Friedrich-Alexander-University Erlangen-Nuremberg

Chair of Multimedia Communications and Signal Processing

Prof. Dr.-Ing. Walter Kellermann

Bachelorthesis

Sound Source Localization with the robot NAO

Markus Bachmann

August 2014

Supervisors: Ph.D. Antoine Deleforge

M.Sc. Roland Maas

Lehrstuhl für Multimediakommunikation und Signalverarbeitung
Universität Erlangen-Nürnberg

Bachelorarbeit für Herrn Markus Bachmann

Nr. 717

Topic: **Sound Source Localization with the robot NAO**

Description:

The problem of localizing one or several sound sources by a microphone array has been extensively studied in the past decades. The most common approach consists in calculating *time differences of arrival* (TDOAs) of the signal between microphone pairs, which can then be mapped to Directions of Arrival (DoAs). This approach has been successfully used in "clean" environments, *i.e.*, with no or little reverberation, little noise, and a direct propagation of the sound from sources to microphones. However, matters become more complicated in real-world environments. In this thesis, we will focus on the specific challenges brought by the emerging field of *Robot Audition*, within the framework of the new European project EARS, which aims at developing auditory capabilities for the humanoid robot NAO.

Sound source localization with the robot NAO is particularly challenging for three main reasons: (i) there is significant noise generated by its CPU's fan, (ii) recordings will be made in a real-world environment, and hence in reverberant conditions, (iii) the shape of NAO's head induces filtering effects, which makes the level and phase differences between microphones frequency-dependent.

The objectives of this bachelor thesis are to implement a baseline sound-source localization method based on TDOA, test it on NAO recordings, and compare these results with implementations of more elaborated methods, such as the MUSIC algorithm and some of its variants. These methods will be implemented in Matlab and tested on recordings made in various conditions (single or multiple sound sources, with various DoAs, reverberation levels, distances…) and compared to a blind source separation-based method developed at LMS. Eventually, real-time C/C++ implementations on the robot NAO could be envisioned.

Supervisors: Dr. Antoine Deleforge, M.Sc. Roland Maas

Professor: Prof. Dr.-Ing. Walter Kellermann

Begin : 01-04-2014

Erklärung

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

_____ _____

Ort, Datum Unterschrift

Contents

Kurzfassung

In dieser Arbeit soll die Fähigkeit des Roboters NAO akustische Quellen zu orten verbessert werden. Hierbei wird ausgenutzt, dass die ausgesendeten Signale abhängig von der Richtung ihres Ursprungs an den Sensoren zu unterschiedlichen Zeitpunkten ankommen.

Zuerst werden verschiedene Modelle, die diese Zeitdifferenzen beschreiben, eingeführt und ihre physikalischen Voraussetzungen erläutert. Danach werden die verschiedenen, hier verwendeten Algorithmen zur akustischen Quellenortung vorgestellt, nämlich eine auf Korrelation basierende Methode sowie Varianten eines Unterraumalgorithmus und eines auf blinder Quellentrennung basierenden Algorithmus. Die Parameter der Modelle werden an die physikalischen Eigenschaften des Roboters NAO angepasst und die gelernten Modelle und die Algorithmen dann bezüglich ihrer Genauigkeit verglichen. Dies erfolgt auf einem neuen, für diesen Zweck erstellten Datensatz.

Abstract

In this thesis the capability of the robot NAO of localizing sound sources shall be improved. Here, it is exploited that the emitted signals arrive at the sensors at different time instants depending on the direction of their origin.
First, different models which characterize those time differences are introduced and their physical premises are explained. Then, the different sound source localization models utilized here are presented, namely, a correlation-based method as well as variants of a subspace algorithm and a blind-source-separation algorithm. The parameters of the models are adapted to the physical properties of the robot NAO and the learnt models and the algorithms are then compared in terms of accuracy. This is done using a new dataset created for this purpose.

List of Abbreviations

ADP	Averaged Directivity Pattern
AVG	AVeraGe of the absolute error
BSS	Blind Source Separation
CT	Computation Time
CTC	Computation Time Coefficient
DN	DOA Normalized
DOA	Direction Of Arrival
GEVD	Generalized EigenValue Decomposition
HUMAVIPS	HUManoids with Auditory and Visual abilities In Populated Spaces
MUSIC	MUltiple SIgnal Classification
MWF	Multi-channel Wiener Filter
OUT	OUTliers
SD	Standard Deviation of the absolute error
SDW-MWF	Speech Distortion Weighted Multi-channel Wiener Filter
SEVD	Specialized EigenValue Decomposition
SFM	Spectral Flatness Measure
STFT	Short Time Fourier Transform
TDOA	Time Difference Of Arrival
TRINICON	TRIple-N Independent component analysis for CONvolutive mixtures

Notation

$(\cdot) * (\cdot)$	Convolution
$\|(\cdot)\|_p$	p-Norm
$\mathcal{E}\{(\cdot)\}$	Expectation of a signal
$\mathrm{tr}\{(\cdot)\}$	Trace of a matrix
$(\cdot)^T$	Transposed
$(\cdot)^H$	Conjugate-complex transposed (hermitian)
$\mathrm{diag}(\cdot)$	Selects the diagonal submatrices
$\mathbf{x} = (x_1,\ x_2,\ ...,\ x_N)$	Vector x
$\mathbf{X}$	Matrix X or matrix/vector X in the frequency domain
$\hat{x}$	Estimate of x
$\mathbb{I}$	Identity matrix

Chapter 1

Introduction

More and more robots populate our world today, most of them in the industrial sector. But recent development focuses on humanoid cognitive robots. At the moment these robots are mostly toys. In the future they might work as shop assistant or home help for the elders.

For humans the sense of hearing, besides vision, plays an important role in the perception of the environment; especially the ability of localizing sound sources as it provides the power of detecting objects which are not necessarily visually detectable because they are either located out of sight or hidden by obstacles. A robot can profit from this ability in the same way. A domestic home robot could move toward a person that calls for help without seeing it as the person could be localized with the auditory system of the robot.

Another aspect of the necessity of sound source localization for robots is explained in [1, 2, 3]. The more the robots look like humans, the more the customers expect them to act like humans. During a conversation, for example, humans typically look at each other from time to time. Not behaving that way can be interpreted as unfriendliness. Thus, the same behavior is expected from robots if they look like humans. The violation of this human behavioral pattern might influence the further collaboration between the human and the robot in a negative way. In this situation sound source localization is able to help out: After localizing the interlocutor, the head of the robot can be turned

toward the sound source, so the robot looks at the dialog partner and avoids leaving an unfriendly impression.

Besides that, the other technical components of the robot can profit from knowing the positions of the sound sources. Whenever an active source is localized, a beamformer can be steered in its direction in order to improve the signal quality of the recorded sounds. This can be utilized, for example, to enhance the word recognition rate of the automatic speech recognition system embedded into the robot.

1.1 Objective

As explained in the previous section, sound source localization is an important part of the auditory system. Thus, some research has already been done in this area. Besides methods based on the cross-correlation of the signals (e.g. [4, 5]) subspace methods have been developed (e.g. [6, 7, 8, 9]). Some of them have already been applied to robots (e.g. [10, 11, 12]). Recently, algorithms based on *blind source separation* (BSS), especially with the *triple-N independent component analysis for convolutive mixtures* (TRINICON) framework [13, 14], have been investigated (e.g. [15, 16, 17]). These methods calculate the *time differences of arrival* (TDOAs) of the signals and use this knowledge to estimate the *directions of arrival* (DOAs) of the emitting sound sources.

The objective of this thesis is the comparison of the method already introduced during the project *humanoids with auditory and visual abilities in populated spaces* (HUMAVIPS), the *multiple signal classification* (MUSIC) procedure and some of its variants and two variants of the BSS-based algorithms using *averaged directivity patterns* (ADPs) for localization on new recordings done with the robot NAO. Here, it is exploited that the signals traveling along the direct path from the source to the sink exhibit the most energy.

But before these methods can be evaluated, an appropriate model for mapping DOAs to TDOAs has to be found.

1.2 Structure of the Thesis

In **Chapter 2** the basics of sound source localization with the robot NAO are explained. Besides the introduction of the signal model utilized in this thesis, a few common methods which are able to calculate the TDOA of a signal depending on a given DOA are presented. Finally, the reader is acquainted with a Wiener Filter which is used to suppress the noise specific to this application.

In **Chapter 3** the different sound source localization algorithms are explained.

The dataset used for the evaluation of the localization procedures and the detection of speech-only segments in these recordings are presented in **Chapter 4**.

In **Chapter 5** the actual comparison is done. After estimating the TDOAs from the dataset using white-noise recordings, the models are learnt using a least squares algorithm and compared in terms of accuracy. The sound source localization methods introduced in Chapter 3 are then evaluated using the appropriate TDOA models determined in the previous step.

Chapter 6 concludes the thesis and gives an outlook to future projects.

