HELMUT SCHMIDT UNIVERSITÄT

Universität der Bundeswehr Hamburg

Whistle Sound Source Localization Using Multiple NAO Robotic Systems

Master Thesis

by

Yuria Konda

Start date:	04. May 2019
End date:	04. November 2019
Supervisor:	Dr.-Ing. Martin Holters
Supervising Professors:	Prof. Dr.-Ing. habil. Udo Zölzer
	Prof. Dr.-Ing. Gerhard Bauch

Abstract

The RoboCup Standard Platform League is a competition for prospective researchers to compete in autonomous robot soccer with the overall goal to contribute to research in the fields of humanoid robotics and autonomous multi-agent systems. According to the rules of this league, implementation is done on NAO robots. Currently, audio signals are only used as indicator for the kickoff in form of a whistle sound. To prevent the detection of false positives from neighboring fields, a whistle sound source localization is designed and implemented. Different methods that are based on the Time Difference Of Arrival (TDOA) are evaluated to obtain the direction of the whistle source using the four microphones attached on the robot's head. To compute a global position of the acoustic source, direction estimates of multiple robots are fed into a multi-agent filter. The resulting algorithm is shown to allow whistles to be localized with an Root Mean Squared Error (RMSE) of 1m in terms of Euclidean distance.

Statement

Hereby I do state that this work has been prepared by myself and with the help which is referred within this thesis.

Hamburg, November 4th 2019

Contents

List of Figures

List of Tables

List of Abbreviations

ALSA	Advanced Linux Sound Architecture
API	Application Programming Interface
CC	Cross Correlation
DFT	Discrete Fourier Transform
FFTW	Fastest Fourier Transform in the West
FFT	Fast Fourier Transform
GCC-PHAT	Generalized Cross Correlation with Phase Transform
GCC	Generalized Cross Correlation
HSU-HH	Helmut-Schmidt-University/University of the Federal Armed Forces Hamburg
IFFT	Inverse Fast Fourier Transform
IMU	Intertial Measurement Unit
PDF	Probability Density Function
PHAT	Phase Transform
PSNR	Peak Signal to Noise Ratio
RMSE	Root Mean Squared Error
RoboCup	Robot World Cup Initiative
SNR	Signal to Noise Ratio
SPL	Standard Platform League
SSD	Signal Start Detection
SSL	Signal Source Localization
TDOA	Time Difference Of Arrival
TUHH	Hamburg University of Technology
UDP	User Datagram Protocol
UNSW	University of New South Wales
WSDE	Whistle Source Direction Estimation
ZCR	Zero Crossing Rate

Chapter 1

Introduction

Assuming robots as forthcoming everyday objects, reaction to acoustic input is one essential step for natural human-robot interaction [2, 3]. From a robots' perspective additional information about the environment helps to manage unknown scenarios. For example, for navigating robots being aware of obstacles beforehand due to acoustic perception is a beneficial ability [4]. An even more tangible case can be seen in conference rooms of business environments where remote participation is commonplace these days. Special features of communication systems like speaker identification and tracking become crucially important to provide smooth operation [5]. In general, Signal Source Localization (SSL) algorithms can be divided into three categories which are based on beamforming, eigenvalue decomposition or Time Difference Of Arrival (TDOA) [5–7].

In this work, the performance of different SSL strategies is evaluated with a focus on whistle sound localization in the context of the *RoboCup Standard Platform League (SPL)*. As stated in [8], beamforming methods are computationally expensive and eigenvalue decomposition is little suitable for signals with small bandwidth. For an application in the RoboCup domain, these constraints are undesirable as this setting requires real time processing of low-bandwidth whistle signals. Therefore, in this work the TDOA method was selected as the most appropriate solution.

A widespread approach to compute the TDOA is to cross-correlate the sensor readings of two microphones to recover the time delay between both channels [9]. By this, the relative direction of a signal source can be obtained from geometric reasoning on a single robot. From multiple such direction estimates a global sound source position can be estimated.

1.1 Motivation

The Robot World Cup Initiative (RoboCup) promotes research on autonomous robotic systems by hosting competitive events for scientists and students from around the world. It provides a platform that focuses on fast intelligent systems and multi-agent collaboration [10]. This initiative encourages young engineers to work in real case scenarios by providing several leagues focusing on different engineering challenges relevant to the field of robotics.

In the RoboCup Standard Platform League (SPL), software for humanoid robots is developed

to play soccer autonomously for scientific purpose. According to the rules of this league, all competitors are constraint to use the same hardware platform and no modifications are allowed. By this means, the SPL provides a competitive algorithmic real world benchmark. Since 2008 commercially available humanoid *NAO* robots by the company *SoftBank Robotics* are the official standard platform in this league. They are equipped with a variety of different sensor like cameras, speakers, microphones, sonars and many more.

Figure 1.1: *V6 NAO robots during game play at the RoboCup Sydney in 2019.*

Within the context of the SPL, for the RoboCup 2019 the *Directional Whistle Challenge* was added as a technical challenge. This challenge acted as initiator for this work. In the SPL, technical challenges cover smaller game independent tasks to test the realizability of concepts and to explore preliminary ideas that might be used to inspire prospective rule changes. By changing the rules every year, the difficulty is increased and the conditions are adapted to reach the level of human soccer as the overarching goal until 2050. According to the current rules of the league [11], whistles are only used for indicating the kick off of a game. However, future considerations exist to use the whistle more frequently to mark game state changes. Therefore, 2019's challenge required participants to localize the position of a whistle blown by a referee with one or multiple robots on the soccer field. Currently, most teams are able to reliably detect whistle sounds. However, the issue occurs that robots detect whistles of neighboring games and therefore start playing ahead of the own game start, which yields a penalty. With increasing importance of the whistle for the communication between the referee and the robots, this becomes an even more pressing issue. By computing the SSL, those whistles that are detected on other fields can be neglected.

As another imaginable case of application, a SSL can be used to improve team communication among the robots. For example, the keeper could use an audio signal to identify itself to other team members. With an SSL in place this signal can be localized, allowing other team members to recognizes if they are about to score an own goal. Potential mistakes and failure of other program components can be corrected by such confirmation behavior between the robots. For this reason, it was taken care of designing an algorithm that is not limited to whistle sounds alone.

1.2 Contribution

In this work, the sixth generation of NAO robots are being used which has four microphones attached to its head. For development of software for this system, an extensive framework is provided by the SPL team *HULKs*, a group of engineering students associated with the Hamburg University of Technology (TUHH). In this software framework, a reliable detector for whistle sounds already exits. This implementation will be utilized for the development of the SSL. The whistle detector of the *HULKs* achieves a true positive rate of 98% and has shown good performance at competitions in the past [12], [13]. In practice, false positives only occur when whistles are blown at other fields as explained above. This emphasizes the importance of a source localization as an useful extension to allow rejection of whistles sounds detected from neighboring fields.

The stages required to realize the SSL can be divided into two steps. First, a single robot Whistle Source Direction Estimation (WSDE) on individual stand-alone systems estimates the relative direction of the sound source by computing the TDOA between multiple microphones. After this, the single robot results are shared among the team via wireless communication in an effort to compute the absolute sound location from the multi-agent data.

As part of this work, three TDOA algorithms are implemented and evaluated to identify an appropriate method to reliably estimate the relative direction on individual robots. The standard Cross Correlation (CC) method is compared with the Generalized Cross Correlation (GCC), as well as a phase difference method which observes the phases of a reference frequency between the microphones. In order to avoid reverberation and observing multi-path propagated signals, focus is laid on the start of the signal which is assumed to be most unaffected by these artefacts. Finally, the objective is to provide a fully functional SSL pipeline within the HULKs framework that can be executed in real-time to estimate the sound location in a global frame using a varying number of robots.

1.3 Outline

This thesis is structured as follows: Chapter 2 covers the theoretical background of this work with regard to the main components of the SSL, namely signal start detection, WSDE and multi-agent filtering; In chapter 3 the implementation details are presented before With the outcome of this, a final conclusion and summary is discussed in chapter 5;

Chapter 2

Prerequisites

For the whistle-sound source localization with multiple robots, some sequential steps needs to interact for a final result. To work with the implementations in chapter 3, the fundamentals are introduced in this chapter on a general basis.

As this work specializes on the localization of a whistle source, this acoustic signal pattern must be detected at first. This work was already done by [12] and thus, the flow of the whistle detection is only briefly explained in section 2.1. For the overall purpose of locating the positions of whistle sound sources, the directions of the whistle sources are obtained first by considering the start samples of the signal. Fundamentals for different approaches of the signal start detection are explained in section 2.2. In this work, the TDOA information between microphone pairs of the robot's head are used to determine a Whistle Source Direction Estimation (WSDE) on each single robot. Section 2.3 introduces the different methods to observe the time delay by different methods in sections 2.4 to 2.6. Due to the low resolution arising from the sample rate and small distance between the microphones, a subsample estimation is stated in section 2.7. After all, the results of the individual robots are filtered by assuming gaussian distribution to produce a sound source position. Section 2.8 describes the formulas for Bayesian Updating in two-dimensional space.

2.1 Whistle Signal

In this work the localization of a whistle-sound source is to be to the fore. Detection of the whistle is done in frequency domain by assuming the whistle sound to be higher than 2kHz and lower than 4kHz. By comparing the mean of the signal between this band with the overall mean of the received signal, a peak arising around the whistle frequency can be detected. For the whistle detection, only one channel of the robot is used and the mean of the whistle band must exceed the threshold multiple cycles in a row. If the team takes action due to the detected signal on individual robots is a team decision.

Further on for this work, the mathematical model of a received whistle signal at one microphone sensor is defined as

$$x_i(t) = s_i(t) + n_i(t) \text{ for } i \in \{0, 1, 2, 3\} \tag{2.1}$$

where $s(t)$ represents the signal and $n(t)$ noise. Both are assumed as real, jointly stationary

random processes.

2.2 Signal Start Detection

One focus of the whistle signal localization is the correct choice of the signal frame, with which the TDOA calculation is done. Assuming that the clearest signal without reverberation and with minimal multipath propagated samples is at the start of a sound signal, the frame to examine is chosen to be at the beginning of a whistle-sound.

By knowing the frequency band of a whistle signal, the start can be detected where these frequencies dominate. Using this indicator only does not always give the desired accuracy, that is why different methods are investigated in this work [14]. In the next subsections, signal start detection using short time energy, zero crossing rate and spectral entropy are subject of discussion. Also, the methods require various computational power. According to the circumstances, the most suitable approach can be chosen. Another point is, that robustness can be increased by considering these methods in combination. As a latter, the consensus of the single methods can be passed as information about the certainty of the computed direction result.

2.2.1 Short Time Energy and Zero Crossing Rate

A common method in signal start and endpoint detection is the evaluation of the short time energy and ZCR.

2.2.1.1 Short Time Energy

The energy

$$E = \sum_{n=1}^{N} E_s(n) \tag{2.2}$$

with the energy spectral density

$$E_s(n) = |x(n)|^2 \tag{2.3}$$

of signal frames with length N are expected to be higher than noise frames and therefore, noise and signal can be distinguished according to [15]. A threshold needs to be specified appropriately dependent on the environment.

2.2.1.2 Zero Crossing Rate

The ZCR of one frame Z needs small computational effort in order to identify a periodic signal in time domain. Its formula is

$$Z = \sum_{n=2}^{N} |sgn(x(n)) - sgn(x(n-1))| \tag{2.4}$$

with the sign function

$$sign(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases}$$

for a discrete signal $x(k)$ of a frame with length N [15]. A higher ZCR is an indication for a periodic signal. To detect the signal start, a threshold is determined dynamically. The ZCR mean of frames which are known to be noise only are averaged with the mean of those frames, that include the whistle signal. The signal start is detected at the point in time, where the ZCR exceeds this threshold.

2.2.2 Spectral Entropy

Entropy provides information about the disorder of a system. From this, one can derive that noise has a high entropy compared to a whistle-sound, which is a highly structured sound signal and a high amount of information accordingly. The spectral entropy of a signal is determined by normalizing the Probability Density Function (PDF) over all frequency components as described in [16]. When $X(n)$ is the Discrete Fourier Transform (DFT) of the sampled signal $x(n)$, the PDF is

$$P(n) = \frac{E_s(n)}{E} \tag{2.5}$$

with eq. (2.3) as the spectral energy density function for $E_s(n)$ and E as the energy. Finally, the spectral entropy results in

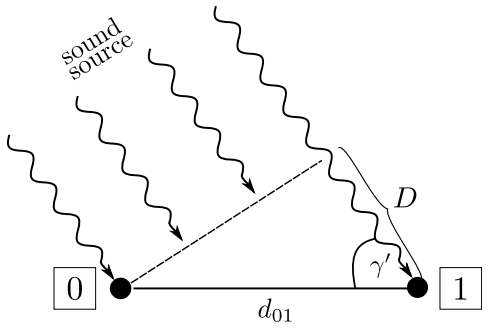
$$H = -\sum_{n=1}^{N} P(n) log_2 P(n). \tag{2.6}$$

Utilizing some a-priori knowledge about the signal, the entropy estimation can be improved. In this work, the frequency of a whistle-sound is known to be between 2kHz and 4kHz from [12], Thus, only the frequency components in the whistle range is considered. Differentiating between noise samples where no signal is present and signal frames, a dynamic threshold can be set to detect the signal start point.

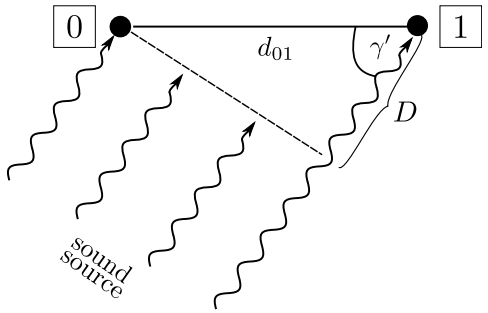
2.3 Time Difference Of Arrival

The direction of a signal source γ' can be determined by the time delay of the received signal. Calculations for the direction of the sound source can be done with a geometrical approach like

in [17]. Figure 2.1 illustrates how a the delay D is introduced by the direction angle of the sound source relative to a vector between channels 0 and 1. Here, the signal arrives first at channel 0. Ideally, the same signal values are measured at channel 1 with delay D. If the delay is zero, the signal is perpendicular to the channels vector. Its value can be D_{max} maximally which in that case delivers the result of the source direction vector being opposite to the *channels vector*. The channels vector is defined as vector directing from one channel to another which in this case starts at channel 0 and points towards channel 1. It is assumed that the distance from the sensors to the sound source is significantly large so that the signal waves proceed parallel which is a necessary criterion for the approach to be valid.



(a) Illustration of TDOA principle where signal comes from front.



(b) Illustration of TDOA principle for the second possible case with equal delay where signal comes from rear.

Figure 2.1: *Picturing both possible sources of the sound when a delay D is measured between channels 0 and 1.*

Specifying the speed of sound c_s being 343m/s in air, the angle γ' can be defined as

$$\gamma' = cos^{-1}\left(\frac{D}{D_{max}}\right) \tag{2.7a}$$

with

$$D_{max} = \frac{f_s \cdot d_{01}}{c_s} \tag{2.7b}$$

where f_s is the sampling rate and d_{01} is the distance between both channels. Not to forget is the ambiguity of the result by observing two channels only. Considering fig. 2.1(a) once more and assuming that the sound source is positioned in front in this case, the same delay can be the result of a sound source from behind. For quick understanding, one can find an illustration of the second possible source directions in fig. 2.1(b).

With the definition of a whistle signal as stated in eq. (2.1), the microphone sensors $channel_0$ and $channel_1$ will output

$$x_0(t) = s(t) + n_0(t) \tag{2.8a}$$
$$x_1(t) = \alpha s(t - D) + n_1(t). \tag{2.8b}$$

Again, D is the delay of x_1 relative to x_0 for what is searched for. As introduced in chapter 1, different methods to detect this delay were implemented and evaluated in this work. In the following sections, the theoretical background of these will be explained in detail.

2.4 Cross Correlation

The CC provides information about the similarity of two signals. Thus, the delay of one signal can be detected where the CC function $R_{x_0x_1}(t)$ is largest. In time domain, the CC of two signals x_0 and x_1 is denoted as

$$R_{x_0x_1}(t) = \int_{-\infty}^{+\infty} x_0(\tau - t)x_1(\tau)d\tau. \tag{2.9}$$

Considering the frequency domain, the function can be transformed into

$$\mathcal{F}[R_{x_0x_1}(t)] = G_{x_0x_1}(f) = X_0^*(f)X_1(f) \tag{2.10}$$

with $\mathcal{F}[x_i(t)] = X_i(f)$ and $X_i^*(f)$ indicating the conjugate complex form. However, the finite observation time of the received signal corrupts the fourier transform [18] and noise of sensors may introduce false peaks in the CC [19]. In frequency domain, the signals $x_0(t)$ and $x_1(t)$ from eq. (2.8) can be expressed as

$$X_0(f) = S(f) + N_0(f) \tag{2.11a}$$
$$X_1(f) = \alpha S(f)e^{-j2\pi fD} + N_1(f). \tag{2.11b}$$

Thus, the CC is

$$G_{x_0x_1}(f) = \alpha|S(f)|^2e^{-j2\pi fD} + N_0^*(f)N_1(f) + S^*(f)N_1(f) + \alpha S(f)e^{-j2\pi fD}N_0^*(f) \tag{2.12a}$$

which will be shortened as

$$G_{x_0x_1}(f) = \alpha\phi_s(f)e^{-j2\pi fD} + \phi_n(f) + \phi_c(f) \tag{2.12b}$$

where

$$\phi_s(f) = |S(f)|^2 \tag{2.12c}$$
$$\phi_n(f) = N_0^*(f)N_1(f) \tag{2.12d}$$
$$\phi_c(f) = S^*(f)N_1(f) + \alpha S(f)e^{-j2\pi fD}N_0^*(f). \tag{2.12e}$$

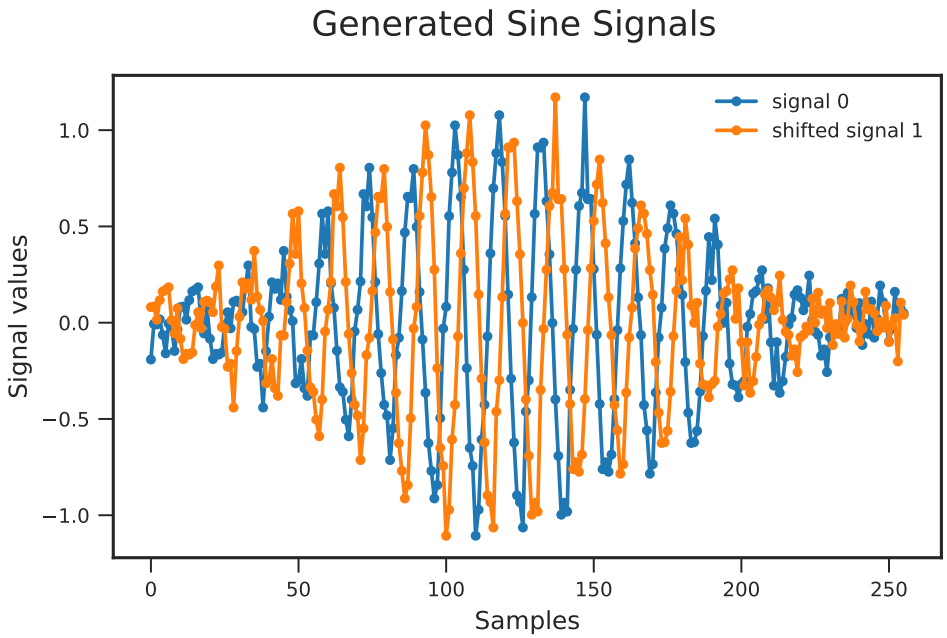
Considering the ideal case where $s(t)$, $n_0(t)$ and $n_1(t)$ are uncorrelated, the terms ϕ_c and ϕ_n disappear and the CC results in

$$R_{x_0x_1}(t) = \mathcal{F}^{-1}[\alpha\phi_s(f)e^{-j2\pi fD}] = \alpha\mathcal{F}^{-1}[\phi_s(f)] * \delta(t - D). \tag{2.13}$$

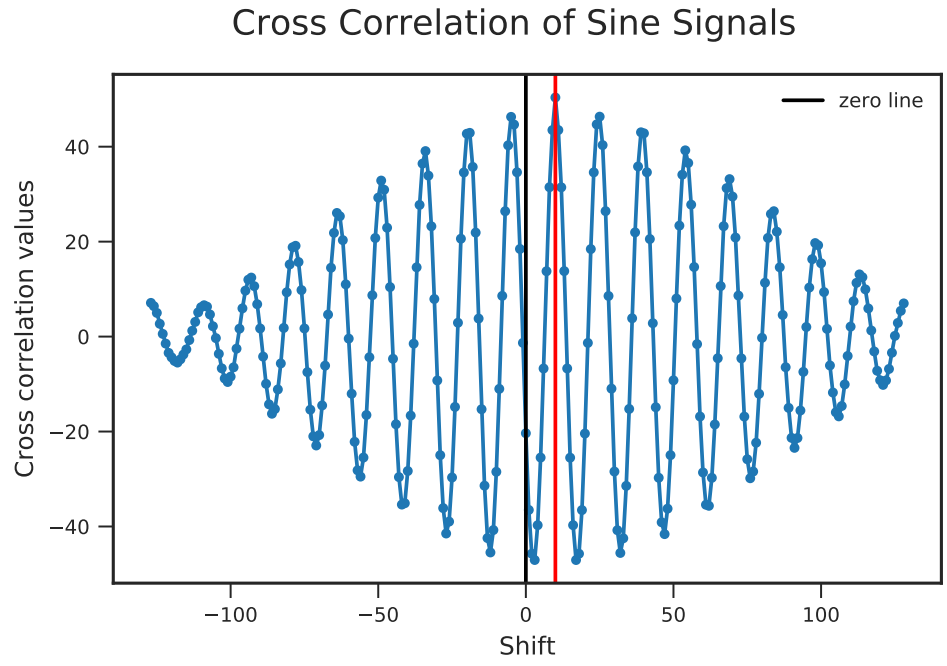
In general, ϕ_c and ϕ_n can neither be neglected nor assumed as uncorrelated to the signal [20], so that they introduce inaccuracies and errors.

