# ROBUST CROSS-CORRELATION-BASED TECHNIQUES FOR DETECTING AND LOCATING SIMULTANEOUS, MULTIPLE SOUND SOURCES

*Hoang Do*

TokBox Inc.,
San Francisco, CA 94107, USA

*Harvey F. Silverman*

LEMS, School of Engineering,
Brown University,
Providence, RI 02912, USA

**ABSTRACT**

Cross-correlation-based methods have been used extensively in the task of locating multiple, simultaneous sound sources in adverse environments. In this paper, we present a low-cost prealignment enhancement to fix the temporal drawback of the cross-correlation functionals by aligning all the microphone signals to ensure they correlate to the same temporal event. We further introduce a new functional, the steered-response power of the minimum-variance distortionless response using the phase transform (MVDR-PHAT), for multiple-source detection and localization. Experimental results using real data of a 10-talker recording in an adverse room show the improvements of the proposed functional and the prealignment enhancement over traditional techniques in detecting and locating simultaneously active talkers.

***Index Terms*—** Microphone arrays, array signal processing, beam steering, position measurement

## 1. INTRODUCTION

Cross-correlation functionals, such as the steered-response power using the phase transform (SRP-PHAT), have been shown to be robust in sound-source localization tasks [1, 2, 3, 4]. For a hypothesized location $\vec{x}_h$ in the search volume, the SRP-PHAT can be computed as the sum of all the phase generalized cross-correlations (GCC-PHAT's) of unique microphone pair signals, $z_u$ and $z_v$ [5]. For a fixed time-frame of $N$ samples, when the time-difference of arrival (TDOA) between the two microphone signals is large compared to $N$ as depicted in Fig. 1, for a single peak event, all the correlation values between $z_u[n]$ and $z_v[n]$ are poor because there is a mismatch of temporal events. The observed mismatch, if repeated in many microphone pairs $\{u, v\}$, will result in an inappropriately low value of the SRP-PHAT as well as other cross-correlation functionals [6]. This mismatch phenomenon is often evident when there is a significant spatial separation between two microphones $u$ and $v$, as happens in a larger room using a large-aperture microphone array. The errors induced by this problem can be eliminated by some natural prealignment. For cross-correlation-based source-localization methods, the computational cost of a brute-force prealignement is large, as the entire computation is required for any hypothesized location. In this paper, we propose a low cost prealignment and demonstrate that using prealignment is beneficial to enhancing multiple sources, which is desired in the problem of multi-source localization or beamforming.

The second contribution of the paper is an introduction of a new cross-correlation functional, the steered-response power of the minimum-variance distortionless response using the phase transform (MVDR-PHAT). This new functional is especially suitable to the task of locating simultaneous, multiple sources as it is capable of enhancing the weak source(s) when the functional beamformer is steered at them, while suppressing the interferences coming from other dominant sources. Experiments were done using a recording of 10 human talkers by a 182-microphone array in a real room with high background and reverberation noise. The results show an improved performance when using the proposed prealignement enhancement and the new MVDR-PHAT over the nominal SRP-PHAT functional.

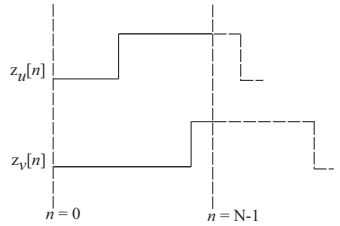


**Fig. 1**. Mismatch of temporal events between microphone signals $z_u$ and $z_v$

## 2. THE LOW-COST PREALIGNMENT ENHANCEMENT

### 2.1. The proposed enhancement

We can prealign the microphone signals to the signal at the hypothesized location $\vec{x}_h$ by shifting the microphone signals by the corresponding time delays in samples, $d_u$, from the location $\vec{x}_h$ to the microphone locations $\vec{x}_u$, $u = 1, ..., M$:

$$\bar{z}_u[n] = z_u[n + d_u], \tag{1}$$

where $\bar{z}_u$ is the prealigned signal of microphone $u$, $n = 0, ..., N-1$, and the time-delay in samples, $d_u$, is defined as:

$$d_u \equiv \left| \frac{||\vec{x}_h - \vec{x}_u||F_s}{C} \right|, \tag{2}$$

where $F_s$ is the sampling frequency, and $C$ is the speed of sound. A brute-force prealignment would require a new set of DFT computations of $\bar{z}_u[n]$ for each and every $\vec{x}_h$. Next, we will present a sequential approach to compute the DFT's for all $\vec{x}_h$, thus, reducing the computational cost significantly.

