MULTI-SOURCE DOA ESTIMATION IN A REVERBERANT ROOM

Andreas M. Ali, Ralph E. Hudson, Kung Yao

UCLA, Los Angeles, CA 90095-1574

ABSTRACT

In a reverberant scenario, phase transformed weighted algorithms are more robust than Maximum Likelihood (ML) because of the insufficiency of the data model to incorporate reverberant information. This transformation has been applied to General Cross-Correlation and Steered Response Power algorithms; the latter has been shown to be more robust. For a multiple known number of sources, both algorithms have problems separating the sources that are close together because of the limitation caused by the resolution. Recently, an approach was made using simple well-known statistical room acoustics to model room reverberation, and another parametric approach called the Approximate Maximum Likelihood was made that was designed for multi-source estimations. By combining these methods, we developed an ML algorithm that is suitable for multi-source target estimates in a reverberant room.

Index Terms— Direction of arrival estimation, Maximum likelihood estimation, Microphones, Architectural acoustics

1. INTRODUCTION

The phase transform (PHAT) method was originally developed purely as an ad-hoc technique to avoid spreading or smearing the delta function containing the time delay information [1]. It is seen as one of the General Cross Correlation (GCC) weights and often called GCC-PHAT. Since then, researchers have shown that GCC-PHAT is more robust against reverberation than the Maximum Likelihood (ML) approach, and recently Gustafsson et al. showed that GCC-PHAT is optimal when using a model that incorporates reverberation [2].

The disadvantage of GCC is that it lacks inclusion of the array geometry information that often is readily available. Furthermore, to obtain the bearing/position estimate, GCC will need an extra multilateration step. With known geometry, we can form the Steered Response Power (SRP) by varying the steering vector over the desired bearings/positions. The estimate is then obtained by selecting the bearing/position that corresponds to the highest SRP. Not only does the SRP method perform better than the two-step GCC with multilateration, when phase transformation is applied, it has been shown to have superior performance over GCC-PHAT [3].

Aside from reverberation, we have developed a multi-source ML estimator based on a parametric model with the assumption that the noise is independent and identically distributed (i.i.d.). Then we apply Fourier transformation to the data, and by Central Limit Theorem, the noise in the frequency domain approaches a Gaussian distribution. In practice, we apply the Discrete Fourier Transform (DFT). This causes an edge effect that can be mitigated if the number of samples is large enough and thus this algorithm is called the Approximate Maximum Likelihood (AML) method [4].

Even in the SRP approach, multi-source bearing/position estimation is possible if there is enough resolution, i.e., the peaks are still separable. Of course, this assumes the knowledge of the number of true sources ahead of time, which we will assume available from this point in our discussion. In AML, the multi-source estimation is performed by looking in a multi-dimensional region, where the number of dimensions corresponds to the number of sources. Hence, even when the sources are very near each other, it is still possible to estimate multi-source bearing/positions. However, the AML suffers from reverberation because of the insufficiency of the model in not incorporating reverberant information.

2. SOURCE AND ROOM REVERBERATION MODEL

In this section we introduce the multi-source data model from [4] and the room reverberation model used in [2]. For a simpler exposition, we will limit the model to a 2-D scenario and far-field sources. For each M wideband sources, the angle of arrivals will be denoted as $\theta = \{\theta_1, \ldots, \theta_M\}$. We adopt angle convention where East points to 0 degrees increasing in a counter-clockwise direction. P sensors in an array, each at position $\mathbf{r}_p = [x_p, y_p]^T$ are assumed to have omni-directional response and are identical. The array centroid position is shown by $\mathbf{r}_c = \frac{1}{P}\sum_{p=1}^{P} \mathbf{r}_p = [x_c, y_c]^T$. We use $\mathbf{r}_c$ as the reference point and define a signal model based on the relative time-delays from this position. The relative time-delay of the mth source is expressed by $t_{cp}^{(m)} = t_c^{(m)} - t_p^{(m)} = [(x_c - x_p)\cos\theta_m + (y_c - y_p)\sin\theta_m]/v$, where $t_c^{(m)}$ and $t_p^{(m)}$ are the absolute time-delays from the mth source to $\mathbf{r}_c$ and $\mathbf{r}_p$ respectively, where v is the speed of propagation. The data