As introduced, the CC gives insight about the similarity of two signals and at peak, they are most alike. Received signals from microphone sensors are digital signals sampled with a certain frequency. The derivations are just as applicable, but transformations into frequency domain are done by DFT. In the case of real data samples with length n and similar DFT size, the shift between the zero index and the peak is the resulting delay D. Zero index is defined as the index of the peak if no shift exists.

Figure 2.2(b) is the outcome of two similar, but shifted sine signals with 3kHz and normally distributed noise shown in fig. 2.2(a). As the second signal is delayed by 10 samples, the peak can be detected where $shift = 10$. One disadvantage of this technique is that for periodic signals the CC also is periodic and the peak is not always easily detectable. Noise and inaccuracies of the Fast Fourier Transform (FFT) then may influence the result what can make the peak unobvious [21].



(a) Generated example sine signals with 3kHz after applying Hann-window and sampled with 44.1kHz. *signal 1* is the same signal as *signal 0* but shifted by 10 samples.



(b) Cross correlation of two generated sine signals with 3kHz.

Figure 2.2: *Graph of CC of generated example signal.*

2.5 Generalized Cross Correlation

As discussed in section 2.4, the CC can bring some error sources in the context of incorrect delay results and inaccuracy. Improvements were done in research by introducing prefilters for the signals which is equal to general frequency weighting as stated in [18]. With certain weightings

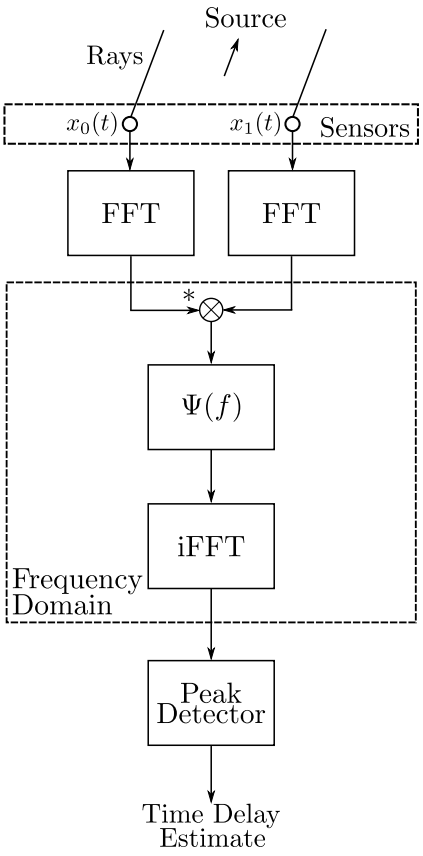


Figure 2.3: *Generalized cross correlation for time delay estimation*

$H_i(f)$ prior to the CC, the peak detection can be rectified by improving the relation between peak and noise or enhancing the accuracy [19]. Figure 2.3 illustrates the process of a GCC with both filters combined as $\Psi(f) = H_0^*(f)H_1$. The figure in appendix A.1 represents the GCC with $H_i(f)$. After transforming the signals $x_i(f)$ into frequency domain, the cross correlated signals are multiplied with the weighting $\Psi(f)$ and transformed back into time domain. The subsequent steps are similar to the CC.

Thus, the GCC is declared as

$$R_{x_0x_1}^{(g)}(t) = \int_{-\infty}^{+\infty} \Psi(f) G_{x_0x_1}(f) e^{j2\pi ft} df. \tag{2.14a}$$

Written-out it is visible how the choice of $\Psi(f)$ impacts the individual segments of eq. (2.12) as

$$\begin{aligned} R_{x_0x_1}^{(g)}(f) = {} & \mathcal{F}^{-1}[\Psi(f)\alpha\phi_s(f)e^{-j2\pi fD}] \\ & + \mathcal{F}^{-1}[\Psi(f)\phi_n(f)] + \mathcal{F}^{-1}[\Psi(f)\phi_c(f)]. \end{aligned} \tag{2.14b}$$

Several variants of the weighting were designed by various researchers with different criteria. They have in common, that they take the characteristics of the received signals into account. Some favor one of both signals, some are designed to suppress the noise and other focus to sharpen the peak as contrasted in [18]. The characteristics of the GCC with Phase Transform (PHAT) most appealed to the task in this work and is chosen as weighting function.

2.5.1 The Phase Transform (PHAT)

The PHAT weighting is known as

$$\Psi^{(P)}(f) = \frac{1}{|G_{x_0x_1}(f)|}. \tag{2.15}$$

For the ideal case that ϕ_n and ϕ_c are nonexistent due to non-correlation, the GCC results in

$$R^{(p)}_{x_0x_1}(t) = \mathcal{F}^{-1}\left[\frac{\alpha|S(f)|^2 e^{-j2\pi fD}}{|G_{x_0x_1}(f)|}\right] = \delta(t-D) \tag{2.16}$$

because $|G_{x_0x_1}(f)| = \alpha|S(f)|^2$. This filter is used regularly in research, due to the characteristic of sharpening the peak what leads in high accuracy [21].

Figure 2.4 demonstrates the result of the GCC-PHAT algorithm with a generated simple Hann-windowed signal as input. Both signals are 3kHz sine signals, whereby the second signal is shifted by 10 samples. The signals are similar to the ones used in section 2.4 and are plotted in fig. 2.2(a). Compared to the CC in section 2.4, the sharp peak is distinct. However, [18] states about the lower robustness of this algorithm.

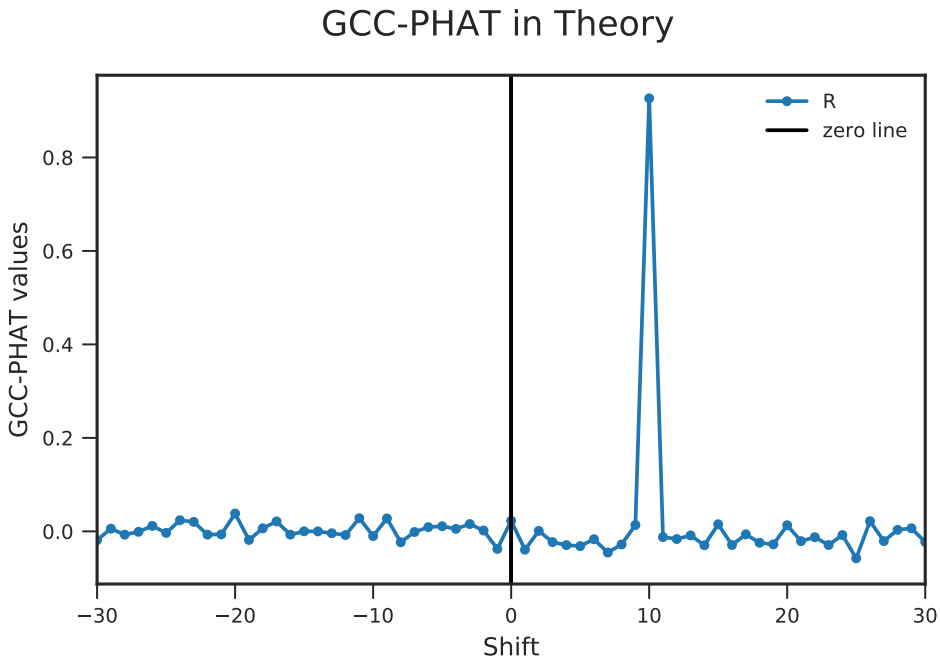


Figure 2.4: *Generalized Cross Correlation with PHAT weighting of two generated 3*kHz *sine signals.*

2.6 Signal Phase Difference

With a different approach to the correlation methods, the TDOA can be detected by observing the phase of one reference frequency f_c. Imaging a single-sinusoidal signal moving from left to right as pictured in fig. 2.5, two distant sensors (*channel 2* and *channel 3*) will receive different

parts of the signal at the same time. Transforming the frames into frequency domain by FFT, the pase of the maximum frequency differ by the delay.

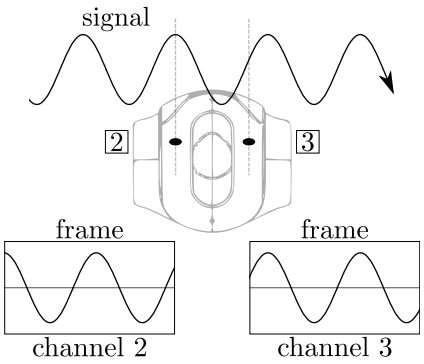


Figure 2.5: *Explanatory illustration of the phase difference method.*

The phase of a signal's reference frequency is easily computable in frequency domain with

$$\phi(f_c) = tan^{-1}\left(\frac{imag(X(f_c))}{real(X(f_c))}\right). \tag{2.17}$$

With the difference of the phases of two channel, the delay in meters is defined as

$$D = \frac{\Delta\phi \cdot c_s}{2\pi \cdot f_c}. \tag{2.18}$$

From that, the direction angle calculation of eq. (2.7a) can be followed. It should be noted that certain requirements needs to be fulfilled to receive a unambiguous result due to signal periodicity. Section 3.2.2.2 covers the conditions that apply for this thesis's hardware.

2.7 Subsample Shift

Considering the case that the sample frequency f_s is set to 44.1kHz and the sound speed is 343m/s, the maximum number of samples between the rear channels is 14. Other neighboring pairs have even less maximum sample differences. This leads to a very low resolution of the direction angle which can be circumvented by either setting a higher resolution or interpolation.

Quadratic interpolation is a well known technique to obtain a floating number shift from a correlation. For this, a parabola $y(x) = a(x-p)^2 + b$ is fitted into the three values of R around the peak of the CC and the peak of the parabola is taken as the more accurate delay. Thus, the subsample delay D_{sub} depends on the maximum value of the correlation y_m and its previous one y_{m-1} and the next value y_{m+1}. Substituting known values and derivations into the parabola function, the subsample delay is defined as

$$D_{sub} = \frac{y_{m-1} - y_{m+1}}{2 \cdot (y_{m-1} - 2y_m + y_{m+1})} \tag{2.19}$$

like in [22]. Figure 2.6 illustrates the CC of two generated sine signals with 3kHz. The second signal is shifted by $\frac{\pi}{3}$ which are 2.449 samples for a sample rate of 44.1kHz. As the plot shows, the peak of the parabola can be determined at an index of 2.446 by quadratic interpolation. In research there are efforts in finding a better approximation function than the quadratic as

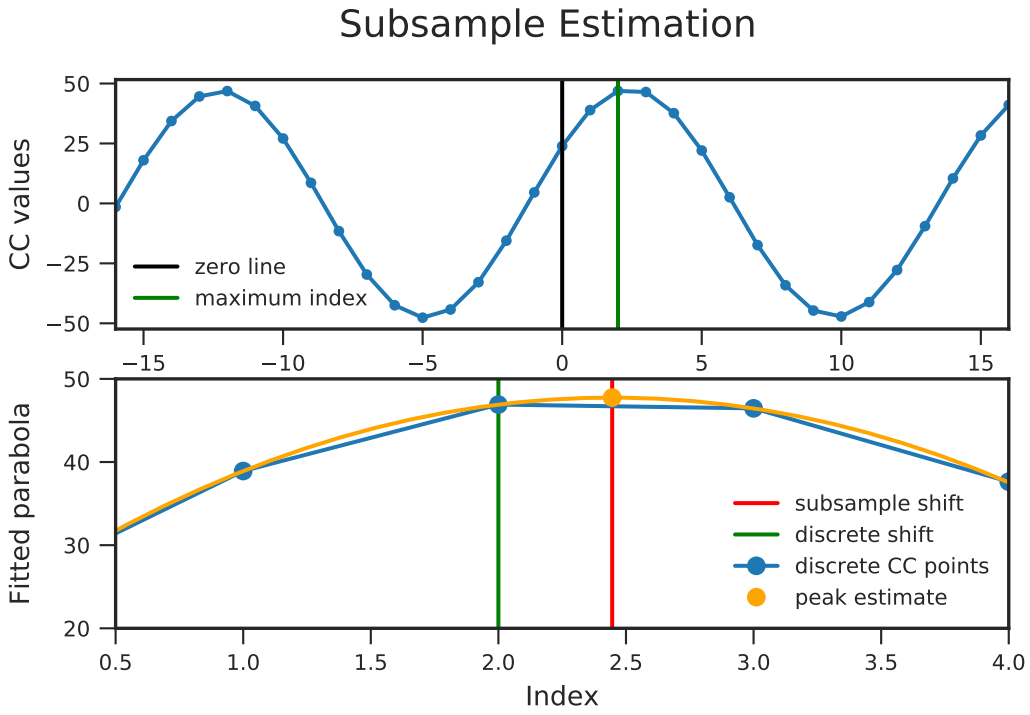


Figure 2.6: *Explanation example of the subsample shift estimation using parabolic interpolation.*

stated in [23] but these are not discussed in greater detail here due to sufficiency.

2.8 Multi-Agent Filter

Assuming gaussian distribution of the single robot results, the multi-agent decision process is done by Bayesian Updating. One dimensional probability density functions of states with variance σ^2 and mean μ are described as

$$\mathcal{N}(x,\sigma,\mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\sigma)^2}{2\sigma^2}}. \tag{2.20}$$

Having two values μ_0 and μ_1 with their variances, the result μ' of the combination of both is

$$\mathcal{N}(x,\sigma',\mu') = \mathcal{N}(x,\sigma'_0,\mu'_0) \cdot \mathcal{N}(x,\sigma'_1,\mu'_1). \tag{2.21}$$

By substitution and conversion, μ' and σ'^2 can be formulated to

$$\mu' = \mu_0 + \frac{\sigma_0^2(\mu_1 - \mu_0)}{\sigma_0^2 + \sigma_1^2} \tag{2.22a}$$

$$\sigma'^2 = \sigma_0^2 - \frac{\sigma_0^4}{\sigma_0^2 + \sigma_1^2}. \tag{2.22b}$$

2.8.1 Two Dimensional Case

Imaging having results from two agents $robot_j$, $robot_k$ that accomplished the WSDE algorithm by computing the TDOA with any method, either the WSDE angles of both robots cross and an intersection point would be found or no final whistle source position result arises. Each of these intersection points $\vec{i_{jk}}$ can be described as whistle source position state $\vec{\mu_{jk}}$ with variance $\boldsymbol{C}_{jk}$ in accordance to eq. (2.20).

Furthermore, the following nomenclature applies:

- $\vec{v}$: vector pointing at $\begin{pmatrix} v_x \\ v_y \end{pmatrix}$ in x- and y-coordinates
- $\boldsymbol{M}$: matrix with elements $\begin{pmatrix} m_{xx} & m_{xy} \\ m_{yx} & m_{yy} \end{pmatrix}$.

Final whistle source position $\vec{\mu}$ of all robots is estimated by iterating over all intersections and updating the result with eq. (2.22).

Exemplary, two robots at position $\vec{p_j}$ and $\vec{p_k}$ with their WSDE results are illustrated in fig. 2.7.