For a specific microphone configuration and search volume, there exists a finite, upperbound $D$ of the time-delays $d_u$, i.e, $d_u \in [0, ..., D]$. For example, in a typical rectangular room, $D$ would approximately be the time delay between two end points of the main diagonal of the room. Next, a look-up table of the phase

values of the prealigned microphone signals for all possible $D+1$ time-delays is proposed.

The DFT of the $d^{th}$-delay, prealigned microphone signal with a framelength of $N$ (starting at sample 0) is,

$$Z_u[r,d] = \sum_{n=d}^{N+d-1} z_u[n] W_N^{(n-d)r} \tag{3}$$

At $d=0$, we have the base-DFT:

$$Z_u[r,0] = \sum_{n=0}^{N-1} z_u[n] W_N^{nr}, \tag{4}$$

where $r \in [0,...,N-1]$ is the frequency sample, and $W_N = e^{\frac{-j2\pi}{N}}$. The $1^{st}$-delay DFT ($d=1$) can be computed from the base-DFT as follows,

$$\begin{aligned} Z_u[r,1] &= \sum_{n=1}^{N} z_u[n] W_N^{(n-1)r} \\ &= \sum_{n=0}^{N-1} z_u[n] W_N^{(n-1)r} - z_u[0] W_N^{-r} + z_u[N] W_N^{(N-1)r} \\ &= W_N^{-r} \left( Z_u[r,0] - z_u[0] + z_u[N] \right) \\ &= W_N^{-r} \left( Z_u[r,0] + A_1 \right), \end{aligned} \tag{5}$$

where $A_1 = -z_u[0] + z_u[N]$ is a real number. Similarly, we can compute the $d^{th}$-delay DFT based on the $(d-1)^{th}$-DFT:

$$Z_u[r,d] = W_N^{-r} \left( Z_u[r,d-1] + A_d \right), \tag{6}$$

where $A_d = -z_u[d-1] + z_u[N+d-1]$, $d = 1,...,D$. Applying the phase transform (PHAT), i.e., removing the magnitude of the DFT to get the unit-magnitude complex spectrum (UMCS):

$$\begin{aligned} P_u[r,d] &= \frac{Z_u[r,d]}{|Z_u[r,d]|} \\ &= e^{j\theta_u[r,d]}, \end{aligned} \tag{7}$$

where $\theta_u[r,d]$ is the corresponding phase angle at frequency $r$ and time-delay $d$. These UMCS in the frequency domain are stored in an array $\mathbf{P}$ of size $M \times \frac{N}{2} \times (D+1)$ (Here only half of the spectrum is needed, i.e., $r = 0,..., \frac{N}{2}-1$). Therefore, for each $\vec{x}_h$, we can compute the time-delays $d_u, d_v$ for each pair $p = \{u,v\}$, look-up the respective UMCS $P_u[r,d_u], P_v[r,d_v]$ from the stored array $\mathbf{P}$. Using these values, we can compute the GCC-PHAT value of a microphone pair $p = \{u,v\}$, $R_p[d_u,d_v]$, in the time-domain [1, 5]. Note that $R_p[d_u,d_v]$ is computed at the zeroth lag because the two microphone signals are already time-aligned. Hence, our **proposed prealignment algorithm** for $Q$ points $\vec{x}_h$ using $K$ pairs is,

