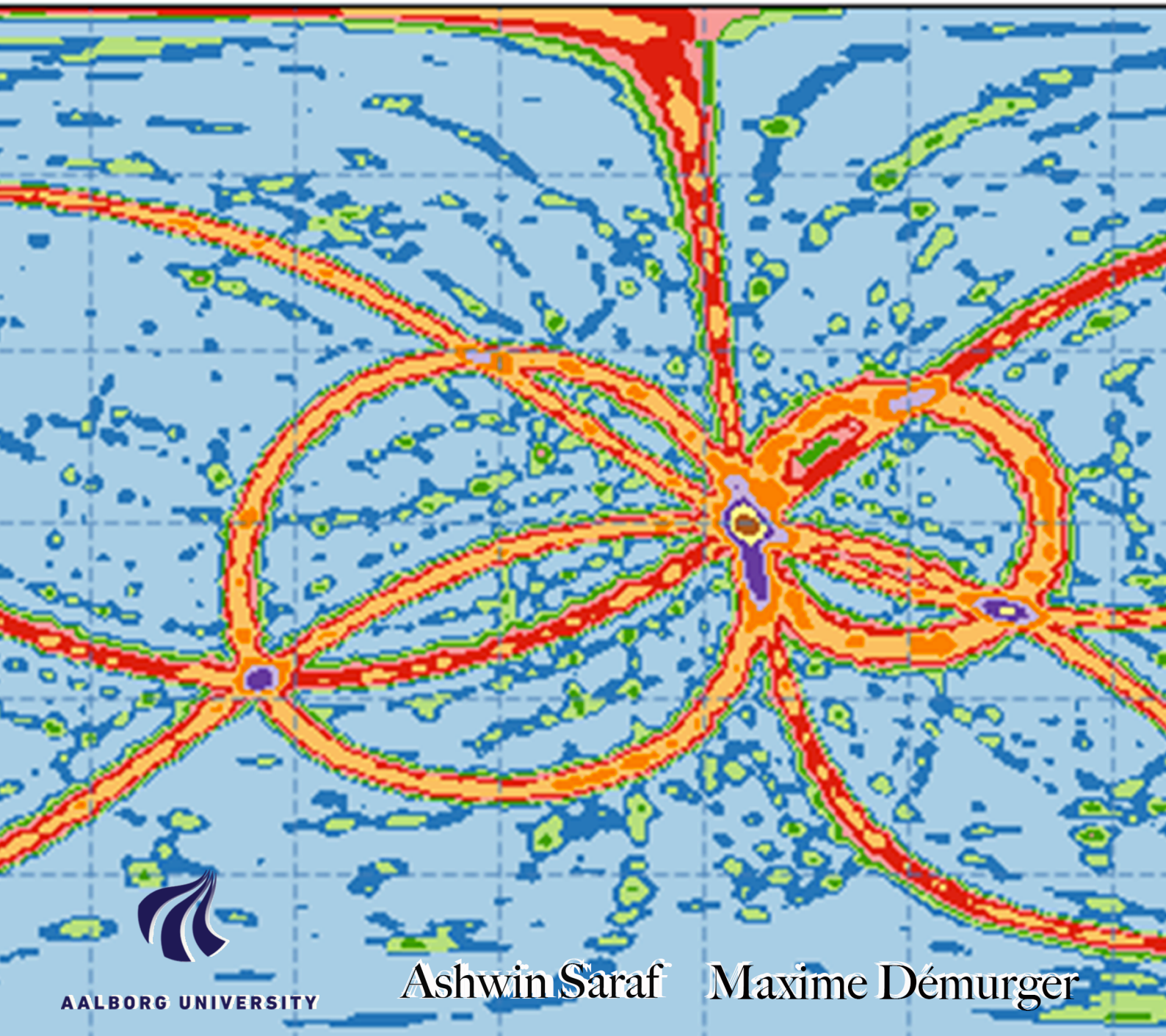
Outdoor Sound Localization Using a Tetrahedral Array



AALBORG UNIVERSITY

Ashwin Saraf Maxime Démurger

This report has been compiled using LaTeX.

Electronics and IT
Aalborg University
http://www.aau.dk

AALBORG UNIVERSITY
MASTER'S THESIS

Title:
Outdoor sound localization using a tetrahedral array

Theme:
Master's Thesis

Project Period:
Fall Semester 2018

Project Group:
AAT10-1062

Participants:
Ashwin Saraf
Maxime Démurger

Supervisor:
Søren Krarup Olesen

Copies: 1

Page Numbers: 85

Date of Completion:
June 7, 2018

Abstract:

The impact of sound on an individual's health can be dramatic when exposed to high sound pressure level (SPL) for an extended period of time. While total SPL can be monitored using just a single microphone, it is a challenging problem to find the major sound contributors in ordinary outdoor environments where multiple loud sources can be present. A wide range of solutions already exist for this purpose, which can compute the sound power map of an environment using microphones arrays. The SRP-PHAT algorithm is notable, as it combines the robustness of beamforming techniques with the accuracy of time difference of arrival (TDOA) based methods. However, the method is designed for sources in a reverberant field and has been optimized for indoor teleconferencing and sound peak detection. Outdoor source localization is still normally done using traditional beamforming on a copious number of microphones (sometimes greater than fifty). The purpose of this thesis then is to apply the SRP-PHAT for outdoor sound localization using a minimal number of microphones. For this purpose, a modified SRP-PHAT algorithm is derived and is shown to be more robust than normal SRP-PHAT. The algorithm, called Min-SRP-PHAT, is presented and experimentally evaluated here.

Preface

This report has been carried out during the Spring of 2018 as an Acoustics and Audio Technology Master's Thesis at Aalborg University by group: 10GR1062.

The group would like to thank Søren Krarup Olesen (Associate Professor, AAU) and Karim Haddad (Research engineer, Brüel & Kjær) for their supervision throughout the project.

The figures in the report are produced by the group unless a source is specified.

Aalborg, June 7, 2018

Ashwin Saraf
asaraf16@student.aau.dk

Maxime Démurger
mdemur16@student.aau.dk

Table of contents

Chapter 1

Introduction

Noise has been a bane to urban life since immemorial time. The first known noise ordinance was passed in Greece where potters, tinsmiths, and other tradesmen, along with roosters, were made to live outside the city walls because of the noise they made. At the time, tinnitus was ascribed to hearing divine sounds, sort of a cosmic music, by Plato and Pythagoras. While philosophers of such mental faculties are highly regarded, one cannot help but wonder their egomania in attributing a disease to divinity. Meanwhile, Hippocrates, a physician, was arguing that tinnitus was, in fact, caused by a prolonged exposure to noise. Centuries passed and noise ordinances piled on, specially with the discovery of the megaphone in the 17^{th} century and the industrial revolution in the 18^{th}. However, it was only late in the 19^{th} century that the first noise measuring device was invented, the Rayleigh Disk, which was a super-light disk suspended in air that rotated in the presence of sound. Unfortunately, it was too delicate to be used outside. Then arrived the First World War, and along it arrived the first time when weapons of war were successfully tracked and destroyed based on the sounds they made. The machines to do such localization were nothing short of comical. Giant rotating wave-guides were used by an operator (or multiple operators) and steered to amplify the sound arriving from a given direction, a prequel to beamforming.

Sound localization technology has matured since then and is now part of our daily life, implemented in countless products such as hearing aids, headsets, mobile phones, laptops, etc, and providing countless solutions. This technology can now be used to solve problems such as accessing the impact of environmental noise on humans. The impact is still being studied by researchers around the world and a reliable, accurate and simple noise monitoring system is yet to be developed which can detect the main noise contributors in an outdoor environment. This thesis tackles the problem of simplicity, in that it aims to develop a robust solution which can function with as few microphones as possible.

Various sound localization algorithms already exist that have been successfully applied for source localization. Traditionally, distinction is made between algorithms using the time-difference-of-arrival (TDOA) of signals between pairs of microphones to find the position of a source, and algorithms using beamforming Steered Response Power (SRP) techniques. The SRP-PHAT algorithm is one of the most robust and widely implemented method and combines the advantages of these two techniques. However, a significant bit of research has been done on implementing SRP-PHAT on speech enhancement systems, whereby speaker identification and teleconferencing in an environment having high background noise and reverberant conditions were needed, and outdoor sound can be appreciably different from this situation. The purpose of this thesis then, is to design a method to adapt the SRP-PHAT

algorithm, to compute the sound map of a multi-source outdoor environment.

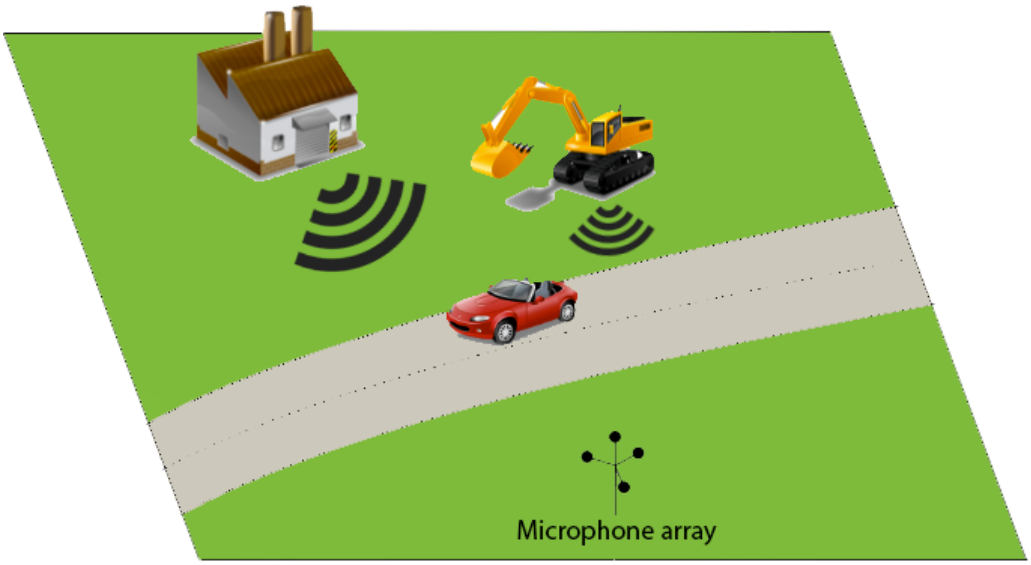


Figure 1.1: Various outdoor sound sources being localized by a microphone array

It should be noted that the primary purpose of this thesis is not to track moving sources in realtime, rather the thesis tackles the problem of 'outdoor sound source locations and their levels' with the constraint of using as few microphones as possible while also being robust to different noise, weather and sound source conditions. The thesis proposes a solution capable of retrieving the positions of multiple sound sources in a variety of scenarios. While previous research most often tackles the problem of single source localization using a linear or circular array, this thesis uses a tetrahedral array to localize the sound source in a 3-dimensional space.
The organization of the thesis is as follows:

- Chapter 2 derives the theory used to analyze the problem and introduces the simulation framework.
- Chapter 3 derives the Min-SRP-PHAT algorithm and compares its robustness and performance in outdoor conditions with SRP-PHAT.
- Chapter 4 contains real world experimental results. Anechoic and outdoor measurements are conducted which investigate the algorithm limits in a variety of scenarios.
- Chapter 5 provides a discussion about the solution and proposes new ideas and further work.

Chapter 2

Direction of Arrival (DOA) Estimation Using a Microphone Array

2.1 Time Difference of Arrival (TDOA)

2.1.1 TDOA of a Pair of Microphones

When using a pair of microphones, sound from a particular source arrives at the two microphones at different times, based on the source distance to the particular microphone. For a pair of microphones located at m_1 and m_2, the time difference of arrival (TDOA) of a sound signal from a source located at s can be defined as:

$$\begin{aligned} T(\{m_1, m_2\}, s) &= \frac{|s - m_1| - |s - m_2|}{c} \\ &= \frac{D_1 - D_2}{c} \end{aligned} \tag{2.1}$$

where c is the speed of sound in the medium and D_1 and D_2 the distance between the source and the microphones at m_1 and m_2. In a two-dimensional space (2D), s can be any point on a hyperbola as shown in Fig.2.1 where the two foci of the hyperbola are the microphone positions. This is because the difference in distances from any point on the hyperbola to the two foci is a constant.

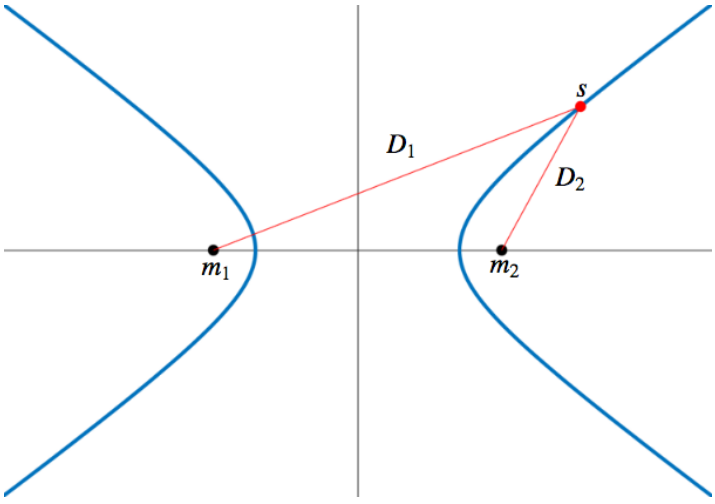


Figure 2.1: A hyperbola (represented in blue), the red dot is any point on the hyperbola, the black dots represent the two foci. For any point on the hyperbola, $|D_1| - |D_2| = constant$

In a three-dimensional space (3D), the TDOA information can be used to locate the source on a two-sheeted hyperboloid $\chi(\{m_1, m_2\}, s)$ such that the microphone positions are its foci. However, the two-sheeted hyperboloid can be approximated to a cone so as to have a much simpler equation for the locus: $\theta = constant$, where θ is the angle of the source to the midpoint of the line segment joining the two microphones as shown in Fig. 2.2.

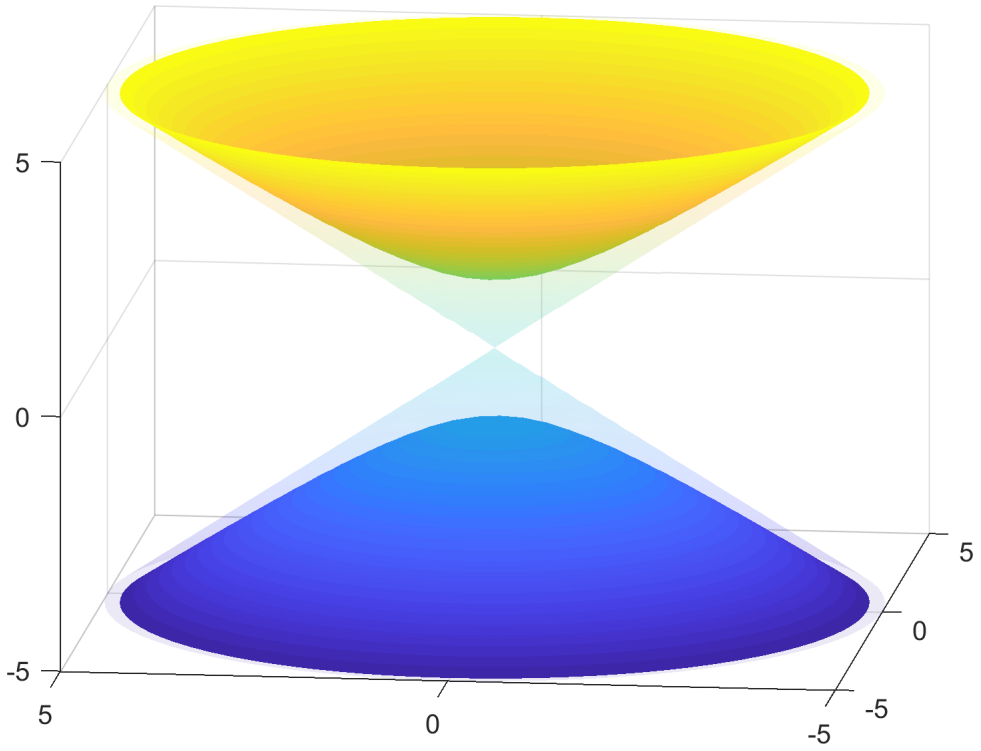


Figure 2.2: A 2-sheeted hyperboloid, overlaid with its cone approximation. Moving the microphones closer together or moving the source further away, effectively moves the tips of the two sheets of the hyperbola closer together, in a relative manner. As the tips get closer, the two sheets of the hyperbola tend towards a cone.

As the source location gets closer to being orthogonal to the midpoint of the line segment joining the two microphones ($\theta = 90°$), the hyperbola gets wider and flatter (more planar) and approximates the cone better. Also as the source gets closer to the line joining the two microphones ($\theta = 0°, 180°$), the hyperbola collapses to a straight line and approximates the cone better. Thus, the error minimizes for broad-side sound source ($\theta = 90°$) and for end-side sound source ($\theta = 0°, 180°$), and maximizes for the midsection ($\theta = 45°, 135°$). The equation for the error, derived by Brandstein [1], is given by

$$\begin{aligned} max\{\theta_{error}\} &\approx \frac{M_{dist}^2}{16R^2} \\ max\{D_{error}\} &\approx \frac{M_{dist}^2}{16R}, \end{aligned} \tag{2.2}$$

where M_{dist} is the distance between the microphones, R is the distance from the source to the microphone pair midpoint and D_{error} is the actual source distance error (the gap between the cone and the hyperboloid).

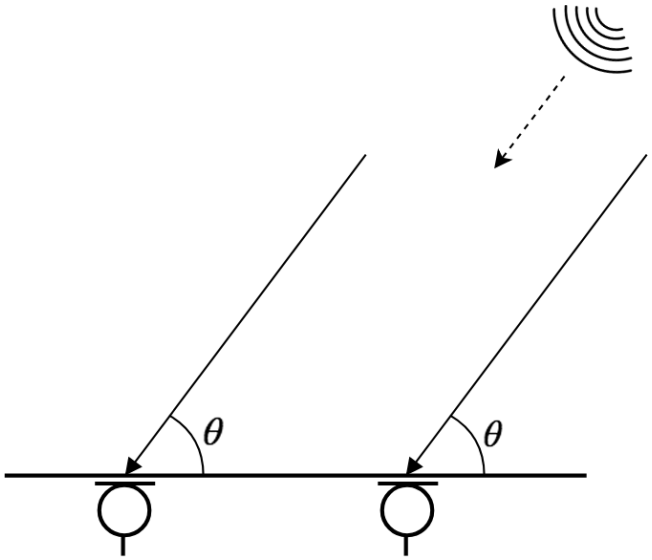


Figure 2.3: With far-field approximation it can be assumed that the sound waves incident on the pair of microphones are parallel (planar incidence). The errors discussed in Eq. 2.2 can then be ignored.

Microphones in a microphone array are usually closely spaced relative to the actual source distance ($R \gg M_{dist}$), so the cone approximation works well. In most scenarios, errors due to noise from other system parameters are greater than the errors associated with this approximation. Thus, given the TDOA information between a microphone pair, the source can be located at a particular direction θ, associated with the cone for that time delay T. The cone approximation is essentially the same as the far-field assumption for a sound source. Thus, in the far-field, DOA estimation is essentially the same as TDOA estimation due to the one-to-one relation between θ and T as shown in Fig. 2.3

2.1.2 TDOA for Multiple Pairs of Microphones

Let's define a three-dimensional Cartesian coordinate system with axes X, Y, Z as shown in Fig. 2.4.

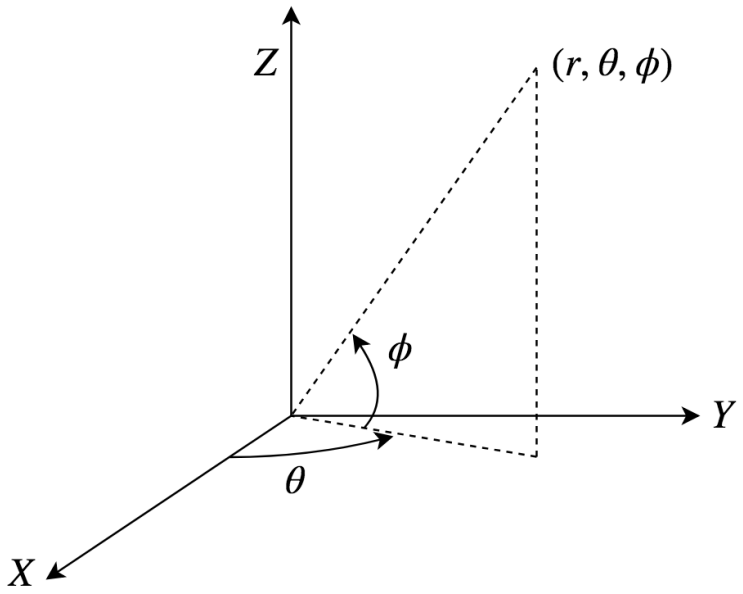


Figure 2.4: Spherical coordinate system used throughout the thesis

A vector $\hat{u}$ in this 3D space can be defined by (r,θ,ϕ), where r is the magnitude of the vector, θ the azimuth and ϕ the elevation as described in Fig. 2.4. For a unit vector $r = 1$, this can by a vector $\hat{a}(\theta, \phi)$ in Cartesian coordinates (x, y, z) by

$$\hat{a}(\theta, \phi) = \begin{bmatrix} cos(\theta)cos(\phi) \\ sin(\theta)cos(\phi) \\ sin(\phi) \end{bmatrix}. \tag{2.3}$$

Suppose sound is travelling along the unit vector $\hat{u}$ and two microphones are placed at positions $\hat{p}_1 = (x_1, y_1, z_1)$ and $\hat{p}_2 = (x_2, y_2, z_2)$. Then $\hat{p}_{12} = \hat{p}_2 - \hat{p}_1$, where $\hat{p_{12}}$ is the distance vector between the two microphones. The projection of this distance vector in the direction of $\hat{u}$ is simply $\hat{a}(\theta, \phi) \cdot \hat{p}_{12}$ and the time it takes for sound to travel between the two microphones is then

$$\hat{t}_{12} = \frac{\hat{a}(\theta, \phi) \cdot \hat{p}_{12}}{c}, \tag{2.4}$$

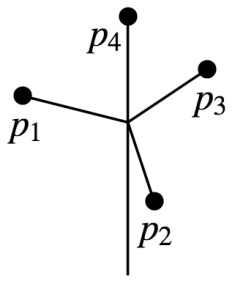
c being the speed of sound.

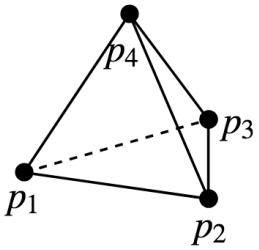
2.1.3 TDOA of a Three Dimensional Array

Given the TDOA between multiple microphone pairs, the source localization problem can be solved by triangulation. For M microphones, located at arbitrary positions in the 3D space, ${}^{M}C_2$ combinations of pairs are possible, giving ${}^{M}C_2$ possible TDOA estimates. However, not all the pairs are linearly independent. Given three microphones i, j and k, the time delays $\hat{t}$ satisfy:

$$\hat{t}_{i,j} = \hat{t}_{i,k} + \hat{t}_{k,j} \tag{2.5}$$

Thus, only M-1 linearly independent combinations of microphone pairs exists. This can be written as the vector $\hat{t} = (\hat{t}_{1,2}.....\hat{t}_{1,M})^T$ in the M-1 dimensional subspace S $\subset {}^{M}C_2$. In order to retrieve a unique and non ambiguous TDOA solution, at least four non co-planar microphones are needed (described later in Sec. 2.4), i.e. a 3D structure. The tetrahedron is a simple spatial structure with four vertices, four faces and six edges (a triangular pyramid). For a regular tetrahedron all the vertices are equally spaced from each other and every face is an equilateral triangle. Two possible tetrahedral microphone array structures composed of four microphones at position p_1, p_2, p_3, p_4 are shown below.





Given a tetrahedral array, only three linearly independent combinations of cross-correlations exist. A Least-Squares approach could be considered, with the three use-able combinations. The 'correct' DOA is then the $\hat{a}(\theta, \phi)$ which minimizes the cost function

$$J(\theta, \phi) = \sum_{j=2}^{4} \left(T_{1j} - \frac{\hat{a}(\theta, \phi)\hat{p}_{1j}}{c} \right)^2. \tag{2.6}$$

2.2 Generalized Cross Correlation Method

The famous Knapp-Carter paper details the generalized cross-correlation method (GCC) for estimation of time delay [2] in free field. For a pair of microphones, m_1 & m_2, separated by a distance, the signals from a source received at time t can be given by

$$\begin{aligned} x_1(t) &= s_1(t) + n_1(t) \\ x_2(t) &= \alpha s_1(t + D) + n_2(t), \end{aligned} \tag{2.7}$$

where $n_1(t)$ & $n_2(t)$ are the noise at time t at the two microphones which are uncorrelated to the signal $s_1(t)$. The microphone m_1 receives the signal $s_1(t)$ first, while the microphone m_2 receives a delayed and attenuated version $\alpha s_1(t + D)$ at time t. The α depends on the microphone relative distance and microphone calibration and within-media factors like absorption. The time delay D depends on the microphone pair relative distance, the speed of sound in the media and the position of the sound source.

Based on the discussions in previous sections, if the value of D can be estimate, the source location can also be. However, depending on source movement and environmental factors, both α and D can change over time. The estimation of D thus can only be made for observations of a finite duration. D can be estimated by computing the cross-correlation of the two signals defined in Eq. 2.8

$$R_{x_1x_2}(\tau) = \mathbf{E}[x_1(t)x_2(t - \tau)], \tag{2.8}$$

Assuming noise to be uncorrelated to each other and the source signal, the cross correlation can be expressed as

$$\begin{aligned} R_{x_1x_2}(\tau) &= \mathbf{E}[\{s_1(t) + n_1(t)\}\{\alpha s_1(t - \tau + D) + n_2(t)\}] \\ &= \alpha \mathbf{E}[s_1(t)s_1(t - \tau + D)] \\ &= \alpha R_{s_1s_1}(D - \tau), \end{aligned} \tag{2.9}$$

The autocorrelation $R_{s_1s_1}(\tau)$ is always maximum at lag $\tau = 0$, therefore $R_{x_1x_2}(\tau)$ peaks at $D - \tau = 0$, i.e. $\tau = D$. So the τ that maximizes the cross-correlation is an estimator for the time delay D. Assuming the processes to be ergodic so that the samples from a finite duration T can be used to estimate the cross-correlation, the estimate can be given by

$$\hat{R}_{x_1x_2}(\tau) = \frac{1}{T - \tau} \int_{\tau}^{T} x_1(t)x_2(t - \tau)dt, \tag{2.10}$$

choosing sample mean as the estimator.

The Fourier transform of Eq. 2.9 gives the cross power spectrum $G_{x_1x_2}(f)$ described in Eq. 2.12

$$R_{x_1x_2}(\tau) \xrightarrow{\mathscr{F}} G_{x_1x_2}(f) \tag{2.11}$$

$$G_{x_1x_2}(f) = \alpha G_{s_1s_1}(f) \cdot e^{-j2\pi f D} \tag{2.12}$$

where $G_{s_1s_1}(f)$ is the auto power spectrum of s_1. The multiplication by $e^{-j2\pi f D}$ in the frequency domain is equivalent to a convolution in the time domain giving.

$$R_{x_1x_2}(\tau) = \alpha R_{s_1s_1}(\tau) \circledast \delta(\tau - D), \tag{2.13}$$

which can be seen as the Inverse Fourier Transform of the signal spectrum spreading the delta function. The way to ensure no spreading takes place is to use a white noise signal. The autocorrelation of white noise is a delta function, in which case convolution with the delay-delta function results in a single peak value. Of course, in any kind of a reverberant field this will never be a single value. This is because the reverberations will have the effect of making the signal add up in a periodic and attenuated manner. However, the peak of the autocorrelation $R_{x_1x_2}(\tau)$ still happens at $\tau = D$, with the spreading having the effect of broadening the peak. If the time delay D is not a single value however, as can be the case in reverberant fields or for periodic signals, the $R_{x_1x_2}(\tau)$ will have multiple peaks. Each broad peak will overlap with the other in an additive or destructive manner making is impossible to detect or distinguish peaks.

$R_{s_1s_1}(\tau)$ in Eq. 2.13 can be expanded to frequency domain to get

$$R_{x_1x_2}(\tau) = \left[\int_{-\infty}^{\infty} \alpha G_{s_1s_1}(f) e^{j2\pi f\tau} df \right] \circledast \delta(\tau - D), \tag{2.14}$$

this cross-correlation $R_{x_1x_2}(\tau)$ is a function that is spread around $\delta(\tau - D)$ according to $G_{s_1s_1}(f)$. This spreading is detrimental to the resolution of the localization results. Also, if the signal itself is non-stationary, like speech signals, this spreading is also unpredictable.

Now we are ready to form a basis for the different GCC weighing methods. If *a priori* signal or noise information is available, the signals $x_1(t)$ & $x_2(t)$ can be pre-filtered to improve the accuracy of estimating the time delay. The method of selection of the pre-filter weights then forms the basis for the different GCC methods.

Suppose, $x_1(t)$ & $x_2(t)$ are filtered through filters $H_1(f)$ and $H_2(f)$, to get filtered signals $y_1(t)$ & $y_2(t)$ respectively, then we have

$$G_{y_1y_2}(f) = H_1(f)H_2^*(f)G_{x_1x_2}(f), \tag{2.15}$$

taking the Inverse Fourier Transform

$$\begin{aligned} R_{y_1y_2}(\tau) &= \int_{-\infty}^{\infty} H_1(f)H_2^*(f)G_{x_1x_2}(f)e^{j2\pi f\tau} df \\ &= \int_{-\infty}^{\infty} \psi(f)G_{x_1x_2}(f)e^{j2\pi f\tau} df, \end{aligned} \tag{2.16}$$

where

$$\psi(f) = H_1(f)H_2^*(f), \tag{2.17}$$

since we can only estimate the cross-power spectra, we can write

$$\hat{R}_{y_1y_2}(\tau) = \int_{-\infty}^{\infty} \psi(f)\hat{G}_{x_1x_2}(f)e^{j2\pi f\tau} df, \tag{2.18}$$

the frequency weights given by $\psi(f)$ can be selected according to the purpose that is wished to be achieved. For example, if the purpose is to maximize the signal-to-noise (SNR) ratio in the signal passed, then the $\psi(f)$ could be selected so that it attenuates the frequencies in the noise spectra. Obviously this requires either a priori knowledge or estimation of the noise spectra. The following sections introduce the different methods of frequency weight selection. Three methods are described here, ROTH, SCOT, PHAT. Of particular interest is the PHAT, direct and improved versions of which have been consistently used to do robust source localization.

2.2.1 ROTH

The frequency weights for ROTH processor are defined as

$$\psi(f) = \frac{1}{G_{x_1x_1}(f)}, \tag{2.19}$$

so we get

$$R_{y_1y_2}(\tau) = \int_{-\infty}^{\infty} \frac{G_{x_1x_2}(f)}{G_{x_1x_1}(f)} e^{j2\pi f\tau} df, \tag{2.20}$$

substituting the value for $G_{x_1x_2}(f)$ assuming uncorrelated noise from Eq. 2.12 we get

$$\begin{aligned} \hat{R}_{y_1y_2}(\tau) &= \int_{-\infty}^{\infty} \frac{\alpha\hat{G}_{s_1s_1}(f)}{G_{x_1x_1}(f)} e^{j2\pi f.(\tau-D)} df \\ &= \delta(\tau - D) \circledast \left[\int_{-\infty}^{\infty} \frac{\alpha\hat{G}_{s_1s_1}(f)}{G_{s_1s_1}(f) + G_{n_1n_1}(f)} e^{j2\pi f\tau} df \right], \end{aligned} \tag{2.21}$$

so now the delta function is spread according to the value of $G_{n_1n_1}(f)$. For frequencies f where $G_{n_1n_1}(f)$ has a high magnitude, the cross-correlation will be suppressed, so that peaks in the frequency regions where n_1 is high disappear. But as can be seen ROTH processor does nothing to improve the high n_2 regions or the spreading around the main peak.

2.2.2 SCOT

The frequency weights for SCOT processor are defined as

$$\psi(f) = \frac{1}{\sqrt{G_{x_1x_1}(f)G_{x_2x_2}(f)}}, \tag{2.22}$$

so this takes care of regions where either n_1 or n_2 might be high solving a possible disadvantage with ROTH.