Every WSDE result is represented as *ray* which consists of the robot position $\vec{p_j}$ and the WSDE angle Φ_j, both in field coordinates. The WSDE angle γ_j is defined relative to the robot and the robot's orientation θ_j is known by its team-message information. Thus, the absolute angle is

$$\Phi_j = \theta_j + \gamma_j \tag{2.23}$$

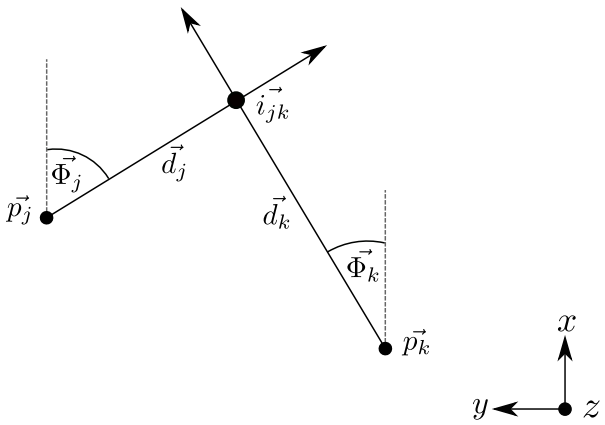


Figure 2.7: *Nomenclature for multi-agent localization algorithm.*

from which the whistle source direction ray can be described as

$$\vec{r_j} = \vec{p_j} + \vec{d_j} = \begin{pmatrix} p_{jx} \\ p_{jy} \end{pmatrix} + \ell \begin{pmatrix} \cos(\Phi_j) \\ \sin(\Phi_j) \end{pmatrix}. \tag{2.24}$$

An position of an intersection point by two rays $\vec{r_j}$ and $\vec{r_k}$ is expressed as

$$\vec{\mu_{jk}} = \begin{pmatrix} \mu_{jkx} \\ \mu_{jky} \end{pmatrix} \tag{2.25}$$

with x- and y-coordinates. If two real numbers u and v exist, so that

$$\vec{\mu_{jk}} = \vec{p_j} + u \cdot \vec{d_j} = \vec{p_k} + v \cdot \vec{d_k} \tag{2.26}$$

a intersection point can be determined by simple geometrical relations. With the given terms and conditions u and v values are computed by dividing the vectors into x- and y-value and solving the equations, resulting in

$$u = \frac{p_{jy} \cdot d_{kx} + d_{ky} \cdot p_{kx} - p_{ky} \cdot d_{kx} - d_{ky} \cdot p_{jx}}{d_{jx} \cdot d_{ky} - d_{jy} \cdot d_{kx}} \tag{2.27}$$

$$v = \frac{p_{jx} + d_{jx} \cdot u - p_{kx}}{d_{kx}}. \tag{2.28}$$

The covariance matrix $\boldsymbol{C}_{jk}$ is obtained in the spirit of an Extended Kalman filter [24] considering the Jacobian matrix of an intersection point $\boldsymbol{J}_i(\Phi_j, \Phi_k)$ and the average angle error ε_Φ over all available measurements. Expressing the direction vector $\vec{d_j}$ of the ray by angle Φ_j as in eq. (2.24), the covariance matrix of the intersection of two rays $\vec{r_j}$ and $\vec{r_k}$ is

$$\boldsymbol{C}_{jk} = \begin{pmatrix} \varepsilon_\Phi & 0 \\ 0 & \varepsilon_\Phi \end{pmatrix} \cdot \boldsymbol{J}_i(\Phi_j, \Phi_k) = \begin{pmatrix} \sigma_{xx}^2 & 0 \\ 0 & \sigma_{yy}^2 \end{pmatrix} \tag{2.29a}$$

with the Jacobian matrix

$$\boldsymbol{J}_i(\Phi_j, \Phi_k) = \begin{pmatrix} \frac{\partial \mu_x}{\partial \Phi_j} & \frac{\partial \mu_x}{\partial \Phi_k} \\ \frac{\partial \mu_y}{\partial \Phi_j} & \frac{\partial \mu_y}{\partial \Phi_k} \end{pmatrix}. \tag{2.29b}$$

The elements of the Jacobian result by derivation of the intersection formula eq. (2.26) with eqs. (2.24) and (2.27) which yield

$$\begin{aligned}
\frac{\partial \mu_x}{\partial \Phi_j} &= \frac{(\cos(\Phi_k) \cdot (\cos(\Phi_j)^2 + \sin(\Phi_j)^2) \cdot ((p_{jy} - p_{ky}) \cdot \cos(\Phi_k) + (-p_{jx} + p_{kx}) \cdot \sin(\Phi_k)))}{(-\cos(\Phi_k) \cdot \sin(\Phi_j) + \cos(\Phi_j) \cdot \sin(\Phi_k))^2} \\
\frac{\partial \mu_x}{\partial \Phi_k} &= \frac{-(\cos(\Phi_j) \cdot (\cos(\Phi_k)^2 + \sin(\Phi_k)^2) \cdot ((p_{jy} - p_{ky}) \cdot \cos(\Phi_j) + (-p_{jx} + p_{kx}) \cdot \sin(\Phi_j)))}{(\cos(\Phi_j) \cdot \sin(\Phi_k) - \cos(\Phi_k) \cdot \sin(\Phi_j))^2} \\
\frac{\partial \mu_y}{\partial \Phi_j} &= \frac{(\cos(\Phi_j)^2 + \sin(\Phi_j)^2) \cdot \sin(\Phi_k) \cdot ((p_{jy} - p_{ky}) \cdot \cos(\Phi_k) + (-p_{jx} + p_{kx}) \cdot \sin(\Phi_k))}{(-\cos(\Phi_k) \cdot \sin(\Phi_j) + \cos(\Phi_j) \cdot \sin(\Phi_k))^2} \\
\frac{\partial \mu_y}{\partial \Phi_k} &= \frac{(\cos(\Phi_k)^2 + \sin(\Phi_k)^2) \cdot \sin(\Phi_j) \cdot ((-p_{jy} + p_{ky}) \cdot \cos(\Phi_j) + (p_{jx} - p_{kx}) \cdot \sin(\Phi_j))}{(\cos(\Phi_j) \cdot \sin(\Phi_k) - \cos(\Phi_k) \cdot \sin(\Phi_j))^2}.
\end{aligned}$$

The one-dimensional formulation eq. (2.22) can be rewritten by defining an updating matrix $\boldsymbol{K}$ as combination of present and incoming covariance

$$\boldsymbol{K} = \frac{\boldsymbol{C}}{\boldsymbol{C} + \boldsymbol{C}_{jk}}. \tag{2.30}$$

It follows the new two-dimensional probability density function for the estimated whistle position with

$$\vec{\mu}' = \vec{\mu} + \boldsymbol{K} \cdot (\vec{\mu_{jk}} - \vec{\mu}) \tag{2.31a}$$

being the latest state along with

$$\boldsymbol{C}' = \boldsymbol{C} = \boldsymbol{C} - \boldsymbol{K} \cdot \boldsymbol{C} \tag{2.31b}$$

as its updated covariance matrix.

Chapter 3

Implementation

In this chapter, the implementation details for the whistle-sound localization are explained briefly. All audio data used in this work are recorded within the HULKs' framework on the NAO robots. Prior to realizing the sound localization on the hardware of NAO robots, a large part of the algorithms were implemented in Python 3.7.4 independently for easier debugging and evaluation.

3.1 NAO Framework

As previously stated, a complete code frame work by the SPL teamHULKs exists to address the NAO robots' hardware in C++. It is implemented in such manner that *Modules* waits for all input values (*Dependencies*) which are necessary to execute their task. Modules can be seen as larger coherent operations that perform specific roles. With the output produced by this module (*Productions*), successive Modules can start to perform their cycle. Due to this sequential process, modules are modifiable independently.

The robots are able to communicate wirelessly via User Datagram Protocol (UDP) to operate as multi-agent system. In the team-message protocol, information about the robot states and knowledge like the believed ball position is exchanged with team mates. By the SPL rules [11] they are limited to one team-message per robot per second. However, for the whistle localization challenge, no particular specification about the communication was defined [25]. In order to localize the sound source with multiple robots, results of multiple single robot whistle localizations are forwarded to the protocol.

3.1.1 Coordinate Systems

Two coordinate systems are introduced to consider the whistle-sound position. One for the relative orientation of the source to the robot's head and another for the absolute position determination in field coordinates.

Figure 3.1 visualizes the coordinate system of the robot. Direction angles γ are specified around the z-axis in mathematically positive direction and range between $-\pi$ and π. The direction

where the whistle source is believed at by a single robot will be described as γ.

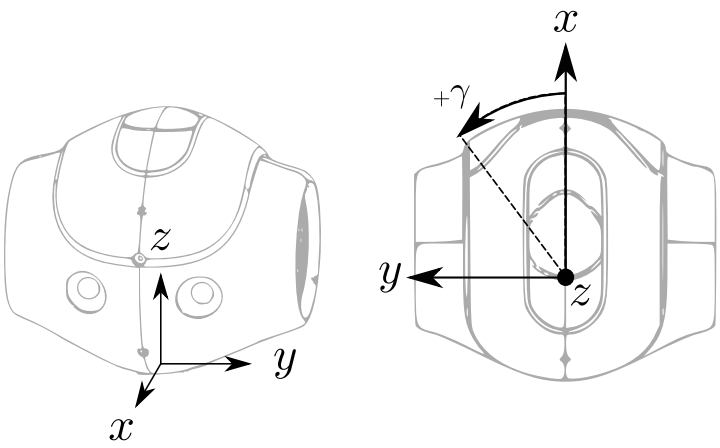


Figure 3.1: *Coordinate System of NAO's head.*

For field coordinates, only the planar case matters in this work. It is defined as illsutrated in fig. 3.2. In game, the robots play into x-direction. The whistle position of the team filter will be described in this coordinate system.

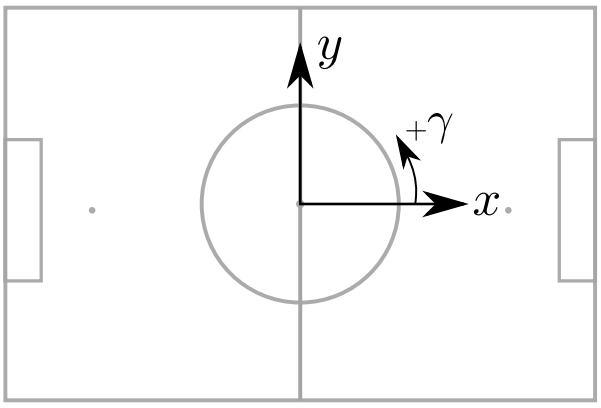


Figure 3.2: *Field Coordinate System.*

3.1.2 Microphones

Four microphones are attached on the NAO's head. Their positions are listed in table 3.1 with respect to the head's coordinate system introduced in section 3.1.1 [26] and outlined in fig. 3.3. As per documentation [1], the operating frequency range is between 150Hz and 12kHz. The sensitivity is indicated with 20mV/Papm3dB at 1kHz.

Prior to this work, there was no need to address more than one microphone on each robot. Thus, receiving data from multiple microphone channels was implemented into the audio interface of the HULKs' framework. Raw microphone data of all channels is captured in interleaved format and unpacked by the `AudioReceiver` module which makes the samples accessible per channel for further usage. The index number of the channels is derived from the order of the interleaved data format straightforwardly. The sampling frequency f_s of the microphones is set to 4.1kHz.

Channel	x [m]	y [m]	z [m]
0	-0.0215	0.0558	0.0774
1	-0.0215	-0.0558	0.0774
2	0.0206	0.0309	0.0986
3	0.0206	-0.0309	0.0986

Table 3.1: *Positions of the microphones on the NAO's head.*

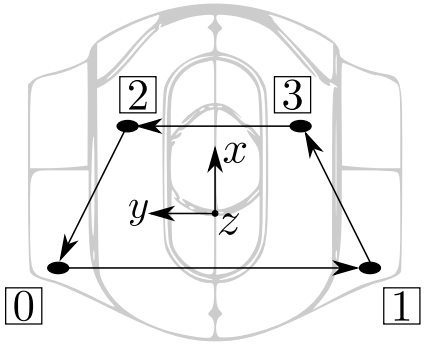


Figure 3.3: *Microphone positions on NAO's head.*

In order to determine the direction of the sound source, the TDOA between two channels is observed in this work. Hereinafter, the right adjacent channel of the *base channel* is defined as *next channel*. This relation is pictured with arrows in fig. 3.3 where the arrowhead points towards the next channel. From the implementation point of view, this order makes the most sense when iterating over all channels.

Base channel	Next channel
0	1
1	3
2	0
3	2

Table 3.2: *Definition of* ***next channel*** *in respect of* ***base channel****.*

Advanced Linux Sound Architecture (ALSA)

Among the addressing of multiple microphones, the audio interface was reworked and utilizes the open source library Application Programming Interface (API) of Advanced Linux Sound Architecture (ALSA) [27]. Using this library, configuration for connecting to the microphones can be realized as desired. Here, the *access type* is set to interleaved format. The data format is set to float 32bit as well as the sampling rate to 4.1kHz.

3.1.3 Existing Whistle Detection

Since the whistle detection was already present before this work, it is introduced only in a short manner. The existing whistle detection is based on the implementation of the SPL team University of New South Wales (UNSW). After 1024 samples are collected in the buffer, these samples are Hann-windowed and then transformed into frequency domain using FFT. A threshold is defined as the mean of the spetrum magnitude, multiplied by a factor of 1.3. Dividing the frequency band between 2kHz and 4kHz into multiple parts, the frequency band is narrowed from both sides until the mean value of one part exceeds the threshold per side. If the mean of the remaining frequency band is larger than a second threshold which is larger than 2.5-times the overall mean, a whistle is detected in this frame. Both factors are parameterized. Whistles need to be detected in at least three frames within the last four cycles until the `whistleFound` state is set to true. Both values are parameterized.

By default, only channel 0 which is located rear left is used for he whistle detection. All transformations into frequency domain are executed with the C++ library Fastest Fourier Transform in the West (FFTW). This open source library provides most common functions for operations in frequency domain and is widely used?.

3.1.4 Whistle Localization Structure

The procedure of the whistle localization in the context of the existing NAO framework can be briefly outlined by the following steps:

1. **Buffer**: the `WhistleDirectionEstimation` module saves microphone data until whistle is detected by the `WhistleDetection` module.
2. **Whistle Source Direction Estimation**: the `WhistleDirectionEstimation` performs the sound source localization algorithm and outputs a believed direction angle and additional information if a whistle was detected. The signal start detection is part of this module.
3. **Send Direction via Team Message**: angular direction result plus additional information are sent to other robots by team-message.
4. **Multi-Agent Whistle Localization**: wait for results of all agents to be present and filter direction information for a final sound position.

3.2 Whistle Source Direction Estimation

Considering only one stand-alone system, i.e. one robot with four microphones, only the direction of a sound source can be estimated. As shortly introduced in section 3.1.4, the `WhistleDirectionEstimation` module is responsible to determine the direction of the whistle-sound detected by the `WhistleDetection`. Before going into detail, the overall structure of the WSDE on a single robot is summarized by means of fig. 3.4.

`WhistleDirectionEstimation` depends on the data type `WhistleData` which is produced by the `WhistleDetection` module and contains two values. A timestamp when the whistle

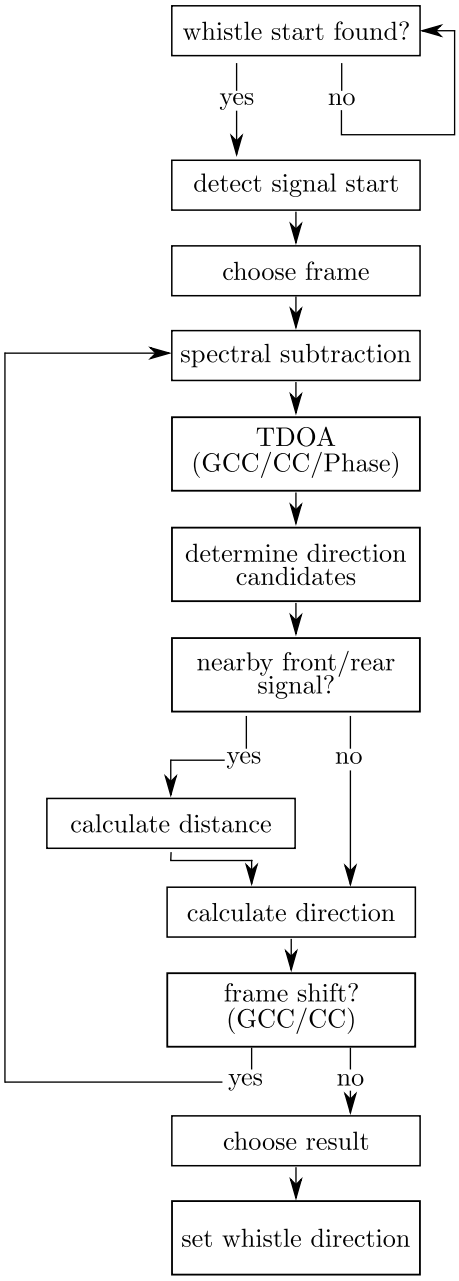


Figure 3.4: *Concept of whistle localization on single robot.*

was last detected and a boolean value if a whistle start is found. Until this value is set to true by the `WhistleDetection`, the `WhistleDirectionEstimation` buffers up to 44100 audio samples per channel. As a whistle needs to be detected for multiple cycles in a short time, one can be sure that the buffer contains a number of whistle samples.

Out of the buffered samples, the signal start is calculated according to the selected signal start detection which will be further described in section 3.2.1.

Frames considered for the delay estimation are chosen in a different way depending on the TDOA method. Detailed descriptions are given in the corresponding method sections 3.2.2.1 and 3.2.2.2. Using TDOA, the delays between two microphone channels are either determined by CC, GCC-PHAT or the phase difference method. As previously stated, there exist various

GCC filter. However, in this work by GCC the GCC-PHAT method is always referenced if not explicitly specified differently. Each delay generates two potential source direction candidates as stated in section 2.3. In the event of candidates indicating a nearby signal from straight forward or backwards, the distance to the source is estimated in addition. More precise explanation is given in section 3.2.4. For a final direction, the mean of the candidates with smallest difference is formed. This means that each candidate of a channel pair is compared to the other resulting candidates of the remaining channel combinations. The combination with the smallest sum of angular error between the candidates is selected as Whistle Source Direction Estimation.

If one of the correlation methods is used, the TDOA is calculated multiple times by shifting the sample selection window to find the most suitable frame for the TDOA estimation. Both the shift range around the start index, as well as the size of the shift are parameterized values. This way, potential start estimation inaccuracies can be corrected and the decision process is optimized. In case of the GCC-PHAT, the PSNR provides information about the certainty of the TDOA estimation what is shown in section 4.3.3. A computation of one frame delivers TDOA values for each channel pair and a mean of all PSNRs from the GCC functions. Comparing the frame shifts, the TDOA results with greatest PSNR mean value is assumed as best performing. The same procedure is done with the CC method but by examining the maximum CC function value.

Regardless of the method, the production of this module is a `WhistleDirection` data type which contains calculated direction outcome in radians and additional information like distance or PSNR.

All transformations into frequency domain and inverse transformations are executed with the FFTW library just as the `WhistleDetection`. For parallel development in Python, the widespread package *NumPy 1.17.0* delivers all fundamental functions for computation in frequency domain.

3.2.1 Signal Start Detection

As mentioned in section 2.2, the detection of the signal start is crucial for the localization. Tests showed that the accuracy of the whistle detection is not sufficient to be used stand alone concerning the start index. Thus, different approaches are evaluated with regard to higher precision and lower computational effort.

The implementation of the different approaches will be presented coupled with an examination of real measurement data. To reduce undesirable effects and demonstrate the simplest form, a sinusoidal signal of 3kHz was recorded with the same circumstances as the whistle-sounds. For the following data, the sound source was placed 2m in front of the robot. In order to find the time point where the signal starts, information about smaller fractions are required. So, the original 44100 samples that were buffered by the `WhistleLocalization` module are divided into several overlapping frames with size 256. The computational effort rises with smaller frame size, but delivers a higher precision in return. In order to perform the FFT most efficiently, the size of one frame should be a power of 2. To compute the entropy, the frames are transformed into frequency domain with the FFT. The ZCR does not require such a transformation. For the reason that the start detection through analyzing the energy requires a fixed threshold which varies depending on the environment strongly, it will not be covered with regard to the signal start detection. In the evaluation section 4.1, the result of the single methods are compared to

each other. For better visualization, the following data is shortened to 2400 samples.

Spectral Entropy

The formula to calculate the spectral entropy of a signal is introduced as eq. (2.6) in the previous chapter. For the entropy information, the signal should not be cleaned previously. The point at which the signal changes from noise to whistle signal was found by computing the finite differences between the sampling points of the entropy. This approach is not real-time capable but introduces a small delay. Figure 3.5 is a plot of the recorded sine signal with the corresponding entropy. According to the frame size, the accuracy of the start index can be increased. However, the frame size is limited by the required number of samples for one FFT and its computational effort. In section 4.1.3, the entropy outcome of a whistle-sound and its derivation is presented for evaluation.

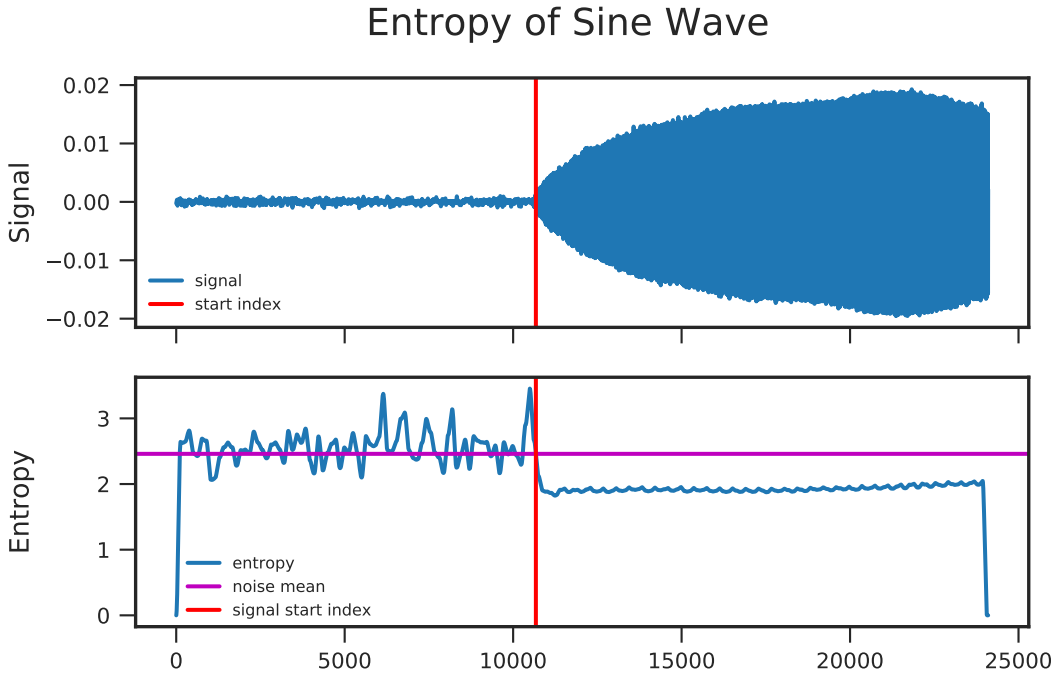


Figure 3.5: *Exemplary entropy of a sinusoidal signal with 3*kHz*.*

Zero Crossing Rate

Alike the other methods, the buffered signal is divided into smaller frames which can be set arbitrarily small for the ZCR where no FFT is necessary. In each frame, the sign changes are counted which only requires simple implementation and is computationally lightweight. An assumption in this work is that the signal is represented at the end of the received data and that only noise exists at the beginning. Using this condition, the ZCR of the first few frames is averaged and defined as noise ZCR mean value. On the other hand, the same can be done with frames at the end of the data to provide a signal ZCR mean. By comparing both means, a dynamic threshold can be defined which can be adjusted depending on the circumstances

optionally. Better results could be obtained by reversing the signal and determining the index where the signal fell below the threshold. Number of noise and signal frames are parameter values which depend on the amount of samples and the size of the frame.