1: Compute the base-DFT ($d=0$) for $M$ microphones as in Eq. 4
2: Compute the $d^{th}$-delay DFT, $d = 1,...,D$ for $M$ microphones according to Eq. 6
3: Compute the UMCS for $M$ microphones as in Eq. 7 and store in $\mathbf{P}$
4: **for** $q = 1 \rightarrow Q$ **do**
5: Calculate $d_u$ according to Eq. 2
6: Look-up $P_u[r,d_u]$ from $\mathbf{P}$ for $u = 1,...,M$
7: Compute the GCC-PHAT values for $K$ pairs
8: Sum-up the GCC-PHAT of $K$ pairs to get the SRP-PHAT:

$$\text{SRP}(\vec{x}_h^{(q)}) = \frac{1}{K} \sum_{p=1}^{K} R_p[d_u,d_v] \tag{8}$$

9: **end for**

### 2.2. Computational cost comparison

The number of hypothesized locations is denoted as $Q$. The computational cost of the brute-force prealignment is $\lambda_1 \approx O(KNQ) + O(MNQ\log_2(N))$, and the cost of the proposed prealignment is $\lambda_2 \approx O(KNQ) + O(MND)$. Although both $\lambda_1$ and $\lambda_2$ are dominated by $O(KNQ)$, the remaining cost of $\lambda_1$, $O(MNQ\log_2(N))$, is about $Q$ times larger than that of $\lambda_2$. The number of hypothesized locations evaluated in the focal volume, $Q$, is often substantially large (a 6m $\times$ 4m $\times$ 1m focal volume implies 24 $\times 10^6$ points using an 1-cm resolution). Hence, the savings obtained by $\lambda_2$ relative to $\lambda_1$ can be significant, (for example in our work $Q$ = 30000, $N$ = 2048, $D$ = 400, and $M$ = 182 yield about 20% saving) see Fig. 2.

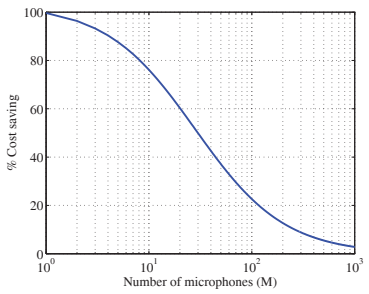


**Fig. 2**. Percent cost saving vs. the number of microphones, M

## 3. THE STEERED MVDR-PHAT RESPONSE POWER

The basic idea of the minimum variance distortionless response (MVDR) beamformer, first proposed by Capon [7], is to retain the signal in the look direction[1] undistorted, while minimizing the power of the signals coming from other directions. This property, from multiple-source localization perspective, is very interesting, since one would like to have a beamforming functional with the capability of spatially suppressing the signals coming from directions (locations) other than the target one. This is important when the target direction (location) is the one of a *weak* desired source. Here, we will quickly derive the MVDR-PHAT functional, a more detailed derivation can be found at [8]. The constituent component of the MVDR is the cross-power spectral density (CPSD) matrix of the microphone signals, $\mathbf{\Phi}_Z$:

$$\begin{aligned} \mathbf{\Phi}_Z(\omega) &\equiv E\{\mathbf{z}(\omega)\mathbf{z}^H(\omega)\} \\ &= \begin{bmatrix} Z_0(\omega)Z_0^*(\omega) & \dots & Z_0(\omega)Z_{M-1}^*(\omega) \\ \vdots & \ddots & \vdots \\ Z_{M-1}(\omega)Z_0^*(\omega) & \dots & Z_{M-1}(\omega)Z_{M-1}^*(\omega) \end{bmatrix}, \end{aligned} \tag{9}$$

The phase transform (PHAT)[1] is a robust weighting function for cross-correlation functionals in reverberant environments.