2.2.3 PHAT

Both SCOT and ROTH suffer from the disadvantage that the value of $R_{y_1y_2}(\tau)$ is spread around the delta function depending on the cross-spectrum $G_{x_1x_2}(f)$. However, the TDOA information is carried only by the phase of the cross-spectrum and not the amplitude. So, setting the weights as

$$\psi(f) = \frac{1}{|G_{x_1x_2}(f)|}, \tag{2.23}$$

we get

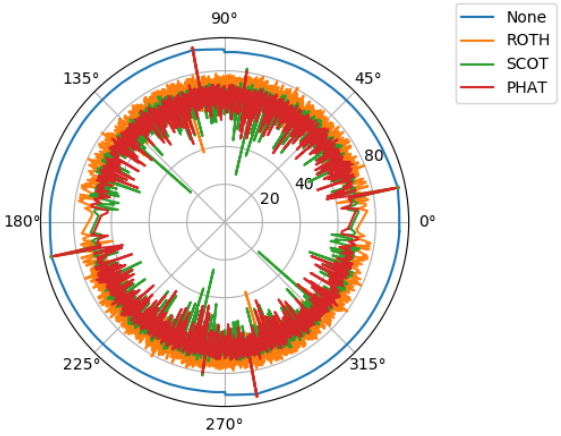
$$R_{y_1y_2}(\tau) = \int_{-\infty}^{\infty} \frac{G_{x_1x_2}(f)}{|G_{x_1x_2}(f)|} e^{j2\pi f\tau} df, \tag{2.24}$$

Now, we have from Eq. 2.12

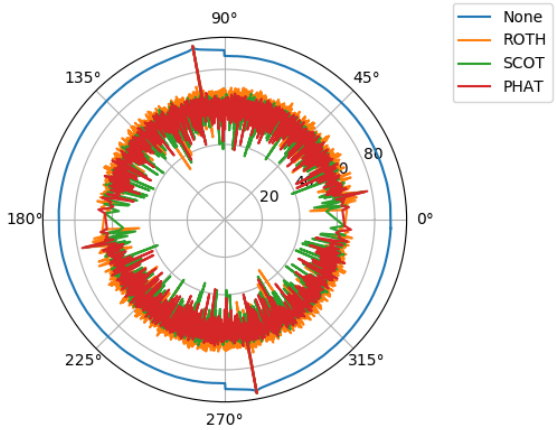
$$\begin{aligned} |G_{x_1x_2}(f)| &= \alpha G_{s_1s_1}(f) \\ \frac{G_{x_1x_2}(f)}{|G_{x_1x_2}(f)|} &= e^{-j2\pi fD}, \end{aligned} \tag{2.25}$$

where the magnitude information is cancelled and only the phase information remains, where D is the delay or the 'phase'. The cross-correlation is

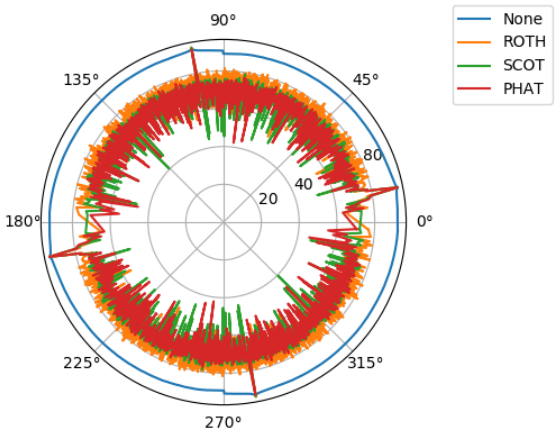
$$\begin{aligned} R_{y_1 y_2}(\tau) &= \int_{-\infty}^{\infty} e^{j2\pi f(\tau - D)} df \\ &= \delta(\tau - D) \circledast \int_{-\infty}^{\infty} e^{j2\pi f \tau} df, \end{aligned} \tag{2.26}$$



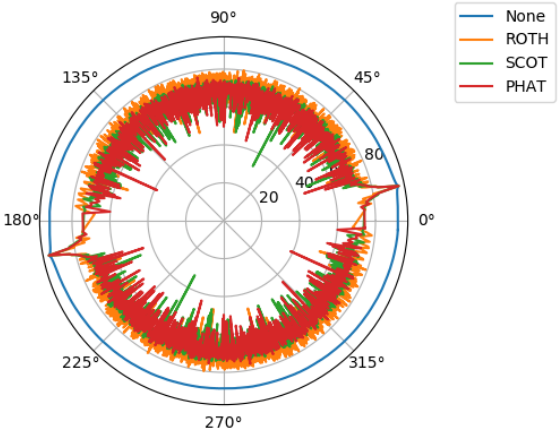
(a) Both sources at 40dB SNR



(b) S1 at 20dB SNR, S2 at 40dB SNR



(c) Both sources at 20dB SNR



(d) S1 at 20 SNR, S2 at 0 SNR

Figure 2.5: Figures compares GCC algorithm with different weighting for localization performance of two sources with various SNRs. The simulations assume two microphones placed 1m apart along the $0° - 180°$ axis. The sampling rate is assumed to be 192kHz and speed of sound is 343m/sec. Two sources playing pink noise at different levels and located at $S_1 : 15°$ and $S_2 : 100°$ are assumed. Uncorrelated white noise is assumed to be present at the two microphones. No interpolation fixing is done. The level of the noise is unchanged but the level of the signal is varied to achieve the different SNRs.

So ideally PHAT weighting gives a cross-correlation value that has no spreading and gives a clean peak at $\tau = D$. However, even though PHAT seems to solve all problems, the method is not without issues. Most of the issues arise from the assumptions made for PHAT. These are itemized below:

- n_1 and n_2 are assumed to be uncorrelated. If that is not the case, the magnitude of $G_{x_1 x_2}(f)$ would not cancel out in Eq. 2.25

- The 'expected' value of $G_{x_1x_2}(f)$ is assumed to be known. In reality it can only be estimated, using $\hat{G}_{x_1x_2}(f)$. In situations where $\hat{G}_{x_1x_2}(f) \neq G_{x_1x_2}(f)$, the cross-correlation in Eq. 2.26 will not be a delta function. This error is magnified even more in regions where $G_{s_1s_1}(f)$ is very low. This has the potential to cause PHAT to provide poor results in low SNR conditions.
- Due to pre-whitening, the actual magnitude of the source is lost. However, if two sources are playing uncorrelated signals at the same time, their relative levels are not lost. This is because, the weight factor (2.23) normalizes the cross-power by the same factor. This is central to the retrieval of actual levels for multi-source localization as will be discussed in detail later.
- The method is developed for only two microphones, so assuming far-field incidence, the TDOA between the two microphones can only retrieve the angle of incidence of the source on the array. Thus, the method needs to be extended in some manner if the accurate source location is required to be known.

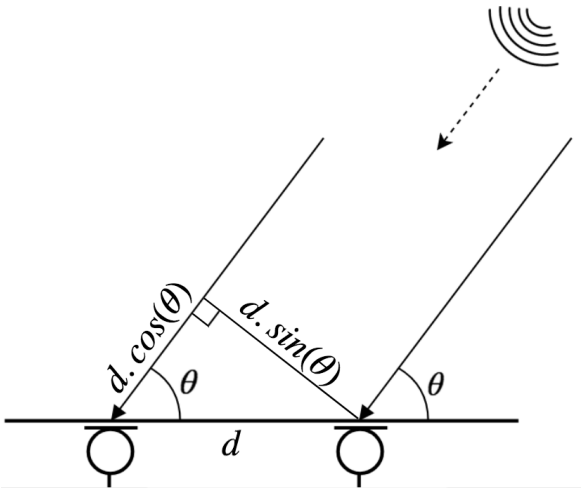


Figure 2.6: Figure represents plane wave incidence on a microphone pair. For broadside incidence ($\theta = 90°$) the time delay is the minimum = 0 between the two microphones. The next time delay allowed is 1/fs, corresponding to travel distance of c/fs. So we have dsin(θ) = c/fs. For endside incidence ($\theta = 0°$) the time delay is maximum = d/c. The next time delay allowed is d/c-1/fs, corresponding to travel distance of d-c/fs, and we have dsin(θ) = d-c/fs.

A practical issue that exists with all GCC methods is the angular resolution of localization. If the signals are recorded at sample rate $f_s = 44.1kHz$, then the minimum time delay allowed is $1/44100$ sec for 1 sample delay. For two microphones placed distance 20 cm apart, the minimum resolution achievable in this time is 2.2° broadside to 16° endside, assuming speed of sound to be 343 m/sec. The resolution is different from broadside to endside because the TDOA to θ is not a linear computation as shown in Fig. 2.6. At 192kHz and 1m microphone distance, the issue is less severe, being 0.1° broadside to 3.4° endside. The issue can be solved by interpolation. Parabolic curve fitting was initially the proposed method to solve it, but was shown to be a biased estimator [3]. Consequently, various interpolation techniques have been developed to overcome this issue [4], [5], [6], [7]. The 2D localization resolution with no interpolation is plotted in Fig. 2.7. Simulations for GCC are shown in Fig. 2.5. It can be seen that the resolution falls the closer we get to end-side (0° and 180°). Also it can be seen that the results are poor if no weights are used. PHAT and SCOT perform quite similarly in the simulations, with PHAT being marginally better. It can be seen that the level difference is maintained between the two sources in the results. However no peaks are visible if the SNR falls to 0 dB.

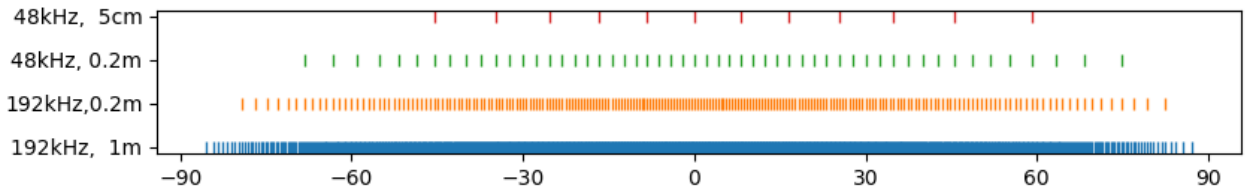


Figure 2.7: Frontal 2D localization resolution for different sample rates and distance between a pair of microphones. The x-axis of the plot specifies the θ in degrees. As can be seen large apertures and high sample rates have a better resolution than lower sample rates and smaller apertures.

2.3 Multiple Pair GCC

GCC methods described previously are designed for a single pair of microphones. Various algorithms have been proposed to extend the GCC algorithm to multiple pairs of microphones. SRP-PHAT approach [8] combines the steered response power (SRP) beamformer method [9] to the GCC approach. Griebel [10] describes a method where the "GCC functions derived from various microphone pairs are simultaneously maximized over a set of potential delay combinations consistent with candidate locations" which can be seen as a special case of SRP-PHAT where the redundant information from additional microphone pairs are utilized. *Okuyama et al.*, in a 2002 study[11] show that when using a spatial array like a tetrahedron, the propagation direction of sound through the array can be determined irrespective of the speed of sound, by using the least-squares approach. This means that for localizing sound sources outdoors, the instantaneous temperature and wind on the microphone array need not be known. However, the technique is employed for a single source localization, i.e peak detection, and is not ideal when the correct magnitude of the source also needs to be determined. Benesty [12] provides a method to fully utilize the redundant information from multiple microphone pairs to make the time-delay estimation (TDE) process more robust against distortion and also improve angular resolution. The method re-derives multi-channel cross correlation (MCCC) to apply linear interpolation on the GCC data to improve the angular resolution of localization. The technique proved robust for indoor speech localization, however, it requires computation of the cross-spectral matrix, leading the algorithm to be inherently narrowband. More recently, Liu and Shen [13] used a motorized robot with four microphones arranged in a cross-formation. The algorithm uses 'de-noising' techniques such as adding a small regularization term to the denominator of the PHAT weight, which can reduce the low SNR issues surrounding PHAT. The low SNR regions can be further penalized by using reliability-weighted RW-PHAT [14], where a-priori SNR information is used to estimate the weight to be multiplied during the PHAT computation. The technique can be seen as an extension to the SRP-PHAT algorithm. Hu[15] provides a method to do eigenvalue decomposition based GCC (ES-GCC) with ES-GCC producing lesser number of outlier locations that GCC-PHAT. Badali [16] compares various localization algorithms using a eight microphone array located on a cube. The authors use hyperbolic intersection on the GCC results from multiple pairs of microphones. They conclude that if a *Direction Refinement* procedure is run, in which first a far-field assumption search is done and the locations are then 'refined' for near field, then the results from SRP-PHAT can be improved. But this procedure might not be relevant for far-field outdoor localization.

2.4 SRP-PHAT

Steered Response Power (SRP) source localization is a method to detect sound source locations using beamforming techniques [9]. SRP is different from TDOA based methods discussed before. While the generalized cross correlation is a simple cross correlation between each pair of microphones and outputs an estimate of the time delay, the SRP method beamforms the space around the array and computes the energy of each location beam. It 'looks' at all possible directions individually (steering) and computes the power of the signal cross correlation in that direction (beamforming). The assumption is that the cross power of the steered microphone signal is maximal in the source position. However, the computational demand for this can rise quite fast (depending on the sample rate and the angular resolution of the beamforming), making it nearly impossible to implement in real time applications. But, its performance in difficult conditions outperforms the TDOA based methods [17]. Since real-time localization is not of primary importance for this thesis, SRP based methods can be applied. However, in the same fashion as the GCC methods proposed to pre-filter the signal before performing the cross correlation, PHAT weighing can also be applied on the beamformed signal. The method, called SRP-PHAT, combines the robustness of the SRP to the accuracy of the PHAT.

2.4.1 Steered Response Power

The SRP method is based on a regular delay-and-sum beamformer, for a given point in space having range ρ, azimuth θ and elevation ϕ with the microphone array, the output of the beamformer is given by

$$y_{\rho,\theta,\phi}(n) = \sum_{m=0}^{M-1} w_m x_m[n + f_{0,m}(\rho, \theta, \phi)], \tag{2.27}$$

where $x_0[n]$ is the signal received at time n, at an arbitrary microphone used as reference, w_m is the amplitude weight for microphone m, and $f_{0,m}(\rho, \theta, \phi)$ is the relative delay between the reference microphone and the m^{th} microphone. When far-field approximation is assumed, the range cannot be computed[1] and the delay-and-sum beamformer output can be rewritten as follows:

$$y_{\theta,\phi}(n) = \sum_{m=0}^{M-1} w_m x_m[n + f_{0,m}(\theta, \phi)], \tag{2.28}$$

For $w_m = 1$ (assuming perfectly omni-directional and equally sensitive microphones), the output power of the beamformer becomes

$$E[y_{\theta,\phi}(n)^2] = \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} R_{x_i,x_j}[f_{i,j}(\theta, \phi)], \text{ for } i \neq j. \tag{2.29}$$

[1] For range computation, the cone approximation cannot be assumed. The delays should be used to compute hyperboloids and not cones. The intersection of the hyperboloids can then be used to compute range. However, it should be remembered that even a small error would lead to large variations in range, as for far-field, small movements in the hyperboloids would cause large movements in range results.

In short, the SRP algorithm can be given by,

- Compute the cross correlations of the signals received for all the microphone pairs.
- Compute, for each angle (ϕ, θ) in the search map, the corresponding set of delays for every microphone pair $f_{i,j}(\theta, \phi)$. So if 1° angular resolution is used, delays for 360*180=64800 angular positions, for each microphone pair, need to be computed.
- For each (ϕ, θ), sum the cross correlation values at the corresponding delays from all microphone pairs. This sum is the output of the SRP beamformer defined in Eq. 2.29[2]

$$S_{SRP}(\theta, \phi) = E[y_{\theta,\phi}(n)^2]. \tag{2.30}$$

2.4.2 Extending PHAT to SRP-PHAT

PHAT can be extended to SRP-PHAT, by simply pre-filtering the cross-correlations before the SRP sum step,

$$R_{x_i,x_j}(\tau) = \sum_{k=0}^{N_f-1} \psi_{ij}(k) X_i(k) X_j^*(k) e^{j2\pi \frac{k}{N_f}\tau} \tag{2.31}$$

where

$$\psi_{ij}(k) = \frac{1}{|X_i(k)X_j^*(k)|} \tag{2.32}$$

2.4.3 Localizing with SRP-PHAT

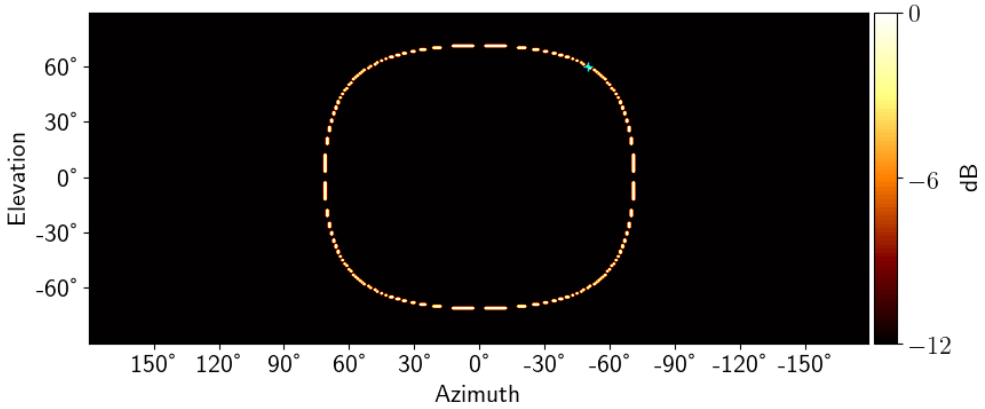


Figure 2.8: SRP-PHAT is run to localize a single point source using two microphones M_1 *and* M_2 (for microphone locations see Appendix E.4). The source can only be localized on a circle. The blue cross in the figure indicates the actual source location. Note that the reason the circle does not appear exactly circular in image is due to the cylindrical projection being used to display the result.

[2] An improvement on the SRP search algorithm was proposed by pre-mapping the relative delays to their corresponding set of locations [17]. Instead of proceeding with a full sequential search in the 3D space, a search on the possible relative delays, where the cross correlation values are above a threshold, is considered. The possible delays between individual microphone pairs are already known based on the array geometry and can be stored in memory. The computational cost gain can be immense depending on the number of microphones. However, the method is not suitable if the whole acoustic map of an environment in required, so it is not detailed further here.

When localizing a point source using SRP-PHAT, if two microphones are used, the only information that can be computed is the angle of incidence of the source on the array. For example, for a source located at (-50°, 60°)[3], the result is a circle around the array where the source might be located, shown in Fig. 2.8. This is because the angle of incidence from every point on the circle, to the mid point of the line joining the two microphones, is the same. This circle is actually the base of the cone discussed in Sec. 2.1.

The results in Fig. 2.8 are displayed using the cylindrical projection technique, such that the entire spherical space around the origin can be shown as a rectangle, with x-axis being the azimuth and y-axis being the elevation. This technique is employed throughout the thesis to give a full picture of the localization results.

A new circle will result for each new microphone pair used to localize the source, as long as the microphone pairs are not all placed in the same line[4]. For example, for three microphones A, B and C placed in an equilateral triangle formation, three circles can be computed (one each for AB, BC and CA). The maximum peak occurs at two locations with (-50°,± 60°) as shown in Fig. 2.9. If a fourth microphone is placed in the same plane as the triangle, the array response will be a combination of circles from six possible microphone pairs (${}^4C_2$). However, the new circles would all pass through the same two locations. For a non co-planar array, e.g. a tetrahedral array, the maximum peak occurs at exactly one point, shown in 2.10.

2.4.4 Some Considerations with SRP-PHAT

Subsidiary Peaks

Even though a tetrahedral array is able to detect a point source to a single maximum peak position, subsidiary peaks can appear in the energy map at DOAs that don't correspond to the true source DOA, since, the cross-correlations values at computed delays are summed by the beamformer (Eq. 2.30). For example, points where only two-five of the circles meet. If multiple sources are localized, these peaks in the SRP-PHAT energy map can add up leading to the detection of a fake source and can also mask real sources. The effect of these subsidiary peaks can be reduced by increasing the number of microphones. This is because, even though more microphone pairs would mean more localization circles, it also means the real peaks would be higher, effectively lowering the noise/subsidiary peak floor. Indeed, solutions in the market exist with even ninety microphones ([18], Fig. 31). However, since, the number of microphones is a constraint requirement for this thesis, this solution is not considered.

Linear Dependency

Since only three pairs out of the six in a tetrahedral array are linearly independent (Eq. 2.5), the localization can also be done considering only three of those pairs. The result in shown in Fig. 2.11. Considering only independent microphones, Eq. 2.30 can be rewritten as,

$$S_{SRP}(\theta, \phi) = \sum_{i=1}^{M-1} R_{x_0,x_i}[f_{0,i}(\theta, \phi)] \tag{2.33}$$

[3]For the purpose of this thesis, locations are designated as (x°, y°), signifying (azimuth, elevation) of the location, respectively, in spherical coordinates.

[4]In the case of a linear array, the multiple circles would overlap completely

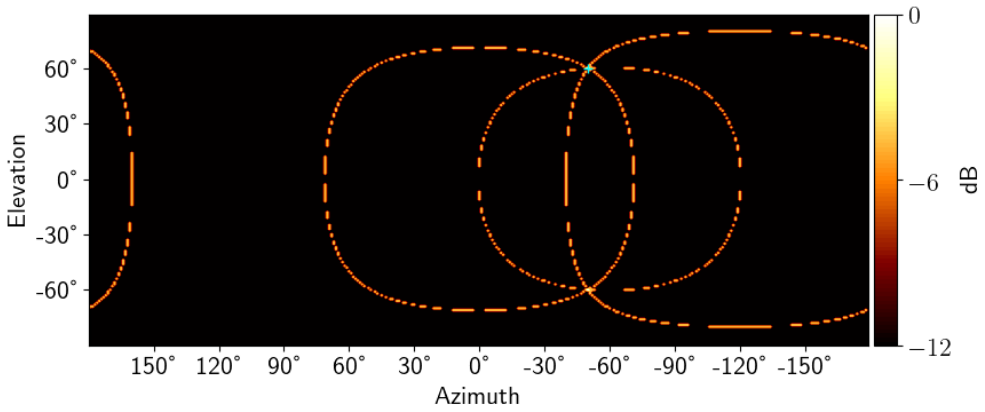


Figure 2.9: SRP-PHAT is run to localize the source with three microphones $M_1, M_2 and M_3$ as described in appendix E.4.

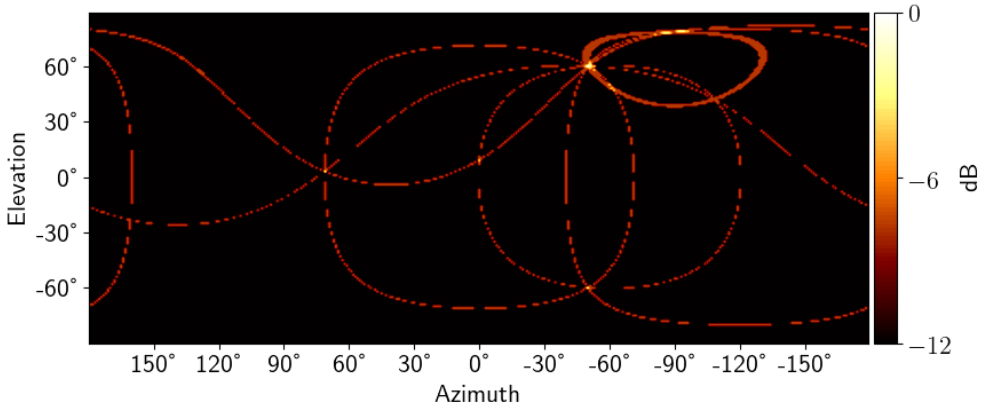


Figure 2.10: SRP-PHAT is run to localize the source with a tetrahedral array ($M_1, M_2, M_3 and M_4$ as described in appendix E.4).

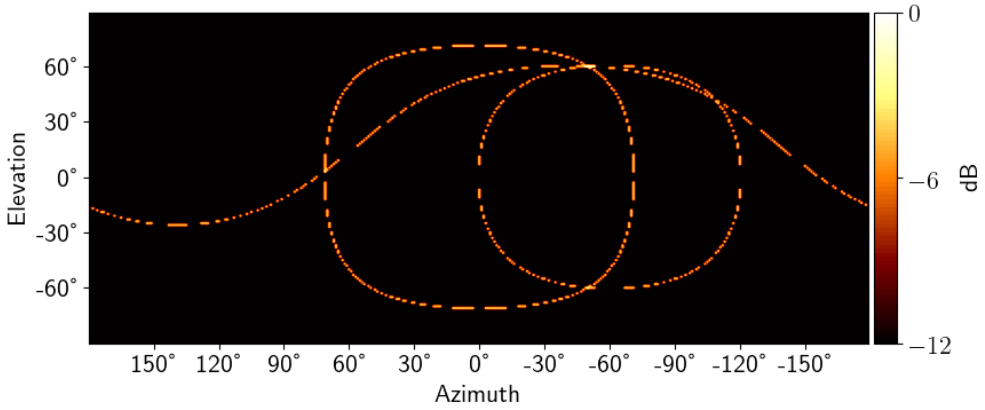


Figure 2.11: SRP-PHAT is run to localize the source with a tetrahedral array but only linearly independent microphone pairs are considered

Note that Eq. 2.5 is only true for no noise conditions. In noisy conditions, there is a potential to gain information by using the redundant microphone pairs. This is because if noise at all microphones

is assumed to be uncorrelated, then even though the noise causes certain microphone pairs to detect a source at a 'sourceless' location, other microphone pairs might detect a lower magnitude at that location. Due to more microphone pairs, the sum of all microphone pairs will be even higher at the real source location, and at other locations, the sum due to the noise will be suppressed. It can be seen in Fig. 2.10 that using all the microphone pairs adds to the overall noise on the map as more pairs can now contribute to the SRP sum, leading to more circles. However, the peak of the true source also becomes higher, due to more localization circles providing power at the source location. This means that even though the noisy floor has noise in more locations, it is of a lower magnitude, leading to a higher achievable dynamic range. For this reason, from here on the localization results considers all possible pair of microphones.

Chapter 3

Minimum-SRP-PHAT

3.1 Introduction

The issues with SRP-PHAT algorithm arise in multi-source localization. As can be seen in Fig. 2.10, even in ideal conditions, the localization result (array response) for a single point source is not a point, but rather a set of intersecting circles. For multiple sources, this leads to the intersection of a multitude of circles from the different sources. Due to the interaction of these multiple circles, it is difficult to distinguish peaks caused by the actual sources from those caused by the array response. The issue is only exacerbated in noisy outdoor conditions. This can reduce the use-able dynamic range of the results. In the case of multiple sources playing at different levels, it can mask the lower magnitude sources. Therefore, it is important to remove the array response from the map. The process of removing the array response from the localization results is commonly referred to as deconvolution. Deconvolution has been applied in several different solutions over the last few decades. CLEAN [19] and CLEAN-SC [20] algorithms apply deconvolution on results using the point spread function[1]. Other methods such as DAMAS [21] or DAMAS-C [22] rely on computing the cross-spectrum matrix (CSM) to solve a set of linear equations and retrieve the location and level of the sources. The computation of the CSM can only be done for a single relevant frequency and as such it is designed for a narrowband algorithm[2]. Note that the SRP-PHAT perform the beamforming in the time domain and the array response varies for each source positions. Therefore, these methods above cannot be applied for SRP-PHAT deconvolution. For the purpose of this thesis, a minimum SRP-PHAT (Min-SRP-PHAT) algorithm is derived, which is described in this section. Simulations are run to elucidate the algorithm performance in various conditions. Real world test are then conducted, whereupon, the algorithm is applied to localize outdoor conditions to compute source locations and levels.

3.2 Theory

SRP-PHAT (Eq. 2.30) is the sum of the cross-correlation values for multiple pairs of microphones at the time-delays corresponding to the beamformed location. Using the far-field assumption, the

[1] Point spread function is the response of the array to a point source. Standard narrow-band beamformers suffer from the issue of side-lobes, where the main source is detected on the main lobe. If the source frequency is higher than the array aperture allows, grating lobes which can be as high as the main lobe can also appear.

[2] Multiple CSMs for a range of frequencies can also be computed, but each individual CSM still corresponds to a single frequency and a single localization result

power received from a single source to all microphone pairs can be assumed to be equal. Then, if, the minimum power between the microphone pairs at each beamformed location is used (instead of summing), peaks which are detected only by a subset of microphone arrays disappear automatically and the deconvolution problem is solved directly. This is because power at positions where circles from all microphone pairs are not present will compute to zero. The Min-SRP-PHAT equation can be rewritten as,

$$S_{min-SRP}(\theta, \phi) = R_{min}[f_{i,j}(\theta, \phi)], \text{ for } i \neq j. \tag{3.1}$$

$$R_{min}[f_{i,j}(\theta, \phi)] = \min R_{x_i,x_j}[f_{i,j}(\theta, \phi)] \text{ with i,j = 0,...,M-1 and } i \neq j. \tag{3.2}$$

Min-SRP-PHAT is equivalent in principle to finding the intersection of multiple cones, since this

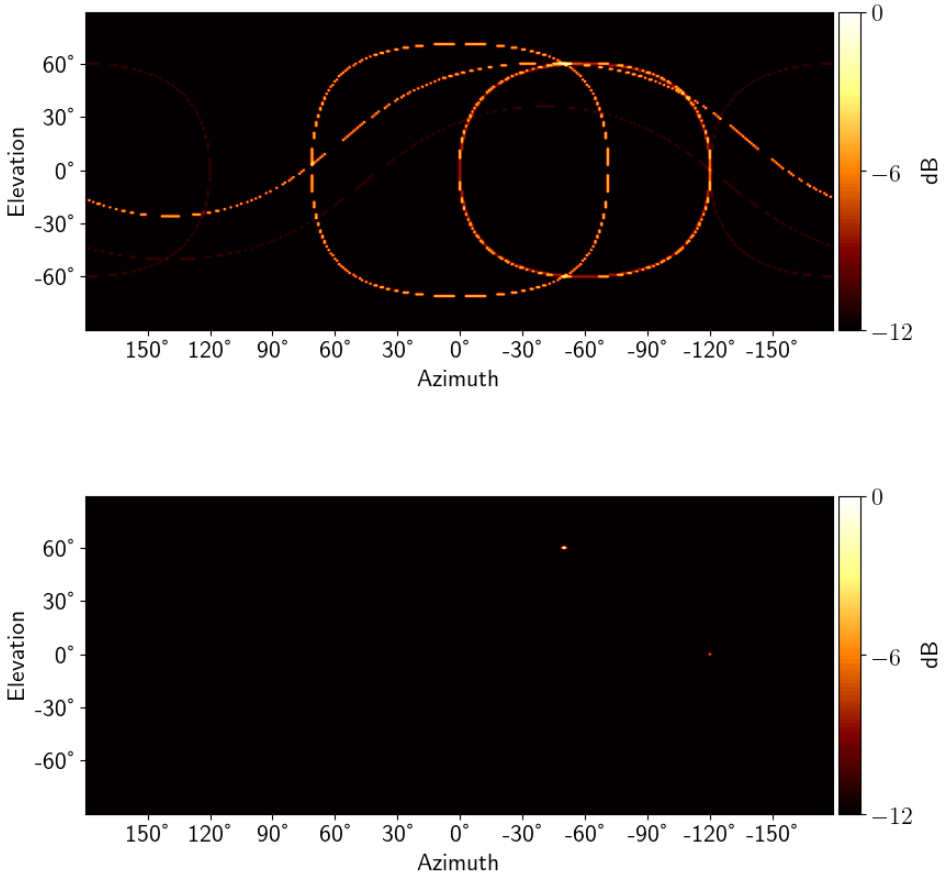


Figure 3.1: Figures depict localization results for sources at $(-50°, 60°)$ having magnitude 0 dB and $(-120°, 0°)$ having magnitude -6 dB, for SRP-PHAT (top) and Min-SRP-PHAT (bottom). In SRP-PHAT, power from the source at $(-50°, 60°)$ affects the result for the source at $(-120°, 0°)$ since they share a localization circle. Min-SRP handles this issue, since the minimum power cone at $(-120°, 0°)$ contains the correct power, thus the higher magnitude cone from $(-50°, 60°)$ is rejected.

method only returns sound sources detected by all independent microphone pairs. The drawback of this method is that in case of localizing point sources, in noiseless conditions, even a minor error in temperature or wind has the potential of not detecting the sound source completely. However, as we shall see later, since outdoor sound sources are usually large, and outdoors environments relatively noisy, the cones from microphone pairs are not sharp circular lines, rather, they are annular. An error in weather conditions would then cause these annular circles to 'smudge' together. The problem is

that this has the potential to underestimate the sound source, both in size and magnitude. However, the advantages of Min-SRP-PHAT are manifold. It removes the subsidiary pseudo-peaks while preserving the relative SPL difference between the sources. The preservation of relative sound levels is important for computing the correct acoustic map of an area. Also, if two sources are located on the same localization cone for a pair of microphones and a normal SRP-PHAT is conducted, both sources would appear higher in magnitude than they actually are. Min-SRP-PHAT fixes this issue. Fig. 3.1 describes this affect. Image sources due to reflection will also have the incorrect power for normal SRP-PHAT due to the same reason. The image source power will increase the main source power and vice versa, as they share two localization circles for a horizontal tetrahedral array. Min-SRP-PHAT takes care of the errors due to reflection (discussed later in Fig. 3.7). With Min-SRP-PHAT, the redundant pair information would always improve results in low SNR conditions. This is because only the lowest power from all possible microphone pairs are used. Thus, the redundant pair information will only result in removal or lowering of results in the non-source positions caused by noise. Fig. 3.2 describes this affect.

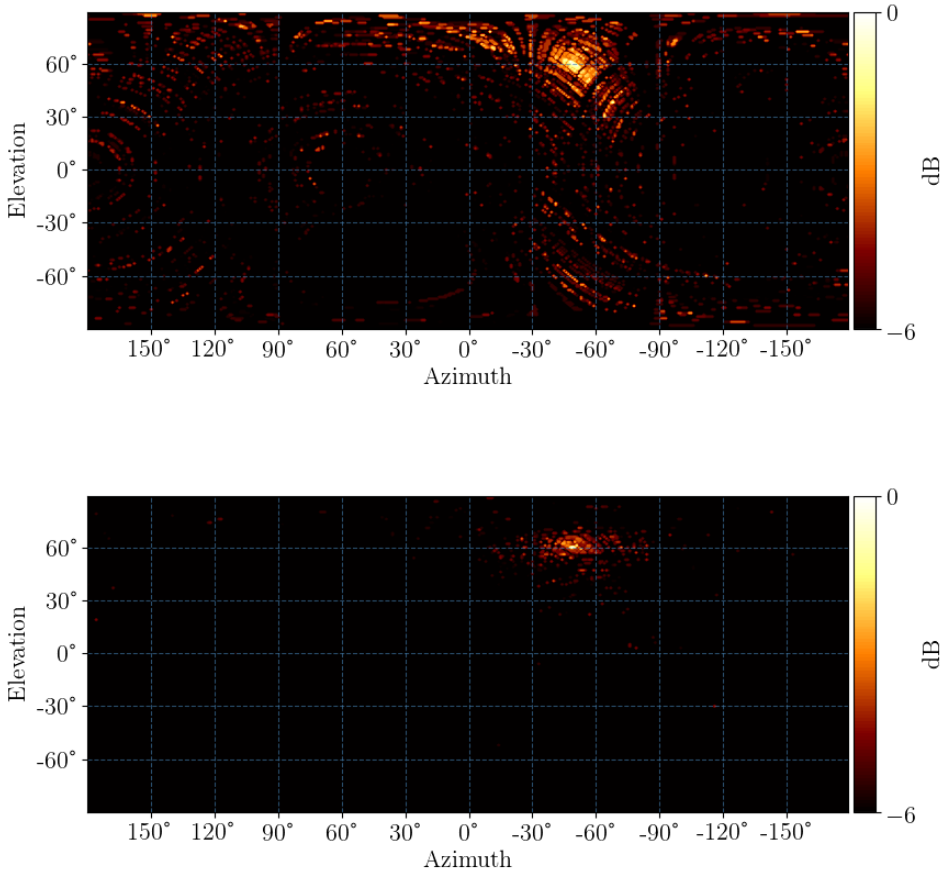


Figure 3.2: Figures depict localization results for a source at $(-50°, 60°)$ for minimum power SRP-PHAT with only independent microphone pairs (top) and minimum power SRP-PHAT with all microphone pairs (bottom). As expected, using all microphone pairs results in removal of some of the incorrect results from the independent microphone pairs (To highlight the differences, the SNR for this simulation is kept at -6dB and the dynamic range has been reduced to 6dB).