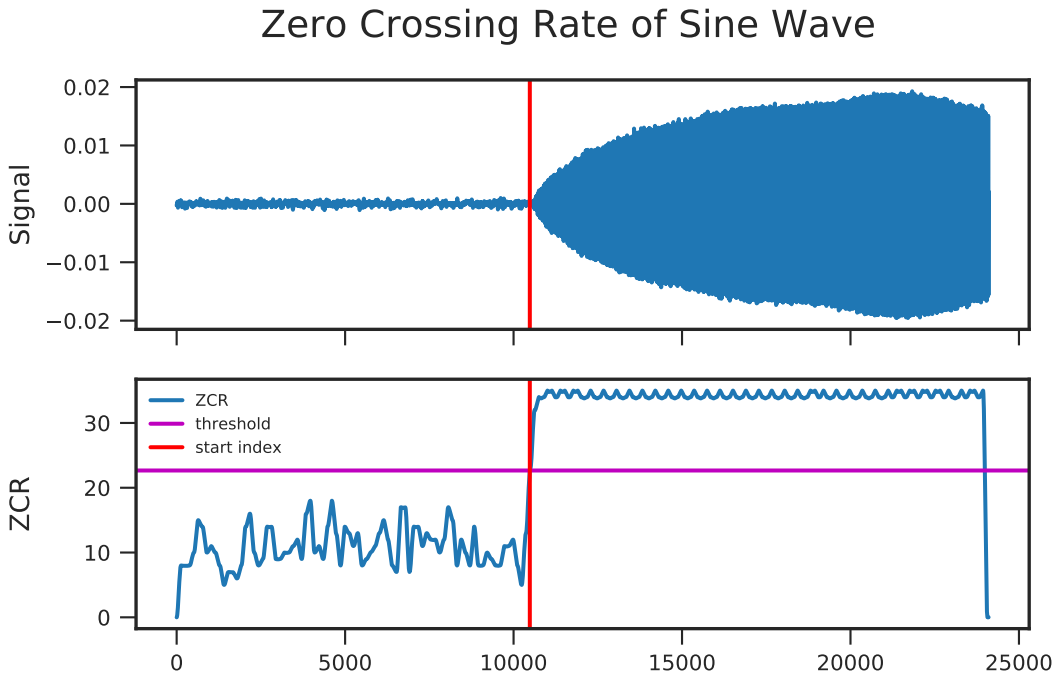


Figure 3.6: *Zero Crossing Rate of a sinusoidal signal with 3*kHz.

3.2.2 Time Difference of Arrival

The TDOA estimation is the main component to identify the whistle source location. Theoretical background to this approach of source localization is given in section 2.3. As stated there, the TDOA of a signal measured between two microphone sensors provides details about the direction of the source. Having four channels attached on a NAO's head, an overdetermined system it given where each channel pair provides TDOA information.

The GCC-PHAT method is a modification of the CC method. Due to their implementation being equal except of the weighting function, they are discussed in section 3.2.2.1 collectively. Section 3.2.2.2 presents the implementation details with its circumstances.

3.2.2.1 Correlation

In theory, Cross Correlation (CC) in time domain is usually illustrated by shifting two signals over each other and recording the similarity for each shift. Similarity in terms of signal processing is measured by the area under the curve of addition. Thus, a peak will arise at that shift where signals are most similar. Having two equal signals, the maximum value of the CC `R` (which in this case is called auto-correlation) arises at the middle of the function. This index in this case is called `zeroIndex` and calculated with `int(length(R))-1`. If one signal is alike

the other but delayed by D samples, the peak will occur at a shift of D next to the `zeroIndex`. This delay, which is directly related to the TDOA, is computed by the CC and GCC in the unit of samples. As sections 2.4 and 2.5 have shown, the GCC is commonly performed in frequency domain due to a performance advantage. For unification, both CC and GCC are implemented in frequency domain.

The samples for the CC are defined by the start index and the frame shift according to section 3.2 and originate from the data which was cleaned by spectral subtraction previously. The frame size in this work is set to 256 samples Hann-windowed prior to the correlation. By zero padding the FFT, resolution can be increased, but is refrained from for now due to higher computational effort and sufficient CC results. For two real signals, the CC can be realized by time-reversing one of the signals first. After transforming both signals into frequency domain by FFT, element wise multiplication is performed.

In the case of GCC-PHAT, each component of the multiplication is divided by the absolute value as the weighting function eq. (2.15) defines. After this, the cross-correlated signal is transformed back into time domain and index of the peak `peakIndex` is found. The delay in samples is then computed by `peakIndex - zeroIndex`. In conformity with the definitions in this thesis, a positive delay `d_01` between `x_0` (signal at channel 0) and `x_1` (signal at channel 1) indicates that the signal was received at channel 0 first.

Subsample Delay

Integer delays only offers a limited number of resulting direction angles. As demonstration, considering the distance between the two rear microphones on the robot's head the maximum sample difference obtains 14 samples. This means that 180° are separated in 14 directions only, whereby this represents the largest distance between neighboring channels. To avoid this low resolution, the subsample shift estimation as in section 2.7 is added to the delay estimation for both CC and GCC. Although the implementation is simple, it yields promising results.

3.2.2.2 Phase Difference

In contrast to the previous methods, this method compares the phases of a specific reference frequency per channel. Section 2.6 covers the theoretical background of this approach. The circumstances that apply on the NAOs and the appropriate implementation is presented here. In order to calculate the phase difference ϕ between two frames, the reference frequency f_c needs to be defined. One can either choose a static reference frequency equal for all measurements or set the reference frequency dynamically according to the sound signal. Both implementations were realized in this work for evaluation with following conditions:

- The reference frequency must be within whistle frequency range (2kHz to 4kHz as proposed in [12]).
- The maximum phase difference between two channels must not overflow π with the selected reference frequency.

Meeting these conditions, a signed phase difference is ascertainable that indicates a distinct direction. **??** lists the maximal feasible reference frequencies (Max. Frequency) meeting the

second requirement. As the table presents, the distance between channel 0 and 1 is too large so that the reference frequency would need to be smaller than 2kHz. For this reason, the phase difference information between this pair is neglected.

Channel Pairs	Absolute Distance [m]	Max. Frequency [Hz]
0 and 1	0,116	1536.75
1 and 3	0,0533	3217.11
2 and 0	0,0533	3217.11
2 and 3	0.0618	2775.08

Table 3.3: *Most feasible frequencies for unambiguous phase difference detection.*

For compatibility with the correlation method implementation, the phase difference is converted to delay samples D_s with

$$D_s = \frac{f_s \cdot \Delta\phi}{2\pi \cdot f_c}. \tag{3.1}$$

Static Reference Frequency

Considering the sensor pair of the two front channels, the maximal feasible reference frequency value is limited to 2775.08Hz as stated in the table 3.3 above. Preliminary tests results have shown, that the quality of the phase difference method relies on the reference frequency to some extent. To realize best performance, a reference frequency between 2.6kHz 2.775kHz should be chosen. The frequency resolution of FFTs is $\frac{f_s}{N}$, depending on the sample frequency f_s and the number of data points N. As the frame size for this method is set to 64 samples, the samples are zero-padded up to 512 samples before transforming them into frequency domain for higher resolution. With this FFT size, f_c can be selected between 2670.1Hz and 2756.25Hz. Frames after the signal start index are chosen, where a whistle is detected in all channels by the same algorithm.

Dynamic Reference Frequency

To define a dynamic reference frequency instead, the frame is chosen by doing a frequency analysis on multiple frames after the start index. The frequency f_m belonging to the maximum magnitude of the signal spectrum is determined. The first frame is chosen where f_m is equal for all channels and within whistle range. For better understanding, the f_m values per frame are plotted in fig. 3.7. Here, one sees that the conditions apply at the beginning of the whistle signal after frame number 174.

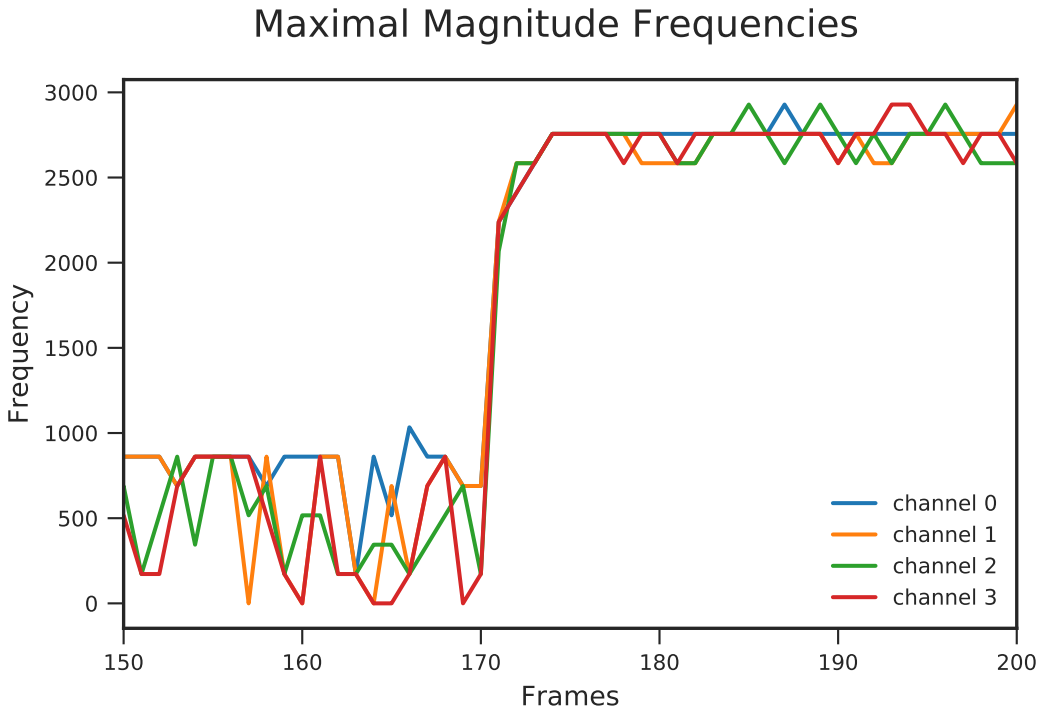


Figure 3.7: *Frequencies of maximum magnitude in signal spectrum per frame.*

3.2.3 Direction Estimation

Delay samples are computed by the TDOA methods introduced in the previous sections. Using eq. (2.7a), one positive and one negative signed angle γ' arise relative to the vector between the channels. Figure 3.8 is used to illustrate the terms and visualize the circumstances for better understanding. The definition of *base channel* and *next channel* stays as introduced in section 3.1.2. Positive delay between two channels implies that a signal source is closer to the base channel than the next channel. For example if the detected delay has the same value as the maximum possible number of samples between those channels, the source direction is equal to the *max-delay vector* in the figure. In this case, γ' is zero. For all smaller delays greater than zero, two direction candidates *candidate 0* and *candidate 1* result in the range of the max-delay vector $\pm\pi$. The same applies mirrored for negative delays.

By this implementation, all candidates are represented in the robot coordinate system. Having four channels, each neighboring channel pair returns two candidate directions. Diagonal channels can be paired as well for the correlation methods. However, this case is not considered profoundly in this work due to the overdetermined system by four pairs. Out of these eight candidates, a final direction angle γ is chosen by computing all combinations and selecting the one with smallest sum of angle difference. During research, different factors like signal strength or smallest start index were tested to include more a-priori knowledge. Due to lack of reliability, no additional signal properties are taken into account.

There exists one exceptional case, where the signal source is detected straight in front or behind the robot and distance can be estimated. If this is the case, the direction angle is corrected to 0 or π. How the distance information is handled is content of the next section.

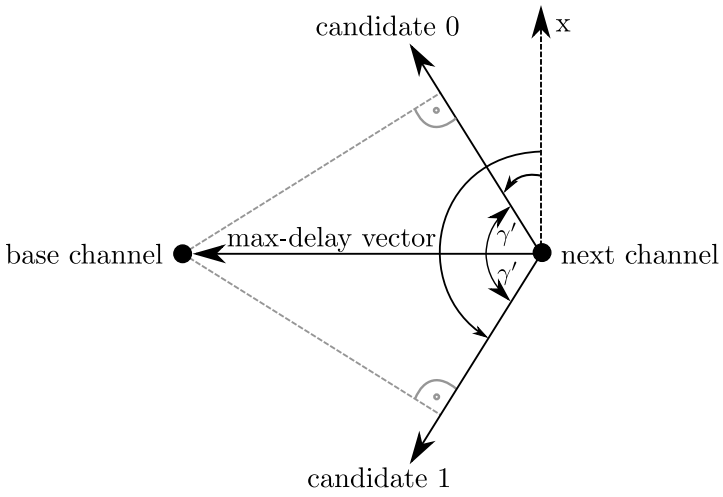


Figure 3.8: *Illustration of the resulting candidates of TDOA implementation.*

3.2.4 Front and Rear Distance

For the particular case when the signal comes straight from the front or behind, the distance of the sound source can be approximated up to a certain range. If $delay_{01}$ (delay between channel 0 and 1) and $delay_{32}$ (delay between channel 3 and 2) are both very small, the signal is detected from the front or rear. Assuming signal waves in one horizontal plane and arriving parallelly, the lateral delays would be larger or equal the maximal sample difference of 5.41 between the front and rear channels on the x-axis. For the case that these lateral delays are smaller, the angle of the sound source in the XZ plane identifiable. Figure 3.9 illustrates the NAO's head from the right with channels 1 and 3.

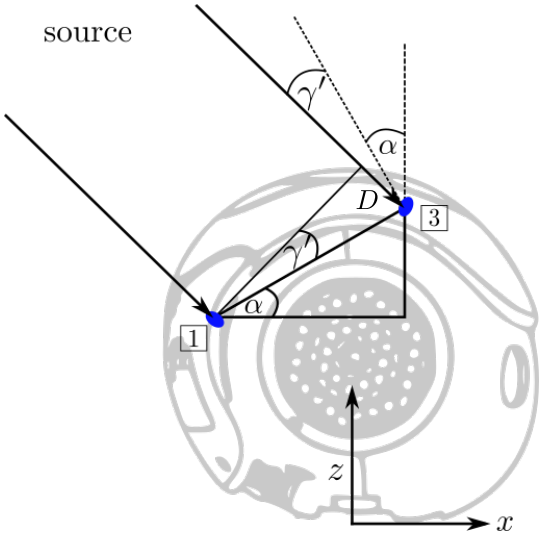


Figure 3.9: *Illustration of arriving sound for sources from near behind. Adapted from [1].*

Assuming, that $delay_{01}$ and $delay_{32}$ are small, the angle of the sound source γ relative to the

Z-axis can be determined with delay D as

$$\gamma = \alpha + \gamma' \tag{3.2a}$$

$$\gamma' = sign(D) \cdot sin^{-1}\left(\frac{D}{D_{max}}\right) \tag{3.2b}$$

whereby

$$\alpha = tan^{-1}\left(\frac{\Delta z_{channel}}{\Delta x_{channel}}\right) \approx 26.73° \tag{3.2c}$$

which is the angle of the orthogonal axis of the plane though front and rear channels.

$\Delta z_{channel}$ and $\Delta x_{channel}$ are the distances between the channels in z- and x-direction in m.

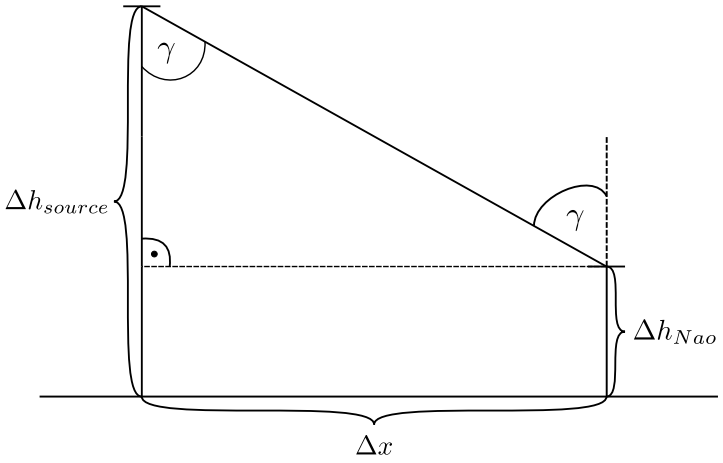


Figure 3.10: *Illustration of distance estimation.*

Knowing that the sound source is ordinarily positioned above of the robot, the distance of the sound source can be approximated with an assumed height Δh_{source} of the source. Δh_{source} differs from referee and thus, can only be guessed. So, the distance in x-direction Δx is

$$\Delta x = (\Delta h_{source} - \Delta h_{NAO}) \cdot \tan(\gamma). \tag{3.3}$$

The distance estimation is triggered, if the direction candidates of $delay_{01}$ and $delay_{32}$ are smaller than ± 10° or larger than ± 170°. Additionally, the lateral delays must be smaller than 5.41 samples as mentioned above.

Restrictions of the front and rear distance measurement differ. For the front case, the maximum angle for an unambiguous distance calculation is $\frac{\pi}{2} - 2\alpha$. Thus, the maximum front distance that can be approximated shrinks to $\Delta x = (\Delta h_{source} - \Delta h_{NAO}) \cdot \tan(\frac{\pi}{2} - 2\alpha)$ according to eq. (3.3). To the rear, the maximal value for γ is bounded by the 5.41 samples that are set as condition. Setting Δh_{source} to 1.5m and Δh_{NAO} to 0.57m for example, the maximal measurable distance to the front is about 0.66m. With the same values the maximal distance backwards is 15.3m in theory. However, measurements in section 4.3.1 show that 7m is the limit for the real case. The difference between front and rear occurs due to the constructional positioning of the microphones being tilted to the back.

3.2.5 SNR

The Signal to Noise Ratio (SNR) is a common value to express the signal power P_{signal} compared to power of the background noise P_{noise}. Conveniently, the buffered audio signal in this thesis always consists of a clean-cut delimitation between signal and noise which is set by the start index. Thus, the SNR which is defined as

$$SNR_{db} = 10 \log_{10} \left(\frac{P_{signal} - P_{noise}}{P_{noise}} \right) \tag{3.4}$$

in decibels can be implemented straightforwardly. Informational content about this measure is investigated in section 4.3.2. Expectations are that the SNR can be fed into the covariance matrix of an incoming result in the Bayesian update process introduced in section 2.8.1.

3.2.6 PSNR

In image processing, the Peak Signal to Noise Ratio (PSNR) indicates the quality of a compressed image. Here in this work, the ratio between the peak of a signal and its noise is related to the GCC-PHAT outcome and called PSNR henceforth. As stated in section 2.5, the most significant characteristic of the GCC-PHAT is the resulting sharp peak which now can be assessed with one value

$$PSNR_{db} = 10 \log_{10} \left(\frac{P_{peak}}{P_{noise}} \right). \tag{3.5}$$

From the implementation view, the power of the correlation peak is divided by the power of the remaining correlation signal. It has to be noted that two adjacent values prior and after the peak are disregarded as they might belong to the peak. A validation if and how much the PSNR and the accuracy of the GCC delay result are linked is done in section 4.3.3.