Appendix A contains additional results on the learnt TDOA models and the evaluation of the localization algorithms.

Chapter 2

Basics

In order to transmit a signal across a medium, the sender must control at least one property of the medium. The receiver, on the other hand, has to be able to measure this property with sufficient accuracy. In this thesis the signal that is sent from the source to the sink consists of sound waves, which can be detected by microphones. The later are able to convert these fluctuations of air pressure into an electric signal, which can be processed by computers. In the following, these sensors are assumed to be ideal and isotropic; i.e. the electric signal is equal to the acoustic pressure at a specific time instant at the position of the microphone. Hence the acoustic and electric signals will not be destinguished in this thesis.

2.1 Signal Model

Whenever a signal is emitted, it spreads across different paths. Assuming that the signal is only transmitted through one medium, these can be classified into the direct path and multiple echo or reverberation paths (see Figure 2.1).

The **direct path** is the shortest connection between the source and the sink. The signal is not reflected at any position along its way. Thus, these sound waves arrive from the same direction where the source is located, the runtime through the main medium is the shortest and the amplitude is the highest as no additional attenuation

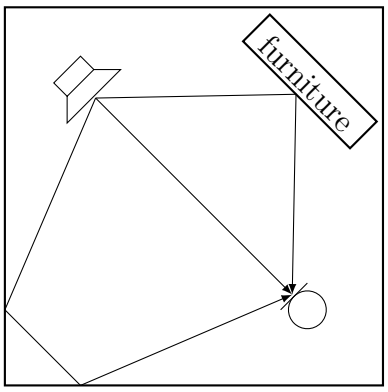


Figure 2.1: Possible propagation directions in an enclosure

besides the one caused by traveling through the main medium occurs.

The sound waves using the **echo paths** for propagation are reflected one or more times on obstacles (e.g. furniture). The attenuation applied to the signal is dependent on the properties of the material of these obstacles. Due to these reflexions the signals traveling along this way typically arrive from a different direction than the signal that uses the direct path.

In most cases it is assumed that the modifications which are done to the signal by the different propagation paths can be modeled by a linear system that depends on the positions of the S sound sources and the microphone and on the enclosure. Dealing with M sensors, the signal $m_q(t)$ with $q \in \{1, 2, ..., M\}$ arriving at the microphone q can be formulated as

$$m_q(t) = \sum_{p=1}^{S} h_{p,q}(t) * s_p(t) + n_q(t) = x_q(t) + n_q(t), \tag{2.1}$$

where $s_p(t)$ denotes the signal emitted by the source p, $h_{p,q}(t)$ the impulse response of the linear system modeling the paths from the static source p to the microphone q and $n_q(t)$ the background noise recorded by the microphone q (cf. Figure 2.2).

In practice, the continuous signals $m_q(t)$ $\forall q \in \{1, 2, ..., M\}$ cannot be captured by the microphones. Instead, these signals are only sampled at discrete time instants with a

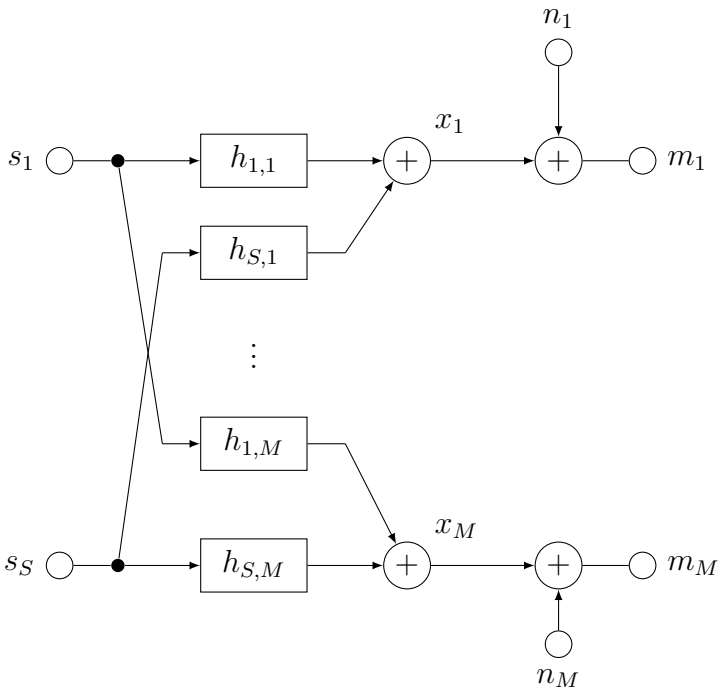


Figure 2.2: Signal model

frequency f_s. Thus, the above signal model can be rewritten as

$$m_q[k] = \sum_{p=1}^{S} h_{p,q}[k] * s_p[k] + n_q[k] = x_q[k] + n_q[k], \tag{2.2}$$

where k denotes the discrete time instant variable. In the following, every algorithm and formula is annotated in the discrete time domain for the above mentioned reason.

2.2 Time Difference Of Arrival

Assuming that the energy of the sound waves traveling along the echo paths is very small compared to the energy of the signal that uses the direct way, the reflected signals can be neglected. Thus, the runtime of the waves from the source p to the sink q can be calculated as

$$k_{p,q} = k_{ref} + \kappa_{p,q}, \tag{2.3}$$

where k_{ref} denotes the time elapsed between the emission at the position of the source and reception at a reference point, which will be a microphone position in most cases, and $\kappa_{p,q}$ the TDOA between the reference position and microphone q when source p emits. Exploiting these time differences $\kappa_{p,q}$ of different microphones, the position of the source can be estimated. For simplicity reasons this thesis focuses on the DOA, i.e. only the azimuth $\phi_{s,p}$ and the elevation $\theta_{s,p}$ of the source p are of interest.

As the results of the sound source localization algorithms are highly dependent on an accurate function $\mathcal{K}(\theta, \phi)$ that converts the DOA (θ, ϕ) to the TDOAs $\kappa = (\kappa_1,\ \kappa_2,\ ...,\ \kappa_M)$ [1], different ways of archiving this mapping are presented in the following.

2.2.1 Free-Field Assumption

The most popular way of calculating the TDOA from the DOA uses the free-field assumption, i.e. the sound waves can travel along straight lines to the different microphones without any obstacles hindering the propagation. Thus, the TDOA can be computed using

$$\kappa_q = \frac{f_s}{c}\left(\|\mathbf{m_q} - \mathbf{s}\|_2 - \|\mathbf{p_{ref}} - \mathbf{s}\|_2\right), \tag{2.4}$$

where $\mathbf{m_q} = (r_{m,q}\cos\theta_{m,q}\cos\phi_{m,q},\ r_{m,q}\cos\theta_{m,q}\sin\phi_{m,q},\ r_{m,q}\sin\theta_{m,q})$ denotes the position of the microphone q, $\mathbf{p_{ref}} = (r_{ref}\cos\theta_{ref}\cos\phi_{ref},\ r_{ref}\cos\theta_{ref}\sin\phi_{ref},\ r_{ref}\sin\theta_{ref})$ the reference point and $\mathbf{s} = (r_s\cos\theta_s\cos\phi_s,\ r_s\cos\theta_s\sin\phi_s,\ r_s\sin\theta_s)$ the position of the source. Inserting these positions leads to

$$\begin{aligned}\kappa_q = \frac{f_s}{c}\Big\{&\sqrt{r_{m,q}^2 + r_s^2 - 2r_{m,q}r_s\left[\cos\theta_{m,q}\cos\theta_s\cos\left(\phi_{m,q} - \phi_s\right) + \sin\theta_{m,q}\sin\theta_s\right]} \\ &-\sqrt{r_{ref}^2 + r_s^2 - 2r_{ref}r_s\left[\cos\theta_{ref}\cos\theta_s\cos\left(\phi_{ref} - \phi_s\right) + \sin\theta_{ref}\sin\theta_s\right]}\Big\}.\end{aligned} \tag{2.5}$$

[1] For the rest of this section κ_q denotes the TDOA between the reference point and the microphone q, which could be caused by an imaginary source with the DOA (θ_s, ϕ_s).

After also applying the far-field assumption ($r_s \to \infty$) to Equation 2.5, the resulting formula is only dependent on the DOA (θ_s, ϕ_s), namely

$$\begin{aligned}\kappa_q = &-\frac{f_s r_{m,q}}{c}\left[\cos\theta_{m,q}\cos\theta_s\cos\left(\phi_{m,q}-\phi_s\right)+\sin\theta_{m,q}\sin\theta_s\right]\\ &+\frac{f_s r_{ref}}{c}\left[\cos\theta_{ref}\cos\theta_s\cos\left(\phi_{ref}-\phi_s\right)+\sin\theta_{ref}\sin\theta_s\right].\end{aligned} \tag{2.6}$$

If the positions of the different microphones are known, like in the case of the robot NAO (see [18, Location]), the reference position $\mathbf{p_{ref}}$ is often set to the origin of the coordinate system specific to the robot, so the second summand is zero and can be omitted.