3.2.1 Source Level Retrieval

For outdoor sound map reconstruction, the actual magnitude of the different outdoor sources is required. PHAT normalization essentially whitens the power signal, so the actual magnitude cannot be retrieved from the SRP-PHAT algorithm. However, since the PHAT division factor is the same for all sources, the relative power levels between sources are maintained[3]. This can be utilized to retrieve the required levels. Various studies have been made on the correctness of the relative source power computed in this manner. In one study, the author compares the error in multi-source power, when instead of summing the SRP-PHAT as is done for regular SRP-PHAT, the powers are computed using geometric (GM) and harmonic means (HM) [23]. As GM and HM give, by their nature, more weight to the lower values, doing GM and HM is essentially a move towards a Min-SRP-PHAT approach. A comparison between the different approaches is described by Fig. 3.3. Three sources located at (-30°, 30°), (-50°, 30°), (-70°, 30°) having source magnitude 0dB, -3dB and -6dB are localized. The choice of locations is arbitrary and is chosen close to each other to be able to zoom into results effectively. As

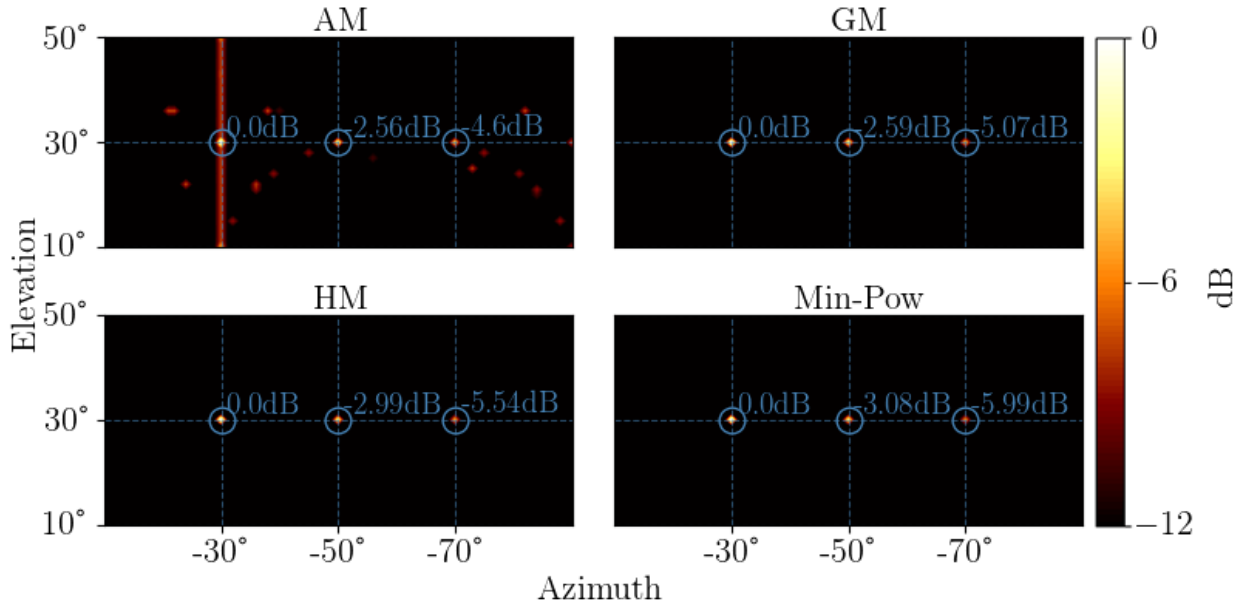


Figure 3.3: Figure compares multi-source localization for AM, GM, HM and MP deconvolution approaches (On many figures in the thesis, axes are only drawn on the left and bottom. This is to allow more space for the plots. The grids can be used to determine the Azimuth and Elevation for every plot).

can be seen, Min-SRP-PHAT provides even more accurate results than the HM based approach. This is because, even in HM the sum of all localization cones is taken, which has the potential to overestimate the source magnitude. The results presented here are for relatively good conditions of +20dB SNR. For worse conditions, the results can be even worse for non min-pow based approaches.

Now, since the relative power levels between sources on the Min-SRP-PHAT map are maintained, the problem of finding the absolute level becomes a trivial one. If the true power at any location on the map is known, the map can be normalized to that power. From the Min-SRP-PHAT results, the peak power location computed has the best SNR, and is thus used for this purpose. The generalized cross correlation value without any weights is computed, for delays corresponding to the minimum

[3]Errors can exist in SRP-PHAT, if two sources share localization cones. This can be due to the sources being located on the same cone of one or more microphone pairs. This can also occur during reflection, when if the tetrahedral array is placed horizontally, 3 cones out of 6 will always be shared between the source and the reflection.

power microphone pair that computes that peak. Note that this power is arriving from the entire cone corresponding to that source location for that microphone pair. If multiple sources correspond to the same minimum cone for the peak source location, this can still result in over-estimating the source power. However, unless information relating to the source spectra is known, more accuracy cannot be derived from using the GCC method.

3.2.2 The Min-SRP-PHAT Algorithm

The Min-SRP-PHAT implementation steps are described in Fig. 3.4. The algorithm is basically the same as the SRP-PHAT algorithm discussed before, with the only change happening at the last step where the minimum power from the localization cones is considered, instead of summing the power from all the cones. The computational complexity of the algorithm and the practical implementation details are discussed in this section. The delays across the microphones for each search location on the search map are computed in advance, stored in memory and fed to the algorithm which correspond to the 'Compute array delays' system block in the figure, therefore this step will not be considered computational load. Upon running the algorithm, first of all the computer reads the stored .wav files into memory. The signal cross correlation between each pair of microphones are then computed. This step is the 'GCC-PHAT' block in Fig. 3.4. The 'SRP' system block is related to the array steering for each of the search location (θ, ϕ). This step looks up the delay table corresponding the (θ, ϕ) for each microphone pair, and saves the corresponding power associated with that delay and that pair into an array (P_{all}). So P_{all} contains 6 power values for each location on the search map. The minimum power then selected from P_{all} the 6 power values for each of the locations (θ, ϕ) and this value is stored in the result array[4].

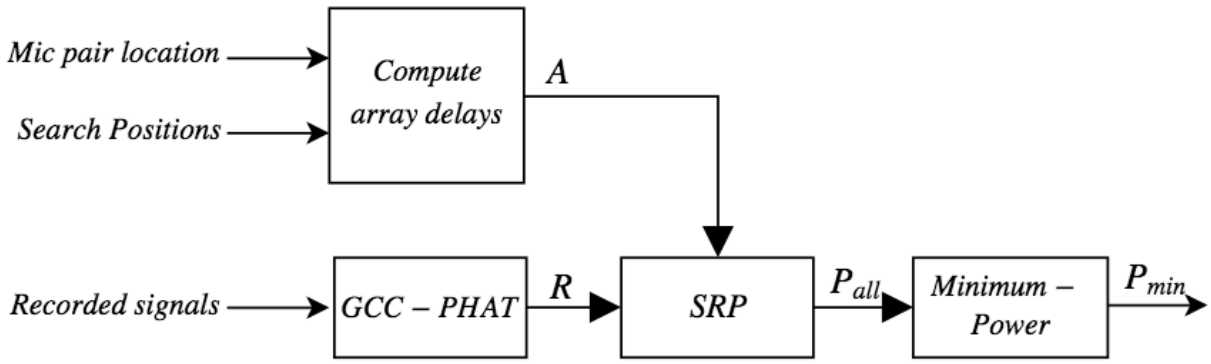


Figure 3.4: Overall localization algorithm

[4]One thing to note is that there is a limit to the granularity of the (θ, ϕ), the achievable angular resolution. The angular resolution of localization is in fact not linear, as shown in 2.6(A delay of one sample does not always correspond to the same change in degree). The SRP-MAP can then be plotted by computing the delays sequentially for a particular angular resolution. However, these delays might be fractional. Since, the values of fractional delays cannot be picked from the cross-correlation array R, these delays are rounded. One way to increase the angular resolution is to increase the sample rate. That way even if fractional delays are encountered, they would be less erroneous. Another way is to apply interpolation to find the value at the fractional delays. For the purpose of this thesis, the interpolation techniques are not considered.

The Min-SRP-PHAT algorithm combines beamforming techniques with cross correlation methods for several pairs of microphones. While the beamforming part does not depend on the size of the input data, it might become a challenge memory wise to store the delay for each search position. The most computationally demanding part of the algorithm is definitely the cross-correlation part. By performing the cross-correlation in the frequency domain, i.e by using the cross-spectrum between pairs of microphones, better averaging of the stationary sources are obtained as well as better efficiency compared to time domain cross-correlation ($\mathcal{O}(n^2)$). The cross-spectrum computation is described in Fig. 3.5.

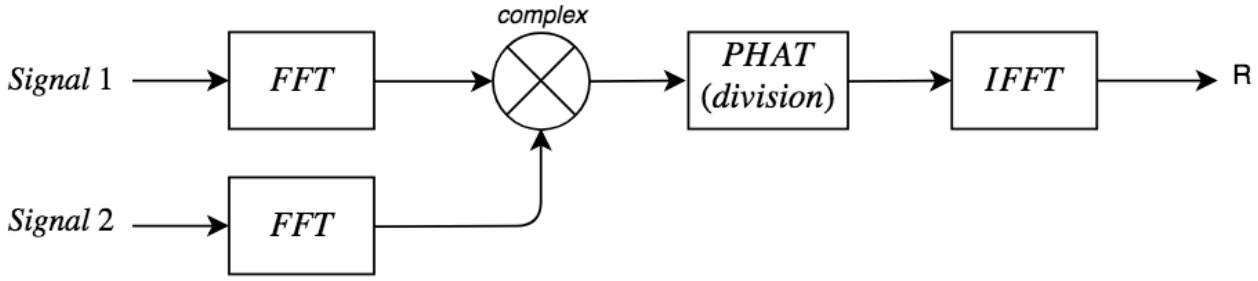


Figure 3.5: Cross spectrum between two signals

Note that not the entire R computed this way is important for the purpose of localization. This is because only a finite number of samples can exist between two microphones, say d. d depends on the array aperture size as well as the sample rate. Delays$d$ could not have been measured by the microphone array. For this reason the R is cropped down to d. Since, the signals arriving at two microphones can be either in front or behind each other, both the first and last d samples of R are taken. The worst-case complexity of the cross-spectrum computation is listed below, with n being the number of input samples (signal length) in the algorithm.

Operation	Worst-case Complexity
FFT	$\mathcal{O}(n \log n)$
Complex Multiplication	$\mathcal{O}(n)$
PHAT (Division)	$\mathcal{O}(n)$
IFFT	$\mathcal{O}(n \log n)$

The SRP block is also computationally heavy, however, it does not scale with the number of samples but rather with the number of locations to look up in the SRP block. Therefore the overall algorithm complexity scales with the FFT complexity $\mathcal{O}(n \log n)$ when n is the number of samples. Memory wise, the function computing the delays at each pair of microphones in the array can be expensive depending on the localization resolution used and the number of microphone pairs. If 1° resolution is used $360 * 180 = 64800$ delays are computed for a pair of microphones. For 6 microphone pairs,$360 * 180 * 6 = 388800$. Using type float64 (8 bytes), the delay table is $360 * 180 * 6 * 8 = 3110400 = 3.11$ MB. However, if high resolution is needed, i.e suppose 0.1° resolution $3600 * 1800 = 6480000$ delays are computed. For 6 microphones,$3600 * 1800 * 6 = 38880000$. Using type float64 (8 bit), the delay table is $3600 * 1800 * 6 * 8 = 311040000 = 311.04$ MB.

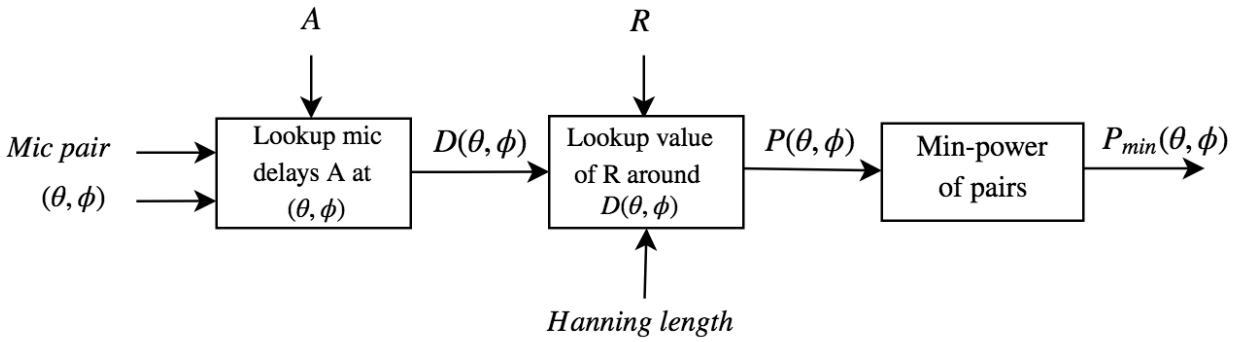


Figure 3.6: SRP block + Minimum Power

3.3 SRP-PHAT vs Min-SRP-PHAT

Comparison between the performance of SRP-PHAT vs Min-SRP-PHAT are done in this section. Simulations are provided to outline the robustness of the SRP-PHAT vs the Min-SRP-PHAT for various environmental aspects and sound source conditions. Effects of practical considerations such as the choice of array aperture size the recording sample rate, and the length of recording are shown. Finally, the robustness of the algorithms for errors in the microphone array placement are given.

3.3.1 Effect of Outdoor Environment

When localizing sound sources, the propagation environment can affect the signals received at the microphones. For outdoor environments, understanding the localization results thus requires knowledge of the physical phenomena at play at the time of the measurement[5]. The effect of ground reflections, temperature and wind are provided here.

Effect of Ground Reflections

Sound received from a far-field sound source is the sum of a plane and a spherical wave component (Eq. A.12). The spherical wave component creates a horizontal ground wave and quickly attenuates with distance. The plane wave component is reflected with the ground (image source), the magnitude of the reflection depends on the acoustic reflection coefficient of the ground material. Rudimentary simulations for a source located at (θ, ϕ) can be made assuming image sound source located at $(\theta, -\phi)$. Fig. 3.7 shows the localization results with a source at (50°,60°) for different ground reflection coefficients. The microphone pairs in the tetrahedral array that are parallel to the ground (horizontal) locate both the source and the image on the same cone. If the array is then placed such that three of its microphones are on the same horizontal plane, three out of the six possible cones will be shared. For SRP-PHAT, this causes the image to be localized at a higher level than it actually is[6].

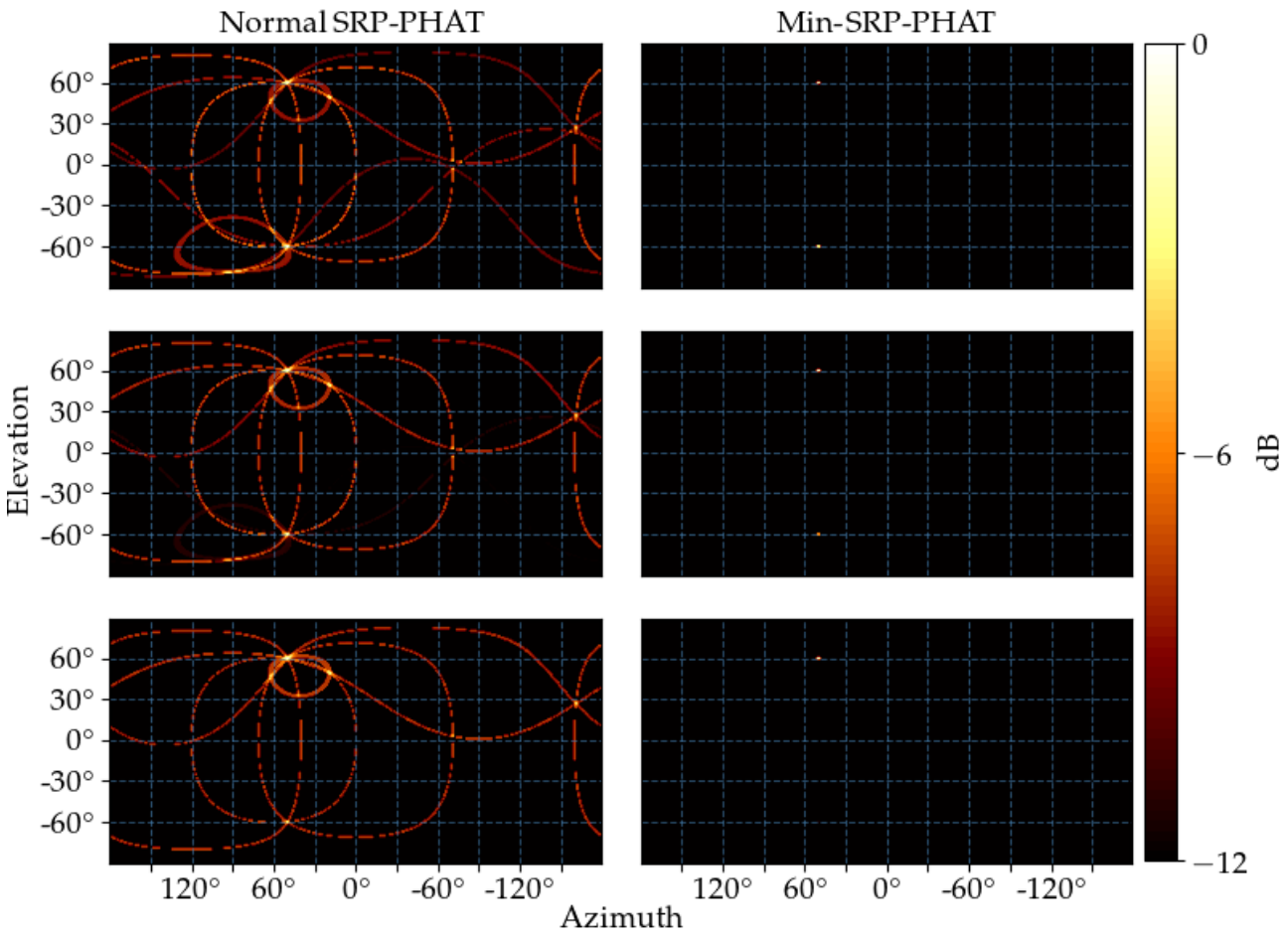


Figure 3.7: Figures depict from top-to-bottom SRP-PHAT localization results for a source at (50°,60°) with ground reflection coefficients (R) of 1, 0.6 and 0.1. For SRP-PHAT, even though the image source should get significantly weaker for R=0.1, it does not as it is supported by the localization cones from the real source. Min-SRP-PHAT, however, is able to detect the source and the image powers correctly.

Effect of Temperature

Temperature affects the speed of sound and thus affects the delay time between the microphone pairs. During measurement, if the speed of sound is assumed to be 343m/sec, this could lead to errors in the localization results. Fig. 3.8 depicts the effect of temperature on localization results for a source at (50°,60°) , where wave files received by the tetrahedral microphone array at temperatures of $0°C$, $20°C$ and $-40°C$ are simulated. Then the localization is run assuming the speed of sound to be 343m/sec in every case. The figure shows zoomed in results around the source location. As can be seen in the figure, if temperature is not considered, it has the effect 'de-focusing' the main peak. If temperature is recorded during measurements, the localization can be run using the correct speed of sound, which would remove this de-focusing issue[7]. For this reason, the temperature is recorded whenever outdoor measurements are done.

[5]For more information on the theory behind outdoor sound propagation are refer Appendix A

[6]One way to mitigate this issue would be to not place the array horizontally, the simulations for this are given in Appendix D.1

[7]Since, the speed of sound is greater at higher temperatures, the number of samples that can fit within the array aperture would reduce. This can also have an effect on the localization results, however, this error is relatively minor. The effect of sample rates on localization results have been discussed later.

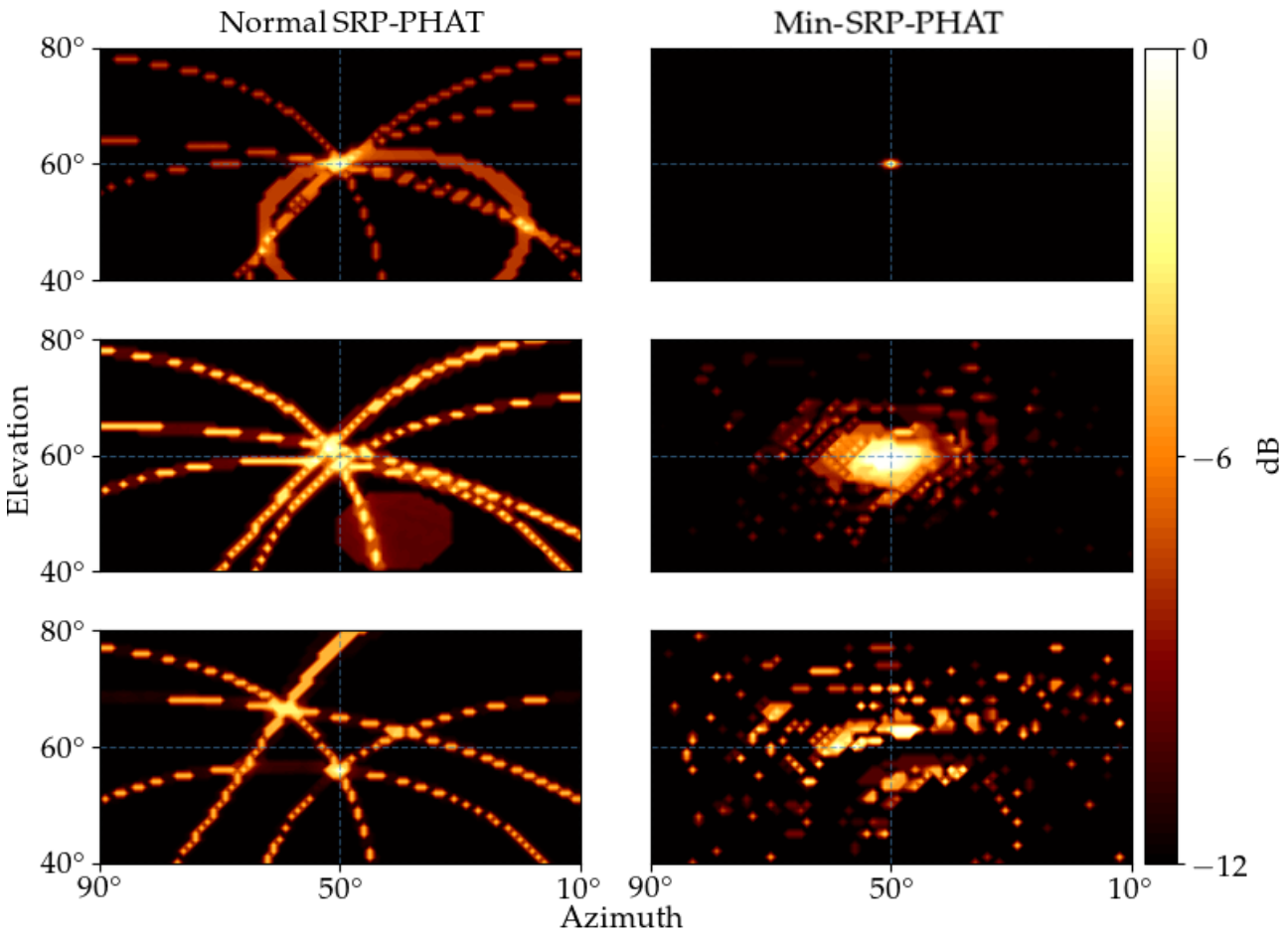


Figure 3.8: Figures depict from top-to-bottom SRP-PHAT localization results for a source at (50°,60°) with at temperatures of $20°C$, $0°C$ and $-40°C$. The extreme temperature of $-40°C$ is chosen to highlight the error.

Effect of Wind

Wind speed effects the speed of sound in the direction of propagation. However unlike temperature, which causes a uniform difference in delays across the different microphone pairs, wind causes the delays to be affected differently depending on where the SRP search is looking and from what direction the wind is blowing. If wind blows perpendicular to the direction of propagation of the sound from the source, then it does not affect the localization. The maximum error happens exactly in and against the direction of the wind. Fig.3.9 depicts the effect of wind on localization results for a source at (50°,60°). It can be seen that when wind blows at 90°, it does not affect the localization results. The magnitude of error when wind blows non-perpendicular to the sound propagation direction depends on the wind speed and the degree of alignment with the wind direction. For SRP-PHAT, an error in wind causes the localization cones to not overlap perfectly. For Min-SRP-PHAT, this causes a lowering of the peaks. This is because the localization circles are annular with peaks in the middle of the annular ring (a 3D torus). A movement in the tori causes the overlap to not happen perfectly, lowering the peaks. This error can be corrected during the SRP search, if the wind is recorded at the time of measurements. However, when doing outdoor measurements, the wind was rarely from a uniform direction. For this reason, the outdoor measurements were only done in relatively low wind conditions (<5m/sec), and no wind correction was applied.

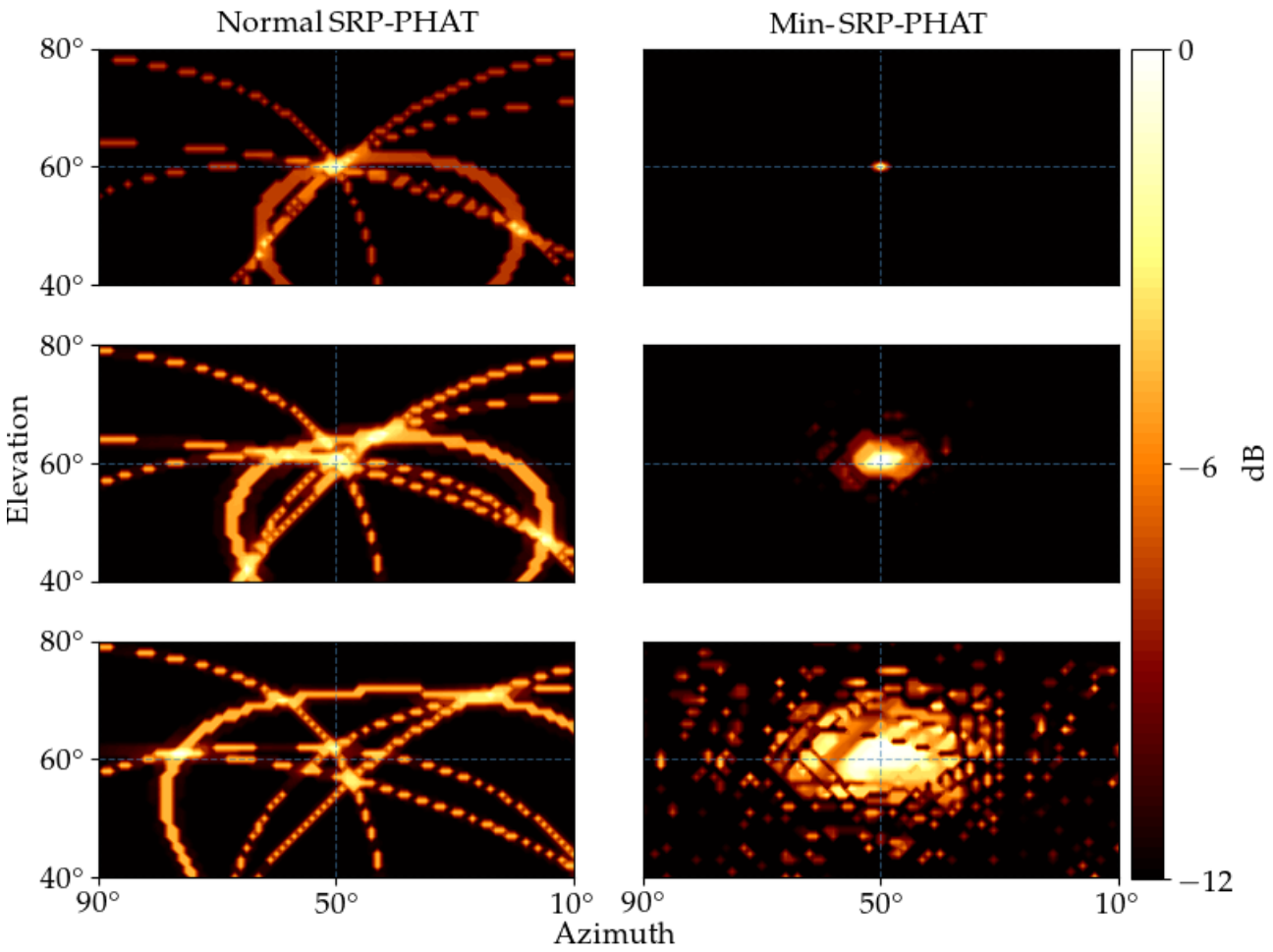


Figure 3.9: Figures depict from top-to-bottom SRP-PHAT localization results with wind of $10m/sec$ blowing 90°, $10m/sec$ blowing 45° and $30m/sec$ blowing 45° to the source sound propagation direction. The extreme wind of $30m/sec$ is chosen to highlight the error.

3.3.2 Effect of Source Conditions

An extremely wide variety of outdoor sound sources exist. They can be spread-out or point sources, narrow-band or wide-band, be uniform across the frequency spectra, or have a lot of low frequency or high frequency content, moving or stationary, constant or transient. To keep the simulations within scope, some outdoor measurements were conducted to realize the major factors that can affect the localization. Based on those measurements, the effect of sound source SNR and the effect of coherence between multiple sound sources were selected. The simulations for those effects are given in this section.

Effect of SNR

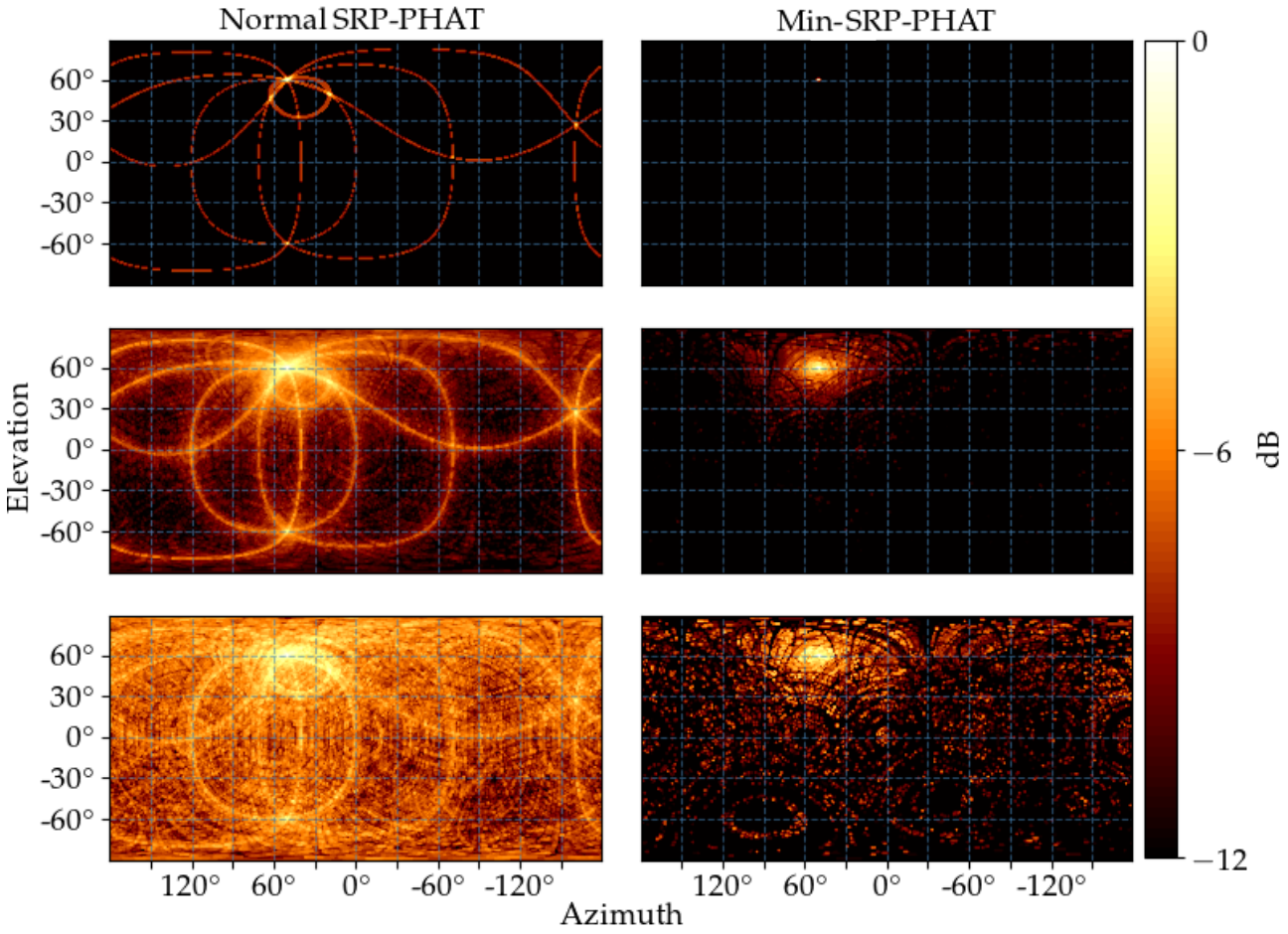


Figure 3.10: Figures depict from top-to-bottom localization results with SNR = 20dB, SNR = 0dB, SNR = -6dB

The source SNR and overall magnitude is an important factor for localizing a source. Even if a source has a high sound level, if the SNR is low, the localization results might be poor. The error due to SNR has been discussed before for GCC comparison for a single pair of microphones, shown in Fig. 2.5, where white noise is used for the simulations and it caused an almost uniform increase of the noise floor across all locations for GCC-PHAT. The same happens for tetrahedral localization, wherein, the noise floor across the entire noise map rises with falling SNR. Fig. 3.10 shows the effect of source SNR on localization, with a point source at (50°, 60°). As can be seen in the figure, the performance deteriorates as the SNR drops. However, understandably, the noise floor is lower for Min-SRP-PHAT, as it clears some of the noise results where not all cones overlap.