received by the pth sensor at time n is then

$$x_p(n) = \sum_{m=1}^{M} s^{(m)}(n - t_{cp}^{(m)}) + w_p(n), \quad (1)$$

for $n = 0, ..., N-1$, $p = 1, ..., P$, and $m = 1, ..., M$, where N is the length of the data vector, $s^{(m)}$ is the mth source signal arriving at $\mathbf{r}_c$. $t_{cp}^{(m)}$ is a real-valued number, and w_p is the zero mean i.i.d noise with variance σ^2/N. We have made the assumption that the signal strength received by all the sensors is the same. This is a good approximation when the distance from $\mathbf{r}_c$ to the source is much larger than the distance from $\mathbf{r}_c$ to the sensors, which we will assume is true.

In a reverberant environment, the measured sensor p signals coming from source m can be modeled as

$$x_p(n) = \int_{-\infty}^{\infty} h(n-\lambda)s^{(m)}(\lambda)d\lambda + w_p(n), \quad (2)$$

where $h(n)$ is the impulse response of the room. The key assumption is that the impulse response can be decomposed as

$$h(n) = d(n) + r(n), \quad (3)$$

where $d(n)$ denotes the direct path, and $r(n)$ corresponds to diffuse sound propagation.

The sound pressure at a microphone is built up from the direct path, plus several waves due to multiple reflections of the original sound. Although these can be computed by solving a wave equation [5], at higher frequencies, the complexity increases to a point where analysis is no longer feasible. To model the high-frequency part of $h(n)$, we will apply the theory of diffuse sound fields. A diffuse sound field is present when the following conditions are fulfilled [5] [2]:

A1 The dimensions of the room are large relative to the wavelength of $s(n)$. For the frequencies of interest (in speech processing we are mainly interested in the band 300 to 3500 Hz), this condition is usually satisfied.

A2 The average spacing of the resonance frequencies of the room must be smaller than one third of their bandwidth. In a room with volume V (in m^3), and reverberation time T_{60} (in seconds), this condition is fulfilled for frequencies that exceed the Schroeder large room frequency $Fs = 2000\sqrt{T_{60}/V}$.

A3 The source and the microphones are located in the interior of the room, at least a half-wavelength away from the walls. Hence, the sound field at a wall-mounted microphone cannot be modeled as diffuse.

Given that conditions A1-A3 are satisfied, we can combine eq. (1) - (3) and take a Discrete Fourier Transform (DFT) to get

$$\mathbf{X}_k = (\mathbf{D}_k(\theta) + \mathbf{R}_k)\mathbf{S}_k + \mathbf{W}_k, \quad (4)$$

where the data spectrum is $\mathbf{X}_k = [X_1(\omega_k), ..., X_P(\omega_k)]^T$, the steering matrix is $\mathbf{D}_k = [\mathbf{d}^{(1)}(\omega_k), ..., \mathbf{d}^{(M)}(\omega_k)]$, the steering vector is $\mathbf{d}^{(m)}(\omega_k) = [d_1^{(m)}(\omega_k), ..., d_P^{(m)}(\omega_k)]^T$, $d_p^{(m)} = e^{-j2\pi k t_{cp}^{(m)}/N}$, the source spectrum is expressed by $\mathbf{S}_k = [S_k^{(1)}, ..., S_k^{(m)}]^T$, and the noise is Gaussian $\mathbf{W}_k \sim \mathcal{N}(0, \sigma^2\mathbf{I}_{N/2})$ due to the Central Limit Theorem. Since the spectrum is taken from a real value, only $N/2$ frequency bins carry information, hence we will keep only the positive frequency bins. We also denote ^{T} as transpose and ^{H} as Hermitian throughout this article.

