# CALIBRATED ACOUSTIC SOURCE LOCALIZATION

*Jacek Dmochowski, Jacob Benesty, and Sofiène Affes*

Université du Québec, INRS-EMT, Montréal, Québec, Canada

## ABSTRACT

The localization of speech sources is critical to the proper functioning of many multiple-microphone devices. Presently, localization algorithms rely on the time-differences-of-arrival (TDOA) to extract the source location from the observed cross-correlation measurements, and view other components (non-direct-path) of the impulse responses as "interference." In this paper, we propose to employ the set of reverberant impulse responses to provide diversity and hence aid in the localization of sound sources in reverberant and diffractive acoustic environments. A relationship between the cross-correlation function of the microphone outputs and the cross-correlation functions of the room impulse responses is established. This relationship is then exploited to yield location estimates which exhibit a significantly lower anomaly rate than the popular SRP-PHAT method as found in an experimental evaluation using real impulse responses.

## 1. INTRODUCTION

Multiple-microphone devices are quickly becoming commonplace in a wide array of modern personal communication devices. The benefits of employing multiple microphones stem from the spatial discrimination abilities of a spatial aperture – speaker localization and tracking, dereverberation, echo cancellation, and noise reduction. In general, to provide speech enhancement using multiple microphones, the location of the desired speaker must be extracted from the array measurements.

The vast majority of source localization algorithms are based on the time-differences-of-arrival (TDOA) across the array. The process of estimating these relative delays is termed *time delay estimation* [1]. By examining the relative delays observed across one or more pairs of microphones, the location is estimated using the known array geometry. The advantage of the relative delay approach is that it requires only the knowledge of the relative microphone positions which do not vary with the acoustic environment. Unfortunately, the presence of reverberation introduces many non-direct path components into the impulse responses to the array. It is well-known that reverberation is the biggest problem obstructing the emergence of a truly robust acoustic source localizer.

Among the relative-delay based source localization techniques, the steered response power (SRP) method [2], [3] represents arguably the most reliable algorithm. When combined with the phase transform (PHAT) generalized cross-correlation weighting [1], the method is known as SRP-PHAT. The SRP (or SRP-PHAT) method belongs to a framework of localization algorithms based on location-parameterized spatial correlation detailed in [4]. There are few localization methods which explicitly model the enclosure impulse responses. The *precedence effect* of the human auditory system has been applied to microphone arrays in an attempt to localize sound in reverberant environments [5]; the idea here is that after a silent period, sound reaches one of the microphones sooner than the others. By carefully examining these onset times across the microphones, the source may be localized. A more recent approach based on the blind identification of the impulse responses to a pair of microphones is proposed in [6], with extensions of the work found in [7]. Once the impulse responses are estimated, the difference (in argument) of the peak values of the impulse responses is designated as the estimate of the relative delay which is then mapped to a direction-of-arrival. Lastly, a recent paper [8] utilizes the assumed *a priori* room impulse response information in a sophisticated manner to map the observed cross-correlation measurements to an estimate of the source location which takes reverberation as a cue.

In this paper, we present a method also based on the impulse response information of the enclosure: that is, the environment is assumed to be calibrated.

## 2. SIGNAL MODEL

Consider an array of $N$ microphones that samples the sound field within an arbitrary enclosure. Assume for now that a single sound source is present and located at $\mathbf{r}$. Denoting the impulse response from the source to the $n$th microphone by $h_{\mathbf{r},n}$, the output of sensor $n$ at time $k$ is then modeled as:

$$\begin{aligned} x_n(k) &= h_{\mathbf{r},n}(k) * s(k) + v_n(k),\ n = 1, \ldots, N, \\ &= \sum_{l=0}^{L_h-1} h_{\mathbf{r},n}(l)\, s(k-l) + v_n(k), \end{aligned} \tag{1}$$

The authors would like to acknowledge the funding for this work provided by the National Science and Engineering Research Council (NSERC) of Canada.

where $s$ is the source signal, modeled as a zero-mean and wide-sense stationary random process, $*$ denotes linear convolution, $L_h$ is the length of the longest impulse response, and $v_n$ is the additive noise at sensor $n$ which is uncorrelated with the source signal.

