Impact of microphone array configurations on robust indirect 3D source localization

Citation for published version:

Digital Object Identifier (DOI):
10.1109/ICASSP.2018.8461786

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
2018 IEEE International Conference on Acoustics, Speech and Signal Processing

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
IMPACT OF MICROPHONE ARRAY CONFIGURATIONS ON ROBUST INDIRECT 3D ACOUSTIC SOURCE LOCALIZATION

Elizabeth Vargas, Keith Brown
Heriot-Watt University
Edinburgh, United Kingdom

Kartic Subr
University of Edinburgh
Edinburgh, United Kingdom

ABSTRACT
Acoustic source localization (ASL) is an important problem. Despite much attention over the past few decades, rapid and robust ASL still remains elusive. A popular approach is to use a circular array of microphones to record the acoustic signal followed by some form of optimization to deduce the most likely location of the source. In this paper, we study the impact of the configuration of microphones on the accuracy of localization. We perform experiments using simulation as well as real measurements using a 72-microphone acoustic camera which confirm that circular configurations lead to higher localization error, than spiral and wheel configurations when considering large regions of space. Moreover, the configuration of choice is intricately tied to the optimization scheme. We show that direct optimization of well known formulations for ASL yield errors similar to the state of the art (steered response power) with 6× less computation.

Index Terms—3D acoustic source localization, microphone array configuration

1. INTRODUCTION
The problem of estimating the 3D position of objects is called localization. There has been tremendous advancement in robust localization of objects using visual features. The use of audio sensing has important advantages such as reliability under poor illumination conditions, relatively inexpensive sensing equipment and the prevalence of signal processing (1D) tools. There have been attempts to use audio localization examples include: in the automotive industry [1], in robotics [2] and in scene understanding [3]. Acoustic source localization (ASL) is typically achieved by leveraging known discrepancies in measurements of the emitted signal at multiple locations. ASL algorithms may exploit differences in time, amplitude or in both time and amplitude.

Some approaches to ASL, such as steered response power [4, 5], directly solve for the most likely position of the source amongst a grid of candidate locations. ‘Indirect’ methods, on the other hand, first estimate the times of arrival (TOA) at the sensors (microphones) or time differences of arrival (TDOA) across pairs of microphones and then use this information to deduce the most likely position of the source via multilateration [6, 7]. Although indirect methods are simpler to express as a least squares optimization [8], the resulting objective function is non-convex and often does not lend itself to analytical solution. Various reformulations of these using weighted least squares, linear correction least squares, constrained least squares, convex constrained least squares [9], total weighted least squares [10] and weight constrained total least squares [11] have been analyzed. Direct methods are believed to be more robust to noise and reverberation [4].

A uniform circular array of microphones [12, 13] along with a ring configuration [14] is a common choice for taking measurements since azimuthal angles to sources are considered more important than elevation. The advantage of acoustic cameras with such arrays is that they can focus on specific targets [15, 16], which is useful for speech processing. The resolution in elevation has recently shown to be improved by using a 2.5D circular array [17]. While there have been a few results on the use of spherical arrays, multiple spheres [18], randomly placed microphones [19, 20] and spiral configurations [21], there is surprisingly little analysis of the impact of the geometric structure of the measurement array on particular optimization algorithms for ASL.

In this paper, we adopt an optimization (sequential least squares programming) approach for indirect ASL. We focus on the core problem of localizing a single source. Other work towards estimating TDOA for multiple sources are directly applicable. Although the objective function we choose is non-linear and non-convex, we show using simulation as well as real data that the method is robust to noise and reverberation. Our experiments verify that it is comparable to SRP for real data while being 6× more efficient to compute. Using this optimization scheme, we study the localization error resulting from different geometric structure for the microphone array. Our results show that circular arrays produce the highest errors (across space) and are therefore least desirable.

2. OBJECTIVE FUNCTION AND OPTIMIZATION
Consider a source at location s that emits an acoustic signal at some arbitrary time t∗. Let the measurements of the emitted sound be recorded by an array of M microphones located at
m_i, \ i = 1, 2, ..., M and the times taken by the signal to travel from s to m_i be t_i. If the distance between the source and the i-th microphone is d_i ≡ ||m_i - s||, then t_i = d_i/c + t^* where c is the speed of sound in air and t^* is not generally known.

