WHY DOES PHAT WORK WELL IN LOW NOISE, REVERBERATIVE ENVIRONMENTS?

Cha Zhang, Dinei Florêncio and Zhengyou Zhang

Microsoft Research
One Microsoft Way, Redmond, WA 98052, USA
{chazhang,dinei,zhang}@microsoft.com

ABSTRACT

Among many existing time difference of arrival (TDOA) based sound source localization (SSL) algorithms, the Phase Transform (PHAT) is extremely popular for its excellent performance in low noise environments, even under relatively heavy reverberation. However, PHAT was developed as a heuristic approach and its working principle has not been completely understood. In this paper, we present the relationship between PHAT and a maximum likelihood (ML) framework for multi-microphone sound source localization. We show that when the environment noise approaches zero, PHAT is indeed a special case of the ML algorithm, which explains its good performance under low noise environments. In addition, we show that as long as the noise stays low, PHAT remains optimal in ML sense even when the room reverberation is heavy, which explains its robustness over reverberation.

Index Terms— Sound source localization, phase transform, maximum likelihood, noise, reverberation

1. INTRODUCTION

Sound source localization (SSL) using microphone arrays has been an active research topic since the early 1990's [1]. It has found many important applications such as human-computer interaction and intelligent rooms. Depending on the application scenario, a number of SSL techniques are popular, including steered-beamformer (SB) based, high-resolution spectral estimation based, time delay of arrival (TDOA) based, and learning based. Among them, the TDOA based approaches have received extensive investigation [1, 2, 3, 4, 5, 6].

Consider an array of P microphones. Given a source signal $s(t)$ and its frequency representation $S(\omega)$, the signals received at these microphones can be modeled in the frequency domain as [5, 7]:

$$X_i(\omega) = \alpha_i(\omega)S(\omega)e^{-j\omega\tau_i} + H_i(\omega)S(\omega) + N_i(\omega), \quad (1)$$

where $i = 1, \cdots, P$ is the index of the microphones, τ_i is the time of propagation from the source location to the i^{th} microphone; $\alpha_i(\omega)$ is a gain factor that includes the propagation energy decay of the signal, the gain of the corresponding microphone, the directionality of the source and the microphone, etc; $N_i(\omega)$ is the noise sensed by the i^{th} microphone; $H_i(\omega)S(\omega)$ represents the convolution between the environmental response function and the source signal, often referred as the *reverberation*. In many existing SSL approaches [8, 1, 6], the reverberation term was ignored for simplicity.

The generalized cross correlation (GCC) based SSL maximizes the sum of weighted cross correlation between each pair of the received signals as:

$$\mathcal{R}(\mathbf{s}) = \sum_{i=1}^{P}\sum_{k=1}^{P}\int \Psi_{ik}(\omega)X_i(\omega)X_k^*(\omega)e^{j\omega(\tau_i-\tau_k)}d\omega. \quad (2)$$

GCC has been investigated widely in literature [8]. While many different weighing functions $\Psi_{ik}(\omega)$ can be applied, the heuristic-based PHAT weighting [8] defined as:

$$\Psi_{ik}(\omega) = \frac{1}{|X_i(\omega)X_k^*(\omega)|} = \frac{1}{|X_i(\omega)||X_k(\omega)|} \quad (3)$$

has been found to perform very well under realistic acoustical conditions [2, 7]. Inserting Eq. (3) into Eq. (2), one gets:

$$\mathcal{R}(\mathbf{s}) = \int \Big|\sum_{i=1}^{P}\frac{X_i(\omega)e^{j\omega\tau_i}}{|X_i(\omega)|}\Big|^2 d\omega, \quad (4)$$

This algorithm is called SRP-PHAT [9], where SRP stands for steered response power. PHAT was first developed by Carter et al. in [10] as an ad hoc technique. Experiments show that PHAT works very well under low noise environments, even when the reverberation of the room is high. Due to its high performance and low computational complexity, PHAT or its variants have since received a lot of attention and been used in a number of systems, such as [1, 2, 3, 7]. Nevertheless, the reason why PHAT works so well in practice has not been fully explored. In [1, 7], the authors showed that a maximum likelihood (ML) approach to sound source localization leads to PHAT under low noise conditions. However, their results are limited to one pair of microphones. It is not clear whether the

1-4244-1484-9/08/$25.00 ©2008 IEEE

same statement is valid in multiple microphone cases, where the direct extension of the ML approach in [1, 7] is in fact suboptimal [11].