[1]The term "direction" implies the far-field assumption usually taken for most beamforming applications

Hence, the PHAT can be applied to Eq. 9 to construct a *phase cross-power spectral density matrix*,

$$\bar{\mathbf{\Phi}}_Z(\omega) \equiv \begin{bmatrix} \frac{Z_0(\omega)Z_0^*(\omega)}{|Z_0(\omega)Z_0^*(\omega)|} & \cdots & \frac{Z_0(\omega)Z_{M-1}^*(\omega)}{|Z_0(\omega)Z_0^*(\omega)|} \\ \vdots & \ddots & \vdots \\ \frac{Z_{M-1}(\omega)Z_0^*(\omega)}{|Z_{M-1}(\omega)Z_0^*(\omega)|} & \cdots & \frac{Z_{M-1}(\omega)Z_{M-1}^*(\omega)}{|Z_{M-1}(\omega)Z_0^*(\omega)|} \end{bmatrix} \quad (10)$$

A well-known solution of the MVDR power spectrum derived from the method of Lagrange multipliers at location $\vec{x}_h$ and frequency $\omega$ [7] is:

$$P_{(\text{MVDR})}^{(\omega)}(\vec{x}_h) = \left| \frac{1}{\mathbf{d}^H(\omega)\mathbf{\Phi}_Z^{-1}(\omega)\mathbf{d}(\omega)} \right|, \quad (11)$$

where $\mathbf{d}(\omega)$ is the steering vector from $M$ microphones to $\vec{x}_h$:

$$\mathbf{d}(\omega) = \left[ a_0 e^{-j\omega\tau_0}, a_1 e^{-j\omega\tau_1}, \ldots, a_{M-1} e^{-j\omega\tau_{M-1}} \right]^T, \quad (12)$$

If a natural prealignement is applied when the beamformer steers to $\vec{x}_h$, the steering vector $\mathbf{d}(\omega)$ becomes a constant $(M \times 1)$ vector of propagation attenuation coefficients:

$$\mathbf{a} = [a_0, a_1, \ldots, a_{M-1}]^T, \quad (13)$$

where $a_u = l_u^{-1}$, $l_u$ is the distance from the location of microphone $u$, $\vec{x}_u$, to $\vec{x}_h$, and $\tau_u$ is the travel time from $\vec{x}_h$ to $\vec{x}_u$. Eq. 10, 11, and 13 allow us to constitute the *steered MVDR-PHAT response power* in the frequency-domain for a hypothesized location $\vec{x}_h$:

$$P_{(\text{MVDR-PHAT})}^{(\omega)}(\vec{x}_h) = \left| \frac{1}{\mathbf{a}^T \bar{\mathbf{\Phi}}_Z^{-1} \mathbf{a}} \right| \quad (14)$$

## 4. EXPERIMENTAL EVALUATIONS

### 4.1. Experimental conditions

A grand recording of 10 human talkers was made (with 10 respective close-talking channels) using 182 microphones from the huge microphone array (HMA) [4] in a $8m \times 8m \times 3m$ room with a $T_{60}$ = 450ms. Fig. 3(a) shows the top view of the room, 10 talkers divided into 3 groups with arrows indicating orientations, and the 182 microphones surrounding the focal volume. The microphone-array data, microphone locations and talker locations are available for download online at `http://www.lems.brown.edu/array/data.html#New_data`

### 4.2. Experimental results

The task is to detect and locate active talkers in each frame using the three functionals: nominal SRP-PHAT, enhanced SRP-PHAT (SRP-PHAT using the proposed prealignment), and enhanced MVDR-PHAT. In order to accurately evaluate the performance, a ground truth of how many talkers and which talker(s) are active in each frame is needed. This ground truth can be established by hand-labeling where each person's speech starts and stops in each of the close-talking channels. Once the close-talking channels are hand labeled, using the travel time in samples from the sources to the microphones, one can determine which talker(s) are active at the far-field microphone channels. A simple probability measure of the average speech activity over all microphone channels is then derived to label if a talker $i$ is active in a frame $f$.