Effect of Coherent Sound Sources

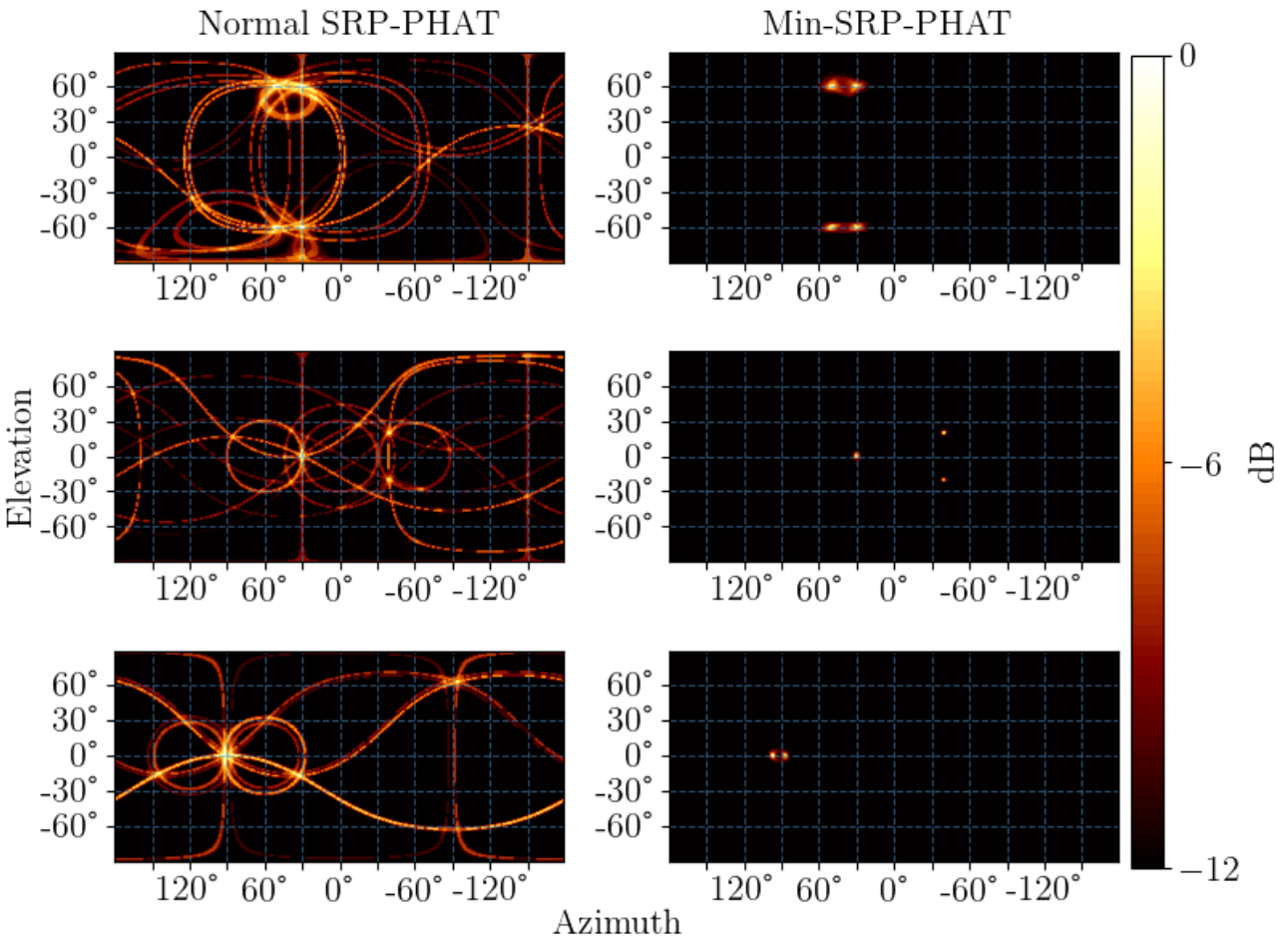


Figure 3.11: Figures depict from top-to-bottom localization results for two sources located at (50°,60°) and (30°, 60°), at (40°, 20°) and (30°, 0°), and at (97°, 0°) and (87°, 0°) respectively. The sources are playing uncorrelated pink noise and the SNR is 6dB.

Fig. 3.11 shows the localization results for two sound sources playing uncorrelated pink noise from different locations. As can be seen, both SRP-PHAT and Min-SRP-PHAT detect the peaks of the source correctly in all the cases. However, if multiple sound sources at different locations play coherently, i.e., the same waveform having a constant phase difference between each other, there is a possibility to detect a pseudo-source corresponding to the phase difference between the sound sources. This is because GCC algorithms inherently depend on the phase difference between the receiving waveforms to do the localization[8]. The localization result for when two sources play the same pink noise are depicted in fig. 3.12. As can be seen in the figure, coherence has an effect on the localization, wherein, pseudo-peaks appear around the main sound source. The magnitude of these pseudo-peaks depends on their proximity and the phase difference. The effect of coherence is also seen later in an outdoor measurement, where multiple loudspeakers were playing music in an outdoor environment.

[8]The same error happens with human ears which causes detection of a stereo image in a two channel loudspeaker setup. Changing the inter-aural time difference causes this phantom image location to shift as well.

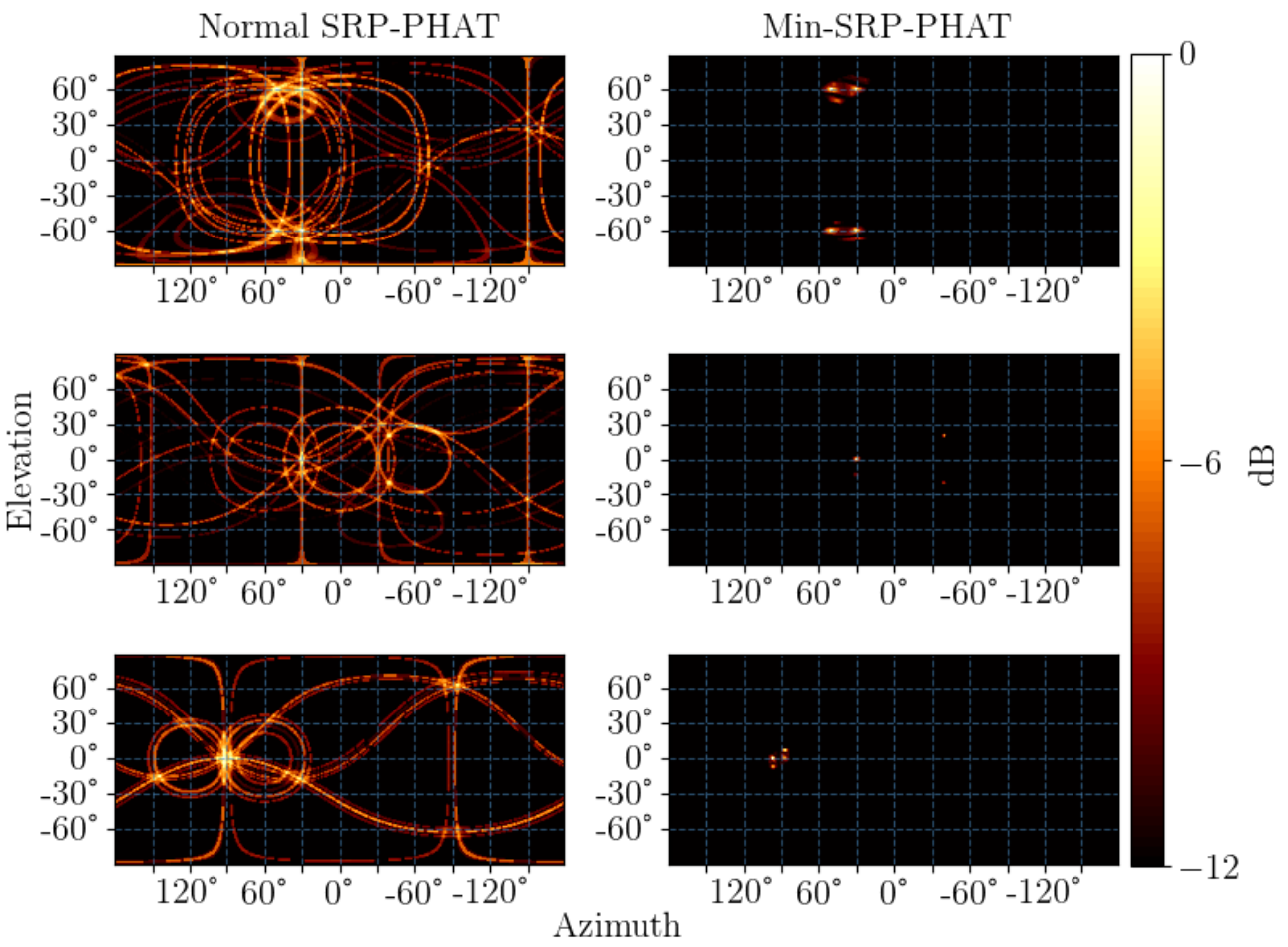


Figure 3.12: Figures depict the localization results for when the sources from the previous figure play coherently. The sources here play the same pink noise. Pseudo-peaks can be seen to appear around the main sound sources.

3.4 Practical Measurement Considerations

Some practical measurement considerations can have a considerable effect on the localization results. For instance, the delay table that is computed for the SRP-MAP has a maximum magnitude, d. This is because the microphone array can measure signal delays only within a finite value. Thus, d is the number of samples that can fit within two microphones of the array. Obviously, d depends on the array aperture and also the sample rate of recording. Higher d values are preferable as that directly translates to a higher achievable resolution on the SRP-MAP (D.2).

Effect of Array Aperture and Sample Rate

As discussed above and shown in Fig. 2.7, reducing the aperture size or sample rate also reduces the angular resolution of localization, which causes a direct degradation in SRP-PHAT performance. Fig. 3.13 depicts the effect of reducing sample rate or aperture size on the SRP-PHAT localization results. Any reduction in the array aperture of the sample rate causes the localization circles to become annular. This is because more θs and ϕs correspond to the same integral delay. These issues can be somewhat alleviated if the algorithm considered fractional delays and interpolation. However, even with interpolation, some data between the integral delays is always lost.

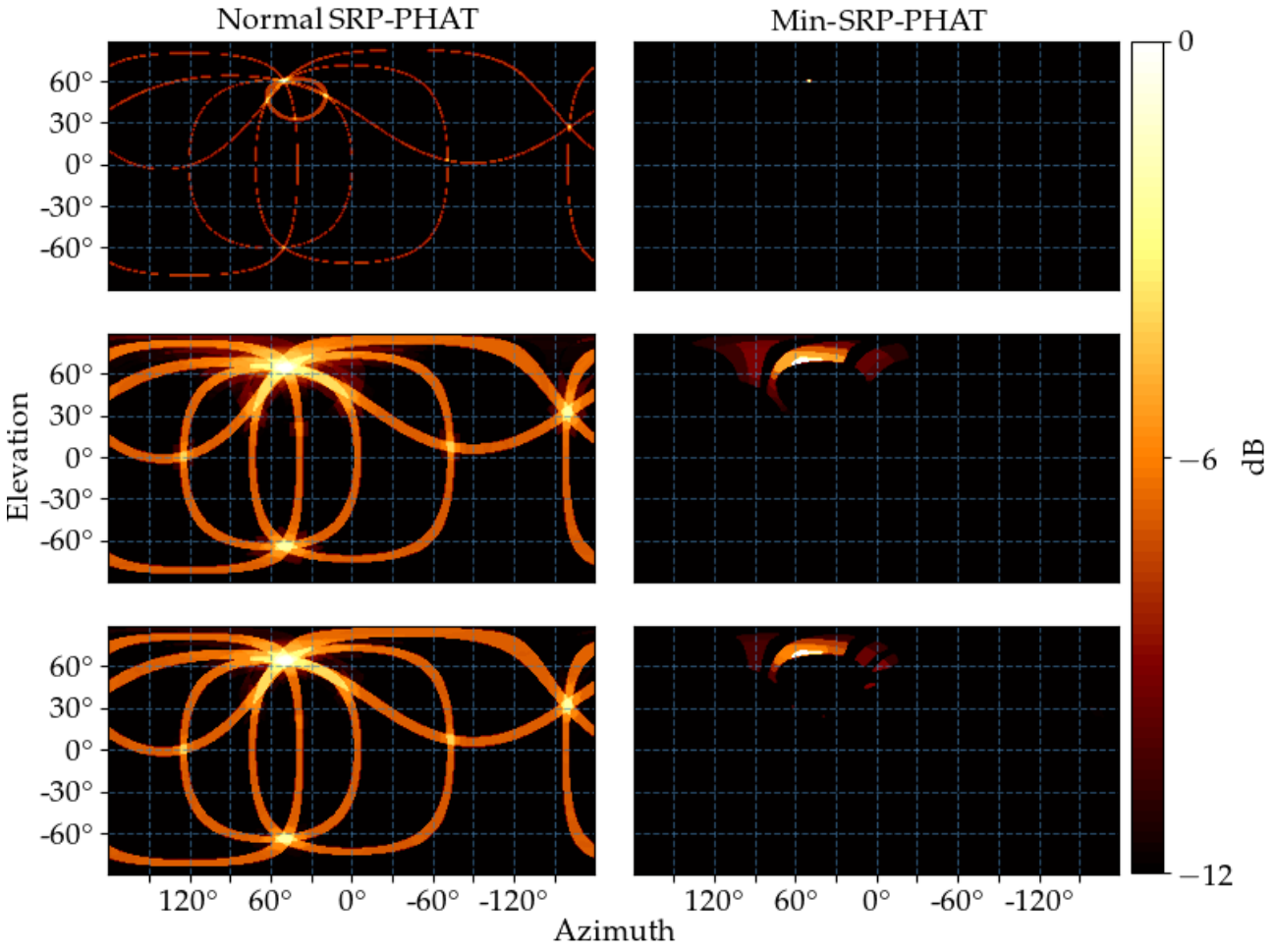


Figure 3.13: Figures depict from top-to-bottom SRP-PHAT localization results with tetrahedral microphone array aperture size of 1m@48kHz, 10cm@48khz and 1m@4.8kHz.

Effect of Audio Recording Length

Noise usually arrives from a random direction. If the measurements are then done for longer recordings, the effect of noise can be reduced. This can be seen as the noise averaging out over time. Thus, longer recordings are preferred when doing outdoor measurements. The obvious downside is that moving sources cannot then be localized, and transient sounds around the sound source with moving parts might also get averaged out. Fig. 3.16

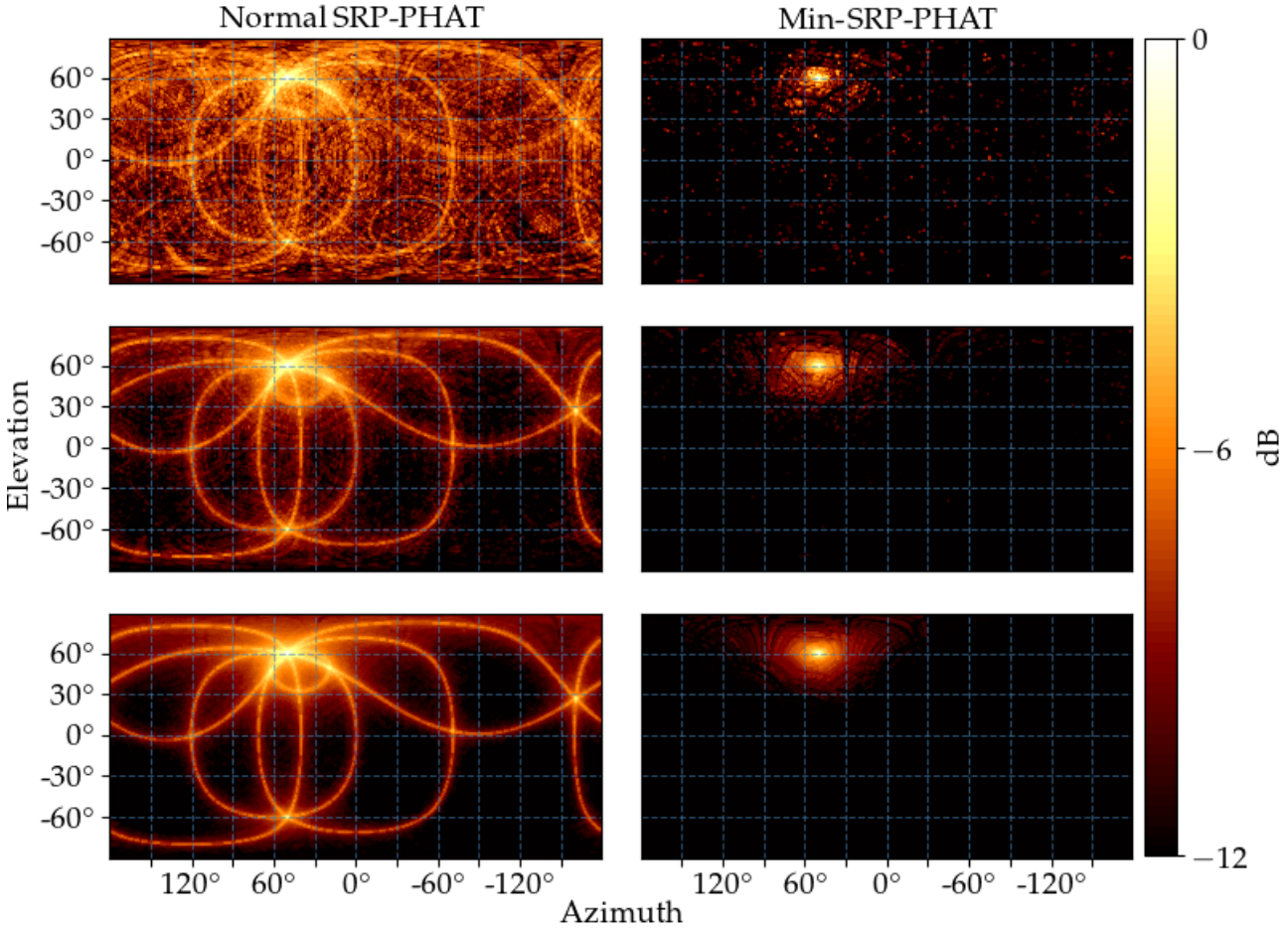


Figure 3.14: Figures depict from top-to-bottom SRP-PHAT localization results with tetrahedral microphone array with recording length of 1sec (top), 10sec (middle) and 100sec (bottom). The SNR here is kept at 0dB to highlight the differences.

Effect of Error in Microphone Position

The microphones in the array could have an error in position, due to structural errors (structural fatigue and sag, thermal expansion/ contraction or just human error). This could lead to an error in localization. Fig. 3.15 illustrates this effect.

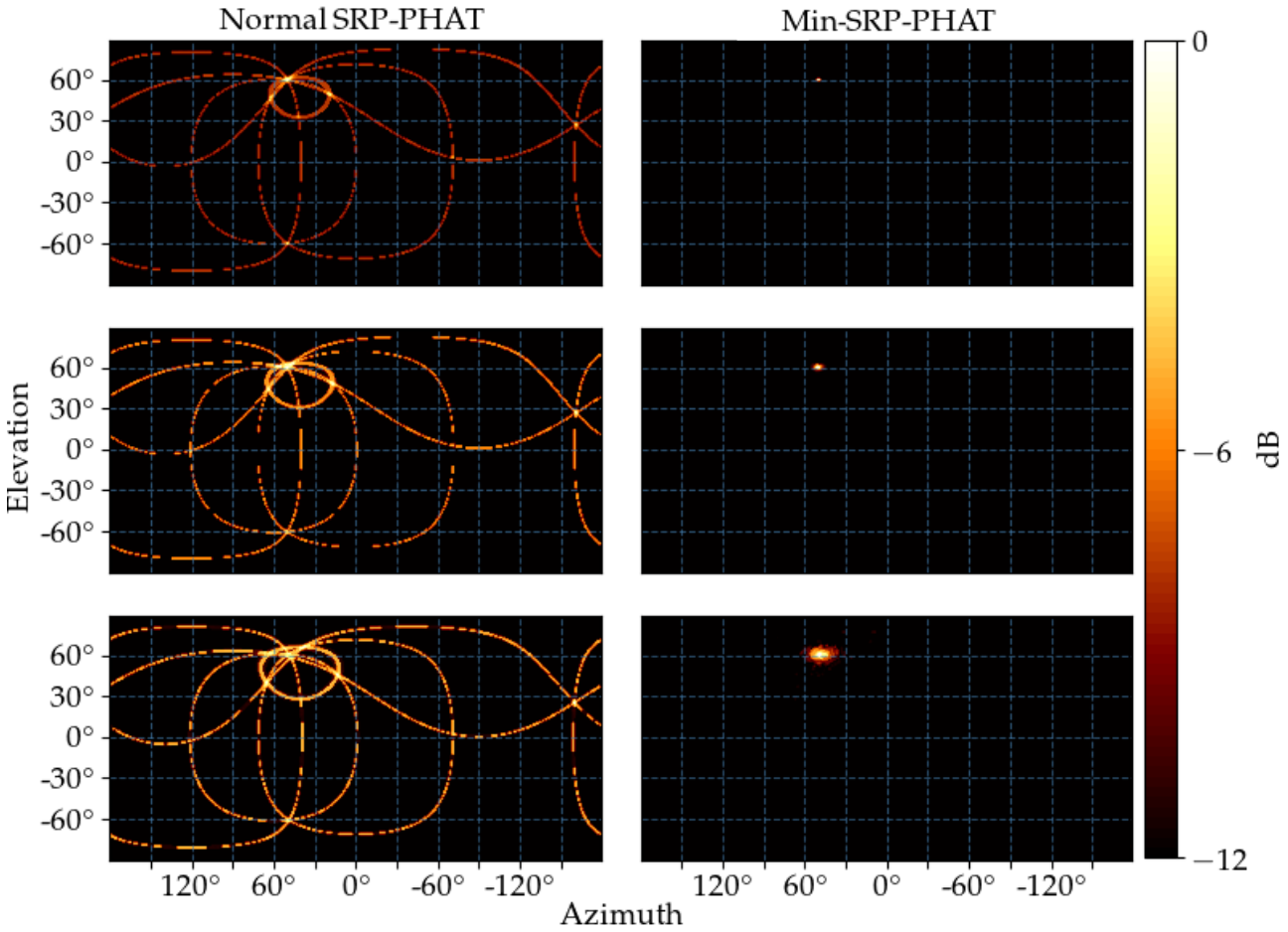


Figure 3.15: Figures depict from top-to-bottom SRP-PHAT localization results with tetrahedral microphone array with no mic error in placement (top), a 1cm placement error in 1 mic (middle), a 2cm placement error in all mics (bottom)

The error is similar the the effect of wind and temperature that we saw earlier. This is because the error in all of these cases is in recording the incorrect time delay between the different microphones.

Effect of Error in Array Tilt

Tilt is a special kind of microphone placement error. This is because an error in the array tilt can be seen as a movement of the source around the axes of the array. This means that an error in tilt does not effect the overlap of the localization circles. However, the localization circles move in such a way that they intersect in a new location. Fig. 3.16 illustrates this effect.

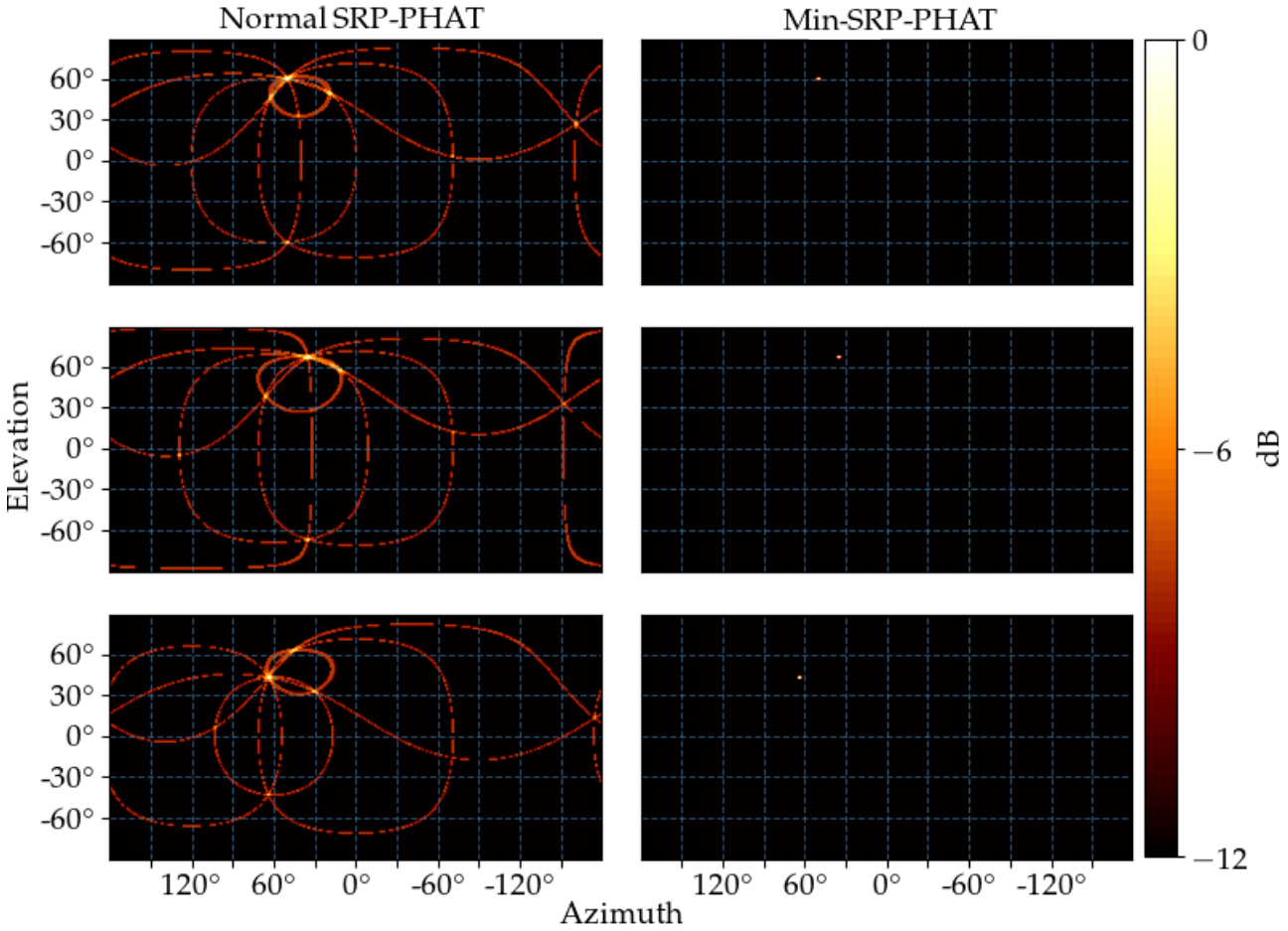


Figure 3.16: Figures depict from top-to-bottom SRP-PHAT localization results with tetrahedral microphone array with no mic tilt error (top), a +10° tilt error (middle) along the x-axis and a -20° tilt error along the x-axis (bottom)

This concludes the simulation section of the thesis. The next chapter will provide the results of some real world measurements to validate the algorithm and test its performance.

Chapter 4

Experimental Evaluation

The simulations conducted in the previous chapter form the basis for real world measurements and experimental evaluation of the Min-SRP-PHAT algorithm, to test its performance and robustness in actual outdoor conditions. To validate the algorithm under ideal, controlled conditions, anechoic experiments were conducted first. After that a series of different outdoor sources and environments were tested. This chapter contains the results of the experiments.

4.1 Microphone Array and Acquisition System

The equipment used for the experiments are listed below.

Item #	Description
	General
1	4 B&K Type 4935 microphones
2	1 B&K Module Type 3050-A-060 interface module
3	Prototype tetrahedral microphone stand
4	PC with MATLAB, Python and B&K Pulse software
5	Relevant cables and wires
	Anechoic Measurements
6	1 B&K OmniPower loudspeaker
7	2 custom spherical speakers
8	1 Pioneer A-656 amplifier
9	1 B&K Type 2270 hand-held analyzer
	Outdoor Measurements
10	1 B&K Module Type 2831 battery module
11	4 B&K microphone ellipsoidal windscreens

A prototype microphone array structure was used to record sound sources, on which four B&K Type 4935 microphones could be placed in a tetrahedral configuration (Fig. 4.1). The four microphones were mounted on the array vertically (pointing upwards). The microphones could be placed between 10cm-1m from each other in discrete 10cm steps. The middle vertical rod of the structure was placed on a tripod. The height of the array could be adjusted by moving the middle rod up-down, and

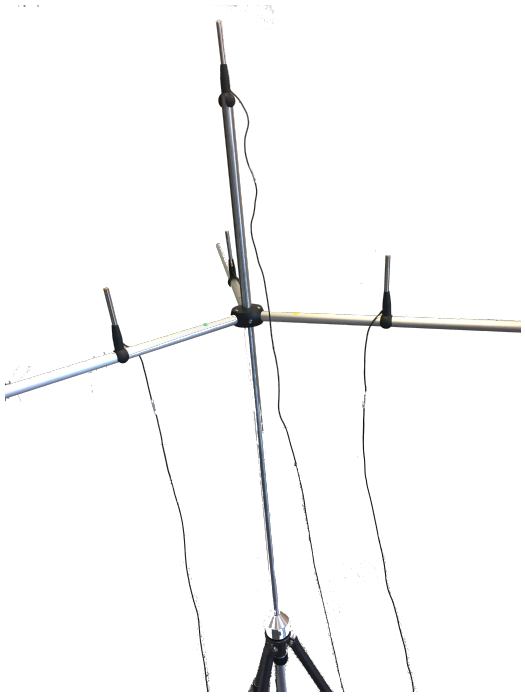


Figure 4.1: Prototype tetrahedral array used for measurements

also by adjusting the tripod height. During the measurements, the array was kept horizontally, such that three of the microphones were on the same horizontal plane. For the outdoor measurements, the array was kept as high as possible, such that the base of the array was $\approx 1.5m$ above the ground. B&K Pulse software suite was used for recording. B&K Module Type 3050-A-060 was used as the main interface sound card and microphones were plugged into it with BNC connectors. During outside measurements, foam microphone windscreens were mounted on the microphones and a B&K battery module was used to power the soundcard. The recordings were converted to 16bit .wav files, recorded with a sampling rate of 131072Hz (the maximum sampling rate available on the system). Two different values of the tetrahedral array aperture were used, 1m and 39.5cm, resulting in two different sizes for the tetrahedral array, large and small. For the large array the achievable angular resolution is between 0.15° - 4.15° whereas fpr the small array the resolution is between 0.38° - 6.6°[1]. Most of the results are provided keeping a 1° resolution for the SRP-search. In the outdoor results, zoomed-in overlaid photos are used to construct the sound map. The resolution is kept at 0.1° in those cases.

4.2 Anechoic Measurements

The purpose of the anechoic measurements is to validate the Min-SRP-PHAT algorithm in a controlled environment. The criteria that are required to be validated are,

- For a single source, the location is computed correctly.
- For multiple sources playing simultaneously, the locations are computed correctly.

[1]The achievable angular resolution are given as a range as for two microphones the angular resolution depends on the location of the source. However, for a particular location, the angular resolution depends on all six microphone pairs for a tetrahedral array. This means that certain locations can have some localization circles which are 'fatty' (annular). However, not all localization circles will be annular, leading to a range on the angular resolution.

- The relative levels of multiple sources are maintained.
- The absolute levels of multiple sources are computed correctly.

4.2.1 Experiment 1: Localizing a Single Sound Source

Pink noise played from a single loudspeaker was recorded with the array[2]. The source was a B&K Omnisource Type 4296 speaker with operating frequency of 100Hz-5kHz. Two separate measurements were done. For the first measurement, the loudspeaker was placed at (90°, 0°). For the second measurement, the loudspeaker position was kept the same and the microphone array was rotated ≈20° around its axis in order to change the relative location (azimuth) of the speaker. The setup is described by Fig. 4.2 where the source positions for the two measurements are shown. In order to approximate plane wave propagation, in the limited space of the anechoic chamber, the array was placed as far away from the source as possible and the array aperture was reduced to 0.395m. Temperature of the lab was 22 °C.

