e.* after smoothing by a window) its frequency representation exists [9]. Thus we can express the received signals of (1) in the frequency domain (2). Using the properties of $S\alpha S$ distributions [10], and assuming that both the noise and signal have the same distribution we can now form the covariation:

$$D_{R_1R_2}(k) = [R_1, R_2]_\alpha \tag{10}$$

$$= [X(k) + N_1(k), X(k)e^{-j\omega_k\tau} + N_2(k)]_\alpha \tag{11}$$

$$\begin{aligned} &= [X(k), X(k)e^{-j\omega_k\tau} + N_2(k)]_\alpha \\ &\quad + [N_1(k), X(k)e^{-j\omega_k\tau} + N_2(k)]_\alpha \\ &= [X(k), X(k)]_\alpha \left(e^{-j\omega_k\tau}\right)^{<\alpha-1>} + [X(k), N_2(k)]_\alpha \\ &= [X(k), X(k)]_\alpha \left|e^{-j\omega_k\tau}\right|^{\alpha-2} \left(e^{-j\omega_k\tau}\right)^* = Be^{j\omega_k\tau} \end{aligned}$$

in which B is a real and positive number, and thus we can again define a smoothed covariation measure:

$$D^w_{R_1R_2} = \frac{D_{R_1R_2}}{|D_{R_1R_2}|} = e^{j\omega_k\tau} \tag{12}$$

As in the PHAT transform case, the peak in the time domain, resulting from the inverse Fourier transform of $D_{R_1R_2}$ will correspond to the delay τ.

It has been shown [11] that an even better measure is the *Fractional Order Correlation Function* defined as:

$$A_{xy} = E\left\{x^{<a>}y^{<b>}\right\} \tag{13}$$

i.e.

$$A_{R_1R_2}(k) = E\left\{R_1^*(k)^{<a>} \cdot R_2(k)^{<b>}\right\} \tag{14}$$

And we can thus define the FLOS-PHAT method as:

$$A^w_{R_1R_2} = \frac{A_{R_1R_2}}{|A_{R_1R_2}|} = e^{j\omega_k\tau} + \varepsilon_k, \qquad a = b < \frac{\alpha}{2} \tag{15}$$

whose inverse Fourier transform will again result in a sharp peak in the time-domain, corresponding to τ.

3. SIMULATION EXPERIMENTS

To test the performance of the above algorithms we must make use of estimation techniques due to the lack of second and fractional lower order statistics. In this case, the statistics of the problem are not available and they are also variable. The algorithm therefore must be fast and able to adapt to new data and statistics. The simple method suggested by this paper is based on the use of blocks of data. All the data used in the experiments described below were obtained using various speech and music signals sampled at typical audio sampling frequencies of 22050 Hz and 44100 Hz; thus a block of data of about 1000 samples introduces a maximum of 0.1 seconds delay. Using overlapping blocks can decrease the delay even more. The algorithm used to obtain the results below can be summarized as follows:

1. A block of 1024 samples is obtained (using a rectangular window function) from each microphone and their FFT is evaluated.
2. The instantaneous second and lower order statistics (in the frequency domain) are found.
3. A weighted-average statistic is obtained *e.g.*

$$\left[C_{R_1R_2}\right]_t = (1-\rho)\left[C_{R_1R_2}\right]_{t-1} + \rho\left[C_{R_1R_2}\right]_t \qquad (16)$$

   The value of ρ, the *adaptation factor*, (where $0 \leq \rho \leq 1$) determines the trade-off between speed of adaptation of the algorithm on new statistics (ρ near 1) versus the accuracy of the algorithm (ρ small).
4. The PHAT or FLOS-PHAT algorithm is applied using the appropriate weighted-average statistic evaluated in Step 3.
5. Repeat.

An important point here is to define the SNR measure used in this paper. Since power is not defined for α-Stable distributions, the conventional definition of SNR can not be used. Two alternative definitions of SNR are used in literature [12]. In this paper we use the *Generalized*-SNR, defined as the ratio of the signal average power to the dispersion of the noise total in the finite interval of interest:

$$\text{GSNR} = 10\log_{10}\left(\frac{1}{\gamma M}\sum_{t=1}^{M}|s(t)|^2\right) \qquad (17)$$

The algorithm overall converges very fast in about five to ten blocks of data (varies with GSNR) and then stabilizes until an outlier appears in the noise. The results obtained were based on a set of Monte-Carlo runs. Each run starts with a "wrong delay" vector of statistics and so the algorithm has to adapt to the statistics of the signal. After the algorithm reaches steady state, data is gathered to form a "hit/miss" performance curve. In total, 4000 values for each point were considered to obtain the curves in Figure 1. The performance tests were all done with a constant $a = b = 0.2$ value and for different GSNR's of 0, 6, 12 and 25 dB. The comparative values of the GSNR and *Effective*-SNR – defined as the average signal power over the average noise power in the finite interval of interest, for the specific data used in Figure 1 – are given on Table 1.