In this paper, we compare the SRP-PHAT algorithm with a TDOA based ML algorithm for multi-microphone sound source localization we developed in our previous work [11]. We show that under the assumption that the environment noise is low, PHAT can actually be derived from our ML-based SSL (ML-SSL) algorithm. Our research lays the ground for two important facts about PHAT: first, PHAT is indeed optimal in ML sense when the noise is low; second, PHAT is very robust to reverberation, because its optimality is independent of the amount of environment reverberation.

The rest of the paper is organized as follows. We briefly review the ML based SSL algorithm in Section 2. Relationship between the ML-SSL algorithm and PHAT is discovered in Section 3. Experiments and conclusions are given in Section 4 and 5, respectively.

2. THE MAXIMUM LIKELIHOOD SSL

Let us start by rewriting Eq. (1) into a vector form:

$$\mathbf{X}(\omega) = S(\omega)\mathbf{G}(\omega) + S(\omega)\mathbf{H}(\omega) + \mathbf{N}(\omega), \tag{5}$$

where

$$\begin{aligned}
\mathbf{X}(\omega) &= [X_1(\omega), \cdots, X_P(\omega)]^T, \\
\mathbf{G}(\omega) &= [\alpha_1(\omega)e^{-j\omega\tau_1}, \cdots, \alpha_P(\omega)e^{-j\omega\tau_P}]^T, \\
\mathbf{H}(\omega) &= [H_1(\omega), \cdots, H_P(\omega)]^T, \\
\mathbf{N}(\omega) &= [N_1(\omega), \cdots, N_P(\omega)]^T.
\end{aligned}$$

Among the variables, $\mathbf{X}(\omega)$ represents the received signals, hence it is known. $\mathbf{G}(\omega)$ can be estimated or hypothesized during the SSL process, which will be detailed later. The reverberation term $S(\omega)\mathbf{H}(\omega)$ is unknown, and we will treat it as another type of noise.

To make the above model mathematically tractable, we assume the combined total noise,

$$\mathbf{N}^c(\omega) = S(\omega)\mathbf{H}(\omega) + \mathbf{N}(\omega), \tag{6}$$

follows a zero-mean, independent between frequencies, joint Gaussian distribution, i.e.,

$$p(\mathbf{N}^c(\omega)) = \rho \exp\Big\{ -\frac{1}{2}[\mathbf{N}^c(\omega)]^H \mathbf{Q}^{-1}(\omega)\mathbf{N}^c(\omega)\Big\}, \tag{7}$$

where ρ is some constant; superscript H represents Hermitian transpose, $\mathbf{Q}(\omega)$ is the covariance matrix, which can be estimated by:

$$\begin{aligned}
\mathbf{Q}(\omega) &= E\{\mathbf{N}^c(\omega)[\mathbf{N}^c(\omega)]^H\} \\
&= E\{\mathbf{N}(\omega)\mathbf{N}^H(\omega)\} + |S(\omega)|^2 E\{\mathbf{H}(\omega)\mathbf{H}^H(\omega)\}
\end{aligned} \tag{8}$$

Here we assume the noise and the reverberation are uncorrelated.

Given the covariance matrix $\mathbf{Q}(\omega)$, the likelihood of the received signals can be written as:

$$p(\mathbf{X}|S, \mathbf{G}, \mathbf{Q}) = \prod_\omega p(\mathbf{X}(\omega)|S(\omega), \mathbf{G}(\omega), \mathbf{Q}(\omega)), \tag{9}$$

where

$$p(\mathbf{X}(\omega)|S(\omega), \mathbf{G}(\omega), \mathbf{Q}(\omega)) = \rho \exp\big\{ -J(\omega)/2\big\}, \tag{10}$$

$$J(\omega) = [\mathbf{X}(\omega) - S(\omega)\mathbf{G}(\omega)]^H \mathbf{Q}^{-1}(\omega)[\mathbf{X}(\omega) - S(\omega)\mathbf{G}(\omega)]. \tag{11}$$

The goal of the proposed sound source localization is thus to maximize the above likelihood, given the observations $\mathbf{X}(\omega)$, gain matrix $\mathbf{G}(\omega)$ and noise covariance matrix $\mathbf{Q}(\omega)$. Note the gain matrix $\mathbf{G}(\omega)$ requires information about where the sound source comes from, hence the optimization is usually solved through hypothesis testing. That is, hypotheses are made about the source source location, which gives $\mathbf{G}(\omega)$. The likelihood are then measured. The hypothesis that results in the highest likelihood is determined to be the output of the SSL algorithm.