## 3. RELATION BETWEEN IMPULSE RESPONSES TO THE ARRAY

Consider the cross-correlation function between microphones $i$ and $j$:

$$R_{x_i x_j}(\tau) = E\left\{x_i(k)\, x_j(k+\tau)\right\}, \tag{2}$$

where it has been assumed that the microphone outputs are jointly wide-sense stationary random processes. Substituting (1) into (2) and ignoring the additive noise term for ease of analysis leads to:

$$R_{x_i x_j}(\tau) = \sum_{l_1=0}^{L_h-1} \sum_{l_2=0}^{L_h-1} h_{\mathbf{r},i}(l_1) h_{\mathbf{r},j}(l_2) R_{ss}(\tau - l_2 + l_1). \tag{3}$$

Assuming that $s(k)$ is a zero-mean white process with variance $\sigma_s^2$, we have

$$\begin{aligned} R_{x_i x_j}(\tau) &= \sigma_s^2 \sum_{l=0}^{L_h-1-\tau} h_{\mathbf{r},i}(l) h_{\mathbf{r},j}(l+\tau) \\ &= \sigma_s^2 R_{h_{\mathbf{r},i} h_{\mathbf{r},j}}(\tau). \end{aligned} \tag{4}$$

From (4), we see that the cross-correlation function of the microphone outputs is a scaled version of the deterministic cross-correlation between the impulse responses to the microphones. Thus, we can expect the cross-correlation functions of the various microphone pairs to exhibit peaks at the peak lags of the deterministic cross-correlation functions of the impulse responses. By comparing the observed cross-correlation functions with those that would be yielded by a source located at given location, we may determine the likelihood that the source is located at that location. This of course requires knowledge of the room impulse responses. Theoretically, the impulse responses of the room may be determined *a priori* using a calibration procedure: the impulse responses from all candidate locations may be measured. Alternatively, acoustic modeling software taking into account the geometry and acoustic properties of the room may generate estimates of the impulse responses.

### 3.1. Non-White Sources

If the source signal is not white, (3) does not simplify to (4); instead, the cross-correlation elements $R_{ss}(\tau)$, $\tau \neq 0$ act as "noise" terms to the relationship of (4). Note that $R_{ss}(\tau) \leq R_{ss}(0)$, $\forall \tau$, and thus, the largest component of $R_{x_i x_j}(\tau)$ is indeed $\sigma_s^2 R_{h_{\mathbf{r},i} h_{\mathbf{r},j}}(\tau)$.

Consider rewriting the cross-correlation function of (2) in the frequency-domain:

$$R_{x_i x_j}(\tau) = \int_{-1/2}^{1/2} G_{x_i x_j}(f) e^{j2\pi f \tau} df, \tag{5}$$

where $G_{x_i x_j}(f)$ is the cross-spectral density (CSD) between microphones $i$ and $j$ at frequency $f$. By definition,

$$G_{x_i x_j}(f) = E\left\{X_i(f) X_j^*(f)\right\}, \tag{6}$$

where $X_i(f)$ is the Fourier transform of $x_i(k)$ and $^*$ denotes complex conjugation. Transposing (1) to the frequency domain,

$$X_n(f) = H_n(f) S(f) + V_n(f), \tag{7}$$

where $S(f)$ and $V_n(f)$ are the Fourier transforms of $s(k)$ and $v_n(k)$, respectively. Substituting (7) into (6) and ignoring the noise term,

$$G_{x_i x_j}(f) = H_i(f) H_j^*(f) G_{ss}(f). \tag{8}$$

Stating that the cross-correlation between the microphone outputs is equal to a scaled version of the cross-correlation between the corresponding impulse responses is equivalent to stating the their cross-spectra are equal up to a scaling factor:

$$G_{x_i x_j}(f) = K H_i(f) H_j^*(f), \tag{9}$$

where $K$ is some constant. However, it is clear from (8) that (9) does not hold for a non-white signal (for a white signal, the relation holds with $K = \sigma_s^2$). It is therefore required to suppress the influence of $G_{ss}(f)$ from $G_{x_i x_j}(f)$.