If the model shall be learnt, one of the sensors has to be chosen as reference and the TDOAs must be computed relative to this microphone; thus κ_r is always zero if $\mathbf{m_r}$ is chosen as $\mathbf{p_{ref}}$.

2.2.2 HUMAVIPS

During the former European project HUMAVIPS some research has already been done in mapping DOAs to TDOAs. According to [4, Section III], the free-field model does not hold well. Instead, the simple model

$$\kappa_q = p_1\theta_s + p_2\phi_s + p_3, \tag{2.7}$$

where p_1, p_2 and p_3 denote the parameters to be estimated, is proposed and learnt using standard linear regression. Some audiovisual recordings of a loudspeaker emitting white-noise attached with a light source have been provided as input to the curve fitting algorithm in order to get a link between the DOA and the TDOA. This audio-visual target has been moved in the visual field of view of the robot.

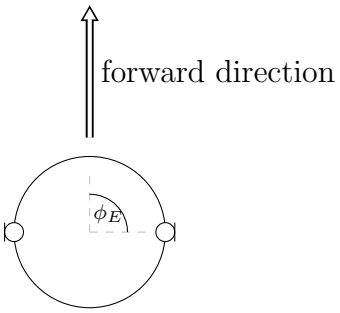


Figure 2.3: Woodworth model

2.2.3 Spherical Head Model

Woodworth Formula

Besides the free-field model discussed in Subsection 2.2.1, the Woodworth model and formula [19] is often employed in (binaural) acoustics (cf. [20, 5]).

In contrast to the free-field model, where no obstacles hinder the sound propagation, the Woodworth formula assumes two antipodal microphones to be mounted on a rigid spherical head at $\phi_E = 90°$ back from the forward direction (cf. Figure 2.3). Therefore, the TDOAs are independent of the elevation θ_s of the sound source [19, Possible binaural cues of direction (pp. 350–352)].

The Woodworth formula takes only one way around the head into account and is given as

$$\kappa_q = \begin{cases} -\frac{f_s a}{c}\left[\pi - \phi_s + \sin\phi_s\right] & \text{if } -\pi < \phi_s \leq -\frac{\pi}{2} \\ -\frac{f_s a}{c}\left[\phi_s + \sin\phi_s\right] & \text{if } -\frac{\pi}{2} < \phi_s \leq 0 \\ \frac{f_s a}{c}\left[\phi_s + \sin\phi_s\right] & \text{if } 0 \leq \phi_s < \frac{\pi}{2} \\ \frac{f_s a}{c}\left[\pi - \phi_s + \sin\phi_s\right] & \text{if } \frac{\pi}{2} \leq \phi_s < \pi \end{cases}, \tag{2.8}$$

where a denotes the radius of the head and c the speed of sound [19, 21].

Note that Equation 2.8 also includes a linear term in the azimuth angle like in Equation 2.7. Therefore, the Woodworth model might be seen as an extension of the model utilized during the project HUMAVIPS.

In contrast to the assumption of the Woodworth formula, the robot NAO is equipped

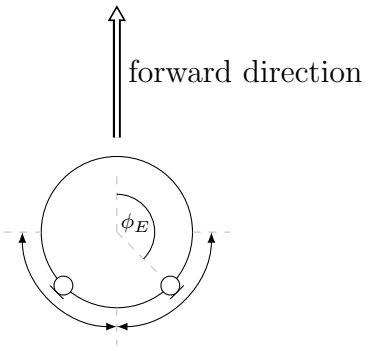


Figure 2.4: Extended Woodworth model

with four microphones instead of only two. For more details on NAO's geometry see Chapter 4 and [22]. Thus, a shifting angle ϕ_{shift} is introduced in order to allow the TDOA estimation based on data recorded by other microphones than the left and the right one. Furthermore, a bias b can be noticed during the TDOA estimation (see Subsection 5.1.1). Hence the modified Woodworth model utilized in this thesis is given as

$$\kappa_q = \begin{cases} -\frac{f_s a}{c}\left[\pi - \tilde{\phi}_s + \sin\tilde{\phi}_s\right] + b & \text{if } -\pi < \tilde{\phi}_s \leq -\frac{\pi}{2} \\ -\frac{f_s a}{c}\left[\tilde{\phi}_s + \sin\tilde{\phi}_s\right] + b & \text{if } -\frac{\pi}{2} < \tilde{\phi}_s \leq 0 \\ \frac{f_s a}{c}\left[\tilde{\phi}_s + \sin\tilde{\phi}_s\right] + b & \text{if } 0 \leq \tilde{\phi}_s < \frac{\pi}{2} \\ \frac{f_s a}{c}\left[\pi - \tilde{\phi}_s + \sin\tilde{\phi}_s\right] + b & \text{if } \frac{\pi}{2} \leq \tilde{\phi}_s < \pi \end{cases}, \tag{2.9}$$

where

$$\tilde{\phi}_s = \phi_s - \phi_{shift}. \tag{2.10}$$

Extended Woodworth Formula

Fortunately, the Woodworth formula has become a recent research topic, so the TDOA estimation can benefit from these results. In [21], the Woodworth model (Equation 2.8) is extended, among other things, to the case of symmetrical sensor angles ϕ_E between 90° and 180° (cf. Figure 2.4) leading to the formula

$$\kappa_q = \begin{cases} -2\frac{f_s a}{c}\left[\pi - \phi_E\right] & \text{if } A\left(\phi_s, \phi_E\right) \text{ is fulfilled} \\ -\frac{f_s a}{c}\left[\cos\left(\phi_s - \phi_E\right) - \cos\left(\phi_s - \phi_E\right)\right] & \text{if } B\left(\phi_s, \phi_E\right) \text{ is fulfilled} \\ -\frac{f_s a}{c}\left[\frac{3\pi}{2} - \phi_s - \phi_E + \cos\left(\phi_s - \phi_E\right)\right] & \text{if } C\left(\phi_s, \phi_E\right) \text{ is fulfilled} \\ -\frac{f_s a}{c}\left[-\frac{\pi}{2} + \phi_s + \phi_E + \cos\left(\phi_s - \phi_E\right)\right] & \text{if } D\left(\phi_s, \phi_E\right) \text{ is fulfilled} \\ -2\frac{f_s a}{c}\phi_s & \text{if } E\left(\phi_s, \phi_E\right) \text{ is fulfilled} \\ 2\frac{f_s a}{c}\phi_s & \text{if } F\left(\phi_s, \phi_E\right) \text{ is fulfilled} \\ \frac{f_s a}{c}\left[-\frac{\pi}{2} + \phi_s + \phi_E + \cos\left(\phi_s - \phi_E\right)\right] & \text{if } G\left(\phi_s, \phi_E\right) \text{ is fulfilled} \\ \frac{f_s a}{c}\left[\frac{3\pi}{2} - \phi_s - \phi_E + \cos\left(\phi_s - \phi_E\right)\right] & \text{if } H\left(\phi_s, \phi_E\right) \text{ is fulfilled} \\ \frac{f_s a}{c}\left[\cos\left(\phi_s - \phi_E\right) - \cos\left(\phi_s - \phi_E\right)\right] & \text{if } I\left(\phi_s, \phi_E\right) \text{ is fulfilled} \\ 2\frac{f_s a}{c}\left[\pi - \phi_E\right] & \text{if } J\left(\phi_s, \phi_E\right) \text{ is fulfilled} \end{cases} \tag{2.11}$$

with

$$A\left(\phi_s, \phi_E\right) = \left(\phi_s \leq \phi_E - \pi\right) \wedge \left(\phi_s \geq \frac{\pi}{2} - \phi_E\right) \tag{2.12}$$

$$B\left(\phi_s, \phi_E\right) = \phi_s \leq \phi_E - \frac{3\pi}{2} \tag{2.13}$$

$$C\left(\phi_s, \phi_E\right) = \left(\phi_s \leq \frac{\pi}{2} - \phi_E\right) \wedge \left(\phi_s \leq \phi_E - \pi\right) \wedge \left(\phi_s \geq \phi_E - \frac{3\pi}{2}\right) \tag{2.14}$$

$$D\left(\phi_s, \phi_E\right) = \left(\phi_s \leq \frac{\pi}{2} - \phi_E\right) \wedge \left(\phi_s \geq \phi_E - \pi\right) \tag{2.15}$$

$$E\left(\phi_s, \phi_E\right) = \left(\phi_s \geq \frac{\pi}{2} - \phi_E\right) \wedge \left(\phi_s \geq \phi_E - \pi\right) \wedge \left(\phi_s \leq 0\right) \tag{2.16}$$

$$F\left(\phi_s, \phi_E\right) = \left(\phi_s \leq \phi_E - \frac{\pi}{2}\right) \wedge \left(\phi_s \leq \pi - \phi_E\right) \wedge \left(\phi_s \geq 0\right) \tag{2.17}$$

$$G\left(\phi_s, \phi_E\right) = \left(\phi_s \geq \phi_E - \frac{\pi}{2}\right) \wedge \left(\phi_s \leq \pi - \phi_E\right) \tag{2.18}$$

$$H\left(\phi_s, \phi_E\right) = \left(\phi_s \geq \phi_E - \frac{\pi}{2}\right) \wedge \left(\phi_s \geq \pi - \phi_E\right) \wedge \left(\phi_s \leq \frac{3\pi}{2} - \phi_E\right) \tag{2.19}$$

$$I\left(\phi_s, \phi_E\right) = \phi_s \geq \frac{3\pi}{2} - \phi_E \tag{2.20}$$

$$J\left(\phi_s, \phi_E\right) = \left(\phi_s \geq \pi - \phi_E\right) \wedge \left(\phi_s \leq \phi_E - \frac{\pi}{2}\right) \tag{2.21}$$

with $\phi_s \in [-\pi, \pi]$ and $\phi_E \in \left[\frac{\pi}{2}, \pi\right]$ (cf. Figure 2.5).