Gustafsson et al. [2] have shown that the coherence properties of $\mathbf{R}(f)$ with $\mathbf{R}(f + \Delta f)$ can be measured by a coherence bandwidth given as $\rho(T_{60}) \simeq \frac{7}{T_{60}}$ Hz. Taking 0.1 second sample duration means, each bin is separated by 10 Hz. Thus, for $T_{60} > 0.7$, every frequency sample can be considered uncorrelated. If this condition applies, we can assume $\mathbf{R}_k$ to be Gaussian random variable $\mathcal{N}(0, \sigma_r^2)$ without resampling.

3. ALGORITHMS

In this section we will review briefly the formulation of SRP-PHAT and then continue to derive ML estimate from (4). Then we will compare their corresponding bearings or Direction of Arrival (DOA) estimates in the next section.

3.1. Review of SRP-PHAT

SRP is based on filter and sum beamforming model represented by

$$J_{SRP}(f) = \sum_{m=1}^{M} G^{(m)}(f)S^{(m)}(f)e^{j2\pi f t_{cp}^{(m)}}, \quad (5)$$

where $G^{(m)}(f)$ is chosen to be $1/|S^{(m)}(f)|$ for $m = 1, \ldots, M$. At a given θ, we can find the corresponding $t_{cp}^{(m)}$ for all m. Then by steering the test θ from 0 to 2π, the SRP-PHAT estimate is achieved by maximizing the following function

$$J_{SRP-PHAT}(\theta) = \sum_{k=1}^{N/2}\left(\sum_{m=1}^{M}\frac{S_k^{(m)}}{|S_k^{(m)}|}e^{\frac{j2\pi k t_{cp}^{(m)}(\theta)}{N}}\right). \quad (6)$$

Note that in (6) we sample the continuous frequency domain, or apply DFT as an approximation. We will assume to have a long enough sample duration to mitigate the edge effect of the DFT.

3.2. Deterministic Source ML Estimate

If we assume $\{\mathbf{S}_k\}_{k=1}^{N/2}$ as deterministic, then the distribution of $\mathbf{X}_k$ will be $\mathcal{N}(\mathbf{D}_k(\theta)\mathbf{S}_k, (\sigma_r^2 tr(\mathbf{S}_k\mathbf{S}_k^H) + \sigma^2)\mathbf{I}_P)$, where $\mathbf{I}_P$ is the $P \times P$ identity matrix and $tr(.)$ denotes the trace. Concentrating on the log-likelihood function with respect to

$Var[\mathbf{X}_k]$, the ML estimator can be obtained by minimizing the following function

$$J_{AMLR} = \sum_{k=1}^{N/2} \log ||\mathbf{P}_k^{\perp}(\theta)\mathbf{X}_k||^2, \quad (7)$$

where the subscript $(.)_{AMLR}$ denotes the Approximate Maximum Likelihood for the Reverberant case (AMLR) algorithm, and $\mathbf{P}_k^{\perp}(\theta)$ denotes the orthogonal complement projection to the subspace of $\mathbf{D}_k(\theta)$

$$\mathbf{P}_k^{\perp}(\theta) = \mathbf{I}_P - \mathbf{D}_k(\theta)\mathbf{D}_k^{+}(\theta), \quad (8)$$

where $\mathbf{D}^{+} = (\mathbf{D}^H\mathbf{D})^{-1}\mathbf{D}^H$ is the Monroe-Penrose pseudoinverse. Note that eq. (7) is very similar to the one derived in [2] and can be seen as the generalization to the multi-source case.

3.3. Non-Deterministic Source ML Estimate

If we assume $S_k^{(m)} \sim \mathcal{N}(0, \sigma_{s,k}^{2(m)})$ is Gaussian and uncorrelated among all m, k and dropping the dependencies, $\{\mathbf{X}_k\}_{k=1}^{N/2}$ will have the following distribution

$$\mathbf{X} \in \mathcal{N}(0, \mathbf{\Sigma}),\ \mathbf{\Sigma} = \mathbf{D}\mathbf{\Lambda}\mathbf{D}^H + (\sigma_r^2 tr(\mathbf{\Lambda}) + \sigma^2)\mathbf{I}_P, \quad (9)$$