To achieve this, the generalized cross-correlation phase transform (GCC-PHAT) pre-filtering is employed. The cross-spectrum of the microphone signals is "whitened":

$$G_{x_i x_j}^{\mathrm{PHAT}}(f) = \frac{G_{x_i x_j}(f)}{\left|G_{x_i x_j}(f)\right|}. \tag{10}$$

Substituting (8) into (10), we obtain:

$$G_{x_i x_j}^{\mathrm{PHAT}}(f) = \frac{H_{\mathbf{r},i}(f) H_{\mathbf{r},j}^*(f)}{\left|H_{\mathbf{r},i}(f) H_{\mathbf{r},j}^*(f)\right|}. \tag{11}$$

The right side of (11) is simply the whitened cross-spectrum of the impulse responses, and thus, the following relationship is established:

$$R_{x_i x_j}^{\mathrm{PHAT}}(\tau) = R_{h_{\mathbf{r},i}, h_{\mathbf{r},j}}^{\mathrm{PHAT}}(\tau), \tag{12}$$

where

$$\begin{aligned} R_{x_i x_j}^{\mathrm{PHAT}}(\tau) &= \int_{-1/2}^{1/2} \frac{G_{x_i x_j}(f)}{\left|G_{x_i x_j}(f)\right|} e^{j2\pi f \tau} df \\ &= \int_{-1/2}^{1/2} \frac{H_{\mathbf{r},i}(f) H_{\mathbf{r},j}^*(f)}{\left|H_{\mathbf{r},i}(f) H_{\mathbf{r},j}^*(f)\right|} e^{j2\pi f \tau} df \\ &= R_{h_{\mathbf{r},i}, h_{\mathbf{r},j}}^{\mathrm{PHAT}}(\tau). \end{aligned} \tag{13}$$

## 4. LOCALIZATION METHOD

For every candidate location $\mathbf{r}$, we measure or compute the set of $N$ impulse responses from that location to the array: $h_{\mathbf{r},n}(k)\ \forall\, n = 1, 2, \ldots, N; k = 0, 1, \ldots, L_h - 1$. For every location $\mathbf{r}$ and each unique microphone pair $(i, j)$, we compute the whitened cross-correlation function $R^{\mathrm{PHAT}}_{h_{\mathbf{r},i},h_{\mathbf{r},j}}(\tau)$ and denote its $p$ peak lags by the set $\kappa_{\mathbf{r},i,j}$:

$$\kappa_{\mathbf{r},i,j} = \arg\max_{\tau}{}^{p} \left\{ R^{\mathrm{PHAT}}_{h_{\mathbf{r},i},h_{\mathbf{r},j}}(\tau) \right\}, \tag{14}$$

where $\arg\max^p$ denotes the $p$ largest values of a set. The $p$ peak lags of the impulse responses from all candidate locations to all microphone are stored in a look-up table.

During runtime, the cross-correlation function between all unique microphone pairs $(i, j)$ are computed for all physically realizable delays:

$$\tau \leq \left\lceil \frac{f_{\mathrm{s}}}{c} d_{i,j} \right\rceil, \tag{15}$$

where $f_{\mathrm{s}}$ is the sampling frequency, $c$ is the speed of sound propagation, and $d_{i,j}$ is the distance between microphones $i$ and $j$. The likelihood of the source being at location $\mathbf{r}$ is then estimated by:

$$\mathcal{L}(\mathbf{r}) = \sum_{(i,j)\in P} \sum_{\tau\in\kappa_{\mathbf{r},i,j}} R^{\mathrm{PHAT}}_{x_i x_j}(\tau), \tag{16}$$

where $P$ is the set of all unique microphone pairs. The location which yields the highest likelihood is chosen as the location estimate:

$$\hat{\mathbf{r}} = \arg\max_{\mathbf{r}\in L} \mathcal{L}(\mathbf{r}), \tag{17}$$

where $L$ is the location space.