3.3 Multi-Agent Source Localization

The final whistle source localization is realized by collecting the all WSDE results of all active robots and combining the directions into one xy-position outcome. Two separate implementation details are covered in the following. First, the actual localization algorithm that concludes the source localization task is set in focus. How the process of the multi-agent decision proceeds including the communication is dealt with in section 3.3.1.

As stated, the actual Bayesian updating filter algorithm is addressed first by paying attention to algorithm 1. With the specification of the filter made in section 2.8.1, the whistle source localization is realized when the results of the individual robots are available. The multi-agent localization algorithm receives an array of rays that represent the WSDE results in the field coordinate system. The whistle-sound position to return consists of the x- and y-position $\vec{\mu}$ and a covariance matrix $\boldsymbol{C}$. For each combination of the rays $R[j]$ and $R[k]$, the intersecting point position $\vec{\mu}_{jk}$ along with its covariance matrix $\boldsymbol{C}_{jk}$ is examined. If such intersection exists, the whistle source position information is updated until all intersections are taken into consideration.

Algorithm 1 Bayesian Updating

```
1:  procedure WhistleLocalization(array < R >)
2:      for j ∈ 0 : R.size() do
3:          for k ∈ j + 1 : R.size() do
4:              μ⃗_jk ← FindIntersection(R[j], R[k])
5:              C_jk ← CalulateCovariance(R[j], R[k])
6:              if first intersection found then
7:                  μ⃗ ← μ⃗_jk
8:                  C ← C_jk
9:              else if intersection found then
10:                 S_jk ← C + C_jk
11:                 K_jk ← C · S_jk^-1
12:                 μ⃗ ← μ⃗ + K_jk · (μ⃗_jk − μ⃗)
13:                 C ← C − K_jk · C
14:     return μ⃗
```

3.3.1 Team Communication

To agree on a whistle position as a multi-agent system by computing the whistle localization process of algorithm 1, the WSDE results of the stand-alone robots must be collected. As introduced in section 3.1, the robots are able to send and receive information from team mates wirelessly. By present implementation, the team-message includes a robot's xy-position information, the orientation as well as the time point of the last detected whistle among further information related to other tasks. Thus, only the angular WSDE and its additional information like distance and PSNR are to be appended to the team-message.

In order to wait for the team mates, a delay of three seconds is implemented before collecting the WSDE results of the other players. By this artificially introduced delay, synchronization issues due to network lags and whistle detection differences can be circumvented. For each team mate available in the network, a ray object is created by considering the last time when the whistle was heard, the robot position with its orientation and the WSDE angle. If the time point of the detected whistle differs more than 4s, it is assumed that the result does not involve the identical whistles and is not counted.

Chapter 4

Evaluation

In this chapter, experimental results are presented while setting focus on different aspects. Before evaluating the performance of the final localization algorithm the WSDE on single robots is examined by first taking a closer look at a representative example in section 4.2. There, the focus is set on the TDOA methods itself for each channel pair before combining them to a robot direction result. Section 4.2.4 presents the validity of the methods with regard to a whole set of data.

In chapter 3, the presence of additional information next to the estimated WSDE on individual robots is discussed. This includes knowledge about SNRs of the channel signals, the PSNR of the GCC functions and a distance approximation if the signal source is detected straight in front or from behind. Section 4.3 analyses if and to what extent these factors are useful for the WSDE or multi-agent whistle source localization.

To examine the whistle-sound source localization of the robots as team, measurements were taken with five Naos on the field of the HULKs' laboratory as specified in section 4.0.1. In section 4.4 the performance of the TDOA methods are presented and compared for all measurements. Depending on the accuracy of the individual direction results, the quality of the team filter is limited.

All measurements presented in this chapter were recorded in the laboratory of the HULKs and are taken from a sound source at a height of 1.5 m above ground. The size of the field used in this work is smaller than the regular SPL field with 7.5m length and 5m width instead of 9m and 6m. This circumstances occur due to lack of space and limited size of the field room in the laboratory. Another deviation to competition conditions arise due to walls being next to field borders. In section 4.0.1 further information about the measurement setup is introduced.

4.0.1 Measurement Setup

Eleven measurement were taken with five robots on the small-sized SPL field of the HULKs laboratory. Figure 4.1 illustrates the positions of the NAO robots and the positions of the whistle-sound sources.

According to these, the x- and y-coordinates of the sound source positions and robots are

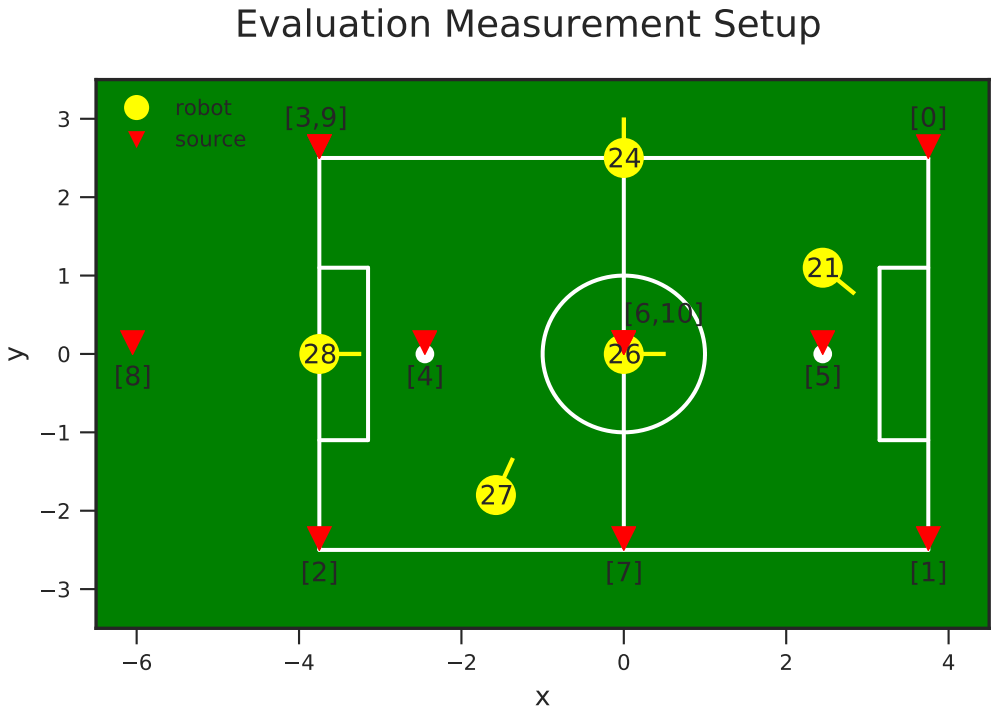


Figure 4.1: *Setup of robots and sound source positions for the evaluation measurement.*

listed in table 4.1 and table 4.2, respectively. Additionally, the positions are named for easier memorability. The orientation θ of the robots are defined relatively to the global x-axis as defined in section 3.1.1. In the following, the measurements introduced in this section are referred to as *laboratory-dataset*.

NAO	x [m]	y [m]	θ [deg]
21	3.75	2.5	-40.2
24	3.75	-2.5	90
26	0	0	0
27	-3.75	-2.5	66.06
28	-2.45	0	0

Table 4.1: *Robot positions of the laboratory-dataset.*

Measurement	Position Name	x [m]	y [m]
0	front left	3.75	2.5
1	front right	3.75	-2.5
2	rear right	-3.75	-2.5
3.9	rear left	-3.75	2.5
4	own penalty spot	-2.45	0
5	opponent penalty spot	2.45	0
6.10	center	0	0
7	center right	0	-2.5
8	behind own goal	-6.05	0

Table 4.2: *Positions of the whistle sources in the laboratory-dataset.*

4.1 Signal Start Detection

The implemented TDOA algorithms requires a smaller number of signal samples at the start of the signal to circumvent multi-path propagated signals and reverberation. In which extend the reflections distorts the delays between the samples will be handled in forthcoming sections.

In order to find a good solution for a highly reliable, accurate but computationally tractable start detection, different approaches were tested profoundly. For high temporal accuracy either the window must be small or is must be shifted with small steps which both implicit a large number of evaluations. The existing whistle detection algorithm of the HULKs presented in section 3.1.3 is computationally intensive and has a low accuracy for small sample selection windows as shown in . Therefore, the goal is to identify a simple algorithm for the signal start detection which is one of the approaches in Section 3.2.1. Ideally, this algorithm should be adaptable to signals other than the whistle-sounds.

Hereinafter, the performance of each method methods is evaluated by analyzing the number of index difference between algorithmically determined start index and manually labelled start index. As a benchmark, the prediction accuracy is evaluated on the eleven laboratory-measurements of all five robots section 4.0.1 to compare the signal start detection algorithms. By the frame size of the FFT being set to 256 samples for the correlation methods, a start index error of at most 256 samples is desired. Therefore, a start index detection result is regarded as failure for errors larger than 256 for the following sections. Having eleven whistle-sounds recorded with five robots, a total number of 55 measurements arise. Taking into account that each robot has four microphones attached, the overall number increase to 220 recorded signals on which the signal start detection algorithms can be tested.

4.1.1 Whistle Detection

First, the accuracy of the whistle detection which is already a part of the HULKs' framework is evaluated in regard of the start index. Here, the start index is defined as the first index of a frame in which the whistle detection found a whistle. With a frame size of 1024, the whistle detection algorithm fails for 81% of the 220 measurements. However, every error is

in the range of 1024 samples which proves that the approximate start of the whistle-sound is detected correctly. With a smaller frame size of 256, the temporal accuracy is improved up to a small failure rate of 9% but also introduces false positive detections. With the results, one can say that the whistle detection is not sufficient for the start detection or is at least not reliable as stand-alone solution. Furthermore, this approach is limited to sounds in fixed and known frequency range.

4.1.2 ZCR

For the evaluation of the ZCR, the frame size is set to 256 samples and number of noise and signal frames are set to 10 frames. In 7% of the 220 channel samples, this method fails. In most cases, the method provides accurate results with small error. Taking all measurements of robot 26 as example, the RMSE amounts 77.06 samples.

Measurement	ZCR Error	Entropy Error	WD Error
0	48.17	129.15	359.25
1	71.66	63.82	260.4
2	85.84	66.31	320.06
3	63.67	46.29	215.35
4	66.9	212.31	210.32
5	79.17	52.99	203.82
6	84.22	43.51	13064.78
7	78.2	54.27	290.59
8	86.06	149.29	118.02
9	88.24	199.33	140.38
10	95.54	290.02	202.32

Table 4.3: *Comparison of the signal start detection methods with a frame size of 512 samples by observing the averaged index error between the channels. Laboratory-measurements on robot no. 26 are selected to show the results exemplary. WD stands for whistle detection and the error are RMSEs between the channels.*

However, there are cases where the algorithm fails. Looking at those cases it could be identified that the errors occur often due to incorrect assumptions that the signal is present at the end of the buffered samples. Some measurements prove that this is not always the case as shown in fig. 4.2. The figure presents a case where the whistle signal ended around 35000 samples. If the start index is determined at the point where the ZCR falls below the threshold searching backwards as stated in section 2.2, the detection fails.

In many cases, only one out of four channels produces an erroneous result. Because the final start index on one robot is set equal for all channel, the failure can be compensated with a simple voting procedure. If the start index is determined at the point where the ZCR falls below the threshold searching backwards as stated in section 2.2, the detection fails.

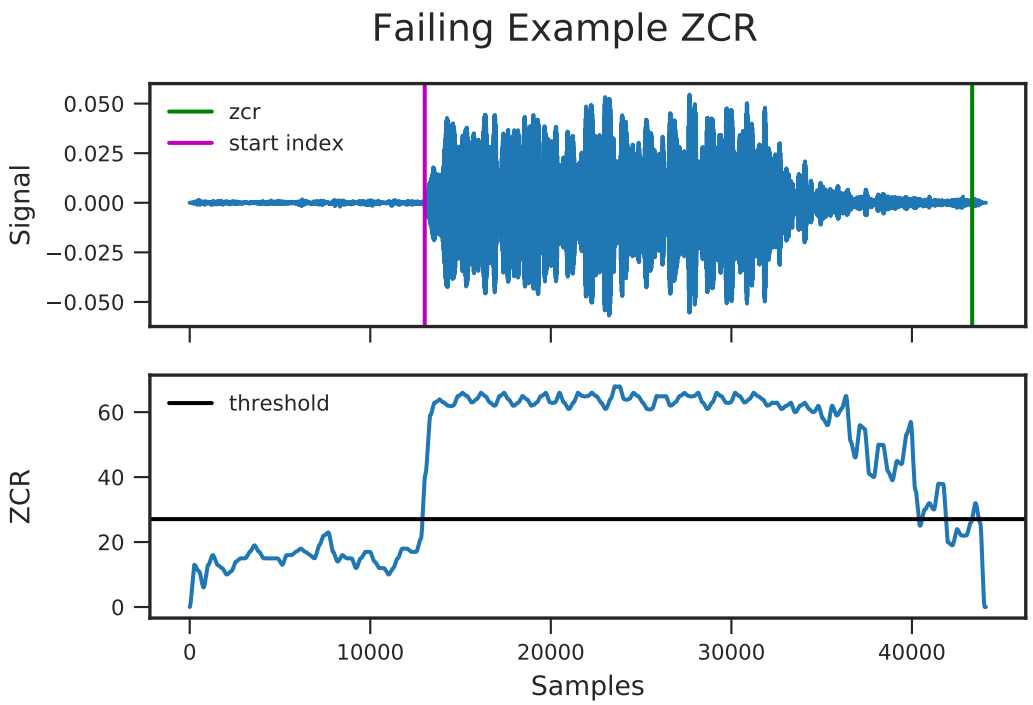


Figure 4.2: *Channel 3 data from measurement 5 of section 4.0.1 for robot number 21. A failing example for the start detection by ZCR is shown.*

4.1.3 Entropy

As discussed in section 2.2.2 the entropy quantifies the amount of disorder in a signal frame. Especially for signals to localize with unknown characteristics, this method can be useful because the only a-priori knowledge must be, that the signal to detect has lower entropy than the background noise. For all measurements this method yielded poorer results than the ZCR method with a failure rate of around 18%. Best results are achieved with a frame size of 512 samples and a step size of 800 samples. However, for records where the whistle ends before the end of the recorded signal it generates more reliable results than the ZCR method. Taking the same measurement as an example for which failure of the ZCR method was discussed earlier, fig. 4.3 shows how the algorithm detects the signal start correctly even though the whistle-sound ended at around 35000 samples. In this measurement, the start index errors of all four channels were smaller than 40 samples.

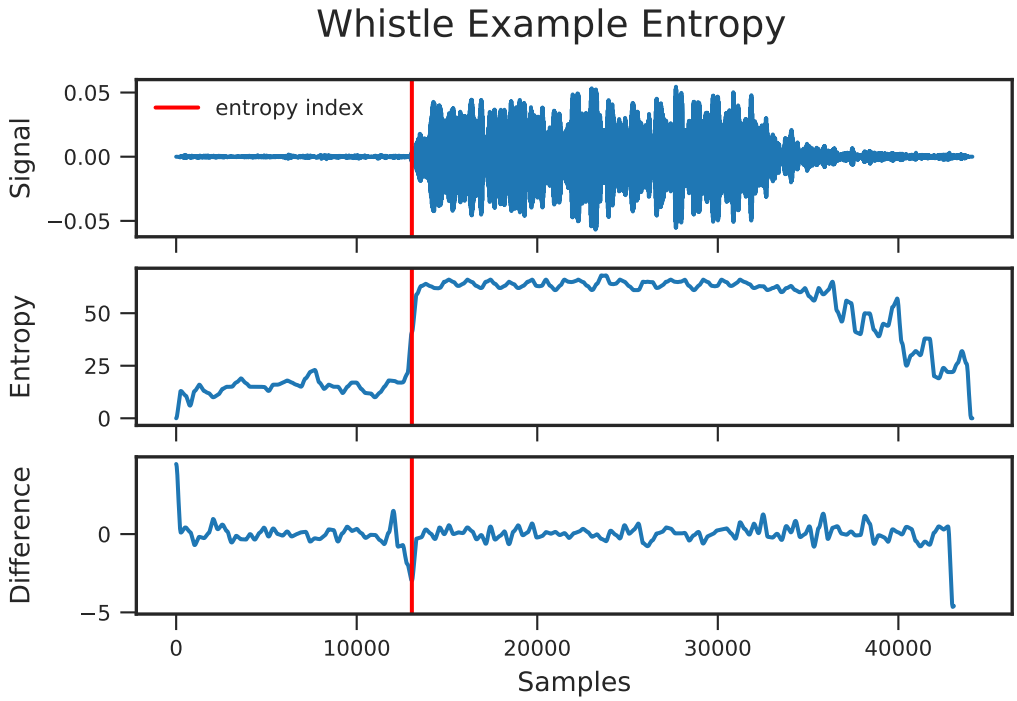


Figure 4.3: *Exemplary result of start index detection by entropy where the ZCR method failed due to fading whistle at the data.*

4.2 Whistle Source Direction Estimation

Before evaluating the performance of the different methods as a whole, one exemplary measurement is utilized to present and analyse the TDOA methods in detail first. In this recording, the sound source is placed at the right front of the robot with 4.5m distance. Hereinafter, this measurement will be referenced to as *demonstration-dataset*. This corresponds to an an angle of -33.7° in robot coordinates. To get an idea about the examined data, according whistle signal samples around the start are plotted in fig. 4.4 for all channels.

As the next sections focus on the performance of the TDOA methods, the start index is set manually.

For the sake of conciseness, throughout the following sections the correlation function $R_{x_a x_b}$ of two signals x_a and x_b (c.f. chapter 2) is denoted as R_{ab}.

4.2.1 Cross Correlation

The CC is a widespread technique to obtain the time delay between two series of samples. To discuss the result of the CC, the belonging correlation functions of the demonstration-dataset are plotted in fig. 4.5. The selection process and implementation correspond to the explanations in section 3.2.2.1. For R_{32} and R_{13} a peak is clearly visible. However, for the other CC the problem of a weak peak arises what was mentioned as downside of the CC in section 2.4.

Section 3.2.3 explains how two possible direction results exists by considering the delay between a channel pair. In Table 4.4, both arising direction candidates are listed for all channel delays.