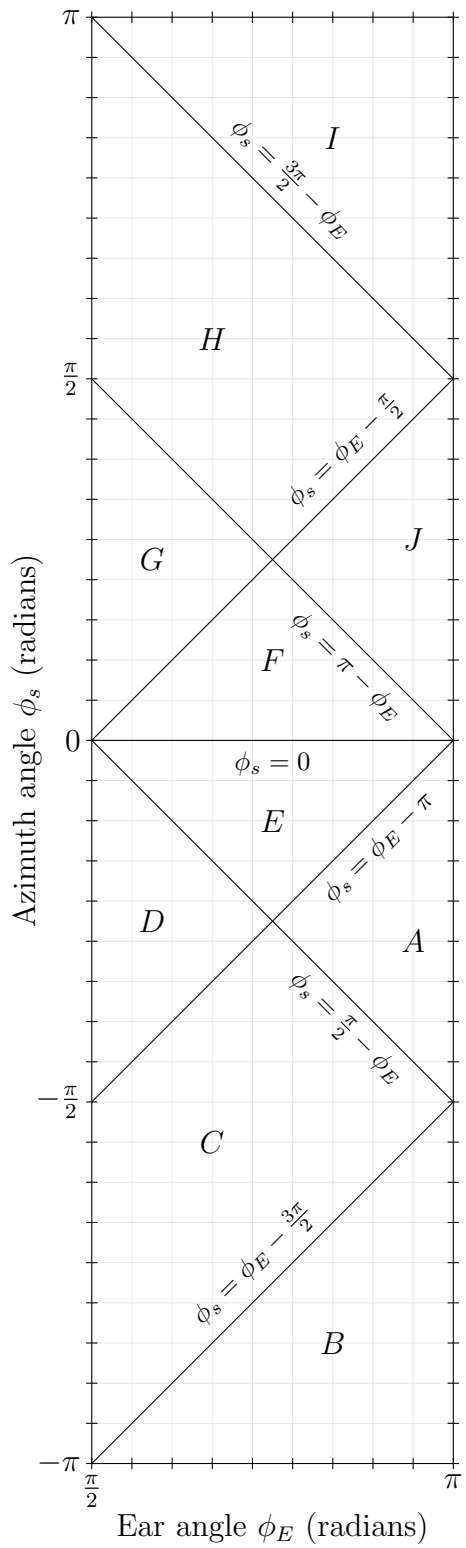


Figure 2.5: Areas of validity of the pieces of the extended Woodworth formula (according to [21, Figure 4])

Like in the case of the pure Woodworth model, the parameters ϕ_{shift} and b are added to Equation 2.11 in order to enable the use of all microphones embedded into the head of the robot NAO, i.e.

$$\kappa_q = \begin{cases} -2\frac{f_s a}{c}\left[\pi - \phi_E\right] + b & \text{if } A\left(\tilde{\phi}_s, \phi_E\right) \text{ is fulfilled} \\ -\frac{f_s a}{c}\left[\cos\left(\tilde{\phi}_s - \phi_E\right) - \cos\left(\tilde{\phi}_s - \phi_E\right)\right] + b & \text{if } B\left(\tilde{\phi}_s, \phi_E\right) \text{ is fulfilled} \\ -\frac{f_s a}{c}\left[\frac{3\pi}{2} - \tilde{\phi}_s - \phi_E + \cos\left(\tilde{\phi}_s - \phi_E\right)\right] + b & \text{if } C\left(\tilde{\phi}_s, \phi_E\right) \text{ is fulfilled} \\ -\frac{f_s a}{c}\left[-\frac{\pi}{2} + \tilde{\phi}_s + \phi_E + \cos\left(\tilde{\phi}_s - \phi_E\right)\right] + b & \text{if } D\left(\tilde{\phi}_s, \phi_E\right) \text{ is fulfilled} \\ -2\frac{f_s a}{c}\tilde{\phi}_s + b & \text{if } E\left(\tilde{\phi}_s, \phi_E\right) \text{ is fulfilled} \\ 2\frac{f_s a}{c}\tilde{\phi}_s + b & \text{if } F\left(\tilde{\phi}_s, \phi_E\right) \text{ is fulfilled} \\ \frac{f_s a}{c}\left[-\frac{\pi}{2} + \tilde{\phi}_s + \phi_E + \cos\left(\tilde{\phi}_s - \phi_E\right)\right] + b & \text{if } G\left(\tilde{\phi}_s, \phi_E\right) \text{ is fulfilled} \\ \frac{f_s a}{c}\left[\frac{3\pi}{2} - \tilde{\phi}_s - \phi_E + \cos\left(\tilde{\phi}_s - \phi_E\right)\right] + b & \text{if } H\left(\tilde{\phi}_s, \phi_E\right) \text{ is fulfilled} \\ \frac{f_s a}{c}\left[\cos\left(\tilde{\phi}_s - \phi_E\right) - \cos\left(\tilde{\phi}_s - \phi_E\right)\right] + b & \text{if } I\left(\tilde{\phi}_s, \phi_E\right) \text{ is fulfilled} \\ 2\frac{f_s a}{c}\left[\pi - \phi_E\right] + b & \text{if } J\left(\tilde{\phi}_s, \phi_E\right) \text{ is fulfilled} \end{cases} \tag{2.22}$$

with

$$\tilde{\phi}_s = \phi_s - \phi_{shift}. \tag{2.23}$$

2.3 Wiener Filter

Every cognitive robot requires a processing unit where the different computations, e.g. sound source localization, can be done. As those units get typically very warm, they require a cooling system so the robot does not overheat while being at work. In the case of the robot NAO, the fan is placed at the back of its head, next to the embedded rear microphone.

The noise produced by the fan and recorded by all microphones is broadband, so only a small amount of the available bandwidth can be used for source localization. In order

to improve the performance of the different sound source localization algorithms, the influence of the fan noise on the microphone signals should be lessened.

A common approach suppressing stationary noise in multichannel signals is filtering the input signal with a *multi-channel wiener filter* (MWF), which minimizes the mean square error between the delayed version of the signal that should be estimated and the estimate itself [23]. A more powerful variant adds speech distortion weighting to the MWF and is derived in [24].

In order to be able to apply the *speech destortion weighted multi-channel wiener filter* (SDW-MWF), a noise-only signal vector $\mathbf{n}_{ref}[k] = (n_{ref,1}[k]\ \ n_{ref,2}[k],\ ...,\ n_{ref,M}[k])^T$ and the input signal vector $\mathbf{m}[k] = (m_1[k],\ \ m_2[k],\ ...,\ m_M[k])^T$ are transformed into the frequency domain utilizing a *short time fourier transform* (STFT). Then the input signal is filtered using

$$\mathbf{M}_{filtered}[b,\mu] = \mathbf{G}^H[b,\mu]\mathbf{M}[b,\mu] \tag{2.24}$$

with $\mathbf{G}[b,\mu]$ denoting the regularized SDW-MWF at a discrete frequency μ and block b given as

$$\mathbf{G}[b,\mu] = \frac{\left(\mathbf{R}_{\mathbf{N}_{ref}\mathbf{N}_{ref}}[b,\mu] + \Delta\right)^{-1}\mathbf{R}_{\mathbf{MM}}[b,\mu] - \mathbb{I}}{\frac{1}{\nu} + \mathrm{tr}\left\{\left(\mathbf{R}_{\mathbf{N}_{ref}\mathbf{N}_{ref}}[b,\mu] + \Delta\right)^{-1}\mathbf{R}_{\mathbf{MM}}[b,\mu]\right\} - M}, \tag{2.25}$$

where $\mathbf{R}_{\mathbf{N}_{ref}\mathbf{N}_{ref}}[b,\mu]$ and $\mathbf{R}_{\mathbf{MM}}[b,\mu]$ denote the correlation matrices of the STFTs of the noise-only signal vector and the input signal vector respectively at frequency μ and frame b (both captured by the microphones), Δ a small regularization parameter and ν the weighting factor controlling trade-off between speech distortion and noise reduction.