**Time of arrival** In the case that the times of arrival at the microphones are measured as ðt_i, we pose the ASL problem as one of jointly determining s and t^* as

\[
O_1 : \arg\min_{s,t^*} \sqrt{\sum_{i=1}^{M} (t_i - t^*)^2}
\]

(1)

**Time Difference of Arrival (TDOA)** Another possibility is to note the difference in measured times between a pair of microphones, \(\tilde{\tau}_{ij} \equiv t_i - t_j\), or TDOA. The literature is rich with methods to estimate TDOA. We choose the popular Generalized Cross-Correlation Phase Transform (GCC-PHAT) [22]. Then, we perform ASL by optimizing [8]:

\[
O_2 : \arg\min_{s} \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{M} (\tilde{\tau}_{ij} - \tau_{ij})^2}
\]

(2)

where \(\tau_{ij} = (t_i - t_j)\).

For both formulations O1 and O2, since we know that the solution is constrained by the dimensions of the room, we supply these constraints as linear inequalities. We solve the constrained non-linear optimization using Sequential Least Squares Programming (SLSQP) which is an iterative procedure. In each iteration, a constrained quadratic programming sub-problem is constructed so that the chain of solutions converges to a local minimum [23]. Each subproblem replaces the objective function with a local, quadratic approximation subject to local affine approximations of the constraints. We used a Broyden-Fletcher-Goldfarb-Shanno (BFGS) approximation to update the Hessian matrix required for the local quadratic approximation and chose the step length using an \(L_1\) test function. The optimizer used to solve each subproblem is a modified version of NNLS [24]. We used the following parameters as inputs to the optimizer: iterations = 1500, accuracy = 1e-20, epsilon = 1.49e-08.

### 2.1. Experiments

We performed experiments using simulation as well as real measurements using an **gfai tech AC Pro Acoustic Camera system** consisting of 72 reconfigurable microphones sampling at 192kHz. We used three different microphone configurations: ring, wheel and spiral. Using each configuration, we measured recorded sounds played by a Bose Soundlink Bluetooth Mobile Speaker II, Model 404600 in five different calibrated positions within a room of size 12m × 7m × 3m. The speaker was positioned, using a tripod, to be on the plane \(y = -0.32\) for all five positions A, B, C, D and E. For each position we acquired three recordings. Fig. 1 illustrates the setup. We repeated the experiments for 4 different audio signals [25]: chirp, gunshot, dogbark and speech.

**Simulation: noisy TOA and TDOA** We tested the robustness of the proposed optimization by evaluating the relative error in localization for different simulated degrees of noise \(\sigma\) in the estimated TOA and TDOA values. To enable comparison across multiple sources locations, we express \(\sigma\) for each source location as a percentage of the time taken for sound to travel from s to the center of the microphone array O. We use a Gaussian model for the noise in simulated TOA \(\tilde{t}_i = t_i + \eta\) and for TDOA \(\tilde{\tau}_{ij} = \tau + \eta\) where

\[
\eta \sim N(0, \frac{\sigma}{100} \frac{1}{c} ||s - O||)
\]

(3)

We measure relative error, expressed as a percentage of the distance from the source to the camera, as the evaluation metric for the accuracy of localization:

\[
\text{error(\%)} = \frac{||s - \tilde{s}||}{||s - O||} * 100,
\]

(4)

where \(\tilde{s}\) is the source location estimated by the optimization.

We compared optimizations for TOA and TDOA with multilateration [7]. Fig. 3 depicts plots of relative localization error (Y-axis) as the noise in the simulation is increased (X-axis). We performed two versions of the experiment: First, assuming that microphone and the sound source are synchronised (\(t^* = 0\) in Fig. 3a), then without that assumption by setting \(t^* = 0.01s\).

**Simulation: microphone configuration** We estimated the localization error at different points in space. Since it would be rather tedious to repeat real measurements over a dense grid of source locations, we obtained this via simulation. For each source position on a dense grid, we estimated the localization errors for three microphone configurations. The three configurations were identical to those used for real measurements with our acoustic camera, consisting of 72
Fig. 2. The plots show simulated relative localization error (Y-axes) for increasing degrees of noise (X-axes) observed at three source locations: P1: (-2,-1,4), P2: (-1,0.5,3), P3: (0.4,0.7,1.05).

Fig. 3. Plots comparing relative localization errors using $O_1$ (TOA), $O_2$ (TDOA) and multilateration [7] (a) when the speaker is synchronized with microphones and (b) when the time of emission is unknown.