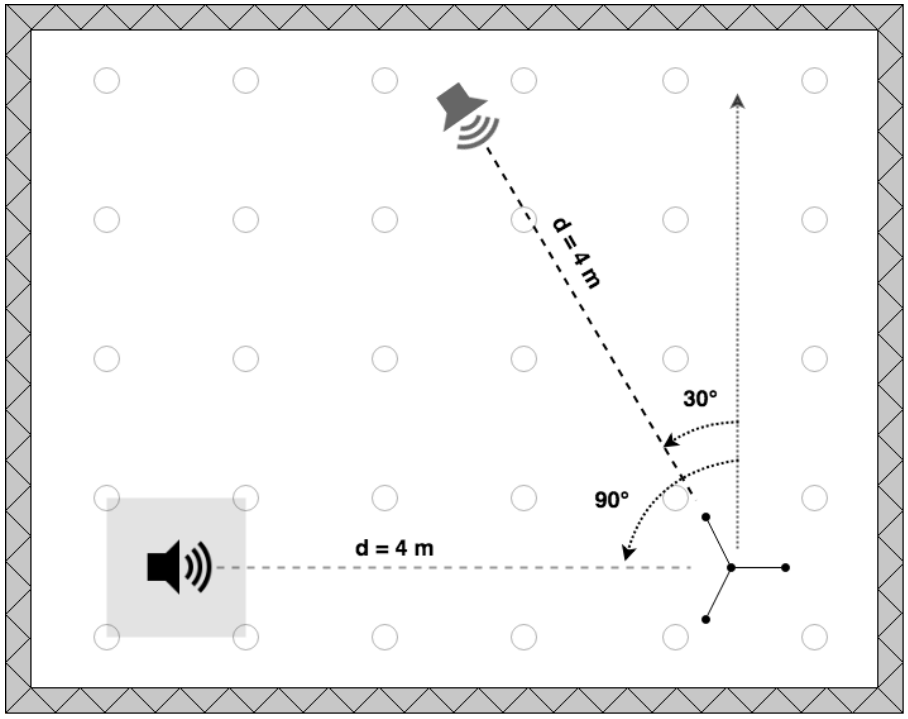


Figure 4.2: Sketch of the experiment.

Results

The energy maps of the SRP-PHAT and Min-SRP-PHAT algorithms are computed and displayed in Fig. 4.3. The source located at (90°, 0°) was localized at (89°, 0°) by both SRP-PHAT and Min-SRP-PHAT. The error of 1° in localization of the speaker is attributed to errors in loudspeaker placement. The source placed at (20 °, 0°) was localized at (23°, 0°) by both SRP-PHAT and Min-SRP-PHAT. The azimuth error in localization of the speaker is again attributed to errors in placement.

[2] A recording of 300Hz sinusoidal wave was also performed to display the delays between the microphones (Appendix. B)

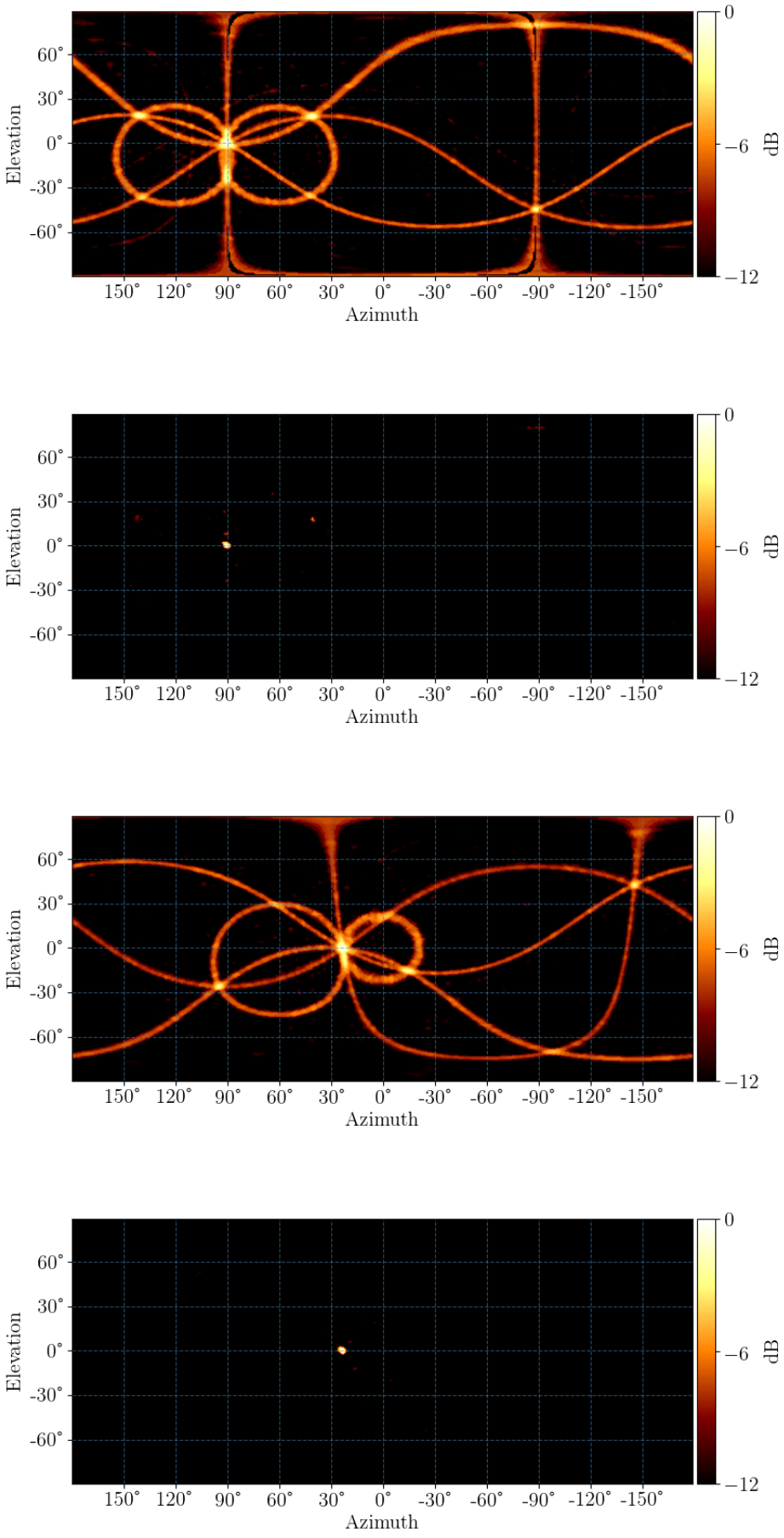


Figure 4.3: Figures depict SRP-PHAT and Min-SRP-PHAT localization for source around (90°, 0°) in an anechoic room (top two), and for source at (≈20°, 0°) in an anechoic room (bottom two).

4.2.2 Experiment 2: Localizing two Sources and Computing their Levels

This experiment was run in the anechoic chamber to validate multiple source localization and retrieval of their relative sound level difference, as well as their absolute levels. Two custom spherical speakers were used to play the source signal and the aperture size of the array was set at 0.395m. One source was placed at (90°, 0°) and played pink noise at 46dB[3], the second source was placed at (130°, 0°) played uncorrelated pink noise at 52dB. The low frequencies (<200Hz) were filtered out in order to accommodate for the speakers used. Temperature of the room was recorded at 22°C. The setup is described in Fig. 4.4 and Fig. 4.5. Fig. 4.6 shows the results.

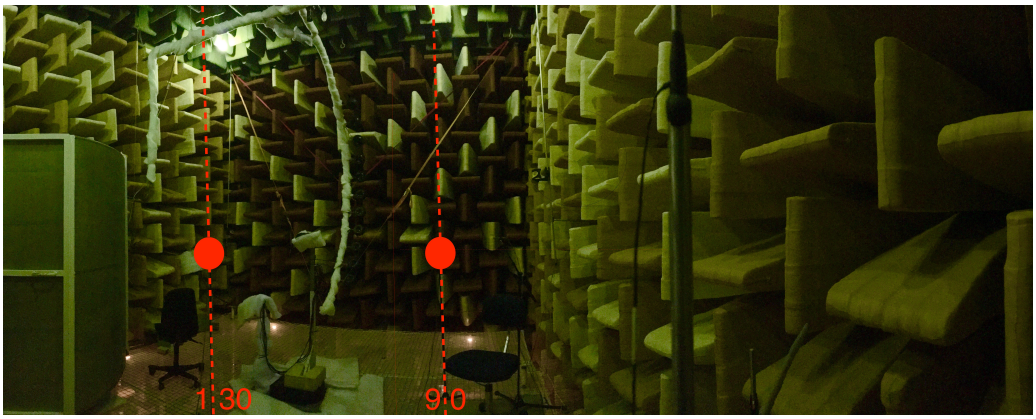


Figure 4.4: Picture of the set up. The anechoic chamber was filled with misc. equipment, therefore the sources have been replaced by red dots for clarity. 90° and 130° azimuth are also drawn on top of the picture.

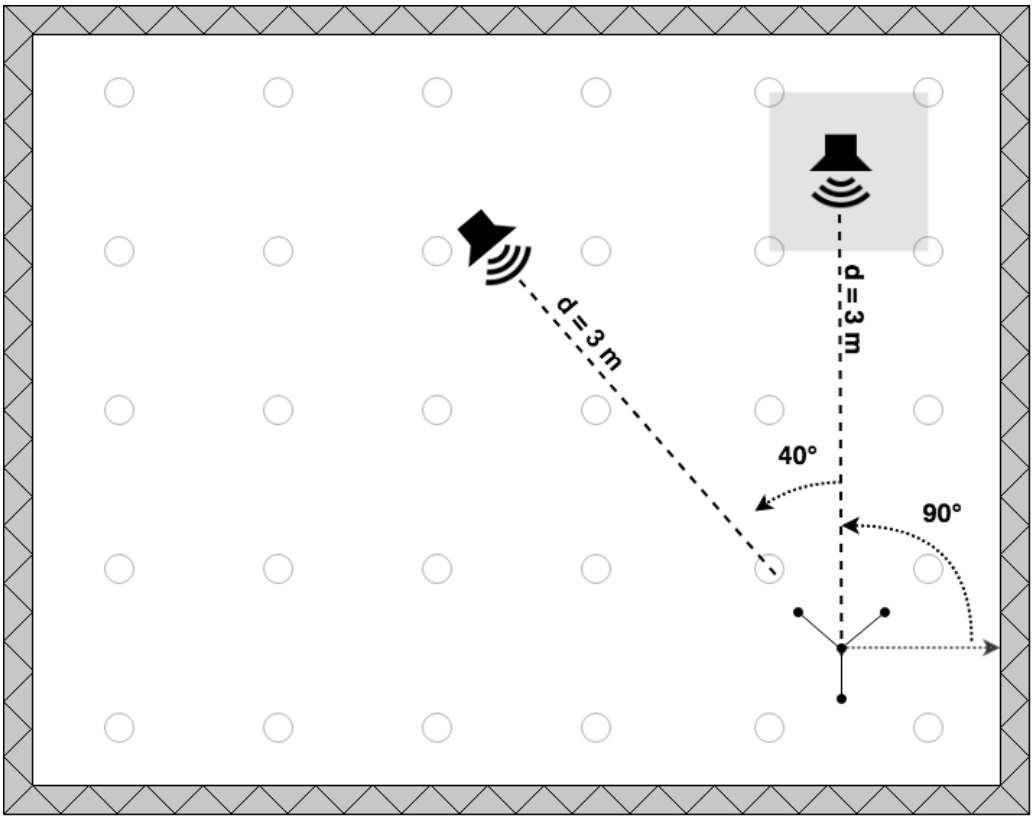


Figure 4.5: Sketch of the experiment

[3]The sound from each speaker was measured individually, using a level meter, at the array location.

Results

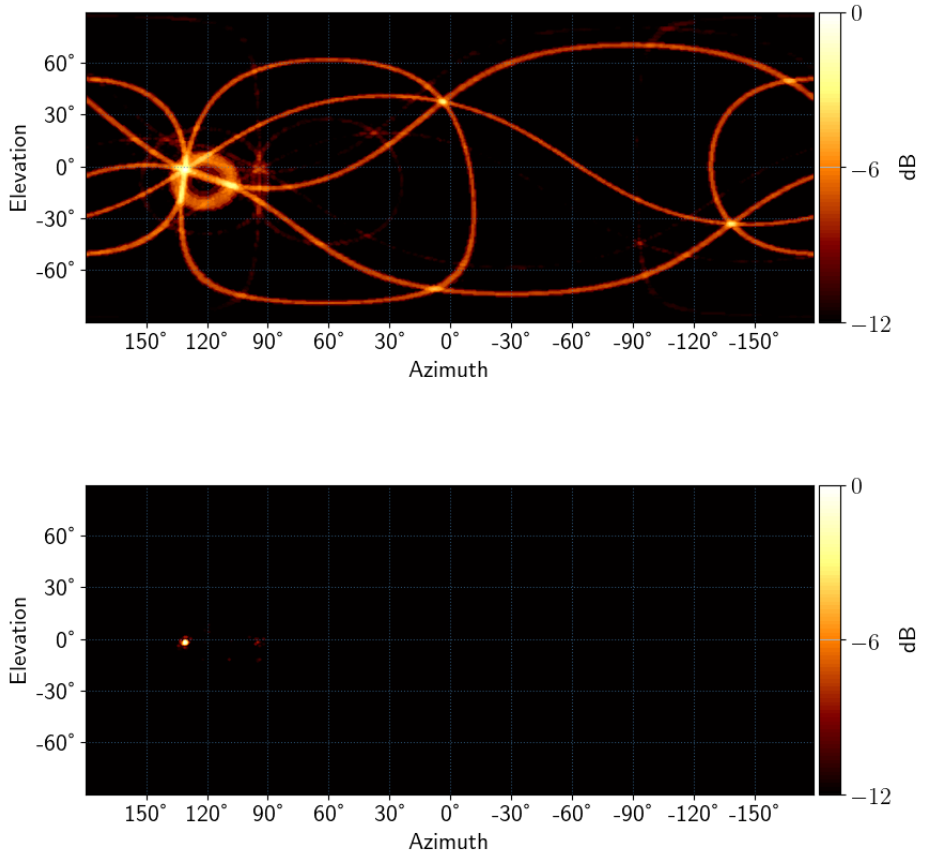


Figure 4.6: Figure depicts SRP-PHAT (top) and minimum power SRP-PHAT (bottom) localization for 2 sources located at (90°, 0°) and (130°, 0°). The sources play uncorrelated pink noise at 52dB and 46dB respectively.

During the experiment, the anechoic chamber was not completely empty and some reflections can be observed at the 12dB dynamic range. Also, since the speakers were relatively close to the array, the cone approximation has larger errors. This can reduce the size of overlap of the multiple cones from the various microphones, or cause them to not overlap at all. This can be seen in the result for SRP-PHAT here, where the cones for the secondary cones barely overlap. Applying Min-SRP-PHAT can then completely hide the secondary source. For this measurement however, the secondary source can be seen. The results from SRP-PHAT showed the secondary source level to be playing at -6.81dB relative to the main source. Understandably, due to the issues discussed here, the results for the Min-SRP-PHAT showed the secondary source level to be -8.18dB relative to the main source. Both SRP-PHAT and Min-SRP-PHAT localized the peak of the main sound source at the same location (130°,-1°). The peak of the secondary source was localized at (95°,-2°) by SRP-PHAT and at (94°,-1°) by Min-SRP-PHAT. This minor difference can be attributed to the fact that the peak at (95°,-2°) was removed by the minimum power algorithm, due to no overlap at that location. The total power received at the array for the delay associated with the peak location was calculated according to Section 3.2.1. It was computed to be 51.57dB. The results were deemed satisfactory and from here on absolute source levels will be shown in the results.

4.3 Outdoor Measurements

In order to test the algorithm in real conditions, several outdoor measurements in various conditions were conducted. Measurements were done in a construction field, in outdoor and indoor concerts, in traffic and in a chalk mining field. The recordings were done at the maximum sample rate available in the system, i.e, 131072Hz. The microphone array aperture unless otherwise mentioned, was set at $1m$, since the sources were always sufficiently far away for plane wave approximation to hold. Photos of the measured outdoor environment were taken using a camera having a known angle of view, such that the azimuth and elevation of the photos was known. Picture overlay is explained in App. E.1 Finally, the localization results were overlaid on the photos.

4.3.1 Single Static Source on a Construction Field

In this experiment, a single construction machine working in a fixed position was recorded by the microphone array. The source was more than 20 meters away. Fig. 4.7 describes the setup. The measurement system was set in the middle of a road, outside the construction field. There was a big office building with a smooth fać ade behind the setup. Temperature was recorded at 23°C, speed of wind was 2m/s from (180°, 0°). Fig. 4.8 displays the full results.

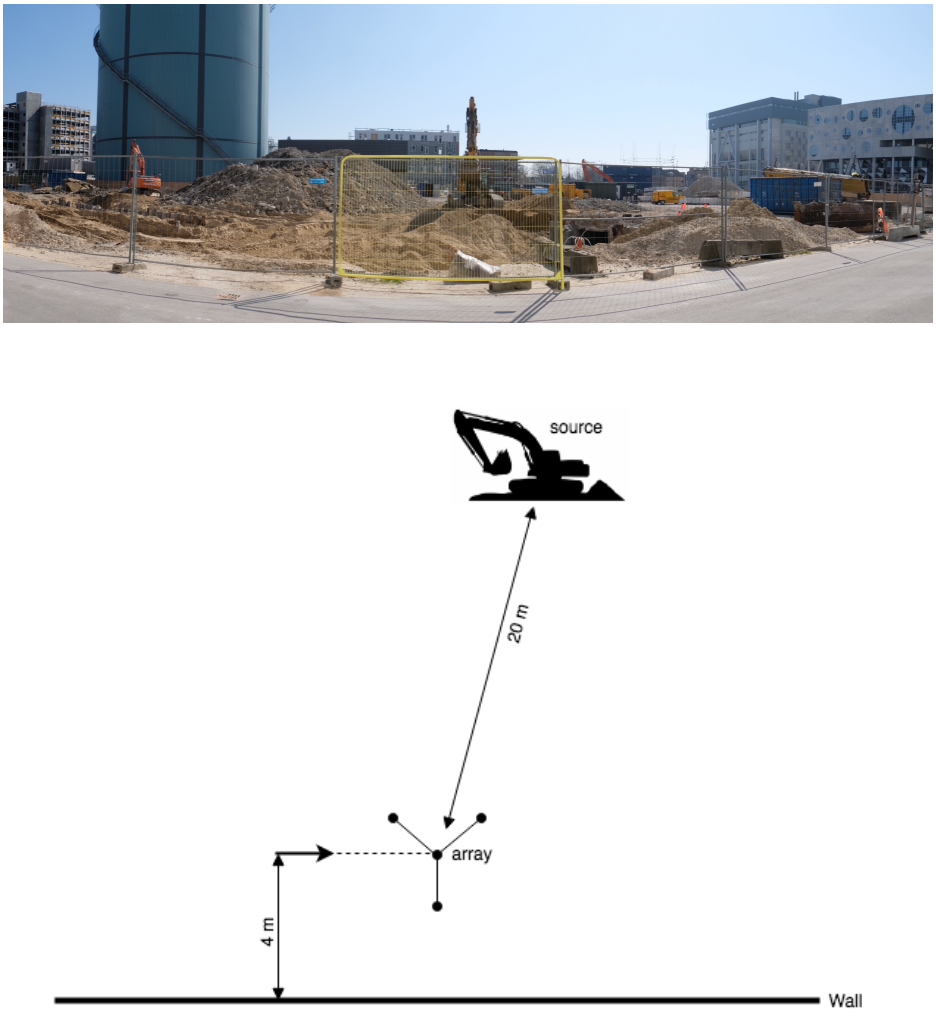


Figure 4.7: Figure shows the picture of the construction field (top) and the top view schematic of the construction field (bottom)

Results

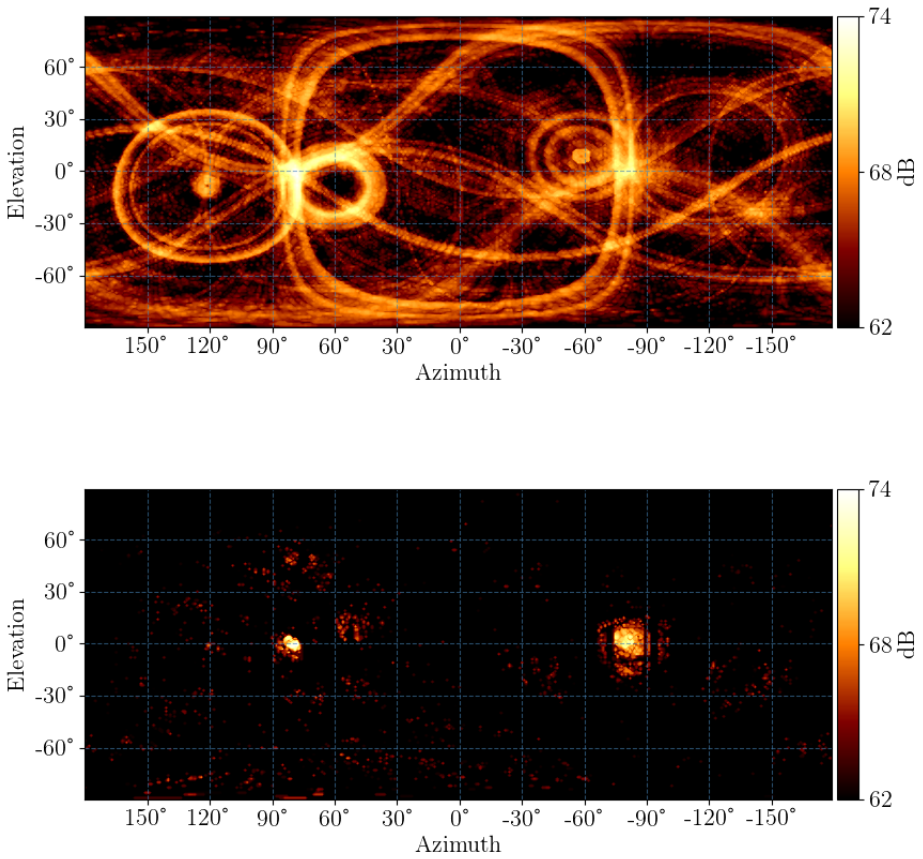


Figure 4.8: Figures depict from top-to-bottom SRP-PHAT and minimum power SRP-PHAT localization for a single construction machine working in a construction field.

As can be seen in the Min-SRP-PHAT result, two clear sources are visible at ≈(±90°, 0°). The source at (90°, 0°) is the actual construction machine and the one at (-90°, 0°) can be attributed to its reflection from the office wall behind the array. The recordings were only 51sec long, the duration during which the machine was in the same location. The overlaid results are shown in Fig. 4.9. The overlay is such that straight infront of the microphone array is the 90° azimuth, with 0° corresponding to the complete right and 180° being the complete left. Negative azimuths signify the back of the array. For elevation 0° is the plane of the base of three microphones of the microphone array, with 90° being straight up and -90° being straight down.

During the measurement, even though the machine itself did not move, the excavator arm moved around the body of the machine emitting noise as it excavated. These sounds appear on the map for the 12dB dynamic range, however they appear at a low magnitude as they are transient and get averaged out. As can be seen the result is fairly noisy. It is indeed difficult to get a clean map for a larger dynamic range in this reverberant field, as more and more reflections from the source become apparent. Other distant sources in the construction field and their own reflections also appear on the results. In addition, other noise sources such as the wind noise or other diffuse reflections might also be appearing on the results, since some results can be seen with relatively high elevation, where no source was present. Taking longer recordings can help reduce these transient sounds from appearing on the results and improve achievable dynamic range.

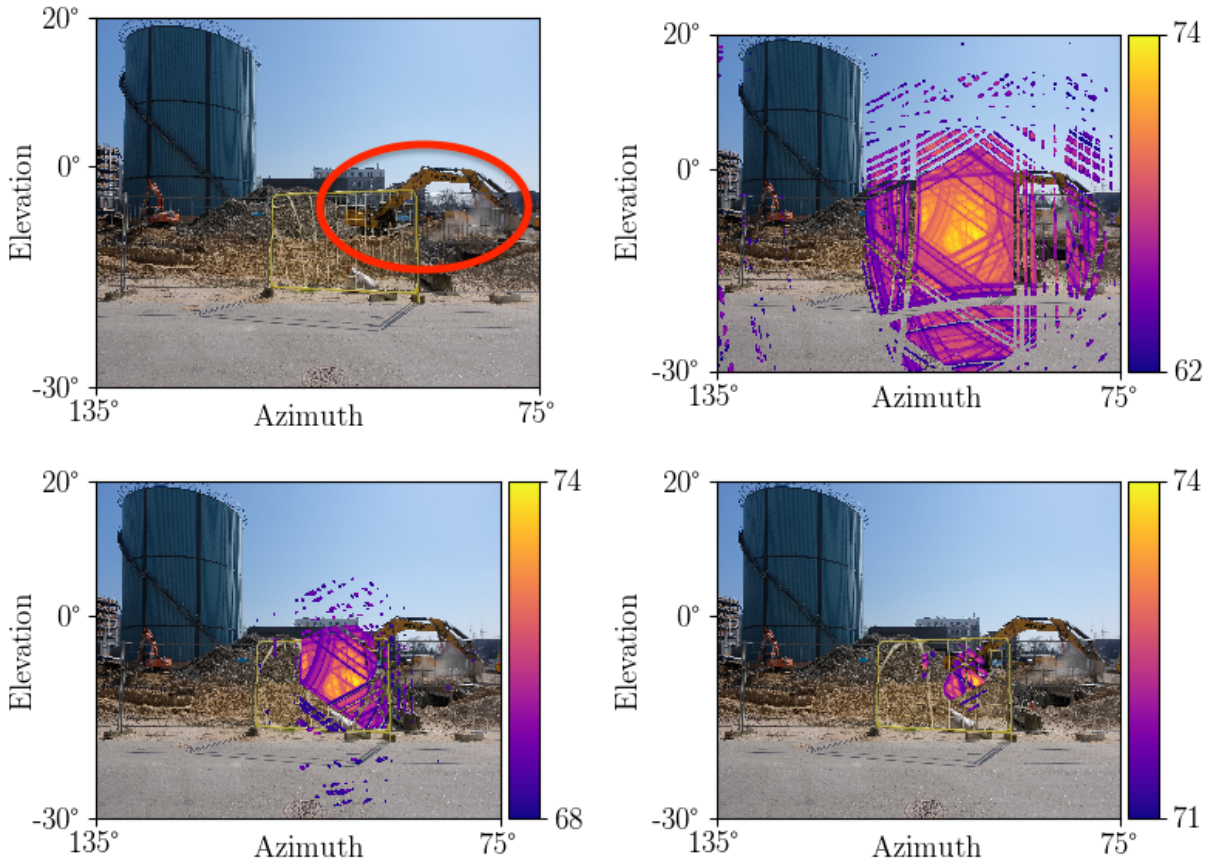


Figure 4.9: Figures depict localization results overlaid on the photo of the measured source with different dynamic ranges (dB). A single construction machine is measured here (top-left) for the duration of 51sec.

4.3.2 Three Static Sources on a Construction Field

In this experiment, 3 distinct noise sources were measured: a tapping machine in a hole in the ground located at (180°, <0°) and two excavators located between (120°, 0°) and (150°, 0°). Fig. 4.10 describes the setup. Fig. 4.11 displays the full results. The overlaid results are shown in Fig. 4.12. As can be seen in the overlaid figure, the source results look elongated. This might be due to the fact that the tapping machine was causing the ground to shake. When doing far field localization, even a small movement in the microphone array can cause large deviations in the localization results. Also, the authors noted that even though the tapping machine was louder, it appears lower on the localization results. This is because of the fact that the sound from the tapping machine was periodic and not constant.

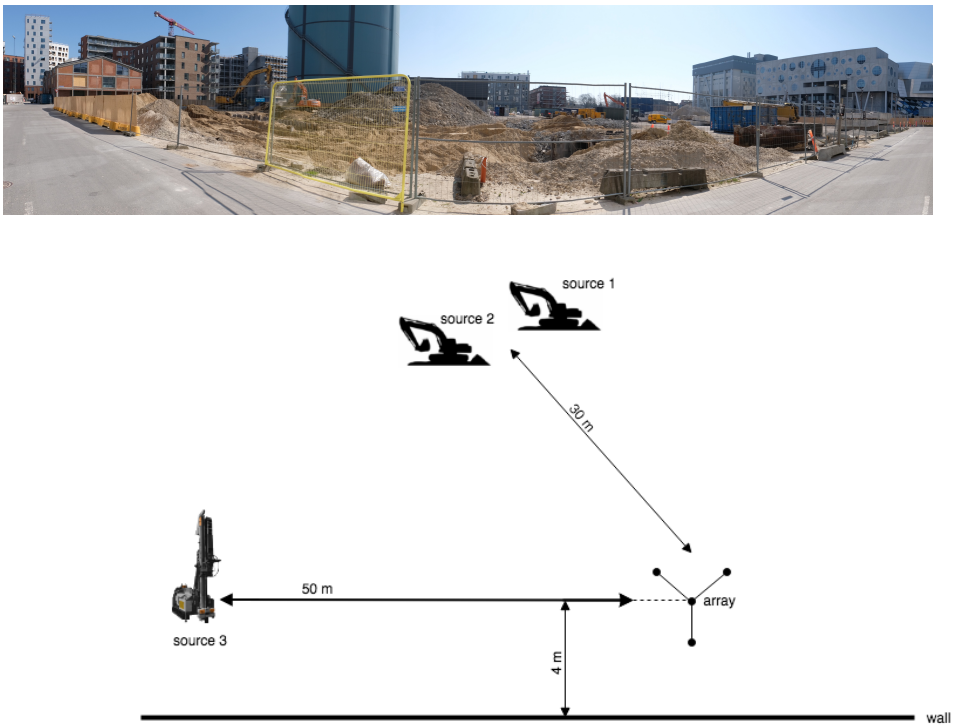


Figure 4.10: Panorama of the construction field at the center point of the microphone array (top) and Top view schematic of the construction field (bottom)

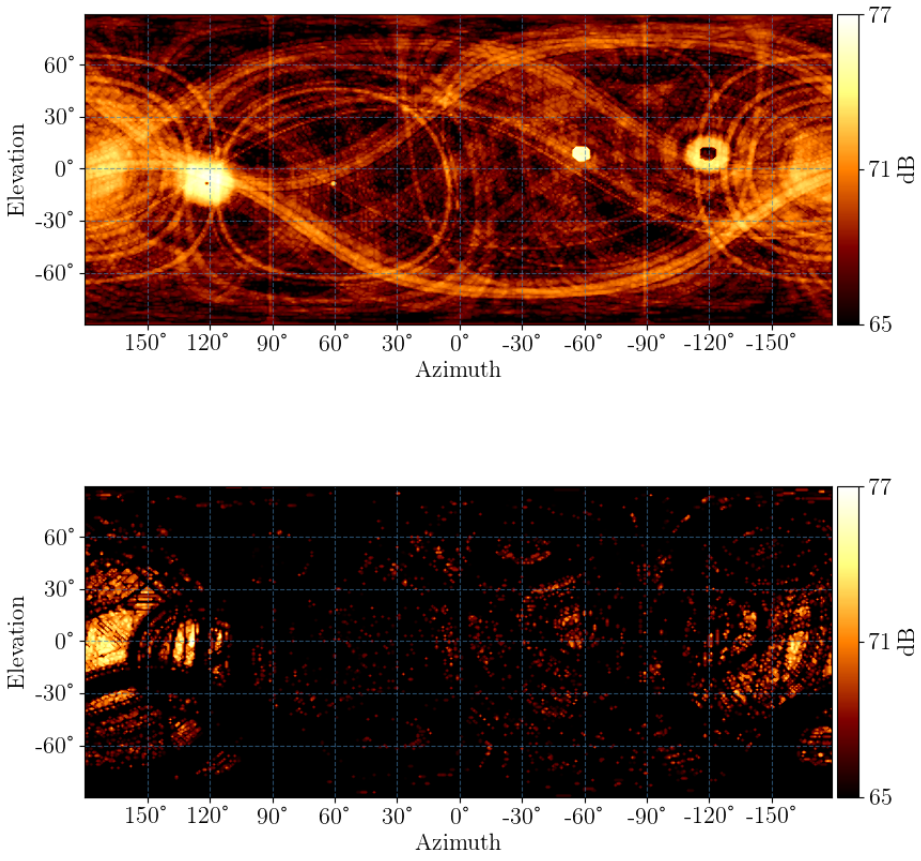


Figure 4.11: Figures depict from SRP-PHAT (top) and Min-SRP-PHAT (bottom) localization for two construction machines located between (120°,0°) and (150°,0°) working in a construction field. A tapping machine is also making sound at (180°,<0°).

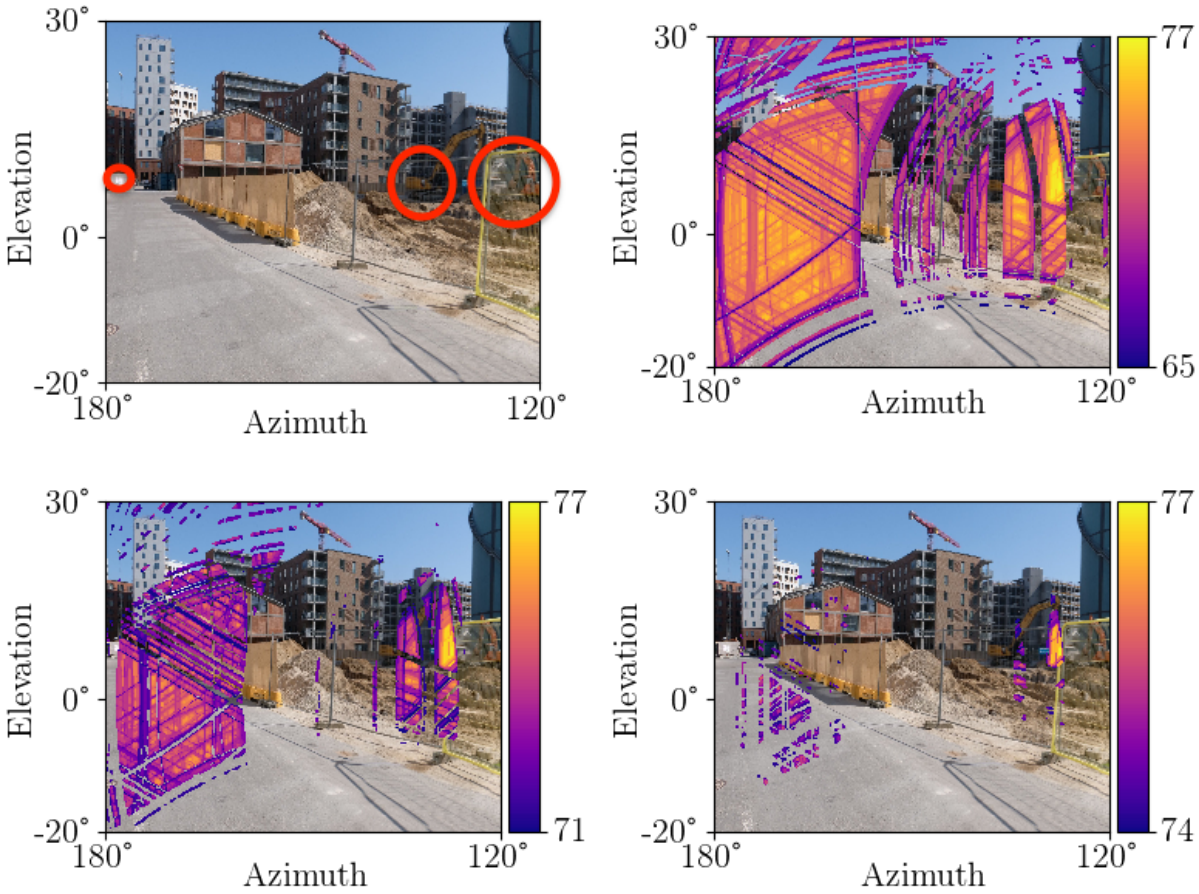


Figure 4.12: Figures depict localization results overlaid on the photo of the measured source with different dynamic ranges (dB). Three construction machines are measured here (top-left) for the duration of 51sec again.