The correlation matrices can be estimated by different procedures. One of them averages the matrix products $\mathbf{R}[b,\mu]\mathbf{R}^H[b,\mu]$ over L blocks, i.e.

$$\mathbf{R}_{\mathbf{RR}}[b,\mu] = \frac{1}{L}\sum_{\beta=L-1}^{0}\mathbf{R}[b-\beta,\mu]\mathbf{R}^H[b-\beta,\mu]. \tag{2.26}$$

Alternatively, the correlation matrix can be obtained by a recursive averaging. Here,

the autocorrelation matrix $\mathbf{R_{RR}}[b, \mu]$ of the signal $\mathbf{r}[b, \mu]$ at frame b is obtained as the weighted sum of the matrix $\mathbf{R_{RR}}[b-1, \mu]$ of the previous block $b-1$ and the estimate of the new matrix $\mathbf{R}[b, \mu]\mathbf{R}^H[b, \mu]$. This can be formulated as

$$\mathbf{R_{RR}}[b, \mu] = \gamma \mathbf{R_{RR}}[b-1, \mu] + (1-\gamma)\, \mathbf{R}[b, \mu]\mathbf{R}^H[b, \mu], \tag{2.27}$$

where γ denotes the forgetting factor.

Chapter 3

Methods

After discussing the basics, the different algorithms utilized and compared for sound source localization with the robot NAO are presented in the following.
All of them share the same basic principle. Every method scores all points on a discrete grid $\mathbb{T} \times \mathbb{P}$ of azimuth and elevation angles by means of their corresponding TDOAs between the different microphones; here, the models discussed in Section 2.2 are employed. The DOAs with the best scores are then chosen to be the estimates of the sound source directions.

3.1 HUMAVIPS

One of the simplest techniques of estimating the TDOA between two microphones employs the similarity of the captured signals. Assuming that the direct path provides most of the signal energy, the sounds recorded by the different microphones are delayed versions of each other. Thus, the highest similarity is achieved if the different signals are translated exactly by this time difference in the opposite direction in the time domain.
In order to use this procedure, the captured noise $n_q[k]$ in the signal model (see Section 2.1) has to be assumed to be zero mean, mutually uncorrelated with all other $n_l[k]\ \forall l \in \{1, 2, ..., M\} \setminus \{q\}$ and not correlated with the signals $s_p[k]\ \forall p \in \{1, 2, ...S\}$.

During the project HUMAVIPS this method has been used [4, Subsection II.A].

At every time instant k when the sound source localization shall take place the similarity of the signals is estimated using the correlation coefficients between the rectangular windows $\mathbf{W}_u[k]$ and $\mathbf{W}_v[k]$, each of length L and ending at the time instants of arrival $k_u = k - \kappa_{max} + \mathcal{K}_u(\theta, \phi)$ and $k_v = k - \kappa_{max} + \mathcal{K}_v(\theta, \phi)$ at the corresponding microphones where κ_{max} denotes the maximum possible TDOA, i.e.

$$corr_{u,v}[k_u, k_v] = \frac{\mathrm{Cov}\left(\mathbf{W}_u[k_u]\mathbf{W}_v[k_v]\right)}{\sqrt{\mathrm{Var}\,\mathbf{W}_u[k_u]\,\mathrm{Var}\,\mathbf{W}_u[k_u]}}. \tag{3.1}$$

In order to get a more reliable estimate, the mean of these correlation coefficients of every possible pairwise channel combination is used for the final scoring

$$\overline{corr}[k, \theta, \phi] = \binom{M}{2}^{-1} \sum_{(u,v)\in\binom{M}{2}} corr_{u,v}\left[k + \mathcal{K}_u(\theta, \phi), k + \mathcal{K}_v(\theta, \phi)\right]. \tag{3.2}$$

As a higher correlation coefficients denotes a higher similarity, the maxima of the correlation coefficients indicate the most probable DOAs of the source. In the single source case the DOA is obtained as [4]

$$\left(\hat{\theta}_S[k], \hat{\phi}_S[k]\right) = \underset{(\theta,\phi)\in\mathbb{T}\times\mathbb{P}}{\arg\max}\, \overline{corr}[k, \theta, \phi]. \tag{3.3}$$

If S sources shall be localized, the S DOAs which correspond to the S largest extrema are chosen as the estimated DOAs; i.e. each of the chosen DOAs corresponds to a maximum, which is larger than the correlation coefficients of all of its neighbors.

3.2 MUSIC

Another very common approach estimating a source position is the MUSIC algorithm. This procedure has been developed by Ralph O. Schmidt in [6] and has already been applied to robots, e.g. [10, 11, 12].

The proposed MUSIC algorithm operates in the frequency domain. Thus, the signal

model is transformed into it using a STFT and can then be formulated as

$$M_q[b,\mu] = \sum_{p=1}^{S} H_{p,q}[b,\mu]S_p[b,\mu] + N_q[b,\mu], \tag{3.4}$$

where μ denotes the discrete frequency variable and b the block index.
In the following, the signals of all M microphones and all S sources are combined to the two vectors $\mathbf{M}[b,\mu] = (M_1[b,\mu],\ M_2[b,\mu],\ ...,\ M_M[b,\mu])^T$ and $\mathbf{S}[b,\mu] = (S_1[b,\mu],\ S_2[b,\mu],\ ...,\ S_S[b,\mu])^T$ respectively. Thus, the signal model can be rewritten as

$$\mathbf{M}[b,\mu] = \mathbf{H}[b,\mu]\mathbf{S}[b,\mu] + \mathbf{N}[b,\mu], \tag{3.5}$$

where $\mathbf{N}[b,\mu] = (N_1[b,\mu],\ N_2[b,\mu],\ ...,\ N_M[b,\mu])^T$ denotes the captured noise in the frequency domain and $\mathbf{H}[b,\mu]$ the $M \times S$ delay and attenuation matrix (see Equations 3.7 and 3.6) that links the signals emitted by the S sources to the recorded data of the microphones [6].

$$\mathbf{H}[b,\mu] = (\mathbf{H_1}[b,\mu], \mathbf{H_2}[b,\mu], ..., \mathbf{H_S}[b,\mu]) \tag{3.6}$$

$$= \begin{pmatrix} H_{1,1}[b,\mu] & H_{2,1}[b,\mu] & \cdots & H_{S,1}[b,\mu] \\ H_{1,2}[b,\mu] & H_{2,2}[b,\mu] & \cdots & H_{S,2}[b,\mu] \\ \vdots & \vdots & \ddots & \vdots \\ H_{1,M}[b,\mu] & H_{2,M}[b,\mu] & \cdots & H_{S,M}[b,\mu] \end{pmatrix} \tag{3.7}$$

The first step of the MUSIC algorithm is the calculation of the $M \times M$ correlation matrix $\mathbf{C}[b,\mu]$ out of the vector $\mathbf{M}[b,\mu]$, i.e.

$$\mathbf{C}[b,\mu] = \mathcal{E}\left\{\mathbf{M}[b,\mu]\mathbf{M}^H[b,\mu]\right\} \tag{3.8}$$

$$= \mathbf{H}[b,\mu]\mathcal{E}\left\{\mathbf{S}[b,\mu]\mathbf{S}^H[b,\mu]\right\}\mathbf{H}^H[b,\mu] + \mathcal{E}\left\{\mathbf{N}[b,\mu]\mathbf{N}^H[b,\mu]\right\} \tag{3.9}$$

$$= \mathbf{H}[b,\mu]\mathcal{E}\left\{\mathbf{S}[b,\mu]\mathbf{S}^H[b,\mu]\right\}\mathbf{H}^H[b,\mu] + \lambda\mathbf{C_0}[b,\mu], \tag{3.10}$$

where no correlation between the recorded signal and the noise is assumed. Premising that the number of sources S is less than the number of microphones M, the matrix $\mathbf{H}[b,\mu]\mathcal{E}\left\{\mathbf{S}[b,\mu]\mathbf{S}^H[b,\mu]\right\}\mathbf{H}^H[b,\mu]$ is singular as the rank of $\mathbf{H}[b,\mu]\ \mathcal{E}\left\{\mathbf{S}[b,\mu]\mathbf{S}^H[b,\mu]\right\}$