In our previous work [11], we have shown that the solution of the above maximum likelihood formulation is to maximize:

$$J_2 = \int_\omega \frac{[\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{X}(\omega)]^H \mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{X}(\omega)}{\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{G}(\omega)} d\omega \tag{12}$$

In the next session, we will show that under certain assumptions, this ML-SSL algorithm can be simplified as the SRP-PHAT algorithm widely used in practice.

3. FROM ML-SSL TO PHAT

We start by examining the combined noise covariance matrix introduced in Eq. (8). The first term in Eq. (8) can be directly estimated from the silence periods of the acoustical signals:

$$E(N_i(\omega)N_j^*(\omega)) = \lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} N_{ik}(\omega)N_{jk}^*(\omega), \tag{13}$$

where k is the index of audio frames that are silent. Note the background noises received at different microphones may be correlated, such as the ones generated by computer fans in the room. In such cases, the above covariance matrix will be non-diagonal.

The second term in Eq. (8) is related to reverberation. It is generally unknown. As an *approximation*, we assume it is diagonal:

$$|S(\omega)|^2 E\{\mathbf{H}(\omega)\mathbf{H}^H(\omega)\} \approx \text{diag}(\lambda_1(\omega), \cdots, \lambda_P(\omega)), \tag{14}$$

with the i^{th} diagonal element as:

$$\begin{aligned}\lambda_i(\omega) &= E\{|H_i(\omega)|^2|S(\omega)|^2\} \\ &\approx \gamma(|X_i(\omega)|^2 - E\{|N_i(\omega)|^2\}) \qquad (15)\end{aligned}$$

where $0 < \gamma < 1$ is an empirical parameter. Eq. (15) assumes that the reverberation energy is a portion of the difference between the total received signal energy and the environmental noise energy. It also assumes that there are many indirect paths, such that when the reflected signals arrive at the microphones, they are largely independent. While it is difficult to justify either assumption theoretically, these two assumptions have been used in the literature [2, 7] and showed very good performance in practice.

The i^{th} diagonal elements of the combined covariance matrix $\mathbf{Q}(\omega)$ can thus be written as:

$$\begin{aligned}\kappa_i(\omega) &= \lambda_i(\omega) + E\{|N_i(\omega)|^2\} \\ &= \gamma|X_i(\omega)|^2 + (1-\gamma)E\{|N_i(\omega)|^2\} \qquad (16)\end{aligned}$$

Due to the computational cost involved in inverting a full $\mathbf{Q}(\omega)$ matrix for each frequency bin ω, in practice we usually assume that $\mathbf{Q}(\omega)$ is diagonal, i.e.:

$$\mathbf{Q}(\omega) \approx \text{diag}(\kappa_1(\omega), \cdots, \kappa_P(\omega)) \qquad (17)$$

Another variable in Eq. (12) is the gain factor $\alpha_i(\omega)$ embedded in $\mathbf{G}(\omega)$. In certain applications, $\alpha_i(\omega)$ can be measured before hand. Otherwise, we may assume it as a positive real number and estimate it as follows:

$$\begin{aligned}|\alpha_i(\omega)|^2|S(\omega)|^2 &= |X_i(\omega)|^2 - \lambda_i(\omega) - E\{|N_i(\omega)|^2\} \\ &\approx (1-\gamma)(|X_i(\omega)|^2 - E\{|N_i(\omega)|^2\}), \\ & \qquad (18)\end{aligned}$$

where both sides represent the power of the signal received at microphone i without the combined noise (noise and reverberation). Therefore, we have:

$$\alpha_i(\omega) = \sqrt{(1-\gamma)(|X_i(\omega)|^2 - E\{|N_i(\omega)|^2\})}/|S(\omega)|, \qquad (19)$$

Given Eq. (17) and (19), Eq. (12) can be simplified as:

$$J_2 = \int_\omega \frac{1}{\sum_{i=1}^P |\alpha_i(\omega)|^2/\kappa_i(\omega)} \left| \sum_{i=1}^P \frac{\alpha_i^*(\omega)}{\kappa_i(\omega)} X_i(\omega) e^{j\omega\tau_i} \right|^2 d\omega \qquad (20)$$