4.3.3 Sport Event with Crowd and PA System

Measurements were performed during a sport competition outside, where two main kinds of sound sources were present. The first one was a distributed PA system as shown in Fig. 4.14, the second one was a crowd at (0°, 0°), however the crowd level was quite low compared to the music level. Fig. 4.15 displays the full results. The overlaid results are shown in Fig. 4.16.

Figure 4.13: Panorama from the center point of the microphone array

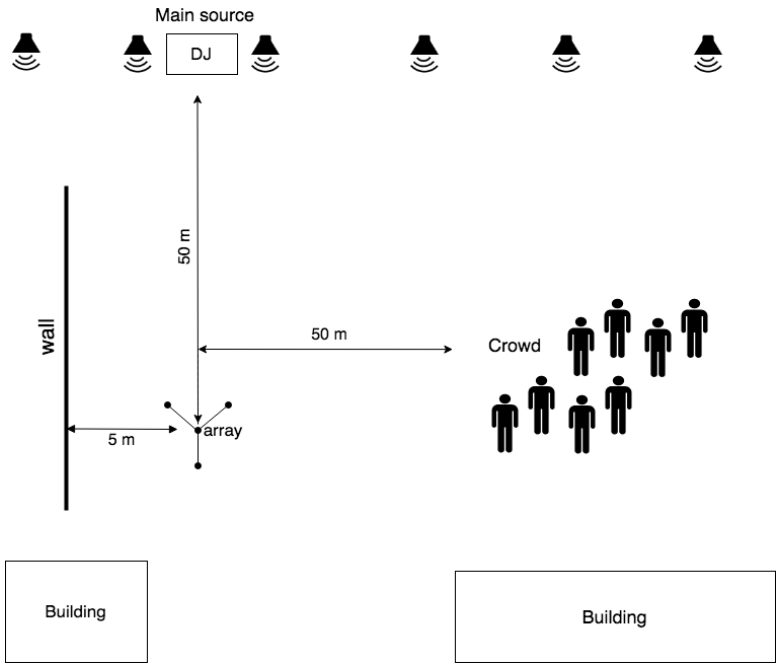


Figure 4.14: Top view of the scenario

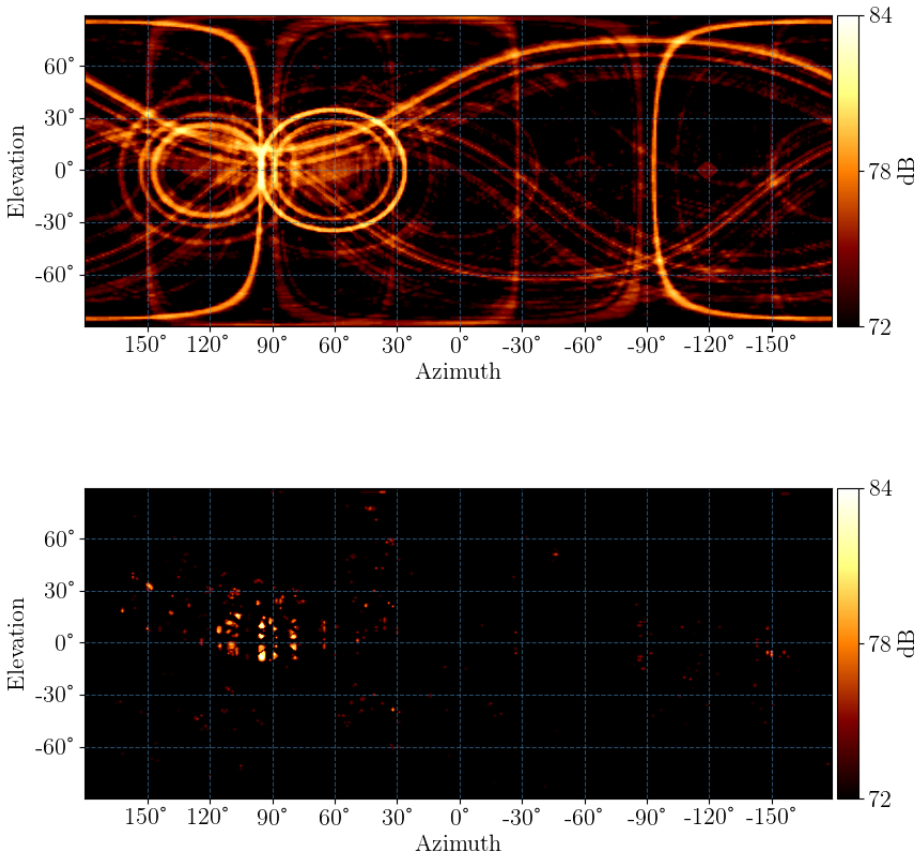


Figure 4.15: Figures depict SRP-PHAT (top) and Min-SRP-PHAT (bottom) localization for a distributed PA system playing outdoors.

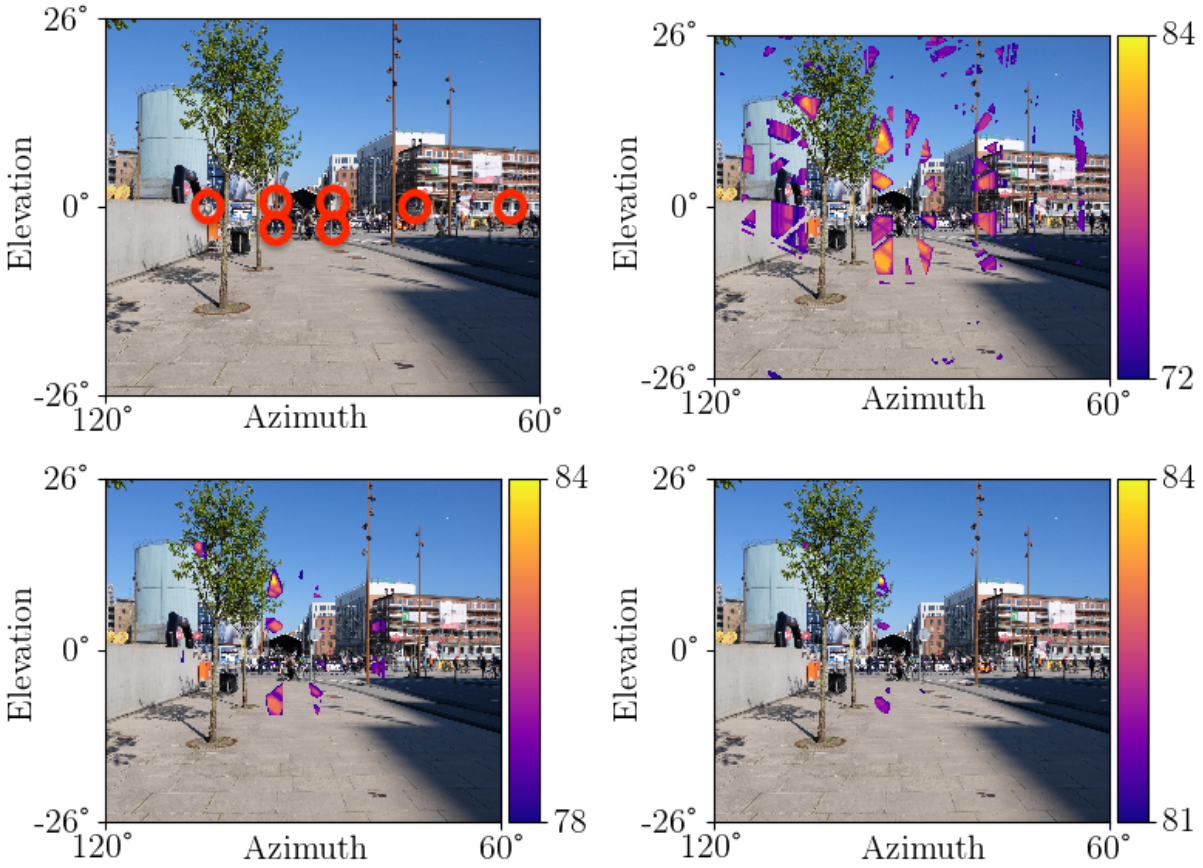


Figure 4.16: Figures depict localization results overlaid on the photo of the measured source with different dynamic ranges (dB). A distributed PA system is measured here (top-left) for the duration of 60sec. The sources were detected incorrectly, with the peak sound source picked up at around 10° elevation. Coherence between the sources is attributed to this error.

As can be seen the results are incorrect. This measurement led the authors to investigate the effect of coherent sources on localization results, the simulations for which have been done before (In Fig. 3.12, the bottom two plots are made with the source locations corresponding to the two main source locations from this scenario, and a similar pattern can be observed in the simulation as well.). It was found that when multiple coherent sources are present, the algorithm can fail as the constant phase difference between the multiple sources can be detected as a pseudo-source.

4.3.4 Outdoor Concert

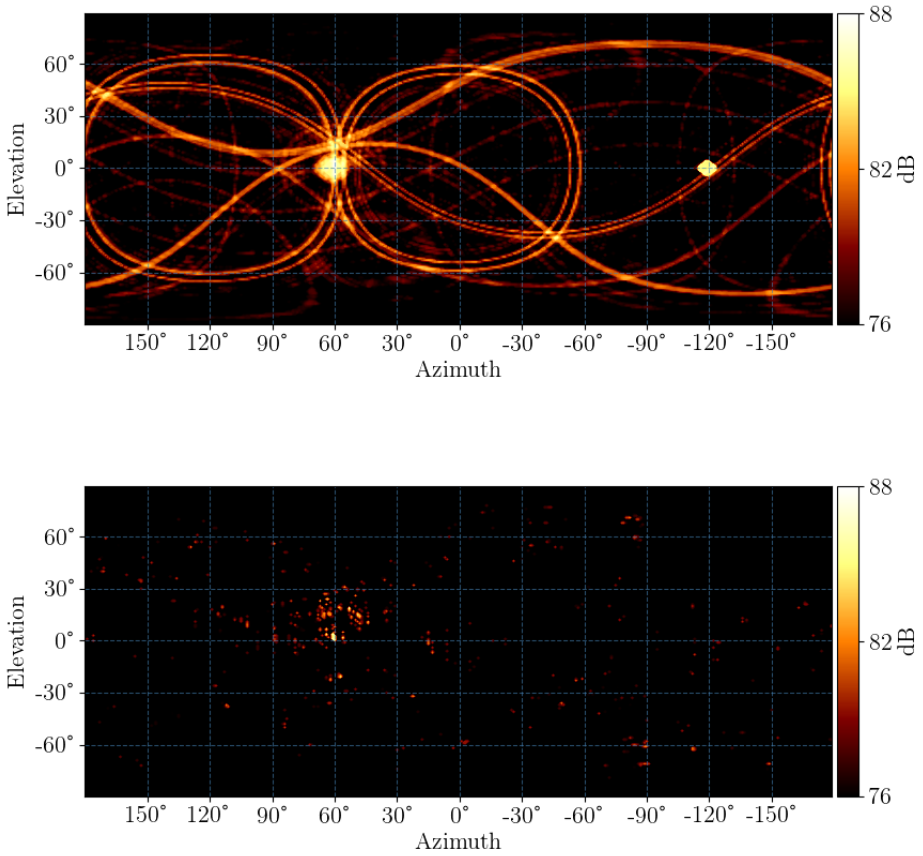


Figure 4.17: Figures depict from SRP-PHAT (top) and Min-SRP-PHAT (bottom) localization for a DJ playing music on a 2.1 channel loudspeaker system. The effect of coherence is visible here as well.

Measurements were performed during an outdoor concert where a DJ was playing music on a PA system. The PA system was composed of two tops and one sub. Fig. 4.17 displays the full results. The overlaid results are shown in Fig. 4.18. The effect of coherence is visible here as well. It is determined that localization of coherent sources is not possible with either SRP-PHAT or Min-SRP-PHAT.

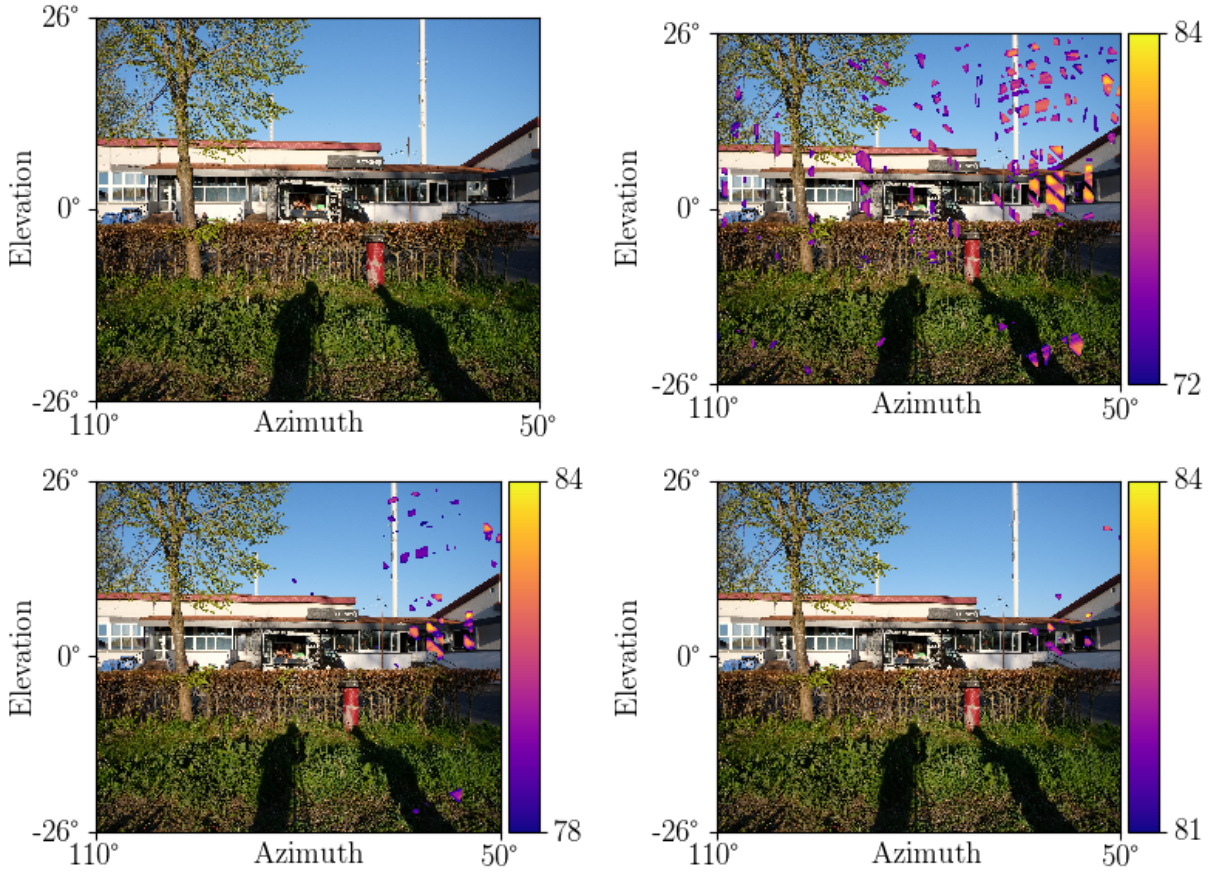


Figure 4.18: Figures depict localization results overlaid on the photo of the measured source with different dynamic ranges (dB). A 2.1 channel PA system is measured here (top-left) for the duration of 60sec. The sources were again detected incorrectly, with the peak sound source picked up at around (54°,6°). The error is again attributed to coherence between the sound sources.

4.3.5 Indoor Concert

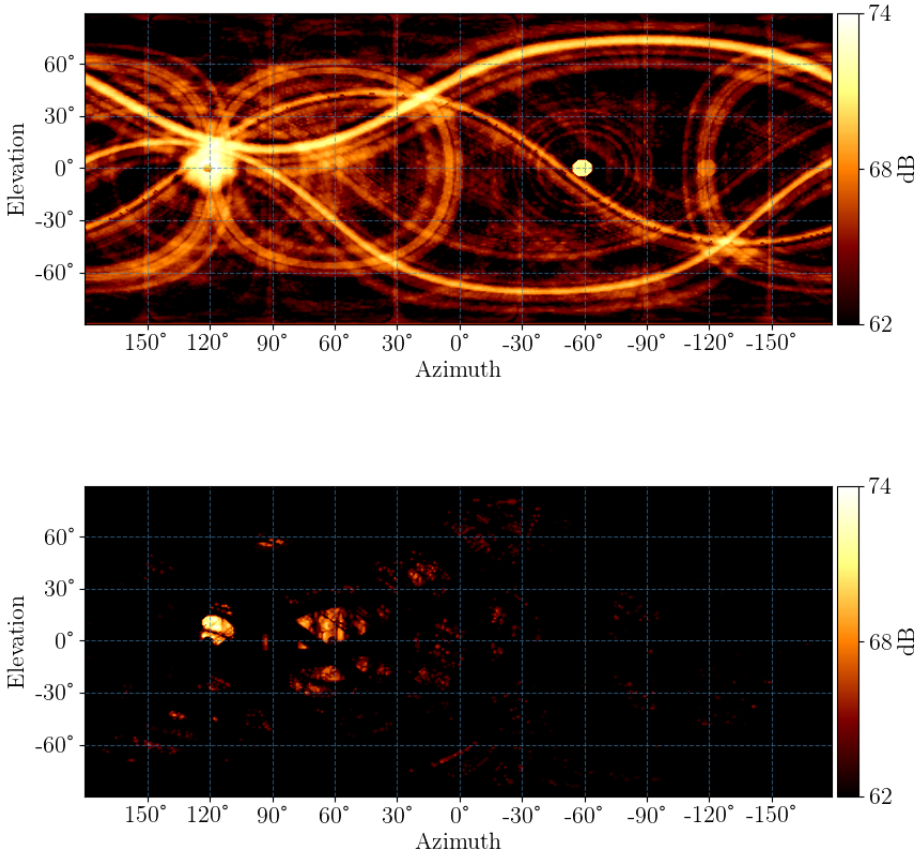


Figure 4.19: Figures depict from SRP-PHAT (top) and Min-SRP-PHAT (bottom) localization for an indoor concert. The measurements were made outdoors so as to test the acoustic insulation of the building.

Outdoor measurements were made of a concert happening indoors. This scenario is interesting as the it is often required to measure the sound insulation of such buildings, so as to determine any leaks. Two main sound sources were present. The door of the building at around (90°, 0°), and a crowd at around (120°, 5°). The event took place at midnight, however the photo was taken during the day so that the overlay is easier to see. Fig. 4.19 displays the full results. The overlaid results are shown in Fig. 4.20 and 4.21.

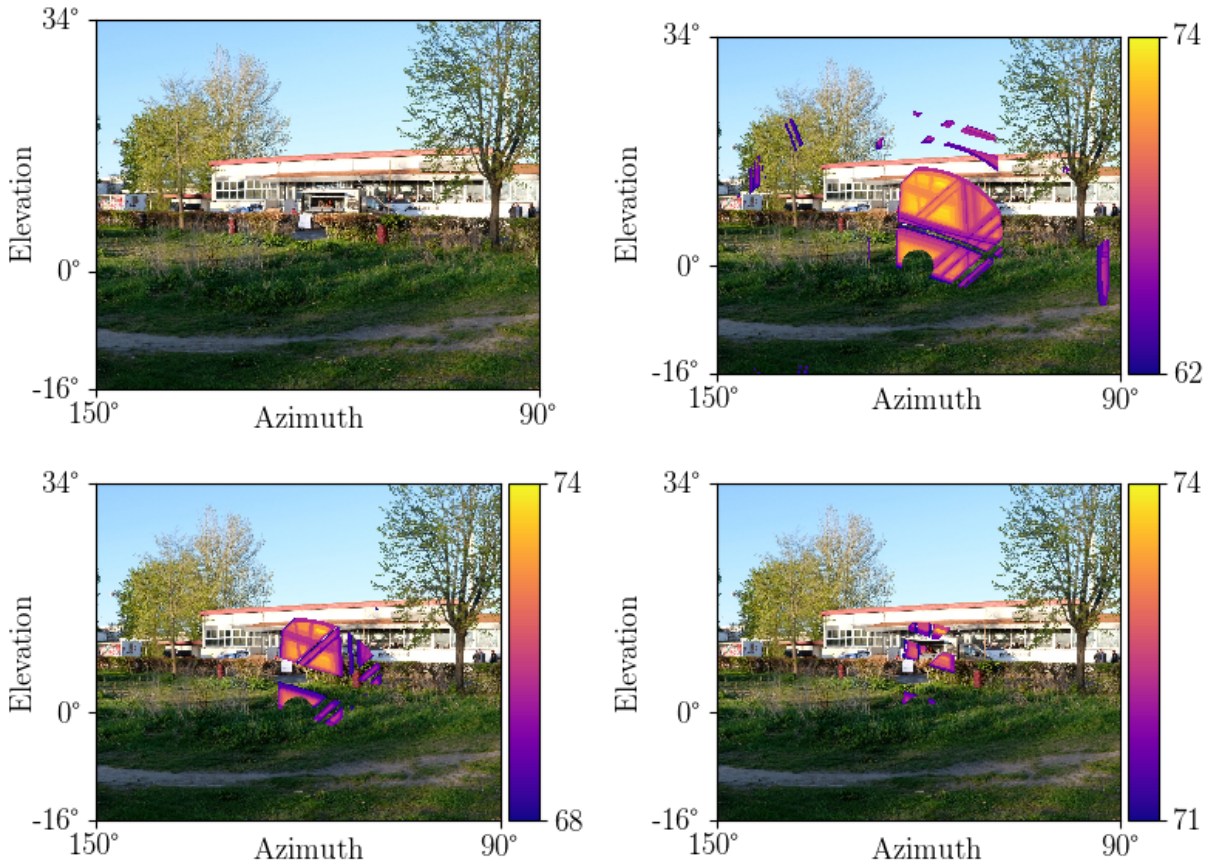


Figure 4.20: Figures depict localization results overlaid on the photo of the measured source with different dynamic ranges (dB). Due to the sound from the crowd being louder than the sound from the concert, the results for the concert are drowned out. Here the results for the crowd are overlaid.

Frequency filtering was applied in Fig. 4.21 to make it possible to separate the two sources which have inherently different frequency spectra and are also playing at very different levels. Direct detection is not possible because the lower magnitude source (the indoor concert) is 20dB below the higher magnitude source (the crowd). As can be seen in the filtered results, more investigations need to be done on the effect of such frequency filtering and the accuracy of the localization as well as the levels detected in this manner.

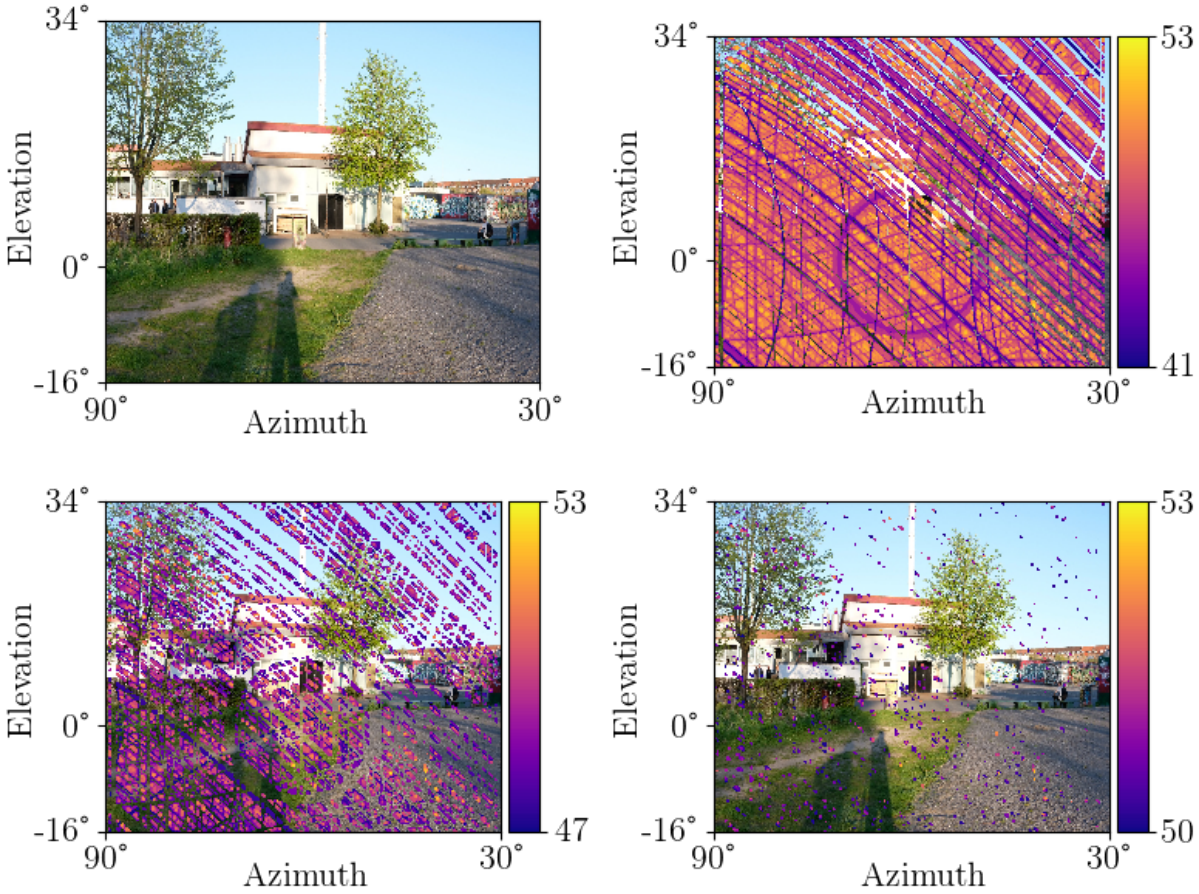


Figure 4.21: Figures depict localization results overlaid on the photo of the measured source with different dynamic ranges (dB). Zero-phase frequency filtering is applied on the signal to low pass the signal at 100Hz (Appendix F). This causes the results due to the crowd to disappear and the localization results from the indoor concert can now be seen.

4.3.6 Roadside Noise

Figure 4.22: Panorama from the center point of the microphone array

Cars passing through a crossing were measured in a close range ($2-10m$). Fig. 4.22 shows a panorama of the measurement taken from the microphone array. Fig. 4.23 shows the results. Even though cars are not stationary, this scenario is interesting because the cross-correlations should still be able to localize the path of the cars. The measurement was done next to a road such that the cars passed from in front of the array with azimuths ranging from 180° to 0° and elevation ranging around ±15°. The recordings were done next to a red light and started as soon as the light turned red. As can be seen in

the results, the localization happens within the azimuth, elevation window described here. Since the cars got louder as they passed from infront of the array, the peaks are detected in the 0° to 60° azimuth range.

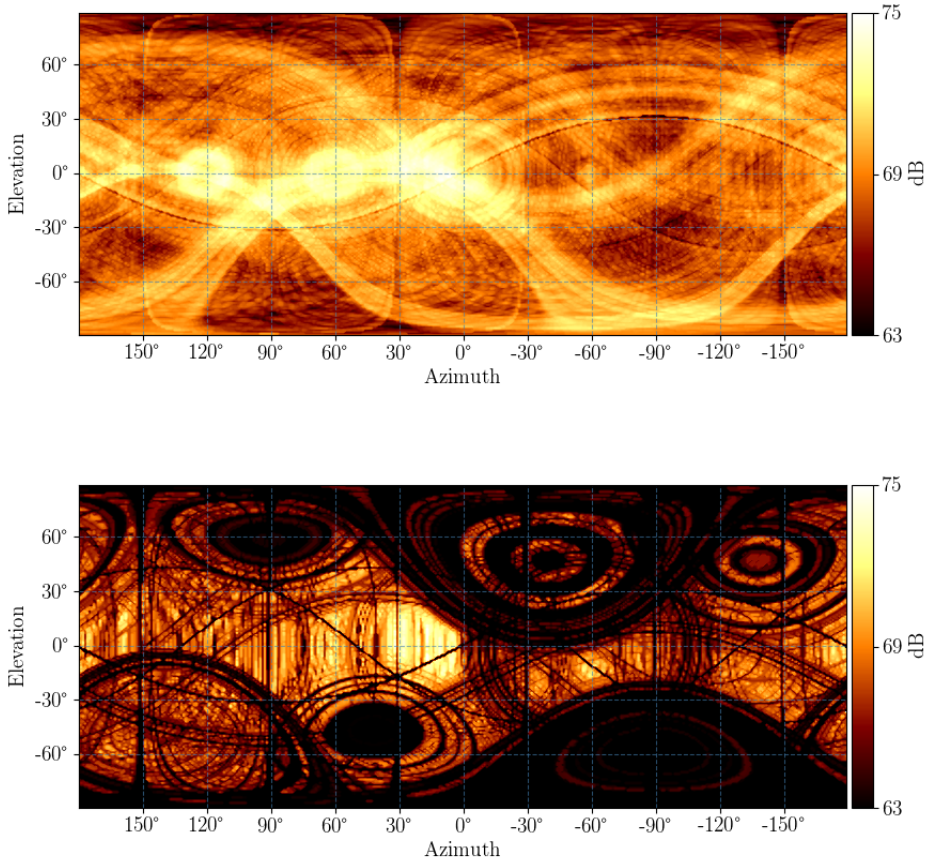


Figure 4.23: Figures depict from SRP-PHAT (top) and Min-SRP-PHAT (bottom) localization for a traffic.

4.3.7 Chalk Mine

A chalk mining machine was measured from a large distance (500m). The same measurement was then run closer (90m). The far distance measurement was conducted from a lookout close to the road, where a lot of traffic noise was present. The near distance measurement was conducted further away from the road, such that the road noise was also blocked by a dirt wall and vegetation.

4.3.8 Lookout results

The full results are given in Fig. 4.24. The overlaid image can be seen in Fig. 4.25. As can be seen in the full results traffic noise was detected between 0°-60° azimuth. This was because the traffic was on a highway to the left of the array. A long recording of 10 minutes was conducted for the measurement, and the chalk mine was localized at the correct location. On the overlaid image it is interesting to note that there appear to be two distinct sound sources in the 6dB and 3dB dynamic range plots. The lower source corresponds to the edge of the pit that surrounds the chalk mine lake, and as such it is an artifact of the reflection from that edge.

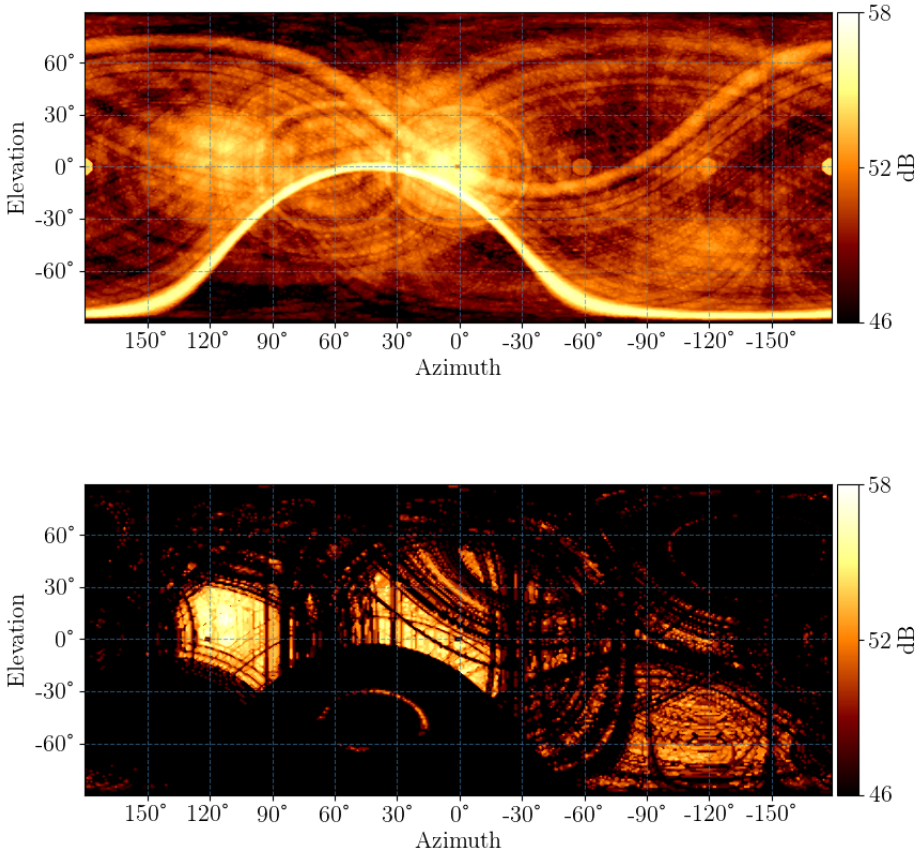


Figure 4.24: Figures depict from SRP-PHAT (top) and Min-SRP-PHAT (bottom) localization for chalk mine measured from a lookout 500m away.