### 4.1. Special Case: Anechoic Environment

In an anechoic environment, the impulse response from the source to microphone $n$ is given by:

$$h_{\mathbf{r},n}(k) = \alpha_{\mathbf{r},n}\delta\left[k - \mathcal{F}_{1,n}(\mathbf{r})\right], \tag{18}$$

where the function $\mathcal{F}_{i,j}(\mathbf{r})$ translates the location of the source to the anechoic relative delay experienced between microphones $i$ and $j$. In (18), microphone 1 is used as the reference. It then follows that

$$R_{h_{\mathbf{r},i},h_{\mathbf{r},j}}(\tau) = \alpha_{\mathbf{r},i}\alpha_{\mathbf{r},j}\delta\left[\tau - \mathcal{F}_{i,j}(\mathbf{r})\right], \tag{19}$$

where

$$\mathcal{F}_{i,j}(\mathbf{r}) = \mathcal{F}_{1,j}(\mathbf{r}) - \mathcal{F}_{1,i}(\mathbf{r}). \tag{20}$$

As a result,

$$\kappa_{\mathbf{r},i,j} = \{\mathcal{F}_{i,j}(\mathbf{r})\} \tag{21}$$

is a singleton and

$$\mathcal{L}(\mathbf{r}) = \sum_{(i,j)\in P} R_{x_i x_j}\left[\mathcal{F}_{i,j}(\mathbf{r})\right], \tag{22}$$

which is precisely the SRP algorithm. Thus, the proposed scheme generalizes the SRP method to reverberant environments.

## 5. EXPERIMENTAL EVALUATION

To evaluate the efficacy of the proposed scheme, an experimental evaluation using data obtained in the Varechoic chamber at Bell Labs was conducted [9]. The chamber's dimensions are 6.6-by-5.85-by-2.75 m. The layout of the chamber's microphones and speaker positions is shown in Figure 1: a set of 31 candidate positions was considered as the location space. The first 3 elements of the Bell Labs uniform linear array were employed in the evaluation. The location of microphone $n, n = 1, 2, 3$, is given by $(2.437 + 0.1n, 0.5, 1.4)$ m. The 60 dB reverberation decay time was chosen to be $T_{60} = 280$ ms [10].

The evaluation employs the chamber's measured impulse responses from each location to each microphone. For every unique microphone pairing $(i, j)$, the cross-correlation function $R^{\mathrm{PHAT}}_{h_{\mathbf{r},i},h_{\mathbf{r},j}}(\tau)$ is computed for each candidate location $\mathbf{r}$. A data set is generated using the impulse responses from location 29: the impulse responses from this chosen location are convolved with a clean speech signal to simulate propagation of speech to the array. Spatially white Gaussian noise with signal-to-noise ratios (SNRs) ranging from 0 to 20 dB is then added to the microphone signals.

The data is then processed in 128 ms frames (sampled at 48 kHz) over a recording of 10 seconds. The proposed scheme is implemented and compared to the SRP-PHAT algorithm [3], a common benchmark for modern localization algorithms. The proposed method is run for $p \in \{5, 10, 15\}$. To evaluate the performance, the number of anomalous estimates is noted – in our case, an anomaly is defined as a selection of a candidate location which does not correspond to the true location (location index 29).

The results are shown in Table I. The proposed scheme offers a significant improvement over the SRP-PHAT technique for all noise levels. Moreover, there seems to be a trade-off in choosing $p$. As $p$ is increased from $p = 1$ (SRP-PHAT), the additional impulse response information aids in localizing the source (i.e., matching the observed microphone cross-correlations to the cross-correlations of the impulse responses across the location space). However, one cannot simply take $p$ to be an extremely high value. This is because such large values of $p$ lead to low-level peaks being included in the sets $\kappa_{\mathbf{r},i,j}$ – by including these small peaks, we are not gaining much information (diversity), but we are making the algorithm vulnerable to large noise energy at these cross-correlation lags.