In order to derive PHAT, let us assume that the signal to noise ratio (SNR) is very high, i.e., $|X_i(\omega)|^2 \gg E\{|N_i(\omega)|^2\}$. We have the following approximations under such a condition:

$$\begin{aligned}\mathbf{Q}(\omega) &\approx \text{diag}(\kappa_1(\omega), \cdots, \kappa_P(\omega)) \\ &\approx \text{diag}(\gamma|X_1(\omega)|^2, \cdots, \gamma|X_P(\omega)|^2) \qquad (21) \\ \alpha_i(\omega) &\approx \sqrt{(1-\gamma)|X_i(\omega)|^2}/|S(\omega)| \qquad (22)\end{aligned}$$

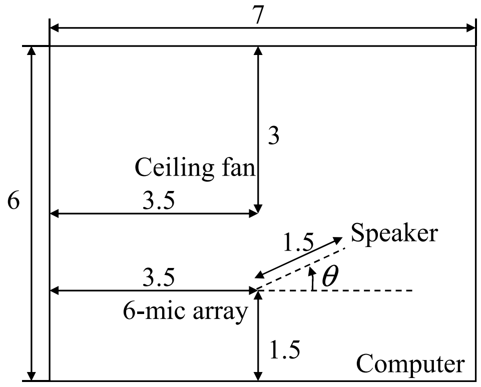

Fig. 1. Top-down view of the virtual room for synthetic experiments.

Inserting into Eq. (20) one obtains:

$$\begin{aligned}J_2 &\approx \int_\omega \frac{\left|\sum_{i=1}^P \frac{|X_i(\omega)|}{\gamma|X_i(\omega)|^2} X_i(\omega)e^{j\omega\tau_i}\right|^2}{\sum_{i=1}^P \frac{|X_i(\omega)|^2}{\gamma|X_i(\omega)|^2}} d\omega \\ &= \frac{1}{\gamma P} \int \left| \sum_{i=1}^P \frac{X_i(\omega)e^{j\omega\tau_i}}{|X_i(\omega)|} \right|^2 d\omega \qquad (23)\end{aligned}$$

which is equivalent to SRP-PHAT (4).

There are two noticeable conclusions that can be drawn from the above derivation. First, when the signal to noise ratio is high, PHAT is a special case of the ML-SSL algorithm, which supports its optimality under low noise environments. Second, in Eq. (23), the reverberation parameter γ is outside the PHAT computation. This indicates that as long as the noise stays low, PHAT remains an optimal solution in maximum likelihood sense regardless the amount of reverberation in the environment. These two conclusions provides strong evidence why PHAT works well in low noise, reverberative rooms.

4. EXPERIMENTAL RESULTS

In this section, we compare the performance of SRP-PHAT (Eq. (4)) and ML-SSL (Eq. (20)) on a synthetic scene, where the noise level and reverberation can be well controlled. A virtual room with size $7 \times 6 \times 2.5$ meters is created, as shown in Fig. 1. A circular 6-microphone array is placed near the center of the room, at $(3.5, 1.5, 1)$. The radius of the microphone array is 0.135 m. A speaker is talking at a distance of 1.5 m from the center of the microphone array, at an angle θ. We introduce two noise sources in the scene. A ceiling fan is mounted in the middle of the room, at $(3.5, 3, 2.5)$, and a computer is located in the corner, at $(7, 0, 0.5)$. The wave signals from the speaker, the fan and the computer are all recordings from the real world. The reverberation effect of the room is

Table 1. Experimental results of SRP-PHAT and ML-SSL accuracy on the synthetic dataset. Cells with bold fonts indicate best performance in the group.

Reverberation = 100 ms

Input SNR	SRP-PHAT		ML-SSL					
			$\gamma = 0.1$		$\gamma = 0.3$		$\gamma = 0.5$	
	<2°	<10°	<2°	<10°	<2°	<10°	<2°	<10°
25 dB	97.6%	**98.9%**	**97.9%**	98.8%	97.9%	**98.9%**	97.8%	**98.9%**
20 dB	92.0%	93.6%	92.8%	94.7%	**93.0%**	**94.9%**	92.7%	94.6%
15 dB	89.0%	91.4%	**91.6%**	**93.9%**	91.5%	93.8%	91.2%	93.7%
10 dB	85.2%	88.8%	**89.0%**	**91.7%**	88.8%	90.9%	88.1%	90.4%
5 dB	76.1%	82.0%	**87.2%**	**90.3%**	85.9%	89.7%	85.2%	89.2%
0 dB	64.5%	71.1%	**81.2%**	**88.0%**	77.4%	84.0%	75.7%	82.9%