where $\mathbf{\Lambda}_k = diag[\sigma_{s,k}^{2(1)}, \ldots, \sigma_{s,k}^{2(M)}]$, and $diag$ denotes diagonalization. Then the ML estimator can be obtained by minimizing the following function

$$J = \sum_{k=1}^{N/2} \left[\log q^{P-M} + Ptr(\mathbf{\Lambda}) + qM + \frac{\mathbf{X}\mathbf{\Sigma}^{-1}\mathbf{X}}{P} \right], \quad (10)$$

where $q = \sigma_r^2 tr(\mathbf{\Lambda}) + \sigma^2$, and the scalar P is the number of sensors in an array (not the projection matrix). Unfortunately (10) cannot be concentrated with respect to σ_s^2, σ_r^2, or σ^2. Because the complexity off estimating simultaneous values becomes prohibitively large, we will not include this algorithm in the comparison.

4. SIMULATION

We simulate the conference room using the Allen-Berkley room reverberant model [6]. For ease of exposition, we will limit the discussion to a two-dimensional room with a reflection coefficient β of 0.9. The room size is 6 by 9 meters and we set the origin at the lower left corner of the room.

The microphone array is a uniform circular array with a radius of 0.2 m with 8 microphone elements. The first source is a male human voice with a dominant spectrum of about 700 Hz, and the second source is a female voice with a dominant spectrum of about 1000 Hz. To generate the source position we follow similar methods like in [2]. First, we fix the distance and the source DOAs, and then generate the placement of the array and the sources randomly in the allowable space as described by condition A3 (see Section 2).

(a) Average DOA (degree)

Sim. #	True	SRP-PHAT	AMLR
1	(60,80)	(61.0,79.0)	(61.6,79.1)
2	(60,75)	(61.0,73.5)	(61.4,74.6)
3	(60,70)	(61.1,37.9)	(60.2,68.9)

(b) RMS Error (degree)

Sim. #	True	SRP-PHAT	AMLR
1	(60,80)	1.5	1.1
2	(60,75)	5.1	1.1
3	(60,70)	26.8	5.8

Table 1. Simulated two source DOA estimate (θ_1, θ_2) with a source-sensor distance of 3 meters for both sources.

For each fixed DOAs, at least 50 different realizations of the sources are generated. Since the statistical properties of the transfer function are independent of the time-instant of observation, the random realization allows the variations to be captured in the simulation. Table. 1 shows the mean value of the estimate and the Root-Mean-Square (RMS) error summary of the simulated angles when fixing the distance of both sources at 3 meters from the sensor array.

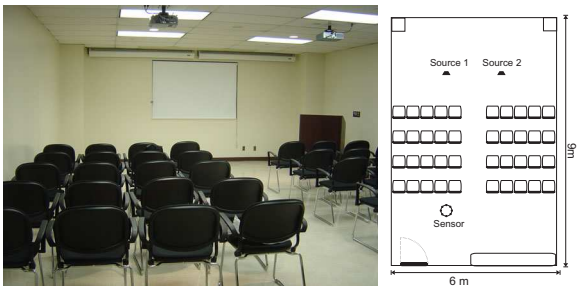
From Table 1 we can observe that when the source angles are close to each other (as in Sim. 3), SRP-PHAT no longer has two distinct lobes, so it picks another lobe that corresponds to the next strongest reverb path. On the other hand, AMLR is able to address the two sources, although with a larger error compared to the first two experiments.

In order to perform a good comparison without having sporadic errors from reverberation, we limit the search to be ± 20 degrees for both sources. This treatment will allow us to compare the true lobe of both algorithms while avoiding strong reverberation errors, which can be mitigated by other means, see for example [7].

Since there is no distinction between the first source and the second source, there are several ways to determine the error computation. We choose the errors that resulted in the smallest sum of the absolute errors among the possible permutations and treat both source errors as a single error vector for RMS computation.