**A. Detection evaluation:** Using the hand-labeled ground truth, a frame where the most talkers (6 talkers: T1, T2, T4, T6, T9, and T10) were simultaneously active is identified. For presentation purposes, using the microphone-array data of this frame, a slice of the grid search (1-cm resolution) through the average height of the talkers is plotted using the three functionals (nominal SRP-PHAT, enhanced SRP-PHAT, and enhanced MVDR-PHAT). Fig. 3(b) shows that the nominal SRP-PHAT functional surface is not smooth, and T1 and T2 are not detected. Fig. 3(c) shows the enhanced SRP-PHAT functional has a smoother surface and enhances the talker peaks. Also, it detects T1 and T2, although these two talkers merge into a large peak. The enhanced MVDR-PHAT functional in Fig. 3(d) shows six clear, enhanced peaks with a smooth background. This indicates the improved performance of the enhanced MVDR-PHAT functional in detecting and enhancing multiple talkers compared to that of the classic nominal SRP-PHAT and the enhanced SRP-PHAT.

**B. Localization evaluation:** The performance of the three functionals in locating multiple sources over 300 frames (7.75 seconds of the ten-talker recording) is evaluated. The multiple-source locationing algorithm employing the three functionals is the region zeroing (RZ) algorithm in [4]. Note that the number of active talkers in a frame was being unknown. A 3-D location estimate was considered "correct" if it was within 20-cm of the measured 3-D location of the true source. The size of this allowance is primarily due to the smaller aperture in the height dimension, $Y$. Also, the talkers participated in the experiment were not completely stationary throughout the experiment, thus, this allowance somewhat compromises their movements. A frame was labeled as "correct" if it completely detected all active talkers labeled in the ground truth, and if **all** the location estimates matched the hand-measured locations of the ground-truth talkers. In all frames that are evaluated, the largest number of active talkers labeled by the ground truth in a single frame was 6 talkers, and the least was 2 talkers. There were 10, 92, 94, 64, and 40 frames having 2, 3, 4, 5, and 6 active talkers, respectively. Errors resulted when an algorithm-derived location did not match the ground-truth ("extra") or when the ground-truth locations were missed ("missed"). Percent correct, missed and extra estimates show how many correct, missed and extra location estimates over all estimates, respectively. Table 1 summarizes the performance of the three functionals over all frames and all estimates. It can be seen that in general, the enhanced MVDR-PHAT is better than the enhanced SRP-PHAT and the nominal SRP-PHAT in 3 out of 4 performance factors (except the $\%$ extra$_{est}$, in which the nominal SRP-PHAT is slightly better, which indicates it is a more "conservative" functional than the other two).

## 5. CONCLUSIONS

In this paper, we introduce a low-cost prealignment enhancement for cross-correlation functionals used in sound-source localization. We also present a new functional, steered-response power of the minimum-variance distortionless response using the phase transform (MVDR-PHAT) for the task of detecting and locating simultaneous, multiple sources. The traditional SRP-PHAT and the newly proposed MVDR-PHAT, when combined with the prealignment enhancement, show an improved performance over the nominal SRP-PHAT in detecting and locating the sources in a challenging experiment using a recording of 10 human talkers in a real room with high background and reverberation noise. A more comprehensive study of this work is being prepared for a journal paper. Although the experimental results are not perfect but can be improved more by using some smoothing filters or tracking algorithms, such as the one described in [9].