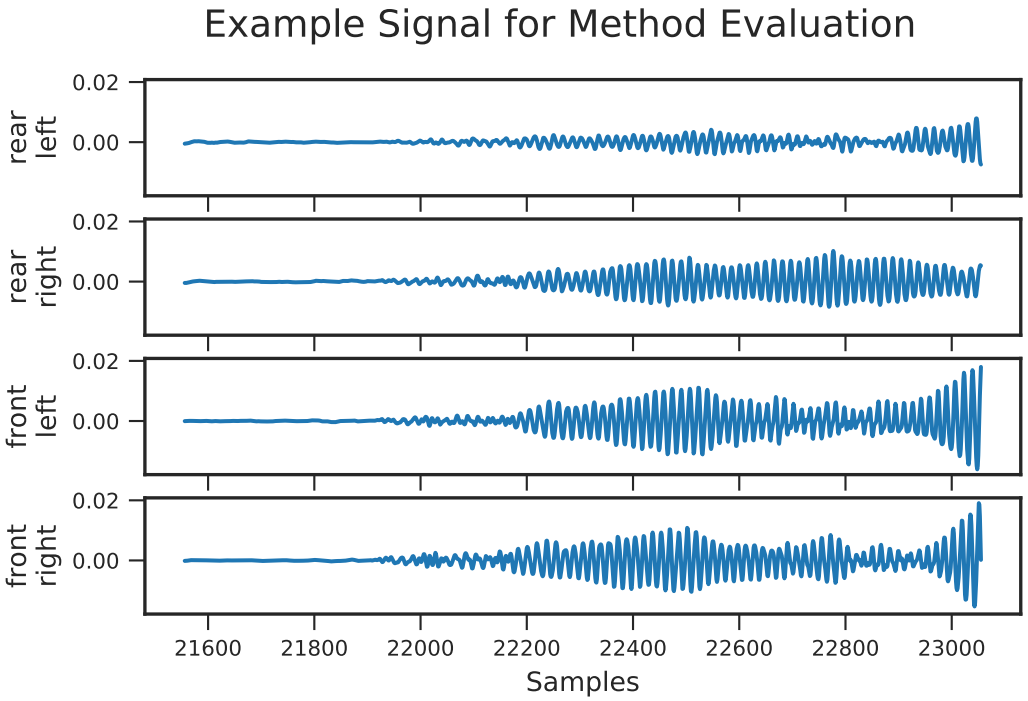


Figure 4.4: *Signal start section of a whistle-sound recorded from front right.*

Base Channel	Next Channel	Delay	Candidate (-)	Candidate (+)
0	1	-8.25°	-144.9°	-35.1°
1	3	-4.59°	-17.4°	78.6°
2	0	9.16	° -30.6°	-30.6°
3	2	3.94°	-150.2°	-29.8°

Table 4.4: *Cross correlation delay results of signal from front right.*

By implementation, the combination of all options with the smallest error is selected as WSDE of one robot. Hence, the algorithm outputs -26.9° what produces an error of 6.8°. The delay between channel 2 and 0 is larger than the maximum delay of 6.85 samples and therefore cut to the maximum sample delay. Besides these, the TDOA between the channel pairs produce one appropriate direction candidate which correctly points to the sound source.

4.2.2 Generalized Cross Correlation

Figure 4.6 presents the GCC result by the GCC-PHAT method of the demonstration-dataset equal to section 4.2.1. The subsample delays for each channel pair and their resulting direction candidates are listed in table 4.5. From this, a final direction of -30.0° is determined resulting in an error of 3.69°. It is apparent that the peaks of the GCC are better to detect than the peaks of the CC.

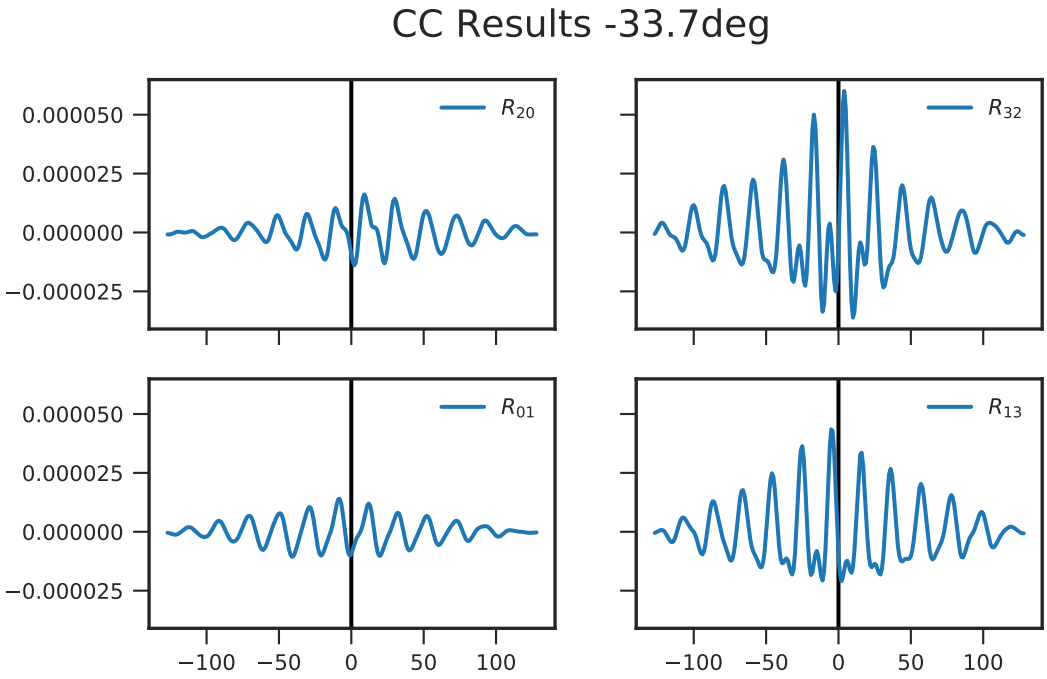


Figure 4.5: *Cross correlation results of signal from front right (-33.7°).*

Base Channel	Next Channel	Delay	Candidate (-)	Candidate (+)
0	1	-8.28°	-144.7°	-35.3°
1	3	-4.09°	-22.8°	84.0°
2	0	7.60°	-30.6°	-30.6°
3	2	4.13°	-148.7°	-31.3°

Table 4.5: *Generalized cross correlation delay results of signal from front right. Two possible direction candidates exists by considering the delay between a channel pair.*

4.2.3 Phase Difference

For detecting the source direction with phase difference, a smaller frame size of 64 samples is defined. In section 3.2.2.2 two variants of this method were introduced that use different strategies to identify a reference frequency. The first version uses a static reference frequency that is fixed a-priori by the user. The second version dynamically estimates a dominant frequency across all four channels. Hereafter, the performance of both variants is discussed.

Static Reference Frequency

In section 3.2.2.2 two different ways to set a reference frequency f_c for the phase difference method were introduced. First, a suitable value for the reference frequency is specified by examining the influence of the chosen value. Therefore, WSDE results with the phase difference

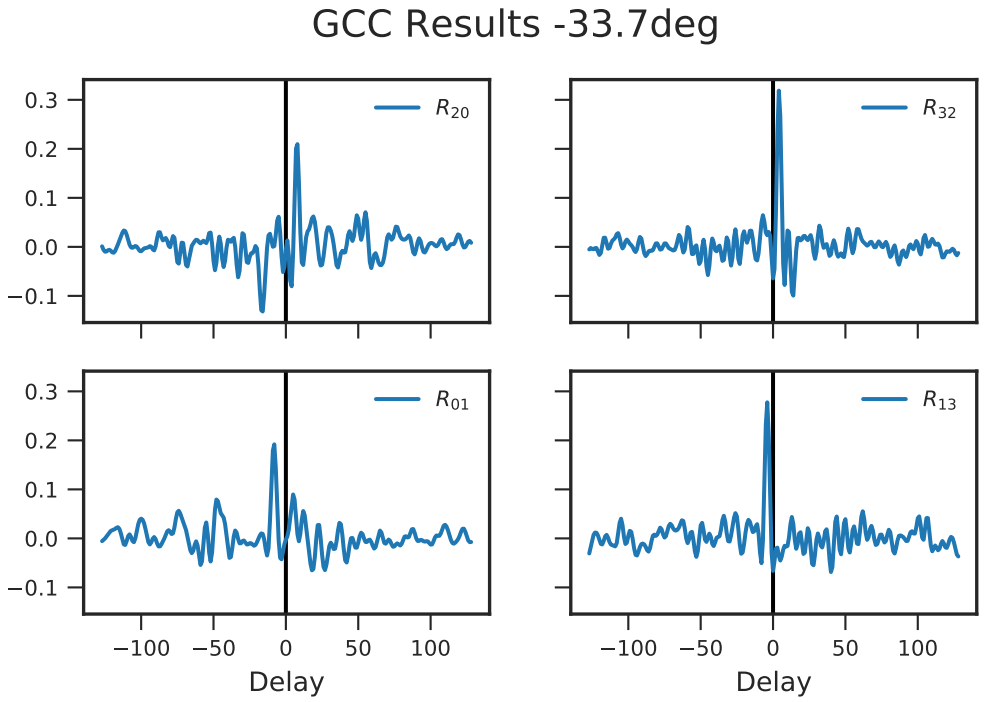


Figure 4.6: *Generalized cross correlation results of signal from front right.*

method are evaluated by setting different values for the reference frequency within whistle range. For this purpose we consider all of eleven measurements of the the laboratory-dataset recorded with robot no. 26 at the center point.

According to section 3.2.2.2, the most feasible frequency is defined as 2775.08Hz by the distance between the channels and the whistle spectrum ranges between 2kHz and 4kHz.

As shown in fig. 4.7 shows, the RMSE is high for frequencies smaller than 2600Hz. With a frequency of 2024.12Hz, error is largest. With this outcome, the fixed frequency is set to 2670.1Hz for further usage of the direction detection by phase method. Limitation exists due to the ambiguity of the signal which is content of section 3.2.2.2.

On the basis of the results, the reference frequency was set to a minimum of 2600Hz for evaluation of the demonstration-dataset. Hence, the reference frequency is 2627.1Hz in the case of a FFT length of 256 samples. Applying the phase difference method with this reference frequency, the final direction estimate computed from the candidates listed in table 4.6 is -29.6° which results in an error of 4.1°.

Base Channel	Next Channel	Phase Difference	Candidate (-)	Candidate (+)
1	3	-79.1°	-26.8°	88.0°
2	0	167.7°	-30.6°	-30.6°
3	2	88.5°	-148.7°	-31.3°

Table 4.6: *Resulting candidates of phase difference method with fixed frequency 2670.1Hz of example measurement from front right (-33.7°) Two possible direction candidates exists by considering the delay between a channel pair .*

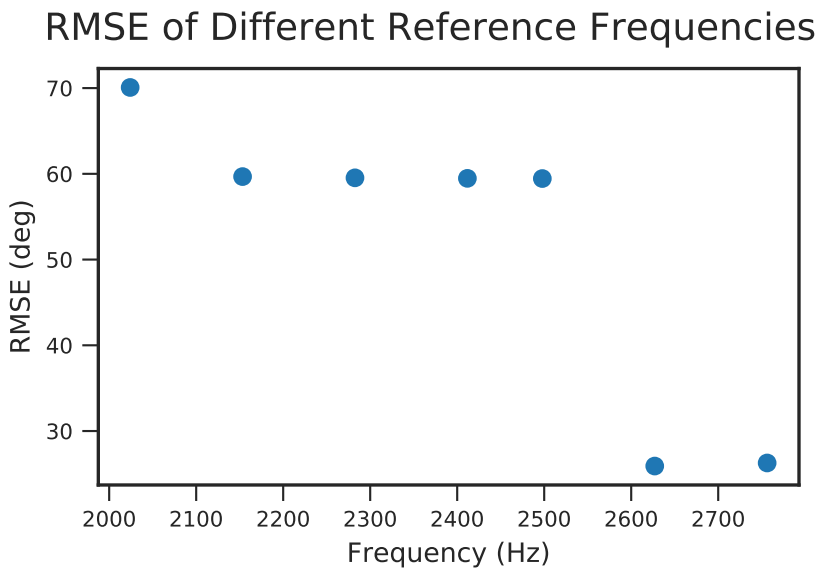


Figure 4.7: *Result of all measurements done with robot 26 to compare different fixed frequency values in whistle range.*

Dynamic Reference Frequency Selection

Another option is to have a nonspecific reference frequency that is computed dynamically without a-priori knowledge. As stated in the implementation chapter, frames are chosen where the frequencies of the maximum amplitudes coincides for all channels. For the running example discussed here, this corresponds to a frequency of 2756.25Hz.

For comprehensibility, the determined frequency information visualized by wave signals with the detected phases and amplitudes in the lower subplot of fig. 4.8. In the upper plot of fig. 4.8 one sees the originally received microphone data before applying a Hann window and transforming it into frequency domain by FFT. The resulting phases and amplitudes are listed in table 4.7. Table 3.3 pointed out that the most feasible frequency of the rear channels 0 and 1 is not in whistle spectrum due to the larger physical distance between the microphones. Thus, the phase difference information is neglected because of the ambiguity of the temporal sequence between the signals. Following the procedure discussed in section 3.2.2.2, the resulting phase difference estimate is -29.2° by combining the candidate direction -17.6°, -30.6° and -39.3° from channels 1, 2 and 3 according to table 4.8.

Channel	Phase [deg]	Amplitude
0	-1.55	0.00144
1	-177.7	0.00287
2	173.4	0.00279
3	-75.0	0.00372

Table 4.7: *Phase and amplitude of frame signals with $f_c = 2756.25$Hz.*

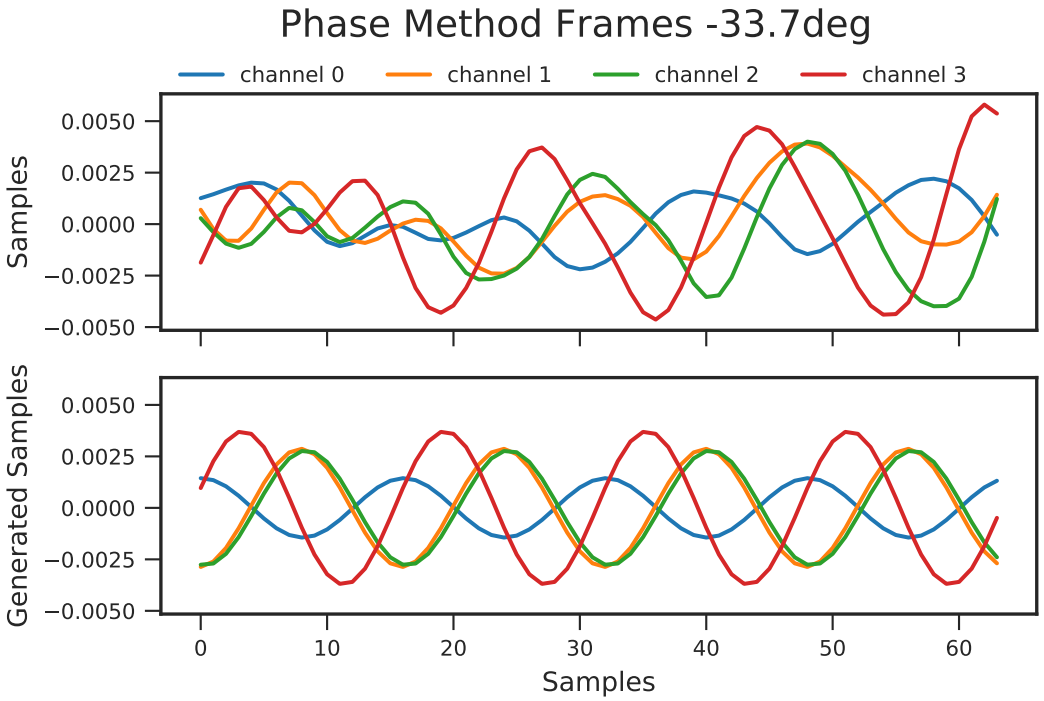


Figure 4.8: *Frames used for the direction detection by phase method.*

Base Channel	Next Channel	Phase Difference	Candidate (-)	Candidate (+)
1	3	-102.7°	-17.6°	78.8°
2	0	173.4°	-30.6°	-30.6°
3	2	113.1°	-140.7°	-39.3°

Table 4.8: *Phase differences and resulting direction candidates of demonstration-dataset with dynamically determined reference frequency.*

Impact of Selected Frame

Not only does the frequency play a major role for the phase method, but also the samples chosen for analysis. Again it is referred to the eleven measurements of the laboratory-dataset recorded on the robot at the center point (no. 26). To evaluate if and how the result changes over time, the selection window is shifted by half the frame size from -5 shift steps to 20. Recapitulating the implementation details for the static reference frequency case in section 4.2.3, the first frame after the start index is chosen in which the whistle detection would count a whistle signal. This frame is represented by zero shift.

In fig. 4.9, the resulting errors per shift for all measurements are plotted, presenting the influence of the selected frame around the signal start. The reference frequency was set to 2627.1Hz due to the outcome that frequencies larger than 2600Hz achieve best results. The graph presents that frames nearest to the signal start reach best results with an RMSE of 25.9°. These results show that the frame position has a significant influence on the prediction accuracy of the direction detection. Therefore, an accurate signal start detection is crucial for the preciseness of the SSL by phase difference.

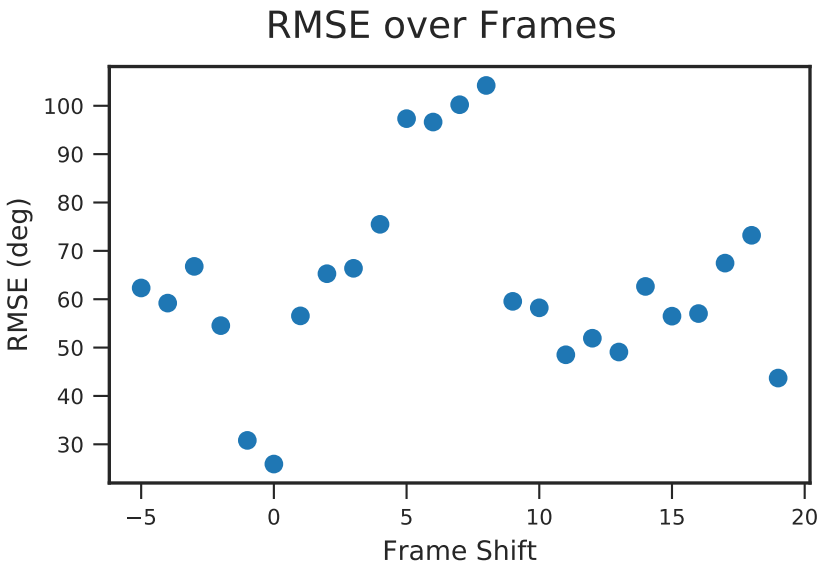


Figure 4.9: *WSDE result errors while shifting the frame over the samples of the laboratory-dataset on robot no 26.*

Phase Method Conclusion

Different options were investigated relating to the phase difference method in the last subsections. One finding is that reference frequency values should be chosen larger than 2.6kHz for best results. Another is that samples at the beginning of the signal are most suitable for this method. An advantage of the dynamically selected reference frequency is the reduction of one parameter. It requires more effort to implement and considering edge cases but does not depend on the whistle detection. Both approaches are valid for this work and result in a similar reference frequency due to the limitations listed in section 3.2.2.2.

4.2.4 TDOA Method Comparison

As the different TDOA methods were discussed profoundly in the last sections, all measurements of the laboratory-dataset are considered here to make a generalized statement about the performance. The WSDEs resulting from here are the inputs of the multi-agent localization filter. Figure 4.10 presents the RMSE considering the direction results of all five robots for each measurement in section 4.0.1. Additionally, the estimated standard deviation of each measurement provides insight into the validity of the single robot results. As one can see, the standard deviation of the relative angle of the GCC method is significantly smaller as compared to the phase method for most measurements. How this influences the reliability of the sound localization is subject of discussion in section 4.4.4.