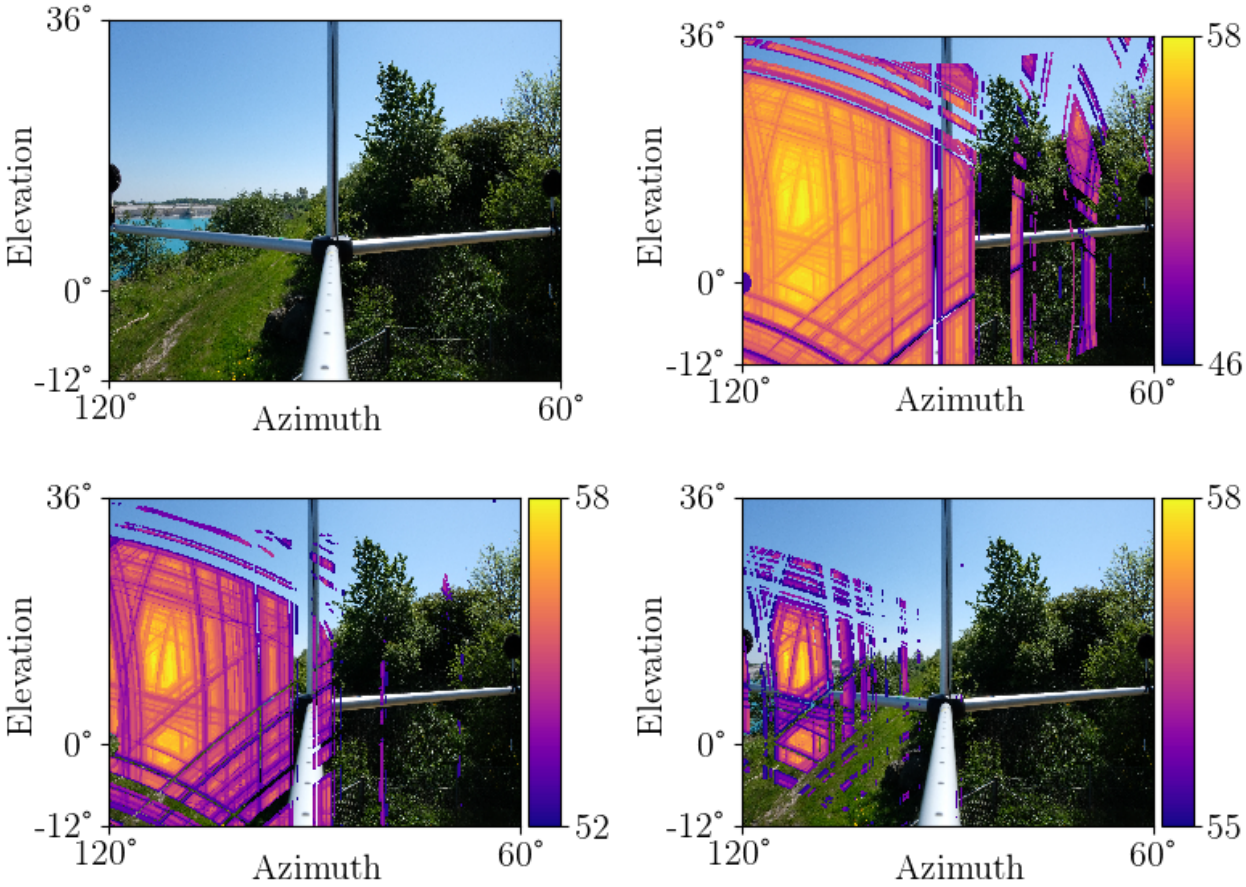


Figure 4.25: Localization results of the chalk mine from a far away lookout close to traffic noise

Close to Mine, Close to Edge

Measurements were run at a distance of 90*m* from the mine. The measurements were made close the edge of the chalk mine lake pit, to remove all the effects of reflection so that a clean line of sight to the mine was achieved. The full results are given in Fig. 4.26. The overlaid image can be seen in Fig. 4.27. It can be seen that the results here are fairly clean. This is because the dirt wall that separated the road attenuated the traffic noise almost completely. Also since the mine was closer, the levels of the mine were higher as well, which helped the results.

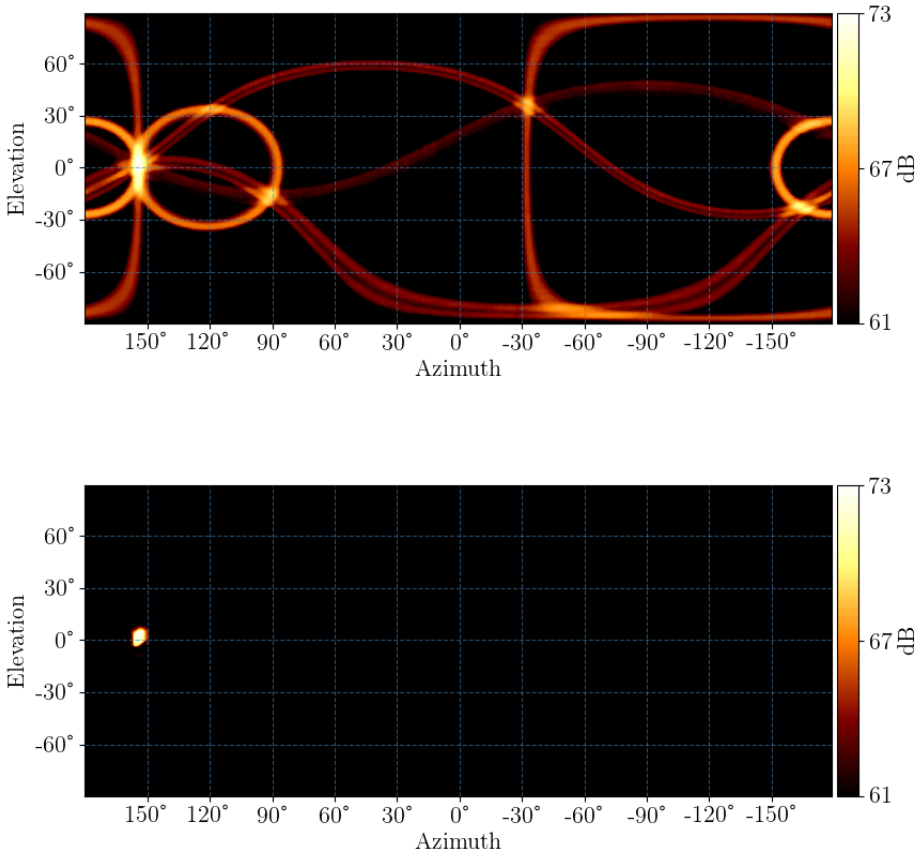


Figure 4.26: Figures depict from SRP-PHAT (top) and Min-SRP-PHAT (bottom) localization for a chalk mine around 90m away. The measurement is made close to the edge of the chalk mine lake pit and a direct line of sight to the chalk mine is achieved.

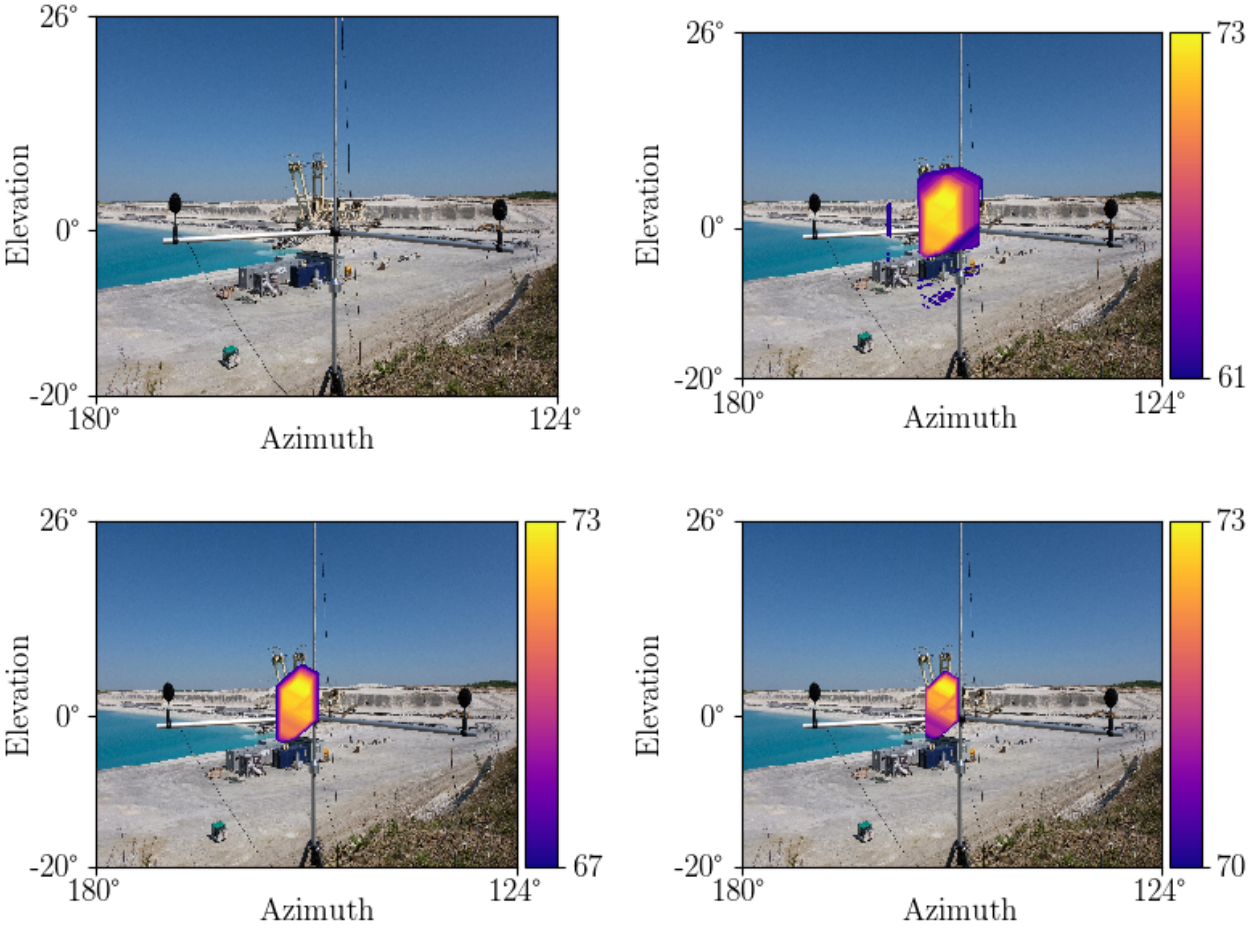


Figure 4.27: Localization results of the chalk mine

Close to Mine, Far from Edge

Another measurement was conducted, this time further away from the edge of the pit. Fig. 4.28 shows the overlaid results. The edge reflection can be seen here again, however, this time, only for the 12dB dynamic range. It is interesting to note that the ratio of the reflection to the direct sound has changed from before when the measurement was done from the lookout. This could be due to the terrain of the edge, or simply because the chalk mine is close enough to have a less diffuse direct sound field from the mine to the array.

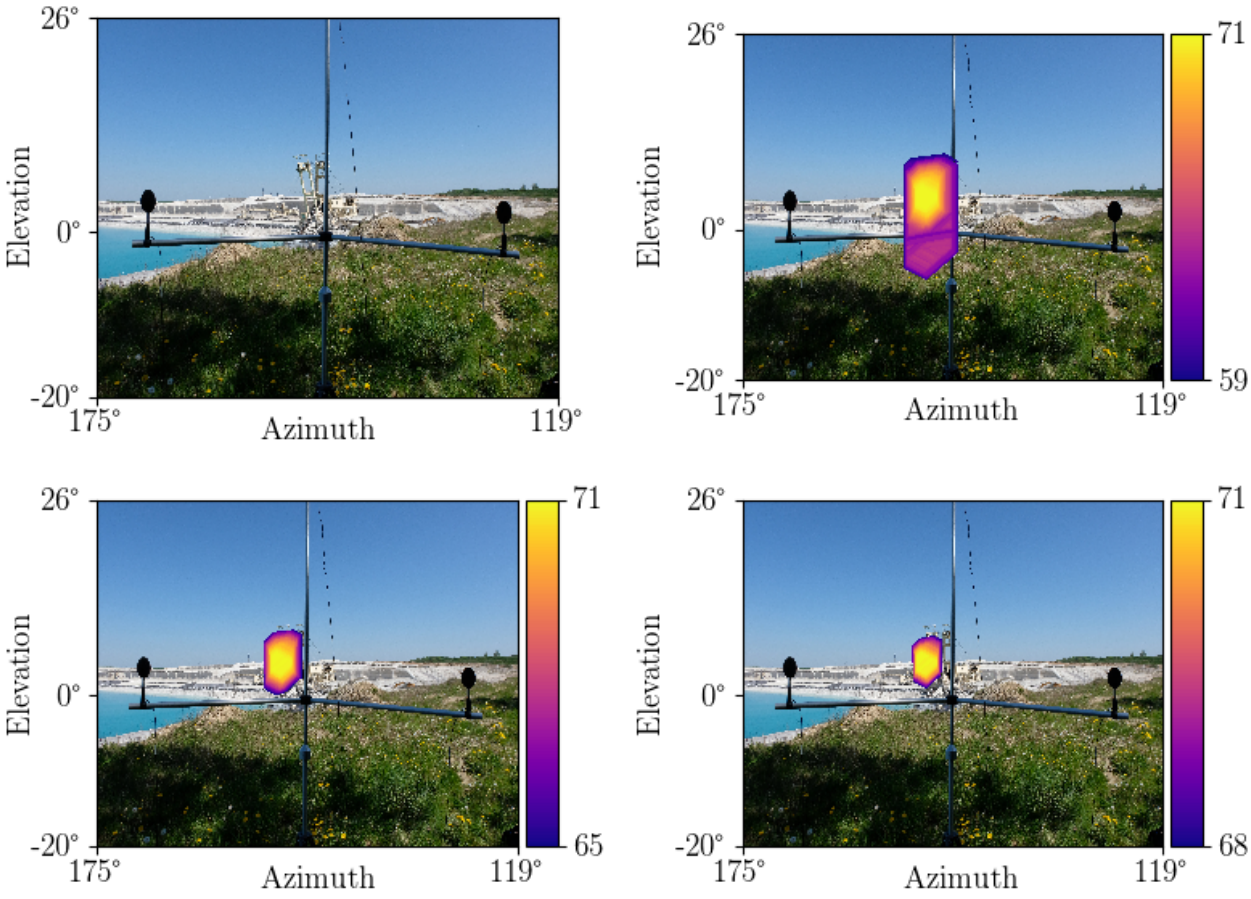


Figure 4.28: Localization results of the chalk mine, close to the edge of the mine pit

This measurement concludes the experimental evaluation section of the thesis. The next section will discuss the conclusion and provide groundwork for future work.

Chapter 5

Discussion & Conclusion

In this thesis, the SRP-PHAT algorithm has been implemented on a tetrahedral array for outdoor source localization and level retrieval. SRP-PHAT simulations were run to determine its performance in ideal conditions and highlight its drawbacks. Based on the analysis, a Min-SRP-PHAT algorithm was developed. The Min-SRP-PHAT algorithm was then compared to the SRP-PHAT algorithm by running various simulations for the different outdoor conditions. Simulations revealed that, other than in the case of multiple coherent sources, the algorithm is fairly robust and provides a cleaner result with a higher achievable dynamic range than the regular SRP-PHAT algorithm. Various anechoic measurements were then conducted to confirm the validity of the simulations. Outdoor measurements were conducted as well as to test the robustness and accuracy of the algorithm in actual outdoor conditions. The measurements results show that the algorithm can indeed be applied for the purpose of outdoor sound map reconstruction, with a higher performance than the regular SRP-PHAT.

One of the drawbacks that was detected was the noise floor of the map depends on the highest magnitude source, as such masking of lower level sources can occur. This means that there is a limit to the dynamic range achievable by this algorithm. The simulations and the real measurements were run for a dynamic range of 12dB or below. This follows the general industry performance for acoustic localization with microphone arrays, where the dynamic range can be anywhere between 3-12dB. Frequency filtering was broached as a solution to unmask spectrally different sources, however more investigation is needed to determine its accuracy.

Future work could include methods to implement the algorithm in real time. This can be done by implementing it on FPGAs/ DSPs. Delays could be stored in memory for various conditions, so that the algorithm reduces to an FFT-IFFT with a delay table lookup. Such a device could also contain an in-build camera with fish-eye lens so that overlaying the 360° images can be done directly.

Overall, the algorithm was able to achieve the goal of localizing major outdoor sound contributors and retrieving their absolute levels.

Appendix A

Outdoor Environment

Various different models have been designed for outdoor sound field received at a receiver by different international standard organizations. The ISO 9613-2 [24] is an international standard model for attenuation of sound when propagating outdoors. The standard uses an empirical method to quantify attenuation in different circumstances. Being empirical is a disadvantage as the model might not fit particular real world scenarios and user discretion is needed when using the model. NMPB-2008 [25] is a French standard model which uses simple engineering methods to model road traffic noise. Over time it has been extended to include other sound sources. Nord2000 [26] and Harmonoise [27] are more advanced engineering models for outdoor sound propagation. Nord2000 was developed in the period 1996-2001 by DELTA (Denmark, project manager, SINTEF (Norway), and SP (Sweden). Harmonoise is a more recent method and is made with a collaboration of various European countries. Nord2000 and Harmonoise are based on a similar approach and often produce quite similar models. Various inconclusive studies have been conducted comparing the two [28],[29]. Eventually, to have a harmonized and coherent approach, a common framework for noise assessment (CNOSSOS-EU) was developed by the European Commission [30] in co-operation with the EU Member States to be applied for strategic noise mapping as required by the Environment Noise Directive (2002/49/EC). CNOSSUS-EU investigates the various existing methods and their advantages and disadvantages. It takes into consideration the accuracy as well as the computational complexities of the various methods. In general, the effect of different factors, that have been explored by these models, are described below.

A.1 Ground Effects

On acoustically hard surfaces such as non-porous asphalt or concrete, ground effects cause sound pressure to approximately double across a wide range of frequency. For porous surfaces, lower frequencies are enhanced while the higher frequencies get absorbed by the ground. When both source and receiver are close to the ground, interference of sound travelling directly from source-to-receiver and sound reflected from the ground causes various ground effects. This interference can be both constructive or destructive. The pressure at a location (x, y, z) due to a sound source can be given as a sum of the direct wave component, P_{dir} and the reflected wave component P_{ref} multiplied with the reflection coefficient R,

$$P(x, y, z) = P_{dir}(x, y, z) + R.P_{ref}(x, y, z), \tag{A.1}$$

Here, $P_{dir} \neq P_{ref}$ as the two might have different propagation path lengths r_{dir} and r_{ref}. We have,

$$P(x, y, z) = \frac{e^{-ikr_{dir}}}{4\pi r_{dir}} + R.\frac{e^{-ikr_{ref}}}{4\pi r_{ref}}, \tag{A.2}$$

For plane waves, the reflection coefficient of sound waves reflecting from the ground at angle ϕ is given by

$$R = \frac{cos(\phi) - \beta}{cos(\phi) + \beta}, \tag{A.3}$$

here β is specific normalized admittance of ground with respect to air. For infinitely hard surfaces $\beta \to 0$ and $R \to 1$. For infinitely soft surfaces $\beta \to \infty$ and $R \to -1$. This can be interpreted as a phase change upon reflection from acoustically soft surfaces, which causes destructive interference and can also be seen as ground absorption. Note that for large distances, $\phi \to 90°$ (grazing incidence), $r_2 \to r_1$ which makes $P_{ref} \to P_{dir}$ causing

$$\begin{aligned} P_{plane}(x, y, z) &= P_{dir}(x, y, z) + \frac{0 - \beta}{0 + \beta}.P_{dir}(x, y, z) \\ &= 0. \end{aligned} \tag{A.4}$$

This predicts a net zero field over large distances irrespective of the value of β. The plane wavefront assumption is the cause of this error. Taking spherical waves, the equation for pressure becomes (Chap. 2 [31])

$$P(x, y, z) = \frac{e^{-ikr_{dir}}}{4\pi r_{dir}} + [R + (1 - R)F(\omega)]\frac{e^{-ikr_{ref}}}{4\pi r_{ref}} \tag{A.5}$$

The $F(\omega)$, known as the boundary loss factor, is given by

$$F(\omega) = 1 - i\sqrt{\pi}\omega e^{-\omega^2}\text{erfc}(i\omega). \tag{A.6}$$

The ω, often called the numerical distance, given by

$$\omega \approx \frac{1}{2}(1 + i)\sqrt{kr_{ref}}(cos(\phi) + \beta) \tag{A.7}$$

and finally the erfc(iω) is known as the complementary error function given by

$$\text{erfc}(i\omega) = 1 - \text{erf}(i\omega) \tag{A.8}$$

where

$$\text{erf}(i\omega) = \frac{1}{\sqrt{\pi}} \int_{-i\omega}^{i\omega} e^{-t^2} dt, \tag{A.9}$$

which is a sigmoid shaped error function. Now by setting

$$P_{plane}(x, y, z) = \frac{e^{-ikr_{dir}}}{4\pi r_{dir}} + R.\frac{e^{-ikr_{ref}}}{4\pi r_{ref}}, \tag{A.10}$$

and

$$P_{sph}(x, y, z) = (1 - R)F(\omega).\frac{e^{-ikr_{ref}}}{4\pi r_{ref}}, \tag{A.11}$$

Eq. A.5 becomes

$$P(x, y, z) = P_{plane}(x, y, z) + P_{sph}(x, y, z), \tag{A.12}$$

here the $P_{sph}(x, y, z)$ contribution is known as the ground wave component. It corresponds to the contribution from the vicinity of the image source in the ground plane. It includes a component known as the surface wave, which propagates close and parallel to a porous ground surface and decays with inverse square root of range. The ground itself impedes sound propagation by a variety of factors. Attenborough [32] created a more detailed 4-parameter model that requires porosity, flow resistivity, tortuosity and pore shape factor for modelling ground impedance on outdoor sound propagation.

A.2 Meteorological Effects

Wind and temperature have different effects on sound propagation. They directly change the speed of sound

$$c_z = c_0\sqrt{\frac{T + 273.15}{273.15}} + u_z, \tag{A.13}$$

where c_z is the speed of sound for temperature T above 0°C, c_0 is the speed of sound for no wind and 0°C, and u_z is the wind velocity in the direction of propagation of sound. They also cause acoustic gradients (varying refractive index) to occur in the atmosphere. Usually, with increasing height, the temperature decreases. This causes sound to travel slower with height. In the absence of wind this causes the sound to refract upwards leading to less sound received at the receiver. Wind speed can increase or decrease the sound speed. Generally, speed of wind increases with height, which causes the sound travelling along the wind to refract downwards. Conversely, if the sound is travelling against the wind, this would cause the sound to refract upwards.

A.3 Atmospheric Absorption

Sound energy converts to heat as it travels through air. The conversion of sound-to-heat in air can happen due to conduction, shear viscosity or by molecular relaxation. The portion of sound absorbed by air becomes increasingly important as distance of propagation increases. For a plane wave, the loudness L at a distance x from a position of known loudness L_0 is given by

$$L = L_0 - k.x, \tag{A.14}$$

where k depends on the humidity, temperature, pressure as well as the molecular composition of atmosphere and is proportional to the square of the frequency. Thus, higher frequencies are absorbed by a far greater magnitude. This causes air to act as a low-pass filter over large distances. Molecular relaxation [34], [33] is an important factor and losses due to oxygen-water vapour molecular relaxation are predominant above 500Hz. The absorption due to this factor is atleast 2 dB/kilometer irrespective of humidity and increases rapidly with frequency. The total absorption below 200 Hz is less than 1 dB/kilometer and decreases with frequency. If the air is extremely dry ($< 10\%$ relative humidity), the oxygen-carbon dioxide relaxation becomes significant and causes an almost constant absorption down from 500Hz to 80Hz of around 2dB/kilometer. The total air absorption as a function of frequency can be seen in Fig. A.1.

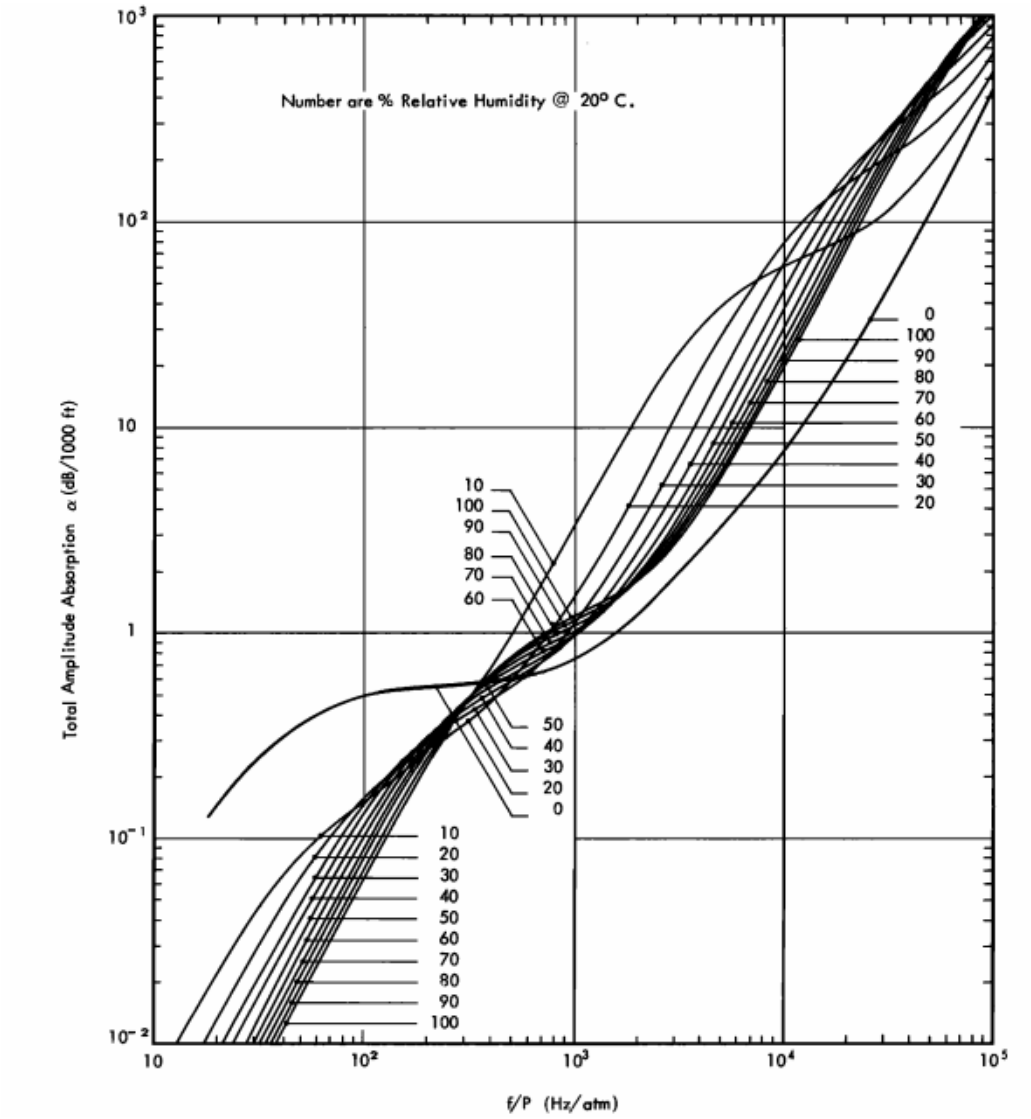


Figure A.1: Total absorption of sound in air as a function of frequency. The curves range from 0 to 100% relative humidity and are for 20°C [33] (Notice that the y-axis units are per 1000ft).

A.4 Other Outdoor Propagation Effects

A.4.1 Spreading Loss

The sound intensity from an omni-directional sound source drops as a function of distance due to wavefront spreading. The intensity I received at distance r from a source with power P, is given by

$$I = \frac{P}{4\pi r^2}. \tag{A.15}$$

This is due to spherical propagation, where the surface of the sphere has area $4\pi r^2$. In logarithm form this becomes

$$\begin{aligned} 10\log(I) &= 10\log\left(\frac{P}{4\pi r^2}\right) \\ L_p &= L_w - 20\log(r) - 11, \end{aligned} \tag{A.16}$$

which means a reduction of $20\log 2 = 6dB$, every doubling of r. This equation assumes uniform omni-directional directivity. For directional sources a Directivity Index DI can be added giving

$$L_p = L_w + DI - 20\log(r) - 11. \tag{A.17}$$

It is important to remember that such a directivity can be inherent to the source or might be induced due to the location of the source. An omni-directional source placed on a perfectly reflecting plane can only propagate sound into a hemisphere, in which case the DI is 3 dB. An infinite line source can be viewed as a linear array of omni-directional point sources. The wavefront spread is cylindrical (surface area = $2\pi r$), which gives

$$\begin{aligned} 10\log(I) &= 10\log\left(\frac{P}{2\pi r}\right) \\ L_p &= L_w - 10\log(r) - 8, \end{aligned} \tag{A.18}$$

The DI is again 3 dB and the reduction is $10\log 2 = 3$ dB, every doubling of r. Highway traffic is modelled in a similar manner, assuming 3 dB drop every doubling of distance.

A.4.2 Diffraction and Barriers

Barriers are sometimes purposefully built to block the direct path from the sound source to the receiver. Sound reaches the receiver either going through the barrier or by diffracting around the top of the barrier. Ground reflections and multi-path-propagation may lead to multiple diffracted wave paths. For a barrier, the ISO 9613-2 [24] provides the following equation for loss due to barrier insertion

$$IL = 10\log\left[3 + \left(C_2\frac{\delta_1}{\lambda}\right)C_3K_{met}\right], \tag{A.19}$$

where $\lambda = wavelength$. The value of C_2 determines if ground reflections are taken care of ($C_2 = 20$) or not ($C_2 = 40$), C_3 is a factor to take care of double diffraction due to a barrier of finite thickness (or two thin barriers placed some distance apart), δ_1 is the difference in distance between the direct

source-to-receiver path and the wave propagation path caused by the barrier, and K_{met} is a correction factor for average downwind meteorological effects. For thin barriers the equation simplifies to

$$IL = 10 \log \left(3 + 40 \frac{\delta_1}{\lambda} \right). \tag{A.20}$$

Over large distances even buildings act like barriers, with the rooftop causing double diffraction. ISO 9613-2 [24] provides a simple empirical method to calculate attenuation due to buildings.

Appendix B

Other Measurements

B.1 Tetrahedral Array Delays for a 300Hz Sine Wave

An experiment to visualize the delays between the microphones in the tetrahedral array is conducted. In order to avoid the pressure field frequency zone of the anechoic chamber and array aliasing, 300 Hz sine wave is used. The array aperture is kept at 0.395m to approximate plane wave better. The sampling rate is 131072Hz.

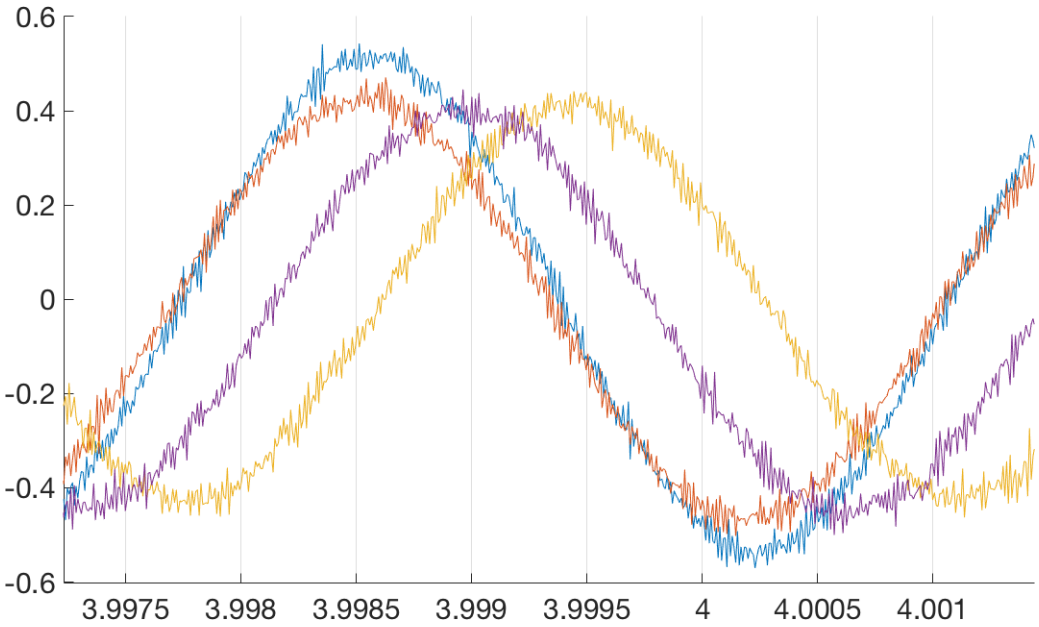


Figure B.1: Sine wave recorded by the four microphones

Delays	Mic 1	Mic 2	Mic 3	Mic 4
Mic 1	0	-6	114	52
Mic 2	6	0	120	58
Mic 3	-114	-120	0	-61
Mic 4	-52	-58	61	0

Table B.1: Sample delay measured between the microphones at 0 incidence (Fs= 131072Hz)

Mic 1 and Mic 2 were faced towards the sound source, with Mic 3 behind and in line with the sound source. Mic 4 being the top microphone. The arrangement was such that the sound source

was approximately at (90°,0°). As can be seen the delay between Mic 1 and Mic 2 is almost zero, indicating source close to 90° azimuth. The combination of these delays can be used to predict the source direction, both in azimuth and elevation. However, if the plane wave approximation does not hold, the combination might result in a null set (No single intersection point of all the cones from the various microphone pairs).