$\mathbf{H}^H[b,\mu]$ is S. Thus,

$$\det\left(\mathbf{H}[b,\mu]\mathcal{E}\left\{\mathbf{S}[b,\mu]\mathbf{S}^H[b,\mu]\right\}\mathbf{H}^H[b,\mu]\right) = \det\left(\mathbf{C}[b,\mu] - \lambda\mathbf{C_0}[b,\mu]\right) = 0 \tag{3.11}$$

holds which denotes the characteristic equation of a generalized eigenvalue problem. In a next step, the correlation matrix $\mathbf{C}[b,\mu]$ is decomposed into a S dimensional signal subspace and a $N = M - S$ dimensional noise subspace by solving the above problem and assigning the S eigenvectors which belong to the highest S eigenvalues to the signal subspace and the remaining N eigenvectors to the noise subspace [6].
The MUSIC pseudospectrum is then calculated $\forall\,(\theta,\phi) \in \mathbb{T} \times \mathbb{P}$ as

$$P\left[b,\mu,\theta,\phi\right] = \frac{\mathbf{A}\left[\mu,\theta,\phi\right]\mathbf{A}^H\left[\mu,\theta,\phi\right]}{\mathbf{A}^H\left[\mu,\theta,\phi\right]\mathbf{E_N}[b,\mu]\mathbf{E_N}^H[b,\mu]\mathbf{A}\left[\mu,\theta,\phi\right]}, \tag{3.12}$$

where $\mathbf{A}\left[\mu,\theta,\phi\right]$ denotes the mode vector from the DOA (θ,ϕ) to the reference point at frequency μ and $\mathbf{E_N}[b,\mu]$ the $M \times N$ matrix whose columns are the noise eigenvectors [6]. The length of the mode vector $\mathbf{A}\left[\mu,\theta,\phi\right]$ which is projected onto the noise subspace is minimal for every DOA for which

$$\mathbf{A}\left[\mu,\theta,\phi\right] \approx \mathbf{H_u}[b,\mu] \text{ for } u \in \{1,2,...,S\}, \tag{3.13}$$

holds as the noise and the signal subspace are orthogonal [6, The Signal and Noise Subspace]. This leads to a large value at the DOA (θ,ϕ) in the MUSIC pseudospectrum. Besides projecting the mode vectors $\mathbf{A}\left[\mu,\theta,\phi\right]$ onto the noise subspace, it is also possible to determine the distance to the noise subspace by calculating the MUSIC pseudospectrum as

$$P\left[b,\mu,\theta,\phi\right] = \frac{1}{1 - \mathbf{A}^H\left[\mu,\theta,\phi\right]\mathbf{E_S}[b,\mu]\mathbf{E_S}^H[b,\mu]\mathbf{A}\left[\mu,\theta,\phi\right]}, \tag{3.14}$$

where $\mathbf{E_S}[b,\mu]$ denotes the $M \times S$ matrix whose columns are the signal eigenvectors. Here, the mode vectors $\mathbf{A}\left[\mu,\theta,\phi\right]$ are assumed to be of unit length so the pseudospectrum exhibits maxima where the distance to the noise subspace is largest [8, Subsection 2.1]. Note that besides the mode vectors $\mathbf{A}\left[\mu,\theta,\phi\right]$, the eigenvectors of the correlation matrix $\mathbf{C}[b,\mu]$ are also of unit length.

Unfortunately, the MUSIC pseudospectrum can only be obtained at one frequency μ at once. As speech is a broadband signal, the pseudospectrum is averaged between the frequency bins μ_{min} and μ_{max}, using

$$P_{broadband}\left[b,\theta,\phi\right] = \frac{1}{\mu_{max}-\mu_{min}} \sum_{\mu=\mu_{min}}^{\mu_{max}} P\left[b,\mu,\theta,\phi\right], \tag{3.15}$$

in order to process most of the frequencies which may contain speech [10].
Like the correlation-based approach employed during the project HUMAVIPS, the largest maxima of the MUSIC pseudospectrum indicate the most probable DOAs. Thus, the estimated DOA of a single source is obtained as

$$\left(\hat{\theta}_s[b], \hat{\phi}_s[b]\right) = \underset{(\theta,\phi)\in\mathbb{T}\times\mathbb{P}}{\arg\max}\, P_{broadband}\left[b,\theta,\phi\right] \tag{3.16}$$

and in the multiple sources case the MUSIC pseudospectrum is search for the S largest extrema [6]. The DOAs at a time index k are estimated as the ones whose frame ends closest to k.

Estimation of the Correlation Matrix

The quality of the localization results is highly dependent on the correlation matrix $\mathbf{C}[b,\mu]$ as it can be discovered in the above section.
For its estimation the same approaches already introduced in Section 2.3 can be utilized. The correlation matrix is either calculated using the average of the matrix product $\mathbf{M}[b,\mu]\mathbf{M}^H[b,\mu]$ over L frames

$$\mathbf{C}[b,\mu] = \frac{1}{L} \sum_{\beta=L-1}^{0} \mathbf{M}[b-\beta,\mu]\mathbf{M}^H[b-\beta,\mu] \tag{3.17}$$

or employing recursive averaging

$$\mathbf{C}[b,\mu] = \gamma\mathbf{C}[b-1,\mu] + (1-\gamma)\,\mathbf{M}[b,\mu]\mathbf{M}^H[b,\mu], \tag{3.18}$$

where γ denotes the forgetting factor.

MUSIC Variants

Since the development of the MUSIC algorithm many variations of it have been suggested, some of them are presented in this thesis.

The common ***specialized eigenvalue decomposition*** **(SEVD)-MUSIC** variation [8, 10] assumes the captured noise $N_l[b, \mu]$ $\forall l \in \{1, 2, ..., M\}$ to be zero-mean, of equal power on each microphone, stationary, temporally and spatially white and independent on the sources. Thus, the noise correlation matrix

$$\mathcal{E}\left\{\mathbf{N}[b, \mu]\mathbf{N}^H[b, \mu]\right\} = \sigma^2\mathbb{I}, \tag{3.19}$$

where σ^2 denotes the power and variance respectively of the noise. The eigenvalue problem is reduced to a standard specialized eigenvalue problem

$$\mathbf{C}[b, \mu]\mathbf{e_i} = \sigma_i^2\mathbf{e_i}, \tag{3.20}$$

which can be solved easily.

In contrast to the above presented MUSIC variation the ***generalized eigenvalue decomposition*** **(GEVD)-MUSIC** approach in [11, 12] reduces the requirements to the recorded noise signal. The noise correlation matrix $\mathcal{E}\left\{\mathbf{N}[b, \mu]\mathbf{N}^H[b, \mu]\right\}$ only has to be a regular matrix, so the eigenvalue problem can be reformulated as

$$\left(\mathcal{E}\left\{\mathbf{N}[b, \mu]\mathbf{N}^H[b, \mu]\right\}\right)^{-1}\mathbf{C}[b, \mu]\mathbf{e_i} = \sigma_i^2\mathbf{e_i} \tag{3.21}$$

and the results can be computed by one of the standard specialized eigenvalue problem solvers.

According to [8, Subsection 2.1], sharper results can be obtained by calculating the broadband MUSIC pseudospectra $P\,[b, \mu, \theta, \phi]$ utilizing the weighted sum

$$P_{broadband}\,[b, \theta, \phi] = \sum_{\mu=\mu_{min}}^{\mu_{max}} \frac{P\,[b, \mu, \theta, \phi]}{\sum_{(\theta,\phi)\in\mathbb{T}\times\mathbb{P}} P\,[b, \mu, \theta, \phi]}. \tag{3.22}$$

This method is referred as ***DOA normalized*** **(DN)-MUSIC** in the following and can be applied to both variants presented above resulting in the DN-SEVD-MUSIC and DN-GEVD-MUSIC algorithm respectively.

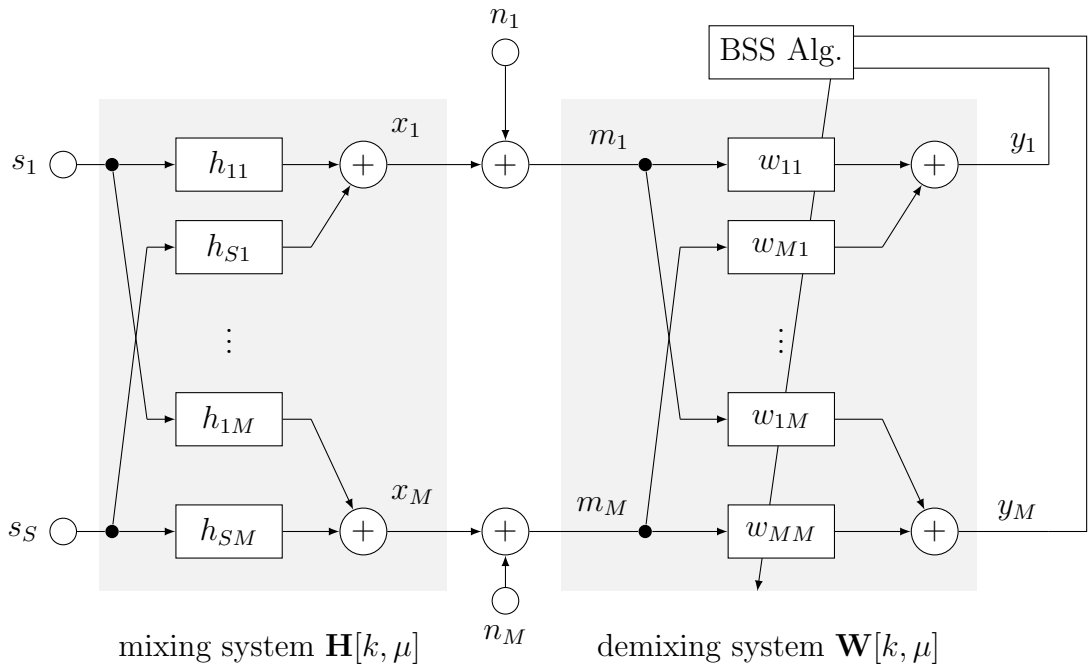


Figure 3.1: BSS setup according to [16, Figure 1]

3.3 BSS-ADP

In contrast to the algorithms presented above, the BSS-ADP method is able to model all impulse responses of the mixing system $\mathbf{H}[b, \mu]$ which characterizes the sound propagation from the sources to the sinks by the use of a continuously adapted demixing system $\mathbf{W}[b, \mu]$ (see Figure 3.1) [13]. The DOAs of the sources can then be extracted out of the estimated impulse responses.