4.2.5 Conclusion

The results in Section 4.2.4 show that the GCC-PHAT algorithm performs best according to the laboratory-measurements. Not only the errors are smallest, but also the low standard deviation displays that the robots agree on the WSDE and little outliers exist. Also in regard to

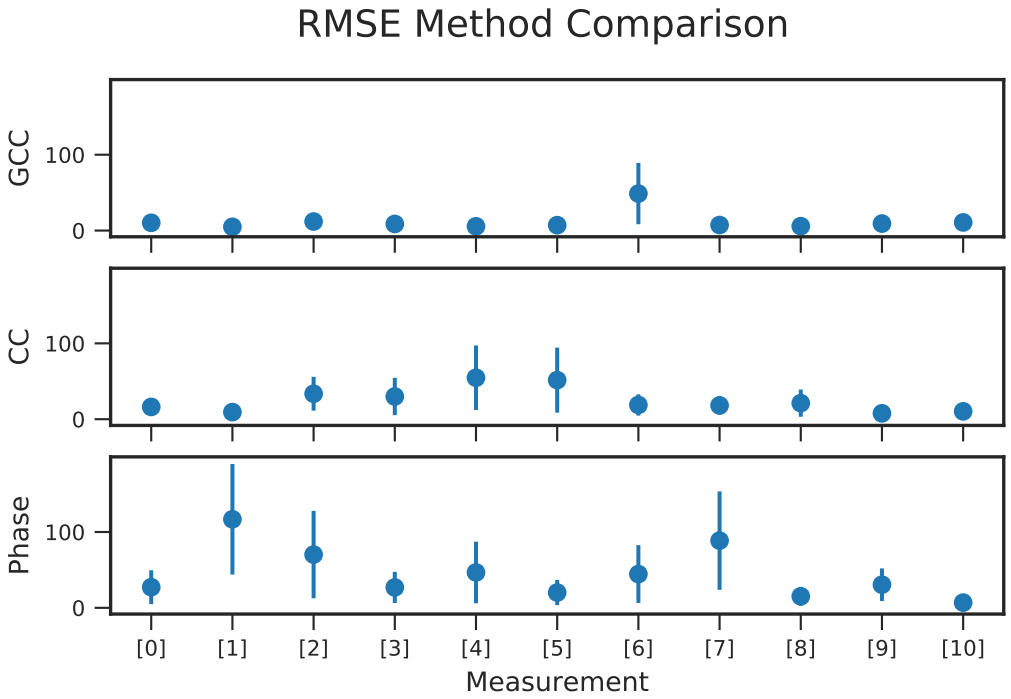


Figure 4.10: *Angular RMSE and standard error of robot results laboratory-measurement of section 4.0.1.*

Method	RMSE	Standard Deviation [deg]
CC	24.67°	18.87°
GCC	11.84°	8.29°
Phase Difference	44.88°	33.94°

Table 4.9: *Method comparison of averaged RMSEs of single robot WSDE results. All laboratory-measurements are considered.*

the multi-agent Bayesian updating filter, unified results of the stand-alone robots are beneficial. Another advantage of the GCC-PHAT method is the presence of an indication regarding the certainty of the measurement. Interpreting the PSNR as such, it can be used to detect outliers or consider results with small PSNR less.

Regarding the computational effort, the phase difference method with static reference frequencies for TDOA estimation accomplishes lowest demand.

4.3 Additional Information

On top of the direction detection by TDOA, additional information can be extracted from the microphone data. This information can be used to improve the single robot result or to feed the team filter with information about the certainty of the WSDE result. As section 3.2.4 has described, the distance of the sound source can be estimated approximately for nearby signals that are aligned with the x-axis of the robot's head. Another intuitive approach is the

inspection of the SNR which is expected to be higher for closer sound sources. In addition, the PSNR of the GCC as defined in section 4.3.3 is examined in detail.

4.3.1 Distance Approximation

Section 3.2.4 stated that the distance to a sound source can be determined approximately if it comes from straight in front or backwards of the robot. Additionally, this calculation is only possible for sound sources that are less than 0.66m in front or 15.3m behind in theory. For this case, a height of the sound source is estimated to be 1.5m. To examine the validity of these assumptions, measurements from the front and back of the robot are collected and evaluated among this range. Table 4.10 lists the true distance as well as the distance estimate for each measurement. For both measurements with zero distance, the orientation of the whistle differs. 180° indicates that the whistle was turned in the opposite direction of the robot. In the other case (0°), the whistle was oriented unidirectional with the robot. The distance is represented in robot coordinates, so that positive distance expresses that the source was placed in front of the robot and oriented towards it and vise versa.

No.	True Distance [m]	GCC Result [m]	CC Result [m]	Phase Result [m]
1	+0.9	∞	∞	∞
2	+0.6	∞	∞	∞
3	+0.3	0.35	0.25	0.13
4	+0.0 (180°)	-0.13	-0.15	-0.23
5	-0.0 (0°)	0.22	0.21	0.02
6	-0.3	-0.15	-0.27	-0.45
7	-0.6	-0.34	-0.50	-0.62
8	-0.9	-0.70	-0.91	-0.99
9	-1.2	-1.00	-1.28	-1.71
10	-1.5	-1.39	-1.59	-2.98
11	-1.8	-1.72	-2.07	-3.33
12	-2.1	-2.16	∞	-3.02
13	-2.4	-2.31	∞	∞
14	-3.75	-3.66	-9.51	-4.15
15	-6.4	-7.35	-7.27	∞
16	-9.8	∞	∞	∞

Table 4.10: *Result of front and rear distance for all methods..*

As the results show, that in many cases the distance can be approximated with sufficiently small error. One can see that the GCC results are erroneous for small distances, but gives a correct approximations that are not completely out of proportion for all measurements except of the edge cases (no. 1, 2 and 16). Compared to this, the CC method performs better for small distances, but fails completely for some measurements. These failures of the CC could be attributed to cases where the lateral delays exceeded the maximum lateral samples to trigger

the distance estimation algorithm. The phase difference method provides most incorrect results. Especially measurement 10 stands out by being double the real value.

From the results one can say, that it is possible to estimate the distance of a sound source by all methods but is mostly reliable for the GCC-PHAT method. Furthermore, the algorithm correctly detects sources that are out of constructionally observable range. However, for real cases one can not rely on the height parameter of the sound source which varies according to the referee's body height. Having this as approximation only, the distance result should be handled with care. Also if a localization algorithm is implemented for sound sources at the same height as the robots' head, the distance estimation becomes unusable.

4.3.2 SNR

Receiving information about the source location by the strength of a signal is an intuitive approach. Ideally, channels and robots that are closer to the sound source receive greater signal data. Section 3.2.5 covers the implementation details about the SNR which is a value to expresses the strength of the signal compared to its background noise. For the purpose of clarification, it is examined if channels with larger SNR are closer to the others on one single robot. Additionally, this examination helps to ensure that the individual channels are not biased towards internal noise of the robot's head.

Therefore, 3kHz sinusoid signals are played back digitally from 0.73m distance and constant volume. 14 Measurements from different angles were taken with one robot. Further evaluation is done by determining the channel with the maximum SNR between the channels. It is expected that the nearest channel to the sound source has maximum SNR. At 85.71% of the measurements this assertion could be evidenced. In general, the SNR seems to correlate in some extend with the distance to the sound source and no channel seems to behave different by nature. To prove this relation for real whistle measurements, the same evaluation is done with the laboratory-measurements. By that, only 54.55% of the maximum SNRs match with the expected channels. This outcome is not sufficient to make conclusions about the signal source direction on one robot.

From the multi-agent perspective, it would be a simple way to predict the sound source position roughly if a relation between received signal strength on one robot and distance to the source exists. For example if multiple potential source positions exist for the multi-agent decision, the SNR information can used as subordinate indicator for the direction. Another point is that information about the uncertainty of a WSDE result can be respected in the Bayesian updating algorithm straightforwardly. Depending on the uncertainty, the covariance matrix of the incoming result can be adjusted so that predictions that are assumed to have higher error have a smaller influence to the posterior estimate. In other words, SNRs of robots closer to the source would be higher and to that effect, the reliability of their results is higher. Thus, a relation between SNR and distance on multi-agent level is examined in addition.

Taking the laboratory-measurements of section 4.0.1, this hypothesis is examined by finding a relation between distance and and the robots' SNR mean over all channel which is now called *robot-SNR*. Because the whistle is not blown equally for all laboratory-measurements, the robot-SNR is scaled by the overall mean of all robots' SNRs for each measurement (named *measurement-SNR*). Figure 4.11 presents the relations between distance of a robot and the robots-SNR relative to the measurement-SNR.

Considering the plot, no straightforward link between both values can be found. Calculating the correlation coefficient ρ between both quantities, its value is -0.2860. This outcome verifies that no simple connection between SNR and distance can be placed.

For the real whistle measurements, no assuring relation between SNR and distance to source can be found neither between the microphone channels on the robots' head nor between the single robots. Consequentially, one must assume that the environmental circumstances like multi-path propagation and reflection have large influence on the signal data. Thus, the SNR will not utilized as additional information for the WSDE algorithm nor for the multi-agent filter.

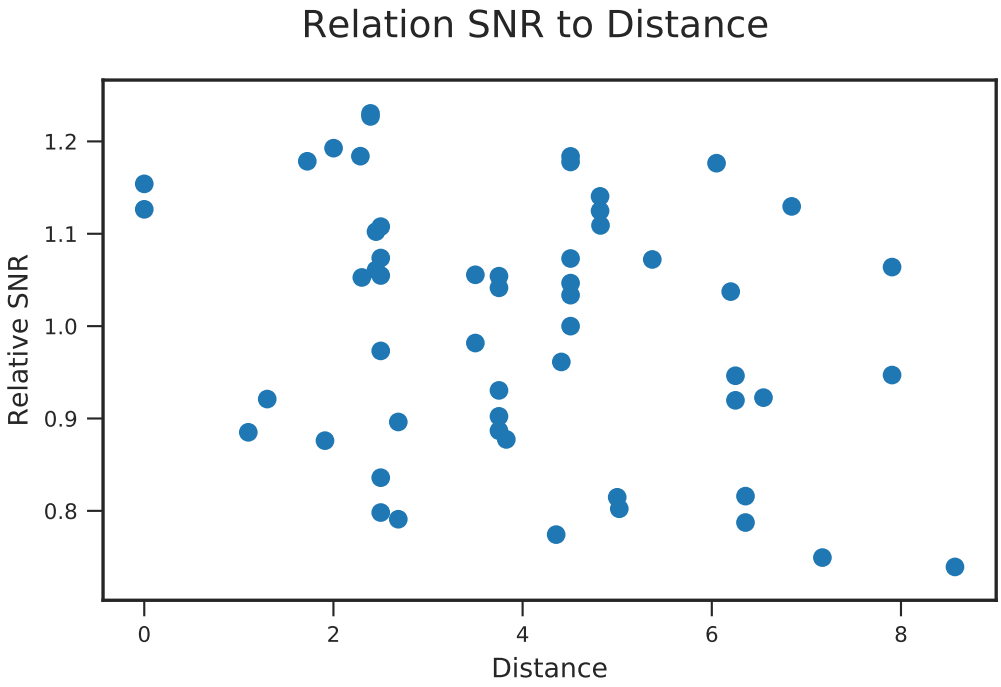


Figure 4.11: *Visualization of relation between SNR and distance.*

4.3.3 PSNR

As referred in section 2.5, the main characteristic of the GCC-PHAT algorithm is the emerging sharp peak compared to CC functions. Figure 2.4 illustrates this characteristic for the ideal case where two similar, but shifted signals with normally distributed noise are input signals of the GCC algorithm. However, with real whistle data the GCC function does not always output a strong peak. This is a known topic and there exists research that inspects how to handle the ambiguity of GCC functions like [28] does. Exemplary fig. 4.12 shows such a case where the GCC does not produce a clear result.

R_{13} and R_{20} represents GCC functions of the same measurement but between different channels. As the plot shows, the GCC between channels 2 and 0 yields an outcome with an easily recognizable peak. In comparison, the peak of the upper plot which is the GCC between channels 1 and 3 is less distinguishable from the remaining part.

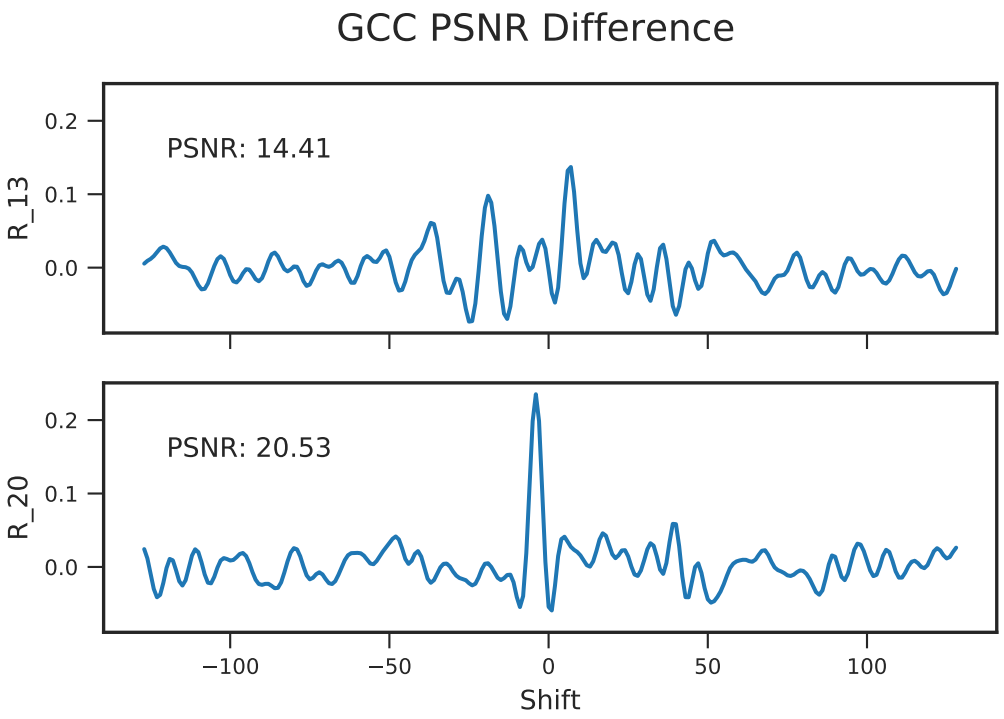


Figure 4.12:

In conclusion, one can assume that the lack of a sharp peak indicates a less valid delay result of the GCC. The next sections covers the validity of this hypothesis by first evaluating if the PSNR can help with the selection of the frames used in the WSDE. On the other hand the informative value contained in the PSNR regarding the overall WSDE on one robot is examined.

Frame Selection

One perspective is to insert the PSNR information into the WSDE of individual robots. In fig. 4.13, the frame to examine is shifted before and after a manually defined signal start index. The robot was positioned at the center point while the whistle is blown at -33.7° with 4.5m distance. The frame size of the GCC is set to the default size of 256 samples and the window shift samples are quarter of the frame size. For better understanding, samples of the front right channel are plotted in the upper graph of fig. 4.13. The second graph shows the RMSE of the robot direction result over the frame shifts. For the lower graph, the mean over the PSNRs of all channels is presented. As a general trend, it can be observed that the error of the predicted direction is low for frame shifts with a high PSNR. This confirms the hypothesis stated early in this section. For shifts smaller than -2, the whistle signal does not intersect with the evaluated frames. Therefore, the prediction error in this regime is high. An important notice is that the PSNR decays as the frame is shifted towards later samples on the whistle signal. This indicates that the implemented GCC-PHAT method is not suitable for arbitrary subsamples of the signal.

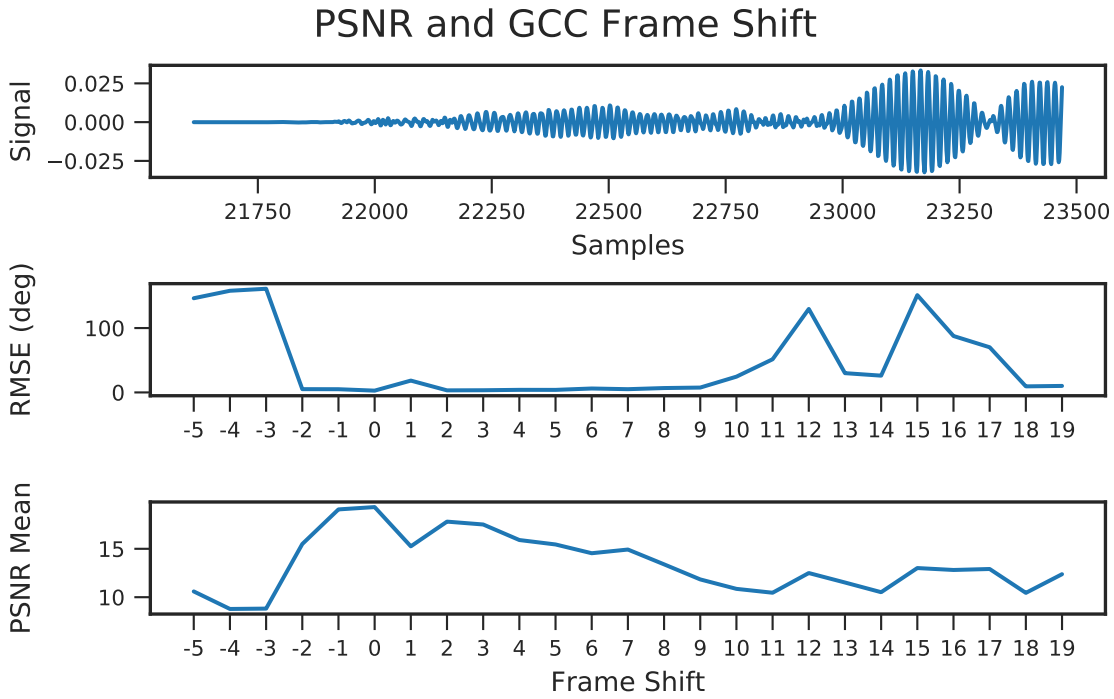


Figure 4.13: *Relation between PSNR and selection of the frame in time. Signal data of the rear left channel is plotted in the upper window. In this measurement, the whistle is positioned at right front of the robot.*

Informative Value

In this section, the PSNR value of one GCC-PHAT calculation is brought into comparison with the error of the direction angle resulting from the delay. In the following, the PSNR is referred to as high if it exceeds 17.5 whereat the PSNR value ranges from 10.1 to 28.8 for measurements in section 4.0.1.

First, for each channel pair the error between true direction of the signal source and the direction candidate with smaller angle error emerging from the GCC delay are put into context. The results are grouped into two classes based on the associated PSNR value. If the PSNR of the GCC is greater than the threshold of 17.5, the angular error of its source direction result is classified as high PSNR. Else, it pertains to the errors with low PSNR. Out of a total number of 220 measurements, 78 correlations reported a PSNR below the threshold. The RMSE of all source directions within this group was 35.77°. Compared to this, the RMSE of the remaining 144 measurements is 15.86° only.

To see the impact on a complete robot result, the same evaluation is done with the final angular errors of the robots' WSDEs. Here, the PSNR value is the mean over all channel pairs' PSNR. The angular RMSE of the direction classified as low PSNR value are larger with 21.15° whereat the error of the high PSNR case is 9.1°.

From the results one can find a relation between PSNR and validity of a computed delay between two channels with the GCC-PHAT method. Within this work the PSNR information is not included for the implementation of the multi-agent decision, but for the WSDE of a single robot. Thus, the frames for the GCC algorithm to compute the delay between the

channels are selected by the highest PSNR mean value in a range around the start index. The implementation is explained in section 3.2 in further detail.

4.4 Multi-Agent Source Localization

After the three TDOA methods CC, GCC and phase difference were evaluated in the preceding sections, the SSL algorithm with a multi-agent system of five robots is evaluated. The remaining part of this chapter focuses on the performance of the SSL with regard to each TDOA method individually. Input to the SSL algorithm are the WSDE of each robot in the team. Based on these results, the multi-agent whistle localization outputs an absolute sound source position in field coordinates. To provide a decoupled result to the signal start detection, the start indexes were set manually. In the following, the laboratory-measurements of Section 4.0.1 are utilized for evaluation of all methods.

4.4.1 CC Method

To determine an overall result, each robot computes a direction prediction from the locally recorded signal using the GCC-PHAT method standing alone. These local direction estimates of individual robots are fed to the team decision filter as specified in section 3.3 which estimates the global sound source position by combining all measurements through Bayesian updates. First, the results of the SSL algorithm are presented with the WSDE calculated by simple CC. The results for the predictions of this method are reported in Table 4.11 lists the error of the localized position in regard to the real sound source position in x- and y-coordinates. Additionally, the angular error in field coordinates is listed. It indicates if the result has a correct tendency.

No.	Measurement	Error x [m]	Error y [m]	Error Abs. Distance [m]	Error Angle
0	front left	0.6	1.39	1.51	8.15°
1	front right	-0.49	1.2	1.3	11.97°
2	rear right	2.32	2.11	3.13	18.28°
3	rear left	1.07	-0.96	1.44	3.71°
4	own penalty spot	1.95	-0.09	1.95	10.44°
5	opponent penalty spot	0.07	0.01	0.07	0.32°
6	center	0.4	-0.01	0.4	1.91°
7	center right	1.06	-0.0	1.06	22.99°
8	behind own goal	1.24	-0.06	1.24	0.77°
9	rear left	-0.04	0.78	0.78	7.22°
10	center	0.03	-0.0	0.03	0.0°

Table 4.11: *Whistle localization results of laboratory-measurements with CC method.*

Over all measurements, the CC predictor has a RMSE of 1.45m in distance and 10.67° angular.