α	1.0	1.2	1.4	1.6	1.8	2.0
GSNR	Effective-SNR					
0	-52.50	-35.80	-24.79	-15.43	-8.18	-2.94
6	-41.44	-25.45	-15.87	-8.40	-1.95	3.06
12	-28.97	-16.59	-7.43	0.13	4.69	9.06
25	-2.01	4.15	11.36	16.29	19.35	22.06

Table 1: Correspondence of GSNR and average *Effective*-SNR for the specific noise of the conducted measurements.

In impulsive noise conditions, the FLOS-PHAT method greatly outperforms the PHAT method, sometimes by as much as 50%. This happens at a GSNR of 12dB, and around the $\alpha = 1.2$ to 1.6 region.

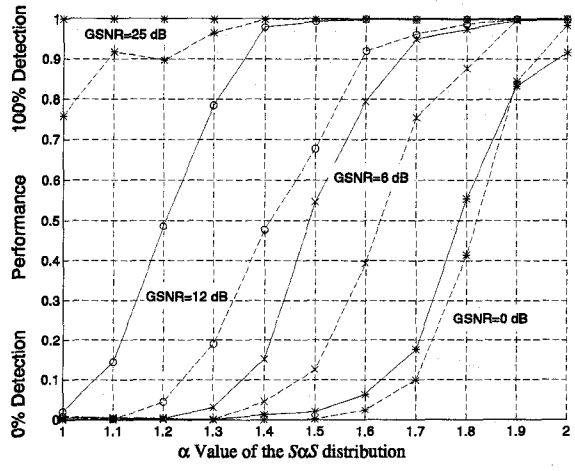

Figure 1: Comparative Performance of the PHAT and FLOS-PHAT methods with $\rho = 0.0125$ and $a = b = 0.2$. Dashed line: PHAT, Solid line: FLOS-PHAT

Although this is the case for heavy-tailed noise, the FLOS-PHAT algorithm fails to outperform the PHAT algorithm only in the Gaussian ($\alpha = 2$) case.

Another way to test the performance of this method is to look at the transient response of the two TDE methods as shown in Figure 2.

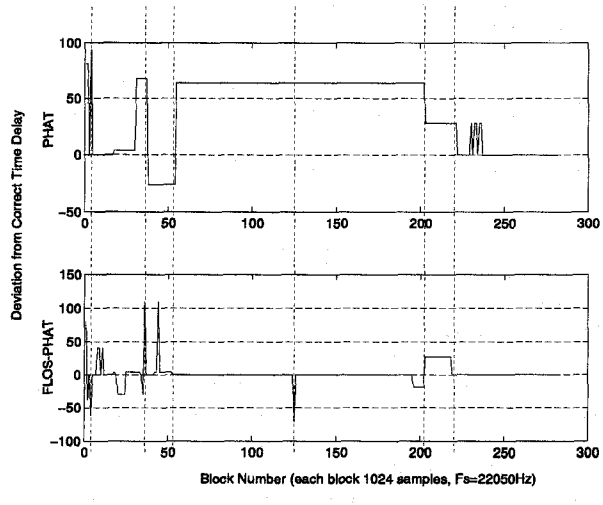

Figure 2: Transient Performance of the PHAT and FLOS-PHAT methods. ($\rho = 0.0125$, $a = b = 0.2$, $\alpha = 1.2$.)

The performance of the PHAT method, based on the time-averaged estimate of $C_{R_1R_2}$, is greatly influenced by impulsive noise. This is due to the fact that an impulsive noise component in the PHAT algorithm is raised to unity power while in the FLOS-PHAT it is raised to fractional power, an operation that limits the effect of the outliers. The FLOS-PHAT method, at severe noise conditions, can

also produce a glitch. However, due to the use of fractional lower order statistics, the significance of the outliers is diminished and thus subsequent estimates are less influenced. It should be noted that the transient response shown in Figure 2 was produced with $S\alpha S$ of $\alpha = 1.2$ which gives an increased number of outliers. This is not a measure of performance, but an indication of the reaction of the two algorithms to the outliers in the noise.

4. CONCLUSIONS

In this paper we have presented a new method for adaptively steering microphone arrays in a wide range of non-Gaussian noise environments. Our method, based on fractional lower order statistics of the measurements, was tested to be better than the second-order based PHAT algorithm, while at the same time adding little computational expense. It is a simple algorithm that gives quite high performance even for small values of α, and can be applied to the "speaker tracking" problem.