**Table 1**. Source localization performance of conventional and proposed estimators with real impulse responses.

| | SRP-PHAT | Proposed $p = 5$ | Proposed $p = 10$ | Proposed $p = 15$ |
|---|---|---|---|---|
| SNR (dB) | $\%_{\text{anomalies}}$ (%) | $\%_{\text{anomalies}}$ (%) | $\%_{\text{anomalies}}$ (%) | $\%_{\text{anomalies}}$ (%) |
| 0 | 50 | 42 | 32 | 35 |
| 10 | 31 | 15 | 13 | 13 |
| 20 | 22 | 4 | 1 | 3 |

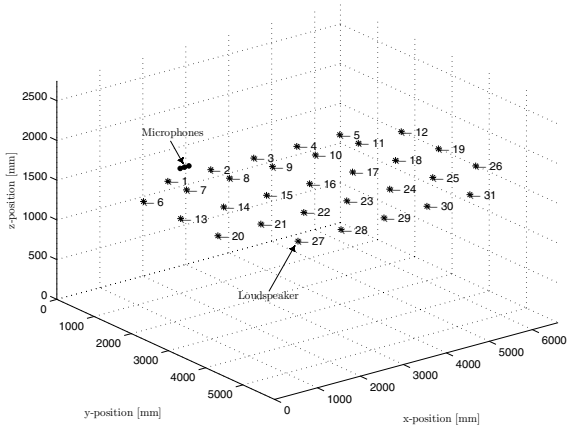


**Fig. 1**. Layout of varechoic chamber [9].

## 6. CONCLUSION

The localization of speech sources is an immensely challenging problem. This difficulty stems from the fact that the current source localization models do not correspond closely to reality. Human speech is not a point source. The vast majority of acoustic environments are not anechoic. Relative delays are not necessarily constant across the frequency band; this occurs with conformal (embedded) microphone arrays which diffract or scatter the incoming sound. Since the acoustic impulse response encapsulates all of this information within its tap values, the ideas presented in this paper take the first step towards mitigating these model inaccuracies.

It is obvious that the *a priori* knowledge of the impulse responses from every candidate location to every microphone is not readily available. The calibration procedure required to measure the various impulse responses may be time-consuming, and is environment dependent. Emerging acoustical simulation software may lighten the tediousness involved in manual calibration. The results presented in this paper demonstrate the benefits of treating reverberation as diversity as opposed to interference for localization applications.

## 7. ACKNOWLEDGMENTS

This work was initially inspired by the ideas presented in [8].

## 8. REFERENCES

[1] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, pp. 320–327, Aug. 1976.

[2] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech and Audio Processing,* vol. 5, pp. 288–292, May 1997.

[3] J. Dibiase, H.F. Silverman, and M.S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications* (M. S. Brandstein and D. B. Ward, eds.), pp. 157–180, Springer-Verlag, Berlin, 2001.

[4] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, pp. 1327–1339, May 2007.

[5] J. Huang, N. Ohnishi, and N. Sugie, "A biomimetic system for localization and separation of multiple sound sources," *IEEE Trans. Instrumentation and Measurement*, vol. 44, pp. 733–738, June 1995.

[6] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Am.*, vol. 107, pp. 384–391, Jan. 2000.

[7] Y. Huang and J. Benesty, "Adaptive multichannel time delay estimation based on blind channel identification," in *Adaptive Signal Processing: Application to Real-World Problems*, J. Benesty and Y. Huang, Eds., Berlin: Springer, 2003.

[8] Z. Fodróczi and A. Radványi, "Localization of directional sound sources supported by a priori information of the acoustic environment" *EURASIP Journal on Advances in Signal Processing*, vol. 2008, Article ID 287167, 14 pages, 2008.

[9] A. Härmä, Acoustic Measurement Data from the Varechoic Chamber, Tech. Memo., Agere Systems, Nov. 2001.

[10] M. R. Schroeder, "New method for measuring reverberation time," *J. Acoust. Soc. Am.*, vol. 37, pp. 409–412, 1965.