Reverberation = 500 ms

Input SNR	SRP-PHAT		ML-SSL					
			$\gamma = 0.1$		$\gamma = 0.3$		$\gamma = 0.5$	
	<2°	<10°	<2°	<10°	<2°	<10°	<2°	<10°
25 dB	**60.1%**	**79.2%**	60.0%	78.8%	59.8%	78.8%	59.9%	78.8%
20 dB	59.4%	78.4%	**60.3%**	**78.9%**	59.7%	78.7%	59.6%	78.6%
15 dB	60.3%	78.0%	**60.4%**	**78.8%**	60.1%	78.5%	59.6%	78.4%
10 dB	58.8%	77.0%	**59.8%**	77.1%	59.5%	77.6%	59.2%	**77.7%**
5 dB	56.3%	**75.5%**	**57.4%**	75.2%	57.2%	**75.5%**	57.1%	75.4%
0 dB	54.5%	74.4%	**56.2%**	74.4%	55.6%	74.8%	55.2%	**75.3%**

added to all signals according to the image model [12]. The noise covariance matrix (Eq. (13)) is computed using silence periods.

The SSL algorithm performs hypothesis testing at 4° intervals in azimuth. The reported results are the average of 10 speaker locations uniformly distributed around the microphone array ($\theta = 0, 36^\circ, ..., 324^\circ$). At each location the signal length is 30 seconds. The analysis window of SSL is 40 ms, overlapping by 20 ms. We sample 100 speech frames from each location and perform SSL on them. Table 1 reports the average accuracy, in terms of percent of the SSL estimates (totally 1000 frames) which are within 2° and 10° of the ground truth angle. To verify the impact of reverberation over the SSL performance, we synthesize rooms with 100 ms and 500 ms reverberation times, as seen in the upper and lower parts of Table 1 respectively.

It can be observed from Table 1 that SRP-PHAT usually performs as good as ML-SSL when the input SNR is high (20 dB or above), but its performance drops significantly when the SNR becomes low. In most indoor (e.g., offices and meeting rooms) environments, the signal to noise ratio is above 15 dB, which explains SRP-PHAT's satisfactory performance in practice.

For the ML-SSL algorithm, the tunable parameter γ does seem to impact the final performance. This is particularly true when the reverberation is low. For instance, in the top table, when the reverberation is low (100 ms), when the input SNR is 0 dB, choosing $\gamma = 0.1$ results in a much better performance than $\gamma = 0.5$. This gap is however not significant when reverberation is high (bottom table, 500 ms). Therefore, for practical applications, using a fixed γ ranging from 0.1 to 0.3 can usually result in satisfactory performance.

5. CONCLUSIONS

In this paper we briefly reviewed the ML-SSL algorithm for multiple microphones, and showed that it degenerates to the popular SRP-PHAT algorithm under the assumption of zero noise, irrespective of the amount of reverberation. This explains the common observation that SRP-PHAT works really well in low-noise reverberative environments.

6. REFERENCES

[1] M. Brandstein and H. Silverman, "A practical methodology for speech localization with microphone arrays," *Computer, Speech, and Language*, vol. 11, no. 2, pp. 91–126, 1997.

[2] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. of IEEE ICASSP*, 1997.

[3] J. Kleban, "Combined acoustic and visual processing for video conferencing systems," Tech. Rep., The State University of New Jersy, Rutgers, 2000.

[4] P. Georgiou, C. Kyriakakis, and P. Tsakalides, "Robust time delay estimation for sound source localization in noisy environments," in *Proc. of WASPAA*, 1997.

[5] T. Gustafsson, B. Rao, and M. Trivedi, "Source localization in reverberant environments: performance bounds and ML estimation," in *Proc. of ICASSP*, 2001.

[6] D. Li and S. Levinson, "Adaptive sound source localization by two microphones," in *Proc. of Int. Conf. on Robotics and Automation*, 2002.

[7] Y. Rui and D. Florêncio, "Time delay estimation in the presence of correlated noise and reverberation," in *Proc. of ICASSP*, 2004.

[8] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, no. 4, pp. 320–327, 1976.

[9] M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. of ICASSP*, 1997.

[10] G. C. Carter, A. H. Nuttall, and P. G. Cable, "The smoothed coherence transform (SCOT)," Tech. Rep., Naval Underwater Systems Center, New London Lab., 1972.

[11] Cha Zhang, Zhengyou Zhang, and Dinei Florêncio, "Maximum likelihood sound source localization for multiple directional microphones," in *Proc. of ICASSP*, 2007.

[12] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *JASA*, vol. 65, pp. 943–950, 1979.