As expected, the FLOS-PHAT algorithm does not outperform the PHAT method under Gaussian ($\alpha = 2$) noise conditions. This observation motivates the use of an adaptive algorithm that will adjust to the noise conditions. This can be achieved, by estimating the noise, as well as the signal, and using the calculated [7] value of α to adaptively change the a parameter of the FLOS. In this analysis, a was kept at the value of $a = 0.2$, which satisfies the requirement that $a + b < \alpha$, but which is at the same time very small for values of α near 2. We are also investigating an improvement, possibly of the form $a = b = \left(\frac{\alpha}{2}\right)^z$ (where $z > 1$).

The comparison in this paper between the PHAT and the FLOS-PHAT is a demonstration of the advantages that the use of α-Stable distributions can offer to audio applications. Further research directions suggested by this method include the development of algorithms based on α-Stable and the study of array signal processing fields, such as radar and sonar. Other methods already under investigation for the improvement of this algorithm include, the use of the the *Least p-Norm Approach* [11] and a modification of the *Maximum Likelihood* or *Hannan-Thompson Window* [6].

Finally, we are currently developing an experimental testbed, which will incorporate microphone-based localization techniques and a novel vision-based tracking architecture. This architecture will use the position estimate obtained by the microphone array measurements (from a number of two-sensor arrays). The vision based algorithm will then integrate cues such as stereo disparity, motion, and color in order to identify the location of the speaker's head within the camera image. In addition, this algorithm allows us to identify lip motion, and thus zoom in on the person speaking. The two algorithms can work in parallel so a wrong measurement from the microphone array can be (under some restrictions) ignored based on the ques received by the vision algorithm.

5. ACKNOWLEDGMENTS

The authors are very grateful to Prof. C.L. Nikias who initiated the study on the subject, and whose knowledge, and support are always available.
We would also like to express our thanks to Dr. H. Neven, and Prof. C. von der Malsburg whose team is currently developing the vision-based algorithm that we will incorporate in our sound source locator.

References

1. Mahieux, Y., Tourneur, G. LE., Saliou, A. "A Microphone Array for Multimedia Workstations," J. Audio Eng. Soc., Vol 44, no. 5, May 1996, pp.365-372.
2. Kuo, S. M., Chen, J. "Multiple-Microphone Acoustic Echo Cancellation System with Partial Adaptive Process," *Digital Signal Processing*, Vol. 3, 1993, pp. 54-63.
3. Brandstein, M.S., Adcock, J.E., Silverman, H.F. "A Practical Time-Delay Estimator for Localizing Speech Sources with a Microphone Array," *Computer, Speech and Language*, Vol. 9, September 1995, pp. 153-169.
4. Brandstein, M.S. "A Framework for Speech Source Localization Using Sensor Arrays," Doctoral Dissertation, Brown University, May 1995
5. Brandstein, M.S., Adcock, J.E., Silverman, H.F. "A Closed-Form Location Estimator for use with Room Environment Microphone Arrays," *IEEE Transactions on Speech and Audio Processing*, Vol. 5, no. 1, January 1997, pp. 45-50.
6. Petropulu, P.A., Nikias, C.L. *Higher Order Spectral Analysis: A nonlinear Signal Processing Framework*, Prentice Hall Signal Processing Series, New Jersey, 1993.
7. Shao, M., Nikias, C.L. *Signal Processing with Alpha-Stable Distributions and Applications*, John Wiley and Sons, New York, 1995.
8. Shao, M., Nikias, C.L. "Signal Processing with Fractional Lower Order Moments: Stable Processes and Their Applications," *Proceedings of the IEEE*, Vol. 81, no. 7, July 1993, pp. 986-1010.
9. Masry, E., Cambanis, S. "Spectral Density Estimation for Stationary Stable Processes," *Stochastic Processes and Their Applications*, Vol. 18, 1984, pp. 1-31.
10. Tsakalides, P., Nikias, C.L. "The Robust Covariation Based Music (ROC-MUSIC) Algorithm for Bearing in Impulsive Noise Environments," *IEEE Transactions on Signal Processing*, Vol. 44, no. 7, July 1996, pp. 1623-1633.
11. Ma, X., Nikias, C.L. "Joint Estimation of Time Delay and Frequency Delay in Impulsive Noise," *IEEE Transactions on Signal Processing*, Vol. 44, no. 11, November 1996, pp. 2669-2687.
12. Tsakalides, P., Nikias, C.L. "Maximum Likelihood Localization of Sources in Noise Modeled as a Stable Process," *IEEE Transactions on Signal Processing*, Vol. 43, no. 11, November 1995, pp. 2700-2713.