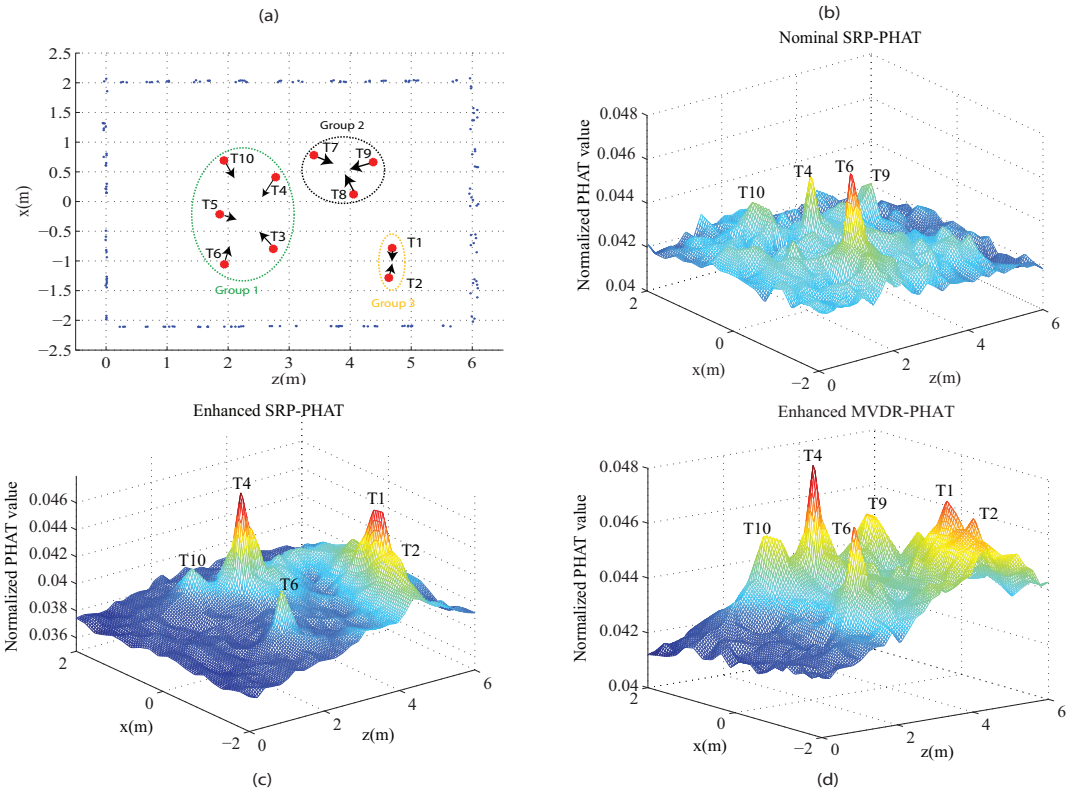


**Fig. 3**. (a) Top view of the room with 10 talkers and 182 HMA microphones; 3-D views of (b) the nominal SRP-PHAT, (c) the enhanced SRP-PHAT, and (d) the enhanced MVDR-PHAT

| Percent | Nominal SRP-PHAT | Enhanced SRP-PHAT | Enhanced MVDR-PHAT |
|---|---|---|---|
| % $\text{corr}_{fr}$ | 36 | 41 | 45 |
| % $\text{corr}_{est}$ | 73.41 | 73.78 | 75.79 |
| % $\text{missed}_{est}$ | 21.40 | 19.62 | 18.40 |
| % $\text{extra}_{est}$ | 5.43 | 6.60 | 5.83 |

**Table 1**. Multiple source localization performance of the three functionals

## 6. REFERENCES


[1] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[2] S. T. Birchfield, "A unifying framework for acoustic localization," in *Proc. of European Signal Processing Conference (EUSIPCO 2004)*, Vienna, Austria, Sept. 2004, pp. 1127–1130.

[3] J. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2510–2526, Nov. 2007.

[4] H. Do and H. F. Silverman, "Srp-phat methods of locating simultaneous multiple sources using a frame of microphone array data," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Dallas, TX, March 2010, pp. 125–128.

[5] J. H. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*, Ph.D. thesis, Brown University, Providence, RI, May 2000.

[6] G. Carter, "Bias in magnitude-squared coherence estimation due to misalignment," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 1, pp. 97 – 99, Feb 1980.

[7] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408 – 1418, Aug 1969.

[8] H. Do and H. F. Silverman, "A robust sound-source separation algorithm for an adverse environment that combines mvdr-phat with the casa framework," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-11)*, New Paltz, NY, Oct. 2011, To appear.

[9] J. Valin, F. Michaud, and J. Rouat, "Robust 3d localization and tracking of sound sources using beamforming and particle filtering," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Toulouse, France, May 2006, vol. 4, pp. 841–844.