Real Data: Comparison with SRP We used optimization scheme $O_2$ to localize a speaker placed in five positions $A \equiv (2.0, -0.32, 0.5)$, $B \equiv (1.5, -0.32, 2.0)$, $C \equiv (0.0, -0.32, 1.5)$, $D \equiv (-1.5, -0.32, 1.0)$ and $E \equiv (-1.5, -0.32, 3.5)$. Fig. 5 plots relative errors (Y-axes) for three different microphone configurations (X-axes) at the chosen five locations (columns). The three rows of plots correspond to results obtained using SLSQP, SRP and Bayesian optimization respectively. Errorbars (standard deviation) are shown with black lines on top of the bars.

2.2. Results and discussion

Microphone configurations Our results overwhelmingly suggests that circular (ring configuration) arrays are worse than spiral or wheel configurations when considering relative localization error over a wide range of positions. Although slightly exaggerated (100% noise), our simulation results (fig. 4) show regions (top view) that are error prone when using circular arrays. This is also true for our real measurements (fig. 5), where the results obtained for position C are worse for ring than for wheel or spiral using any of the three localization techniques. The yellow bars in the first row show that the errors observed with real data correspond to errors obtained with about 10% noise in our simulation.

Comparison with multilateration Our experiments revealed that both optimization strategies $O_1$ and $O_2$ result in lower relative errors than state of the art multilateration [7]. This is particularly true when the time of emission of the signal is unknown and when the emitter is not synchronized with the microphones ($t^* \neq 0$). When $t^* = 0$, we observed that our implementation of the multilateration algorithm has similar accuracy to optimizing $O_1$ (TOA). Our proposed approach to optimizing $O_2$ (TDOA) has the least relative errors and remains unaffected by $t^*$.

Comparison with SRP A common criticism that is faced by indirect methods is that the optimization is not as robust as direct methods such as SRP. However, our results (Table 1) show that our localization error is comparable to SRP but is more efficient. For this comparison, we used an efficient implementation of SRP that leverages stochastic region contraction [5] and a naive implementation of our optimization in python. Just as with their method, the accuracy of the proposed optimization may also be traded for performance.

Accuracy vs performance One way to approximate the localization is to modify the nested summation in $O_2$ to only consider some of the microphone pairs. We studied convergence plots of localization error for different source positions, as the number of microphone pairs is increased from just 1 pair to all pairs ($C_3^2$). We observed that the error generally drops below 10% for 100 mic pairs (see Table 1 for the corresponding computation times), except for the dogbark signal. Figure 6a plots relative error averaged across spatial locations for all four test signals using only 100 microphone pairs.

Bayesian optimization We tested a Bayesian optimizer with $O_2$ as its loss function ($\kappa = 1$). This took an order of magni-
Fig. 4. Relative error percentages visualized as heatmaps obtained using simulations, at 100% noise, for a 2m x 2m room. 100 estimates were averaged for the error estimate at each grid position. The insets show the distributions of errors as histograms.

Fig. 5. Localization Error using SQLP and simulation with 10% noise (top row) SRP (2nd row) and Bayesian Optimization with exploitation (3rd row).

We tested with various degrees of the \( \kappa \) parameter to trade-off exploitation versus exploration, speculating that the poor performance was due to the presence of multiple local minima. However, the plot (fig. 6b) shows that exploitation (\( \kappa = 1 \)) performs better than exploration (\( \kappa = 10 \)) in most cases. The number of iterations and tolerance were set so that optimizer converged to the reported solutions, suggesting that the problem is not due to multiple local minima.

Limitation One of the drawbacks of indirect localization achieved by minimizing \( O_2 \) is its dependency on the estimated TDOA values. Although our results show that GCC-PHAT is accurate enough to yield localization errors comparable to SRP, the former performs worse when dealing with signals with repeating patterns such as the barking of a dog (red bar in fig. 5). Interestingly, our localization was more robust to reverberation (when the source was placed at room boundaries) than to repetitive macro-structures. Perhaps using full signal correlation matrices, as adopted by spectral estimation techniques, would resolve this problem.

Fig. 6. (a) Errors (real data) for four signals (colored bars) across spatial locations using 100 mic. pairs. (b) Exploitation (\( \kappa = 1 \)) vs exploration (\( \kappa = 10 \)) for dogbark (blue) and speech (purple) for spiral configuration.
3. REFERENCES