Although the mixing system can also be approximated in the time domain, the estimation of the DOAs using ADPs is done in the frequency domain. Thus, the input signals are transformed into it before further processing using a filterbank. As in the previous section, b denotes the frame variable in the following.

As shown in Figure 3.1, the BSS output is given as

$$\mathbf{Y}[b, \mu] = \mathbf{M}[b, \mu]\mathbf{W}[b, \mu], \tag{3.23}$$

where $\mathbf{Y}[b, \mu] = (Y_1[b, \mu],\ Y_2[b, \mu],\ ...,\ Y_M[b, \mu])$ denotes the output channel vector and

$\mathbf{M}[b,\mu] = (M_1[b,\mu],\ M_2[b,\mu],\ ...,\ M_M[b,\mu])$ the input channel vector.
The demixing system is approximated as the minimum of the cost function imposed by the (assumed) signal properties non-stationarity and non-whiteness using a steepest descent method [14]. Thus, the update equation of the demixing system at block b is given as

$$\mathbf{W}[b,\mu] = \mathbf{W}[b-1,\mu] - \alpha\mathbf{\Delta W}[b,\mu], \tag{3.24}$$

where α denotes the step size parameter and $\mathbf{\Delta W}[b,\mu]$ the update term of the demixing system at frame b and frequency μ [13]. The later can be calculated as [14]

$$\mathbf{\Delta W}[b,\mu] = (\operatorname{diag}\mathbf{R_{YY}}[b,\mu])^{-1} \cdot (\mathbf{R_{YY}}[b,\mu] - \operatorname{diag}\mathbf{R_{YY}}[b,\mu]) \cdot \mathbf{W}[b-1,\mu], \tag{3.25}$$

where $\mathbf{R_{YY}}$ denotes the output correlation matrix estimated as

$$\mathbf{R_{YY}}[b,\mu] = \mathbf{Y}[b,\mu]\mathbf{Y}^H[b,\mu] \tag{3.26}$$

and $(\operatorname{diag}\mathbf{R_{YY}}[b,\mu])^{-1}$ is calculated element by element as

$$\mathbf{R}^{-1}_{\mathbf{YY}ii}[b,\mu] = \frac{1}{\rho\mathbf{R_{YY}}_{ii}[b,\mu] + (1-\rho)\,\sigma_i^2[b,\mu] + \delta_i}, \tag{3.27}$$

where $\mathbf{R_{YY}}_{ii}[b,\mu]$ denotes the element of the matrix $\mathbf{R_{YY}}[b,\mu]$ at row i and column i, ρ a weighting factor, $\sigma_i^2[b,\mu]$ the variance of the output channel i which is computed as the normalized inner product of $\mathbf{R_{YY}}_{ii}[b,\mu]$ and δ_i a regularization parameter.
After approximating the inverse mixing system, the DOAs of the sources can be extracted out of it. For this task the magnitude square response of each output q of the demixing system to the mode vector $\mathbf{A}\,[\mu,\theta,\phi]$ from the DOA (θ,ϕ) at frequency μ is calculated as

$$B_{\mathbf{W}_{:q}}\,[b,\mu,\theta,\phi] = \left\| \sum_{p=1}^{P} \mathbf{W}_{p,q}[b,\mu]\mathbf{A}_p\,[\mu,\theta,\phi] \right\|_2^2. \tag{3.28}$$

In order to exclude unreliable responses because of low resolution at very low frequencies for small sensor apertures and spatial aliasing at high frequencies, the responses are

averaged over all outputs and between the frequency bins μ_{min} and μ_{max}, i.e.

$$\overline{B}_{\mathbf{W}}[b,\theta,\phi] = \sum_{\mu=\mu_{min}}^{\mu_{max}} \sum_{q=1}^{P} B_{\mathbf{W}_{:q}}[b,\mu,\theta,\phi]\,. \tag{3.29}$$

Additionally, the influence of unwanted side lobes is reduced by applying a nonlinear transformation $g(\cdot)$ to the averaged response $\overline{B}_{\mathbf{W}}[b,\theta,\phi]$ [16]

$$\tilde{B}_{\mathbf{W}}[b,\theta,\phi] = g\left(\frac{\overline{B}_{\mathbf{W}}[b,\theta,\phi] - \min \overline{B}_{\mathbf{W}}[b,\theta,\phi]}{\max\left\{\overline{B}_{\mathbf{W}}[b,\theta,\phi] - \min \overline{B}_{\mathbf{W}}[b,\theta,\phi]\right\}}\right). \tag{3.30}$$

According to [17], BSS can be considered "as a set of adaptive nullbeamformers". Thus, the averaged magnitude square response is scanned for the smallest extrema in order to estimate the DOAs of the sources. This leads to

$$\left(\hat{\theta}_S[b], \hat{\phi}_S[b]\right) = \underset{(\theta,\phi)\in\mathbb{T}\times\mathbb{P}}{\arg\min}\, \tilde{B}_{\mathbf{W}}[b,\theta,\phi] \tag{3.31}$$

in the single source case. The averaged magnitude square responses are searched for the S smallest minima when S sources emit. The estimation of the DOAs at a specific time instant k is done the same ways as with the MUSIC algorithm, i.e. the DOA at time instant k is chosen as the one whose frame ends closest to k.

The BSS-ADP approach is known to work very well on two input and output channels. If more than two sensors are available, a pairwise full-channel implementation is also possible besides the full-channel one. In contrast to the later, the adaptation of the demixing system and the computation of the magnitude square response $B_{\mathbf{W}_{:q},v}[b,\mu,\theta,\phi]$ is done for each channel pair v in this case. Before averaging over the outputs and frequencies, the mean of all pairwise magnitude square responses is computed as

$$B_{\mathbf{W}_{:q}}[b,\mu,\theta,\phi] = \frac{1}{V}\sum_{v=1}^{V} B_{\mathbf{W}_{:q},v}[b,\mu,\theta,\phi]\,, \tag{3.32}$$

where V denotes number of channel pairs. Afterward, Equation 3.29 is applied to this result. All further calculations are the same as in the full-channel case.

Chapter 4

Dataset

The algorithms presented in the previous chapter will be tested on a new dataset which is introduced in the following.

The recordings are done in the LMS AudioLab (Room G 0.25) at the Friedrich-Alexander-University Erlangen-Nuremberg. Besides an absorbing reverberation mode, this room also offers a reflecting one. In the absorbing room configuration the panels on the walls are covered with cloth. This reduces the effect of the reverberation and leads to a T_{60} value of 190 ms. If the turned panels are installed, their wooden surface causes a more reflective setup and increases the T_{60} value to 510 ms. The room geometry with the installed panels is depicted in Figure 4.1.

The robot NAO (version 4) is kept in the standard "standing" position during the recordings which are done at a sampling frequency of 48 kHz using the four microphones embedded into the robot's head. The positions of these sensors are detailed in Table 4.1 relative the end of the head transform of the robot, which is located in a height of 0.4596 m [25, Length overview]. The corresponding coordinate system is shown in Figure 4.2.

In the following, all positions and angles are annotated with an accuracy of ± 1 cm and $\pm 1^\circ$ respectively.