B.2 Length of Recording

The effect of the length of recording has been simulated before. The effect was also checked on a real world measurement. The results used here are from the Chalk Mine, close range and close to edge case.

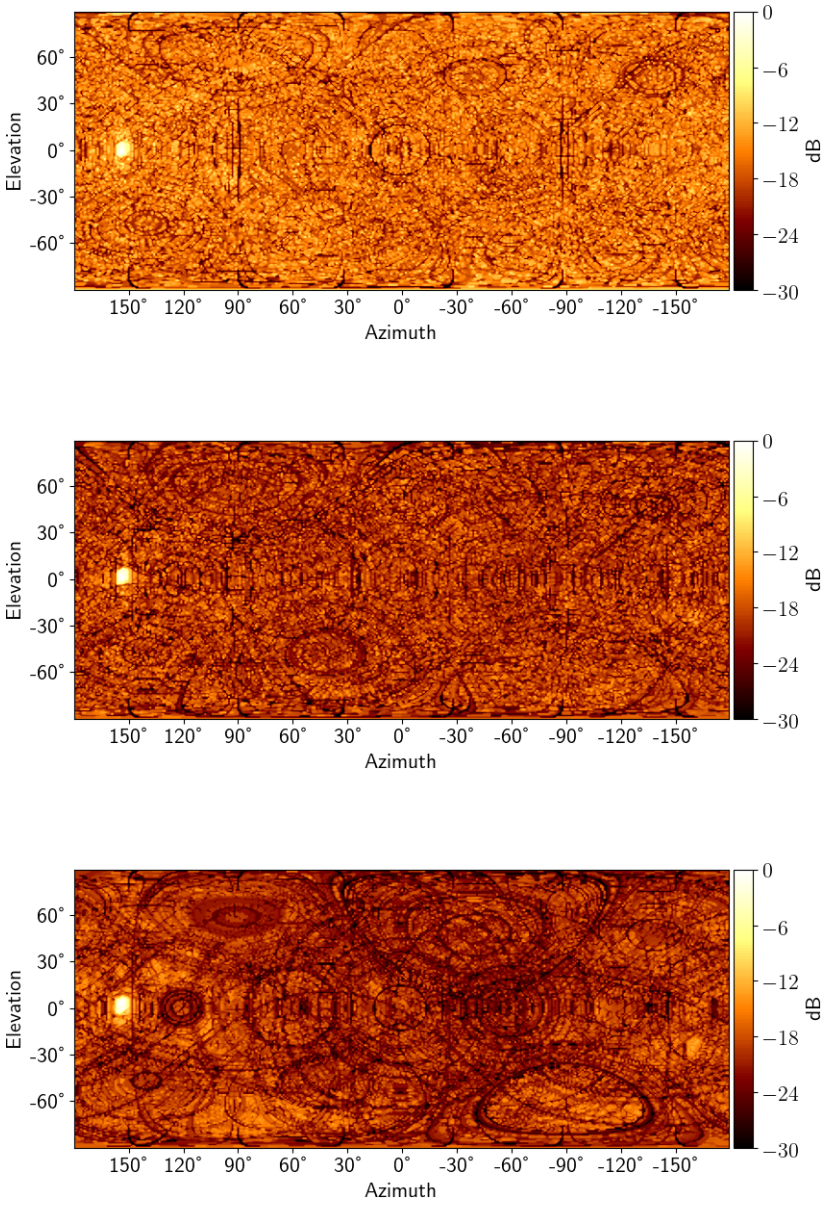


Figure B.2: Figures depict from SRP-PHAT localization results, for recording length 1sec (top), 10sec (middle) and 120sec (bottom). Relative dB are plotted here

Appendix C

Other Deconvolution

C.1 Product-SRP-PHAT

A simple deconvolution approach could be to penalize sources which are only detected by a subset of the microphone pair combinations. This could be done by taking a product and not a sum in Eq. 2.33.

$$S_{SRP}(\theta, \phi) = \prod_{i=1}^{M-1} R_{x_0,x_i}[f_{0,i}(\theta, \phi)] \tag{C.1}$$

This way, if a peak is caused by a single localization circle, the cross-correlation values from other microphone pairs would be close to zero, and thus would scale the false peak down. The localization results from this are given in fig. C.1. The drawback of using product-SRP-PHAT is that the sound level difference between the different sound sources is lost. In normal SRP-PHAT, the array magnitude response at a particular azimuth and elevation is averaged over all microphone pair combinations. Then the level difference between 2 sources is maintained. In product-SRP-PHAT this would not be the case. However if it is assumed that a particular source will have similar magnitude response for all microphone pairs (which is not a strong assumption in far-field), then taking source power $P_{SRP} = {S_{SRP}}^{1/M}$, the level difference can be maintained. Fig. C.2 depicts the results of product-SRP-PHAT after level correction.

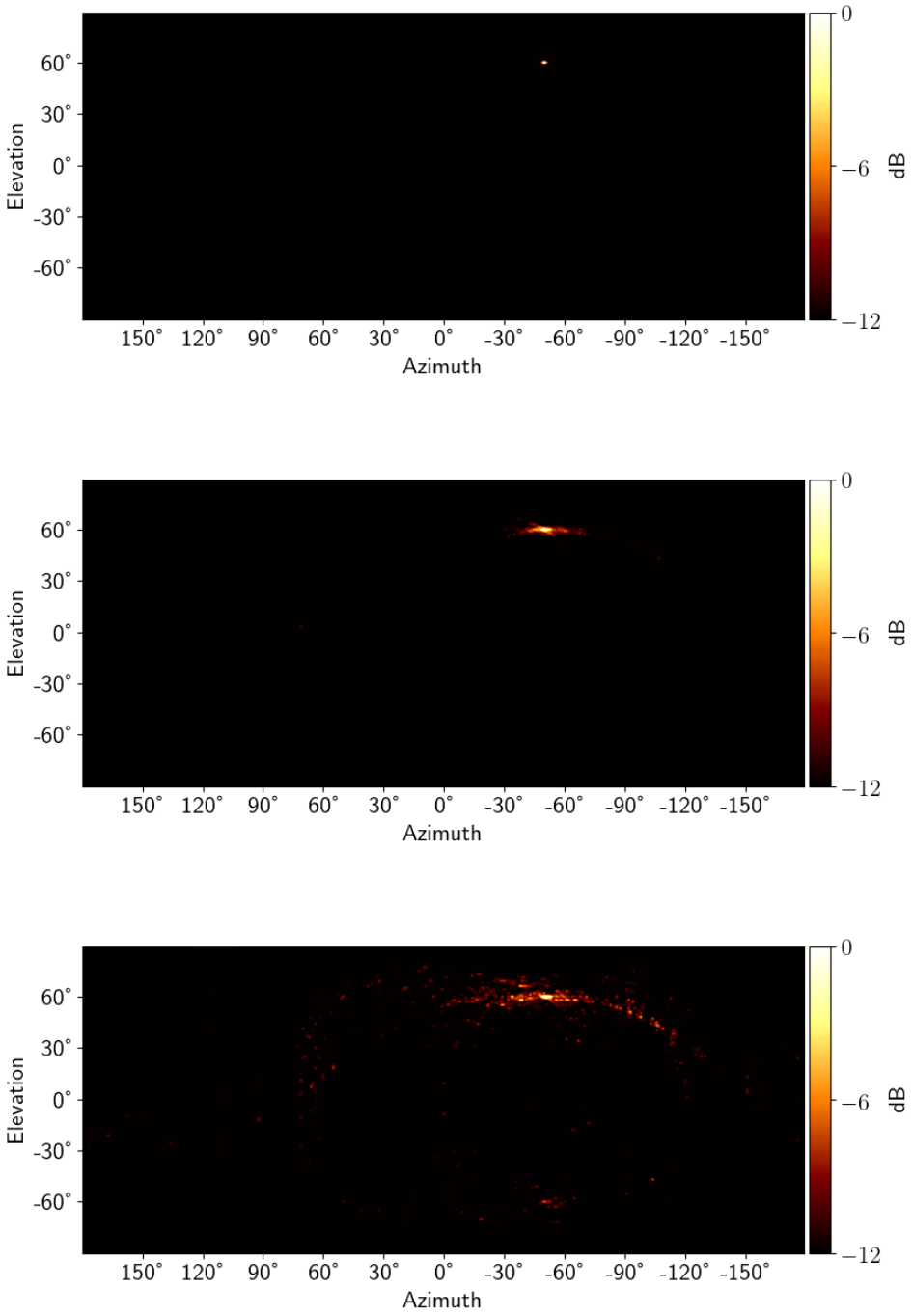


Figure C.1: Figures depict from top-to-bottom product-SRP-PHAT localization results with SNR = 20dB, SNR = 0dB, SNR = -10dB

C.2 Threshold SRP-PHAT

For product SRP-PHAT, if one of the cones is really high is magnitude, the algorithm might not be able to penalize a wrong location enough by simple multiplication with low power from other microphone pairs. This means that if one of the microphone pairs detects a very low power at a particular location, it should have a higher priority when deciding the power at that location. One way to achieve this could be to sort the power for the different cones at each location, and then divide the values with each other. If the numbers are quite different in magnitude, the division could then cross a certain pre-defined

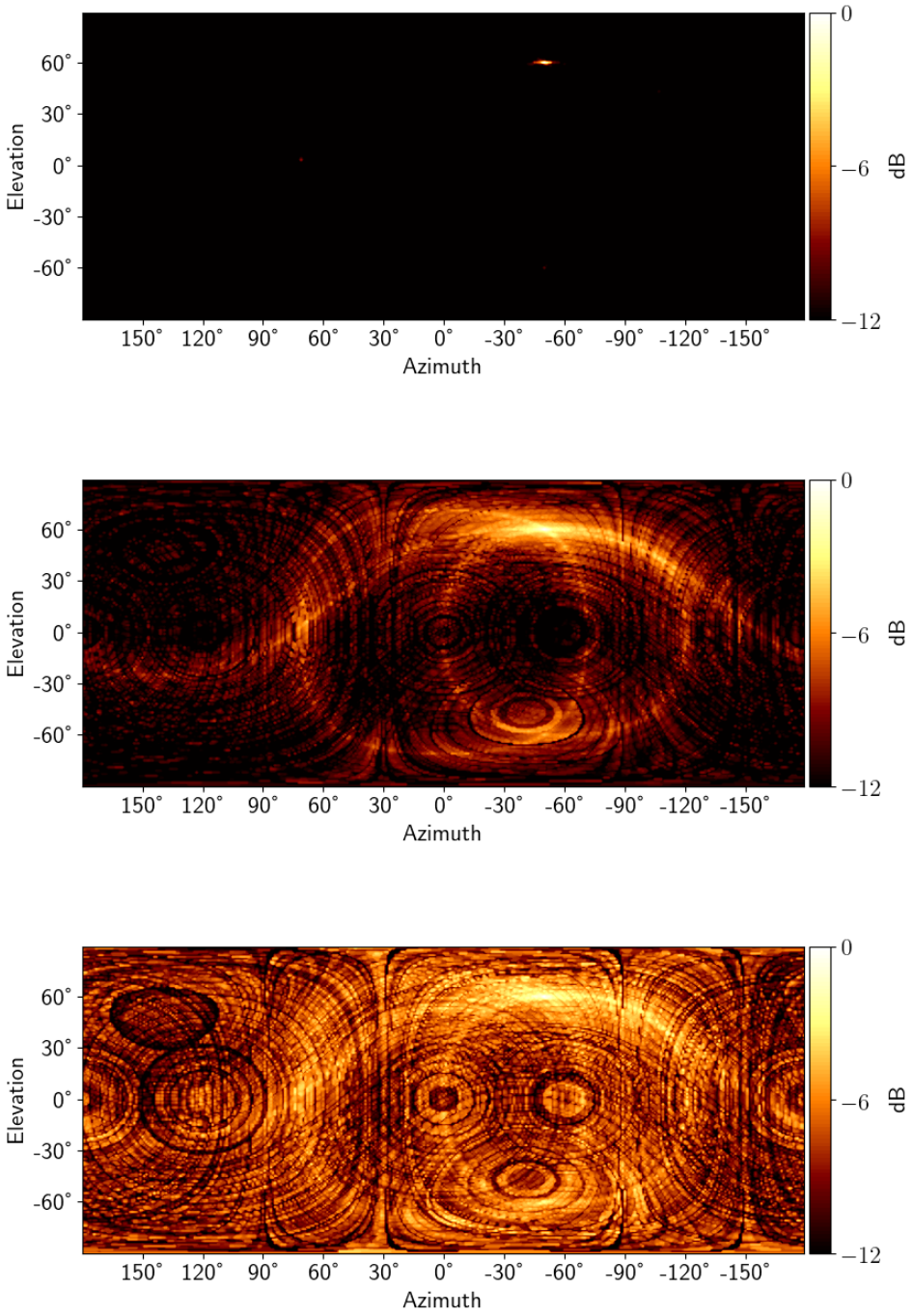


Figure C.2: Figures depict from top-to-bottom level corrected product-SRP-PHAT localization results with SNR = 20dB, SNR = 0dB, SNR = -10dB

threshold, and the power at that location could be set to zero. However, while this methodology could work for a single source, for multiple sources playing at different levels, and which share a localization cone, this would lead to the masking of the lower magnitude source.

Appendix D

Other Simulations

D.1 Effect of Array Tilt on Ground Reflections

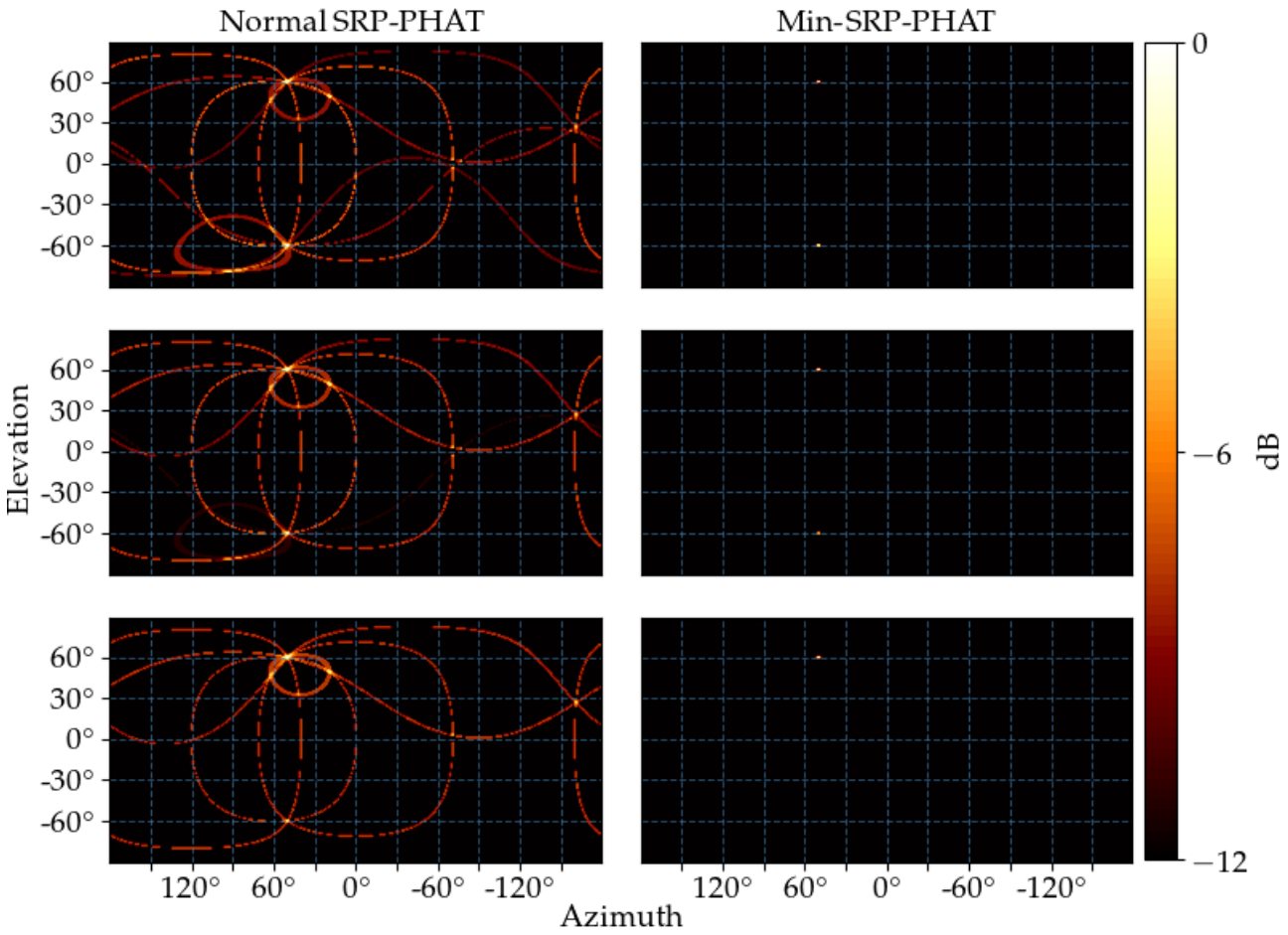


Figure D.1: Figures depict from top-to-bottom SRP-PHAT localization results with ground reflection coefficients (R) of 1, 0.6 and 0.1. The array has been tilted 30° along the X-axis and 15° along the Y-axis. This causes the ground image source to stop sharing cones with the real source and the correct relative power level between the image and real source is maintained for both normal SRP-PHAT and Min-pow SRP-PHAT

Keeping the tetrahedral array horizontal[1], causes ground image source and the main source to

[1] Such that three of its microphones are at the same height

share 3 out of 6 localization cones. This can cause magnitude errors when localizing with normal SRP-PHAT. This issue can be solved by tilting the microphone array as can be seen in Fig. D.1. As an effect of tilting the tetrahedron 30° along the X-axis and 15° along the Y-axis, the localization cones stop overlapping.

D.2 Effect of Angular Resolution of Localization

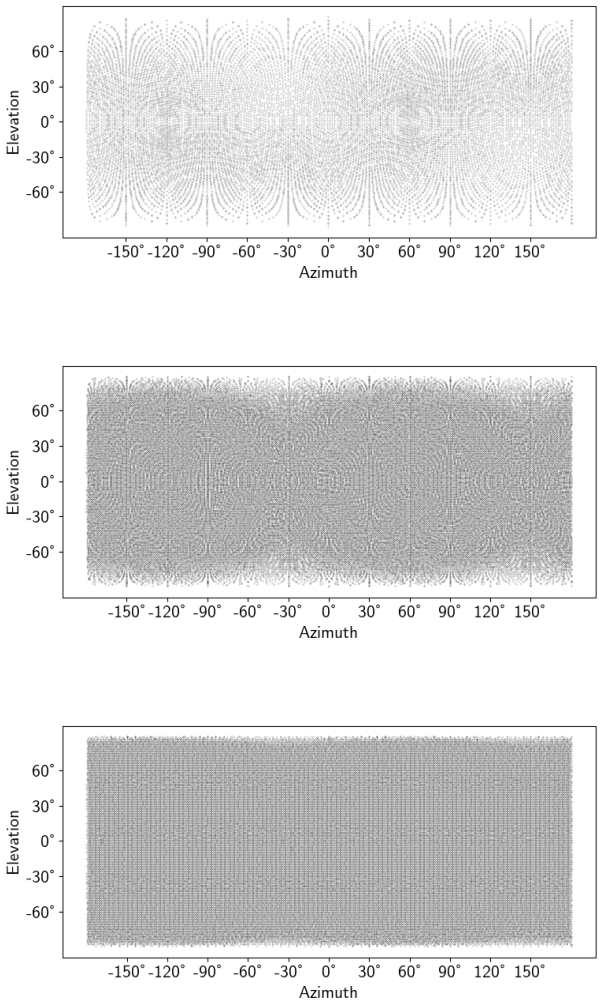


Figure D.2: Figures depict from top-to-bottom SRP-PHAT angular resolution of localization with sample rate of 12kHz, 48kHz and 192kHz.

Similar to the 2D angular resolution due to a single pair (Fig. 2.7), the tetrahedral array has its own 3D angular resolution depending on the sample rate and the array aperture size. Fig. D.2 describes the angular resolution of a tetrahedral array of aperture 1m for different sample rates. If the SRP search space is 360° by 180°, and the resolution of search is 1°, there will be some angular locations where the delay in sample between a microphone pair is fractional. For simplicity, this fractional delay is rounded for all plots in this thesis. This means that certain locations will contain duplicate data from the nearest integral delay location to themselves. Obviously, the resolution improves for higher sample rate.

Appendix E

Practical Details

E.1 Field of View Calculation

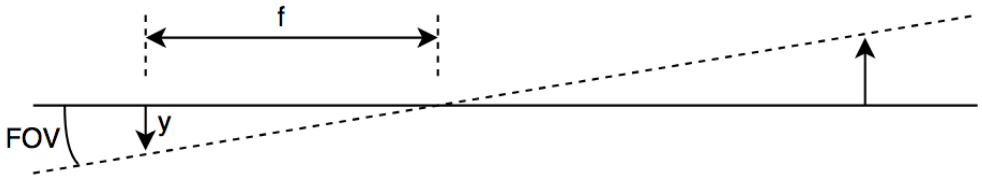


Figure E.1: Field of view

In order to overlap the map to the picture, the field of view (FOV) is calculated.

$$FOV = 2 * arctan(\frac{y}{f}) \tag{E.1}$$

where f is the focal length of the lens and y is the height/ width of the sensor. In some cameras, there is a crop factor which needs to be multiplied to y. In the case of the FUJIFILM xt-20, the crop factor is 1.5 and sensor dimension is $23.5 * 15.6$mm. For a focal length of $16mm$ The horizontal FOV is 44° and the vertical FOV is 33°. Therefore the horizontal span is 88° and the vertical span is 66°. Some photos used in the projects are panoramas, the range of the panoramas is $-44°$ to 224° in azimuth and $\pm 33°$ in elevation. Some overlay photos are cropped manually from panoramas to show relevant results. In that case, the azimuth and elevation of the cropped photo is determined linearly from the panorama.

E.2 Generation of White Noise

White noise is generated using the inbuilt python function RandomState() from the numpy library. The function generates a sequence of random numbers using the Mersenne Twister pseudo-random number generator algorithm.

E.3 Generation of Pink Noise

Pink noise is generated by first generating multiple chunks of 1 sec long white noise using the inbuilt python function RandomState() from the numpy library. Once the white noise is generated, the chunks are convolved with a pink filter of the requisite frequency range to convert the noise to pink.

E.4 Microphone Information

The sensitivities of the microphones used for the tetrahedral array are tabulated below. These sensitivities are only relevant for the outdoor measurements. For simulations all microphones are assumed to have equal sensitivity.

Mic	Sensitivity
M_1	6.694 mV/Pa
M_2	5.863 mV/Pa
M_3	5.743 mV/Pa
M_4	5.696 mV/Pa

The tetrahedral array was used with two different apertures, 1m and 39.5cm, resulting in two different array configurations, large and small. The (x,y,z) placement of the 4 microphones for the configurations are given below. The origin of the coordinate system signifies the viewpoint of the array, i.e, the (azimuth,elevation) shown in the results are relative to the origin on these co-ordinates. For the large aperture,

Mic	Position (m)
M_1	(0.5, 0, 0)
M_2	(-0.5, 0, 0)
M_3	(0, -0.866, 0)
M_4	(0, -0.433, 0.7071)

And for the small aperture,

Mic	Position (m)
M_1	(0.1975, 0, 0)
M_2	(-0.1975, 0, 0)
M_3	(0, -0.3421, 0)
M_4	(0, -0.171, 0.2793)

For simulations where less than 4 microphones are used, the microphone used are

- 2 Microphones: Only M_1 and M_2
- 3 Microphones: M_1, M_2 and M_3

E.5 Calibration of audio files

Using the BK Pulse software suite, the recordings were saved as 16bit .wav files. The pulse system saves the wav files as a factor of the actual Pascal values. Thus returning to the true pascal values requires converting the .wav files to 64bit and then dividing the wav files by the scaling factor. This factor is saved at the end of the .wav files so that future retrieval is easy.

E.6 Further information

The audio recordings and the scripts used in this project are available upon request from the authors.

Appendix F

Filter Design

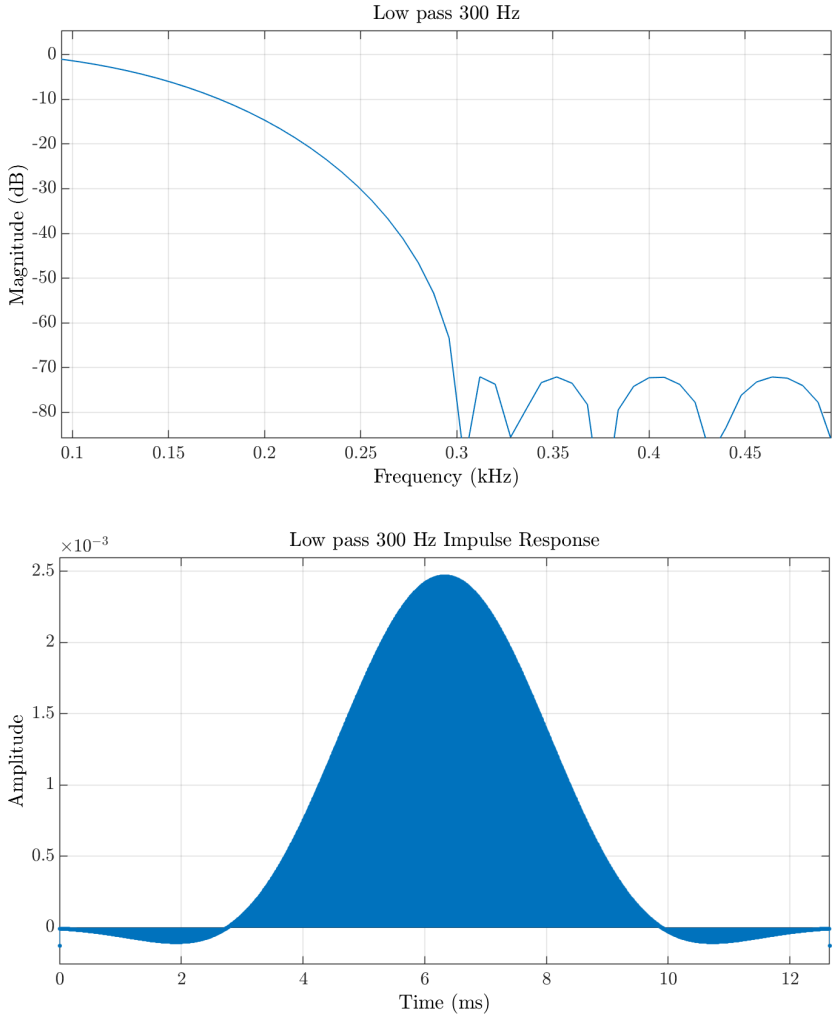


Figure F.1: Figures depict the frequency response and impulse response of a 300Hz low pass filter with stop band $80dB$, stop band frequency $300Hz$ and pass band frequency $100Hz$

Filtering the recorded signal can be useful to isolate certain frequencies of the signal as well as unmasking sources with different frequency content or different levels. The min-power SRP-PHAT algorithm uses the PHAT transform, which whitens the magnitude of the signal in order to only use the phase information to localize. For this reason a zero-phase filter is considered in our specific case

so that the phase remains unaffected. The low pass filters are designed using Matlab Filter Designer. Design specification for the low pass is $fs = 131072Hz$, minimum order, FIR equi-ripple. The filter order is 1659. Figure F.1 shows the frequency and impulse response of the filter. In order to zero-phase filter the signal, the signals are filtered once, then inverted and filtered again. The Matlab function filtfilt is used for this purpose.

Bibliography

[1] Michael Shapiro Brandstein. "A Framework for Speech Source Localization Using Sensor Arrays". AAI9540732. PhD thesis. Providence, RI, USA, 1995.

[2] C. Knapp and G. Carter. "The generalized correlation method for estimation of time delay". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24.4 (Aug. 1976), pp. 320–327. ISSN: 0096-3518. DOI: 10.1109/TASSP.1976.1162830.

[3] R Boucher and J Hassab. "Analysis of discrete implementation of generalized cross correlator". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 29.3 (1981), pp. 609–611.

[4] Giovanni Jacovitti and Gaetano Scarano. "Discrete time techniques for time delay estimation". In: *IEEE Transactions on signal processing* 41.2 (1993), pp. 525–533.

[5] Michael S Brandstein and Harvey F Silverman. "A practical methodology for speech source localization with microphone arrays". In: *Computer Speech & Language* 11.2 (1997), pp. 91–126.

[6] Lei Zhang and Xiaolin Wu. "On cross correlation based-discrete time delay estimation". In: *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*. Vol. 4. IEEE. 2005, pp. iv–981.

[7] Sakari Tervo and Tapio Lokki. "Interpolation methods for the SRP-PHAT algorithm". In: *Proc. of 11th IWAENC* (2008).

[8] Joseph Hector DiBiase. *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. Brown University Providence, 2000.

[9] Hamid Krim and Mats Viberg. "Two decades of array signal processing research: the parametric approach". In: *IEEE signal processing magazine* 13.4 (1996), pp. 67–94.

[10] Scott M Griebel and Michael S Brandstein. "Microphone array source localization using realizable delay vectors". In: *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*. IEEE. 2001, pp. 71–74.

[11] Fukutaro Okuyama et al. "A study on determination of a sound wave propagation direction for tracing a sound source". In: *SICE 2002. Proceedings of the 41st SICE Annual Conference*. Vol. 2. IEEE. 2002, pp. 1102–1104.

[12] Jacob Benesty, Jingdong Chen, and Yiteng Huang. "Time-delay estimation via linear interpolation and cross correlation". In: *IEEE Transactions on speech and audio processing* 12.5 (2004), pp. 509–519.

[13] Hong Liu and Miao Shen. “Continuous sound source localization based on microphone array for mobile robots”. In: *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE. 2010, pp. 4332–4339.

[14] J-M Valin, François Michaud, and Jean Rouat. “Robust 3D localization and tracking of sound sources using beamforming and particle filtering”. In: *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. Vol. 4. IEEE. 2006, pp. IV–IV.

[15] Jwu-Sheng Hu, Chia-Hsing Yang, and Cheng-Kang Wang. “Estimation of sound source number and directions under a multi-source environment”. In: *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. IEEE. 2009, pp. 181–186.

[16] Anthony Badali et al. “Evaluating real-time audio localization algorithms for artificial audition in robotics”. In: *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. IEEE. 2009, pp. 2033–2038.

[17] Jacek P Dmochowski, Jacob Benesty, and Sofiene Affes. “A generalized steered response power method for computationally viable source localization”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.8 (2007), pp. 2510–2526.

[18] Mehdi Batel et al. “Noise source location techniques-simple to advanced applications”. In: *Sound and Vibration* 37.3 (2003), pp. 24–38.

[19] JA Högbom. “Aperture synthesis with a non-regular distribution of interferometer baselines”. In: *Astronomy and Astrophysics Supplement Series* 15 (1974), p. 417.

[20] Pieter Sijtsma. “CLEAN based on spatial source coherence”. In: *International journal of aeroacoustics* 6.4 (2007), pp. 357–374.

[21] Thomas F Brooks and William M Humphreys. “A deconvolution approach for the mapping of acoustic sources (DAMAS) determined from phased microphone arrays”. In: *Journal of Sound and Vibration* 294.4-5 (2006), pp. 856–879.

[22] Thomas Brooks and William Humphreys. “Extension of DAMAS phased array processing for spatial coherence determination (DAMAS-C)”. In: *12th AIAA/CEAS Aeroacoustics Conference (27th AIAA Aeroacoustics Conference)*, p. 2654.

[23] Thomas Padois et al. “On the use of geometric and harmonic means with the generalized cross-correlation in the time domain to improve noise source maps”. In: *The Journal of the Acoustical Society of America* 140.1 (2016), EL56–EL61.

[24] Organización Internacional de Normalización. *ISO 9613-2 : Acoustics - Acoustics - Attenuation of Sound During Propagarion Outdoors: Parte 2, General method of calculation*. pt. 2. ISO, 1996.

[25] Guillaume Dutilleux et al. “NMPB-ROUTES-2008: the revision of the French method for road traffic noise prediction”. In: *Acta Acustica united with Acustica* 96.3 (2010), pp. 452–462.

[26] Birger Plovsing and J Kragh. “Nord2000”. In: *COMPREHENSIVE OUTDOOR SOUND PROPAGATION MODEL. DELTA Report, Lyngby* 31 (2000).

[27] Jérôme Defrance et al. “Outdoor sound propagation reference model developed in the European Harmonoise project”. In: *Acta Acustica united with Acustica* 93.2 (2007), pp. 213–227.

[28] Naveen Garg and Sagar Maji. “A critical review of principal traffic noise models: Strategies and implications”. In: *Environmental Impact Assessment Review* 46 (2014), pp. 68–81.

[29] Gunnar Birnir Jónsson and Finn Jacobsen. “A comparison of two engineering models for outdoor sound propagation: Harmonoise and Nord2000”. In: *Acta Acustica united with Acustica* 94.2 (2008), pp. 282–289.

[30] Stylianos Kephalopoulos, Marco Paviotti, and Fabienne Anfosso Ledee. *Common noise assessment methods in Europe (CNOSSOS-EU)*. 2012.

[31] Keith Attenborough, Kai Ming Li, and Kirill Horoshenkov. *Predicting outdoor sound*. CRC Press, 2006.

[32] Keith Attenborough, Imran Bashir, and Shahram Taherzadeh. “Outdoor ground impedance models”. In: *The Journal of the Acoustical Society of America* 129.5 (2011), pp. 2806–2819.

[33] LB Evans, HE Bass, and LC Sutherland. “Atmospheric absorption of sound: theoretical predictions”. In: *The Journal of the Acoustical Society of America* 51.5B (1972), pp. 1565–1575.

[34] HE Bass, LC Sutherland, and AJ Zuckerwar. “Atmospheric absorption of sound: Update”. In: *The Journal of the Acoustical Society of America* 88.4 (1990), pp. 2019–2021.