4.4.2 GCC Method

As the GCC method provides the best results for the WSDE results, more precise steps of the multi-agent source localization algorithm are presented here. For further clarification, measurement 1 of the laboratory-measurements is selected as example. Figure 4.14 illustrates the result of the relative direction estimates γ_i of the individual robots listed in table 4.12 for this example. For figs. 4.1 and 4.14, robot positions are marked by yellow dots where a short yellow line indicates each robot's orientation. The arrows in fig. 4.14 represent the local direction estimates γ_i as predicted by each robot. Finally, the true position of the sound source is marked with a red star while the joint position estimate over all robots is visualized by a cross.

NAO	γ_i	Abs. Error
21	-26.22°	3.71°
24	-133.77°	9.32°
26	-30.19°	3.50°
27	-75.26°	1.71°
28	-15.90°	2.53°

Table 4.12: *Resulting direction estimates of the individual robots with GCC-PHAT method for a whistle-sound signal in the right front corner of the playing field.*

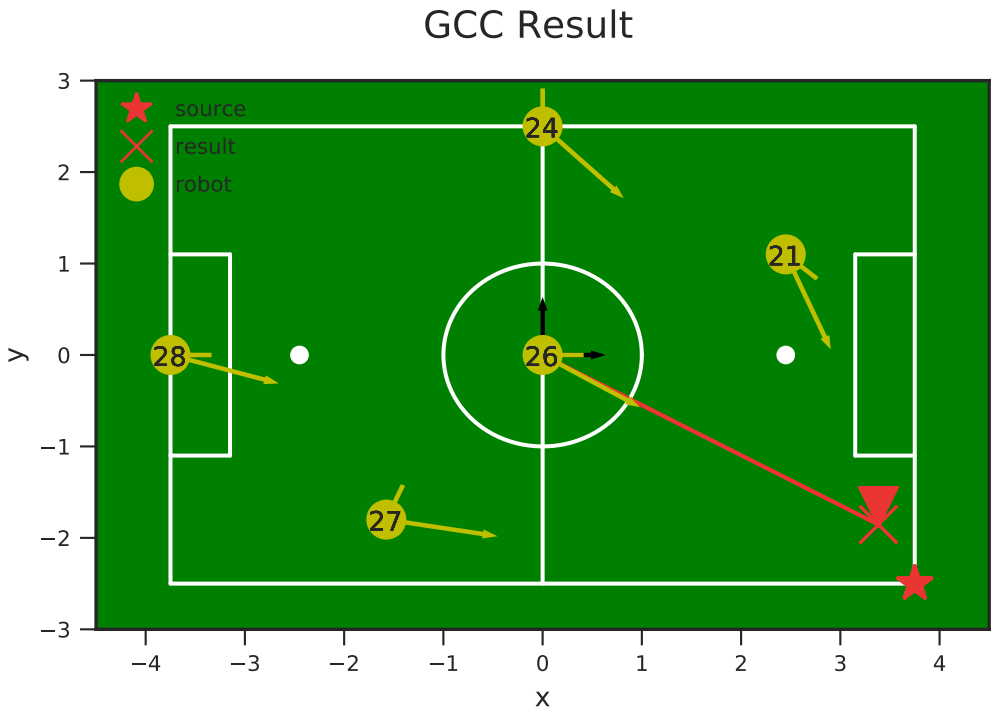


Figure 4.14: *Team whistle localization result with GCC-PHAT method.*

The final result and its corresponding errors are listed in table 4.14.

	Result	Error
Position x [m]	3.38	-0.37
Position y [m]	-1.85	0.65
Angle	33.18°	1.57°
Distance [m]	3.85	0.74

Table 4.13: *Whistle localization result of measurement 1 with GCC-PHAT method.*

Table 4.14 shows the distance and angle errors for all laboratory measurements in section 4.0.1. The RMSE of the localized source positions in distance being 0.87m and angular RMSE being 5.07° one can say that the GCC-PHAT algorithm works well for whistle-sound source localization.

No.	Measurement	Error x [m]	Error y [m]	Error Abs. Distance [m]	Error Angle
0	front left	1.31	1.06	1.68	1.45°
1	front right	0.13	0.06	0.15	1.57°
2	rear right	0.59	0.43	0.73	0.42°
3	rear left	0.54	0.47	0.72	9.09°
4	own penalty spot	0.27	0.0	0.27	0.01°
5	opponent penalty spot	0.15	0.14	0.21	3.18°
6	center	0.41	-0.02	0.41	2.67°
7	center right	0.39	0.02	0.39	8.98°
8	behind own goal	1.84	-0.01	1.84	0.14°
9	rear left	0.58	0.52	0.78	9.89°
10	center	0.03	-0.0	0.03	0.0°

Table 4.14: *Whistle localization results for all laboratory-measurements with GCC-PHAT method.*

4.4.3 Phase Method

Finally, the performance of the phase method is evaluated. In this experiment, the reference frequency is set to a minimum of 2700Hz due to the result that reference frequency larger than 2600Hz obtain best results. The results of phase method for this experiment are shown in table 4.15. The RMSE of the position estimate is close to the prediction accuracy of the CC method with 1.33m. The RMSE of global direction estimate of 74.8° shows a significantly worse performance than the other two methods. However, it must be noted that these angular error mainly arise from measurements 6 and 10. Both measurements are taken at the center point of the field. Since the absolute position is in an acceptable error range, these angular results will

be treated as outliers for the error calculation. Thus, the angular RMSE of the phase method without measurements 6 and 10 is 11.69°.

No.	Measurement	Error x [m]	Error y [m]	Error Abs. Distance [m]	Error Angle
0	front left	-1.07	-0.98	1.45	4.13°
1	front right	0.21	0.27	0.34	4.27°
2	rear right	0.22	1.39	1.41	16.25°
3	rear left	1.26	-0.02	1.26	11.19°
4	own penalty spot	0.16	-0.04	0.17	0.99°
5	opponent penalty spot	-0.42	0.19	0.47	5.46°
6	center	-0.32	0.08	0.33	166.25°
7	center right	-0.28	1.76	1.78	20.82°
8	behind own goal	2.37	0.27	2.39	4.23°
9	rear left	2.04	0.29	2.06	24.89°
10	center	-0.32	0.0	0.32	180.0°

Table 4.15: *Whistle localization results for all measurements in section 4.0.1 with phase method.*

4.4.4 Conclusion

The SSL algorithm is tested with all TDOA methods in regard to the laboratory-measurements. Table 4.16 summarizes the results briefly by the absolute distance RMSE of all measurements.

Method	Abs. Distance RMSE [m]
CC	1.45
GCC	0.87
Phase Difference	1.33

Table 4.16: *Summarized performance of the multi-agent SSL according to the TDOA methods.*

Comparing the SSL results with three different TDOA methods, the GCC-PHAT algorithm performs best. As expected, the CC method yields poorer results what underlines the statements about the CC method at the beginning of this work in Section 2.4. Assessing the CC and phase difference results of the SSL, the resulting outcomes in table 4.16 could lead to the conclusion that both are equally valid.

Recollecting the WSDE results in fig. 4.10 one sees that the single WSDE results of the CC are more precise regarding the error and standard deviation. Small deviation means in this case that the individual robots agree on the direction roughly. The more outliers exist in the measurement, the more does the deviation increase. Having all WSDE results of five robots

available for the SSL, the outliers can be neglected by good filtering. However, with less number of robots included in the multi-agent SSL the accuracy of each single WSDE becomes more important. With this in mind, the results of the simple CC approach is more reliable than the phase difference method.

Finally with the given results from the laboratory-measurements, one can state that the Generalized Cross Correlation with Phase Transform (GCC-PHAT) algorithm yields the most accurate results for the Whistle Source Direction Estimation (WSDE) outcome of individual robots. Through the high reliability of the single results with this approach, the Bayesian updating filter of the multi-agent system produces appropriate whistle source position estimates that differ less than 1m from the real source in average.

Chapter 5

Summary and Conclusion

The objective of this work was to find an approach to localize a whistle-sound source with multiple NAO robots. Each of these robots offers four microphones attached on their head. In the scope of this work, stationary sound sources and motionless robots were considered. This issue arises in the context of RoboCup SPL which is a competition where humanoid robots play soccer autonomously. In this games, whistle-sounds are used to initiate kickoffs at the current state of the rules. Due to parallel games at neighboring fields, other games' whistles can be heard and the risk exists that the robots begin to move illegally. To circumvent those situations by neglecting whistle-sounds from other games, a Signal Source Localization (SSL) algorithm was to be designed and implemented.

The beginning of this work covers the theoretical principles used for the multi-agent SSL which has been divided into three major parts. Based on this, the implementation details of each of this components are proposed and the different strategies are evaluated.

First, different signal start detection algorithms were examined with regard to accuracy. Using this algorithm, appropriate subsignals were selected for the following processes with the assumption that the beginning of a whistle signal is most unaffected by reverberation and multi-path propagation. Evaluation of the recorded data confirmed this assumption. As accuracy and low computational effort are important on a real-time constraint system, different approaches were compared to detect the signal start. Best performance could be obtained by considering the ZCR.

Having appropriate samples of the whistle signal, the relative direction of a whistle source is computed by each robot of the multi-agent system. In order to solve this task, the Time Difference Of Arrival (TDOA) method was selected as solution method which is popular a approach for acoustic signal source localization. The fundamental concept of TDOA algorithms is to obtain direction information about acoustic signals by observing the time delay between separate microphones. Three different approaches were evaluated to achieve a stable algorithm for the delay estimation. The Cross Correlation (CC) and Generalized Cross Correlation with Phase Transform (GCC-PHAT) approaches are based on cross-correlation theory in terms of signal processing. In comparison, the phase difference method performs spectral analysis on the signal to attain TDOA information. This Whistle Source Direction Estimation (WSDE) is of significant importance which is why a wide range of examination was done with real measurement data. According to the results, the Generalized Cross Correlation with Phase Transform (GCC-PHAT) algorithm performs best by providing most accurate results for the

WSDE which offers the most precise whistle-sound source position estimations in consequence. Evaluating WSDE measurements recorded by the multi-agent system of five robots, the angular RMSE by all robots was 11.84°.

Finally, a multi-agent decision is executed on a Bayesian updating filter depending on the WSDE results of the single robots. For the measurement data recorded within this work, an average distance error between estimated position and real source location of less than 1m could be achieved.

With the outcome of this work, further research topics arise to solve more complex SSL challenges. Using the ability to localize signals other than whistles, differentiation of multiple sounds and separation of their positions becomes possible. Furthermore, consideration of moving sources and robots can be interpreted as the next essential step for the usage in real case scenarios. Especially when adopting the SSL for verification of other robots' information like position and orientation, advancement towards these issues are important. In another perspective, the multi-agent behavior can be improved with various concepts. This includes processes like multi-modal filtering or the detection of outliers. Also, more a-priori knowledge like the prior position of the sound source can be included.

It can be concluded that wide-ranging progress and novel applications can be expected regarding SSL in the field of autonomously operating robotics.

Appendix A

Anhang 1

A.1 Alternative Figure GCC

Figure of GCC with filters before the correlation.

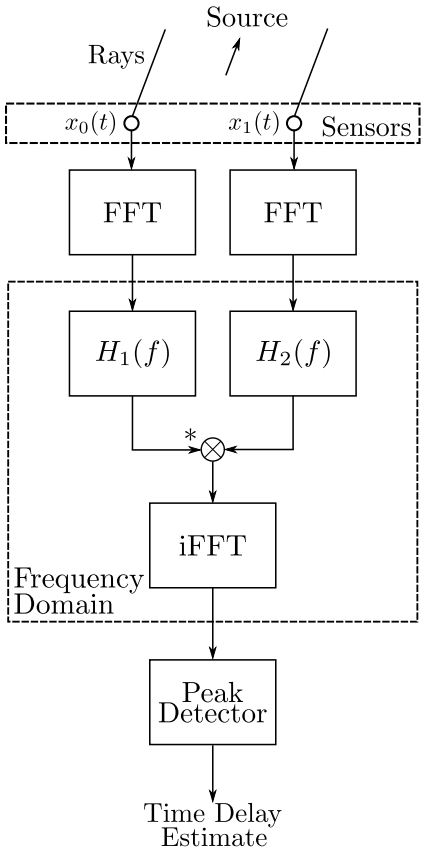


Figure A.1: *Generalized cross correlation for time delay estimation.*

A.2 GCC Method Frame Shift

Link between PSNR and correctness of the whistle source direction detection on one robot is charted. Data of all measurements of section 4.0.1 on robot 26 are used.

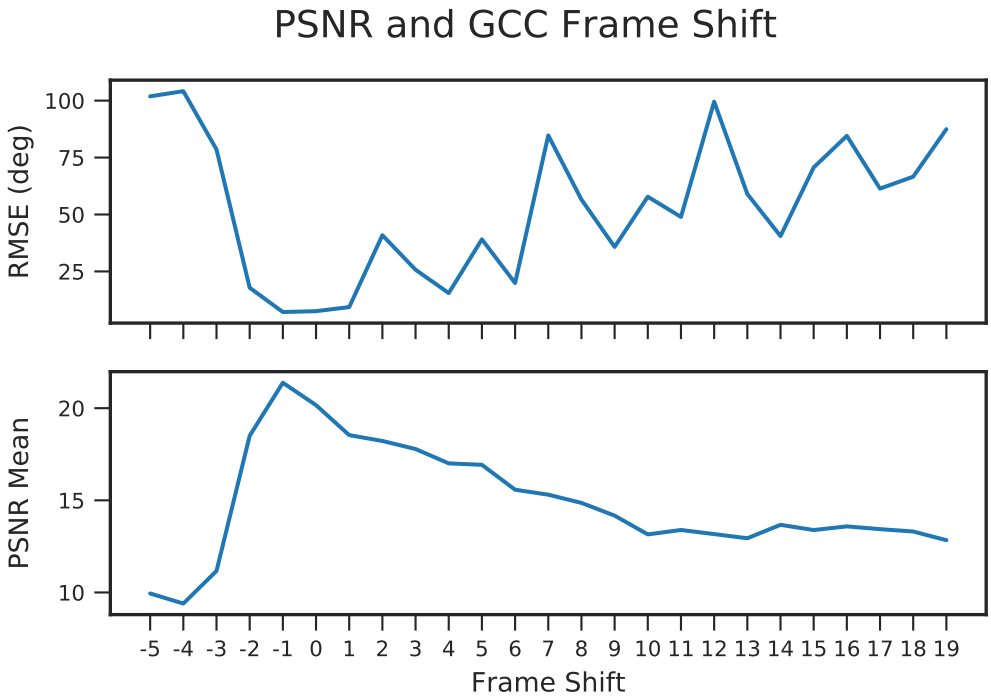


Figure A.2: *Frame window shifted around start index. All measurements of section 4.0.1 are utilized for robot at center point.*

List of Software

Name	Version	URL
Python3	3.7.4	https://docs.python.org/3/
NumPy	1.17.0	https://numpy.org/
ALSA	1.17.0	https://www.alsa-project.org/
HULKs Framework	2019	https://github.com/HULKs/HULKsCodeRelease

Table A.1: *Utilized Software.*

Bibliography

[1] (2019) Nao - technical overview. [Online]. Available: http://doc.aldebaran.com/2-1/family/robots/index_robots.html#all-robots

[2] A. Badali, J. Valin, F. Michaud, and P. Aarabi, "Evaluating real-time audio localization algorithms for artificial audition in robotics," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2009, pp. 2033–2038.

[3] F. Michaud, D. LÃ©tourneau, P. Lepage, Y. Morin, F. Gagnon, P. GiguÃ¨re, E. Beaudry, Y. Brosseau, C. CÃ´tÃ©, A. Duquette, J.-F. Laplante, M.-A. Legault, P. Moisan, A. Ponchon, C. RaÃ¯evsky, M. Roux, T. Salter, J.-M. Valin, S. Caron, and M. Lauria, "A brochette of socially interactive robots." 01 2005, pp. 1733–1734.

[4] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216 – 228, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921889006001576

[5] M. S. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," 1996.

[6] B. Van Den Broeck, A. Bertrand, P. Karsmakers, B. Vanrumste, H. Van hamme, and M. Moonen, "Time-domain generalized cross correlation phase transform sound source localization for small microphone arrays," 09 2012.

[7] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, pp. 384–91, 02 2000.

[8] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A practical time-delay estimator for localizing speech sources with a microphone array," pp. 153 – 169, 1995. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230885700095

[9] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, April 1997, pp. 375–378 vol.1.

[10] H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda, and E. Osawa, "Robocup: The robot world cup initiative," 1995.

[11] R. T. Committee. (2019) Robocup standard platform league (nao) rule book. [Online]. Available: https://spl.robocup.org/wp-content/uploads/downloads/Rules2019.pdf

[12] A. Hasselbring, “Implementierung und evaluation einer pfeifendetektion fÃ¼r den nao-roboter,” TUHH, 2017.

[13] (2019) Gamecontroller statistics from the robocup 2019. [Online]. Available: https://spl.robocup.org/wp-content/uploads/2019/07/RoboCup2019Statistics.pdf

[14] W. Zhou, Y. Ling, Y. Zhang, and W. Wu, “Time difference calculation based on signal starting point detection,” in *2015 7th International Conference on Modelling, Identification and Control (ICMIC)*, Dec 2015, pp. 1–5.

[15] T. H. Zaw and N. War, “The combination of spectral entropy, zero crossing rate, short time energy and linear prediction error for voice activity detection,” in *2017 20th International Conference of Computer and Information Technology (ICCIT)*, Dec 2017, pp. 1–5.

[16] J.-l. Shen, J.-w. Hung, and L.-s. Lee, “Robust entropy-based endpoint detection for speech recognition in noisy environments,” in *Fifth international conference on spoken language processing*, 1998.

[17] J. . Valin, F. Michaud, J. Rouat, and D. Letourneau, “Robust sound source localization using a microphone array on a mobile robot,” pp. 1228–1233 vol.2, Oct 2003.

[18] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, August 1976.

[19] J. Hassab and R. Boucher, “Optimum estimation of time delay by a generalized correlator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 373–380, August 1979.

[20] ——, “A probabilistic analysis of time delay extraction by the cepstrum in stationary gaussian noise,” *IEEE Transactions on Information Theory*, vol. 22, no. 4, pp. 444–454, July 1976.

[21] L. Chen, Y. Liu, F. Kong, and N. He, “Acoustic source localization based on generalized cross-correlation time-delay estimation,” *Procedia Engineering*, vol. 15, pp. 4912–4919, 12 2011.

[22] I. Cespedes, Y. Huang, J. Ophir, and S. Spratt, “Methods for estimation of subsample time delays of digitized echo signals,” *Ultrasonic Imaging*, vol. 17, no. 2, pp. 142–171, 1995, pMID: 7571208. [Online]. Available: https://doi.org/10.1177/016173469501700204

[23] L. Svilainis, K. Lukoseviciute, V. Dumbrava, and A. Chaziachmetovas, “Subsample interpolation bias error in time of flight estimation by direct correlation in digital domain,” *Measurement*, vol. 46, pp. 3950–3958, 12 2013.

[24] S. J. Julier and J. K. Uhlmann, “Unscented filtering and nonlinear estimation,” *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, March 2004.

[25] R. T. Committee. (2019) Robocup standard platform league (nao) technical challenges. [Online]. Available: http://spl.robocup.org/wp-content/uploads/downloads/Challenges2019.pdf

[26] (2019) Nao 6 - preliminary marketing datasheet. [Online]. Available: https://robotics.ostechnology.co.jp/wp-content/themes/ostRobots_1811/_pdf/NAOV6_Datasheet_EN.pdf

[27] (2019) Alsa project - the c library reference. [Online]. Available: https://www.alsa-project.org/alsa-doc/alsa-lib/index.html

[28] H. Zhu, Z. Li, and Q. Cheng, "Sound source localization through optimal peak association in reverberant environments," in *2017 20th International Conference on Information Fusion (Fusion)*, July 2017, pp. 1–6.