5. EXPERIMENT

We use a conference room at UCLA Engineering IV building with a dimension of $6 \times 9 \times 2.6$ meters, Fig. 1. The array center is at (2,2) meters, and its 90 degrees point to position (2,9). The first and second sources are male and female voices respectively, as described in the simulation section. The first source is placed at (2,7), (3,7) and (3.5,7) meters, while the second source is held fixed at (4,7) meters. All the sources and microphones are elevated to 1.4 meters above ground.



(a) Photograph (b) Floor Plan

Fig. 1. Conference room used for the experiment

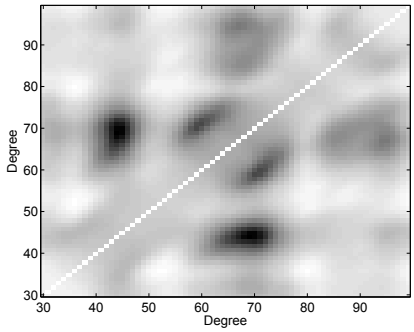


Fig. 2. AMLR likelihood function for Exp. 3 from 30 to 99 degrees. The darkest point is the most probable.

In each experiment, we run both the SRP-PHAT and the AMLR algorithm. Since we have information on the array position and the speaker is more likely to be at the front of the room (where the source positions are), we can reduce the search space from 0-360 degrees to 50-110 degrees. This way we eliminate strong reflections that cause both algorithms to report erroneously. Table 2 reports the estimates.

The SRP-PHAT at Exp. 3 only report one angle since it can only find one peak because the two sources are too close to each other. If we did not restrict the search space in Exp. 1, the SRP-PHAT second estimate would be 45.1 degrees, and in the Exp. 3, the AMLR second estimate would be 44 degrees. Fig. 2 shows the likelihood map of the AMLR in Exp 3. It is interesting to see that there are two main dominant lobes (the likelihood is symmetric) corresponding to the two possible source locations. The false one at (71, 44) degrees is formed because there is a strong signal coming at 44 degrees. By observing the room geometries, we know this angle is the reflection coming from the right wall. Hence it is possible to use an algorithm, such as the image model that finds the true source based on the room information, see for ex. [7].

6. CONCLUSION

We have investigated a room reverberation model using a well-known statistical room acoustic and developed a multi-source ML DOA estimator. We further considered two source models and found the ML estimate for the deterministic model and called this algorithm AMLR. Meanwhile, the non-deterministic model has complexity that grows rapidly because there are many parameters that have to be simultaneously estimated.

Exp. #	True	SRP-PHAT	AMLR
1	(90,68.2)	(89.3,66.8)	(89.5,66.8)
2	(78.7,68.2)	(85.3,68.3)	(74.8,66.5)
3	(73.3,68.2)	(69.5)	(70.9,59.6)

Table 2. DOA estimate in degrees (θ_1, θ_2)

From the comparisons both in simulations and in the physical experiments, we have shown the AMLR method that is able to distinguish sources when the SRP-PHAT method failed, although we also observed an increase in estimation errors when sources are near each other. When the sources are farther apart, both algorithms are comparable in performance.

The only disadvantage of AMLR is that it has relatively high computational complexity. While SRP-PHAT just needs to perform a single dimensional search, AMLR requires M dimensional search.

7. REFERENCES

[1] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," vol. 24, no. 4, pp. 320–327, August 1976.

[2] T. Gustafsson et al., "Source localization in reverberant environments: modeling and statistical analysis," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 791– 803, Nov. 2003.

[3] Joseph Hector DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*, Thesis of doctor of philosophy, Brown University, 2000.

[4] J. Chen, R. Hudson, and K. Yao, "Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field," vol. 50, no. 8, pp. 1843–1854, August 2002.

[5] H. Kuttruff, *Room Acoustics*, Wiley, 1973.

[6] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," vol. 65, no. 4, pp. 943–950, April 1979.

[7] P. Bergamo et al., "Collaborative sensor networking towards real-time acoustical beamforming in free-space and limited reverberance," *IEEE Trans. on Mobile Computing*, vol. 3, no. 3, pp. 211–224, July 2004.