First, reference recordings in both room configurations are done with no active sources so the sensors only capture the background noise, i.e. noise produced by the CPU fan of

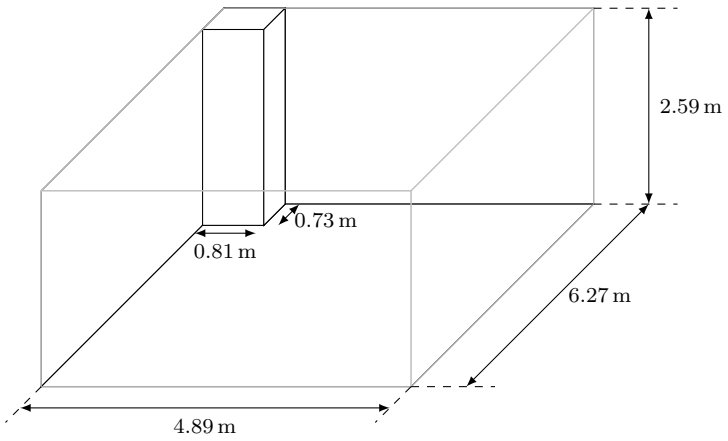
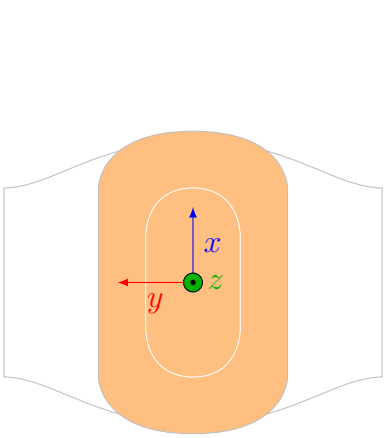


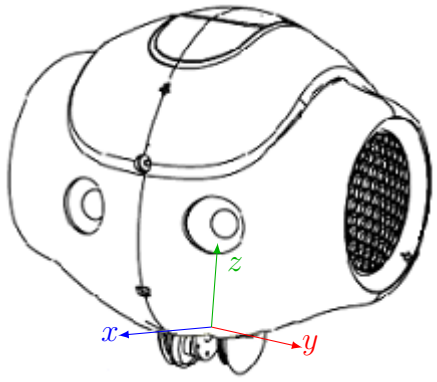
Figure 4.1: LMS AudioLab room geometry

Microphone	Channel	X (m)	Y (m)	Z (m)
Left	1	-0.0195	0.0606	0.0331
Right	2	-0.0195	-0.0606	0.0331
Front	3	0.0489	0.0	0.076
Rear	4	-0.046	0.0	0.0814

Table 4.1: Positions of the microphones embedded into the head of the robot NAO relative to the end of the head transform according to [18]



(a) Top-down view



(b) 3D-View (modified version of [18, Location])

Figure 4.2: Coordinate system of the robot NAO according to [18, Location]

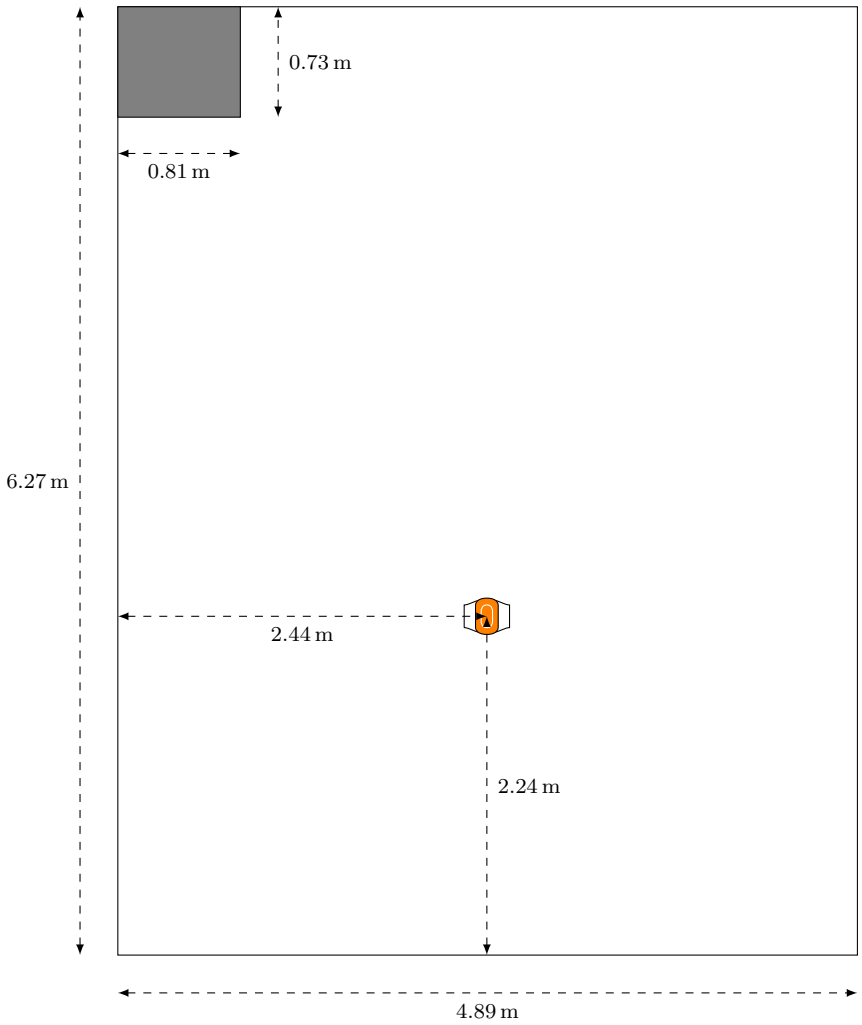


Figure 4.3: Room setup of the reference recordings

the robot and the room background noise. These will be provided as input to the noise correlation matrix estimation needed by the SDW-MWF presented in Section 2.3 and the GEVD-MUSIC variants (cf. Section 3.2). The geometry of this setup is depicted in Figure 4.3.

In order to do the recordings which will be later utilized for the comparison of the different sound source localization methods introduced in Chapter 3, a Genelec 1029A loudspeaker (see [26] for details) is placed at azimuth angles from $-90°$ to $90°$ in $10°$ steps toward the robot. The distance between the z-axis of the robot and the loudspeaker membrane amounts 200 cm. In the case of the recordings in the absorbing

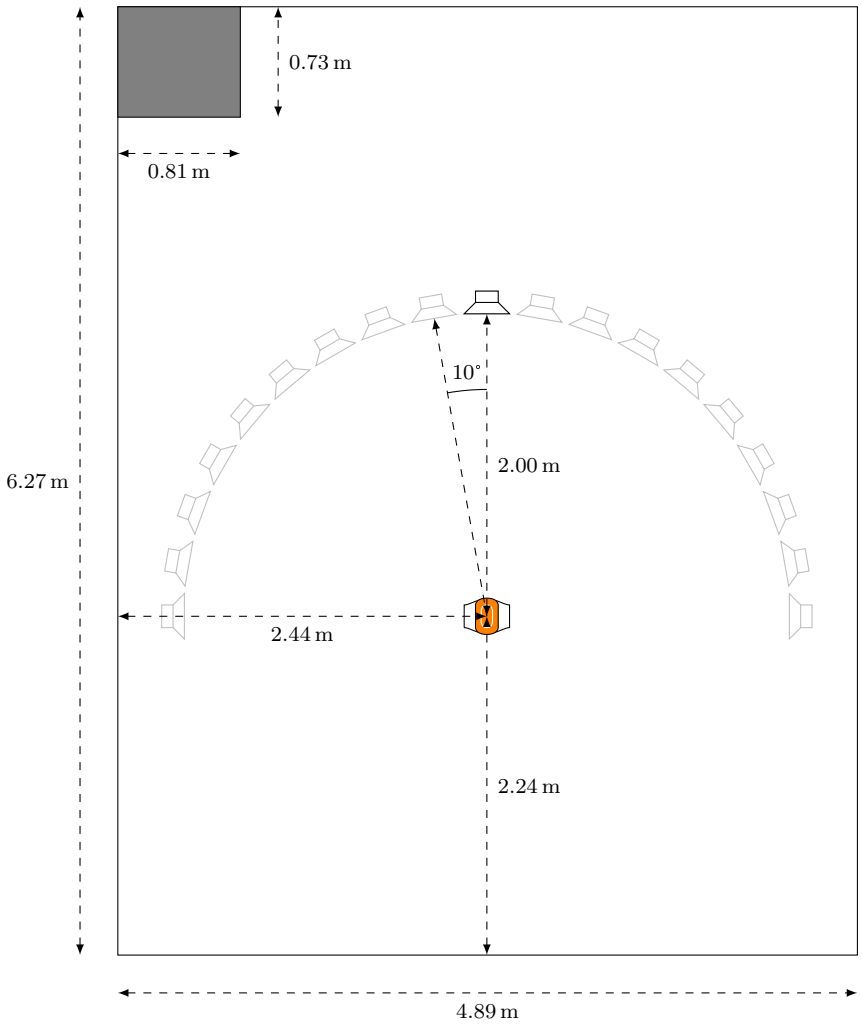


Figure 4.4: Room setup of the single source recordings

room configuration the loudspeaker is located at a height of 93 cm and 143 cm. The recordings in the reflecting room configuration are only done with a loudspeaker at a height of 143 cm. At each of these positions 5 s of a white-noise signal and 5 consecutive random utterances from the TIMIT database [27], which contains male and female speech segments, emitted by the loudspeaker are recorded. The geometry of this setup is depicted in Figure 4.4.

4.1 Estimation of the Speech Activity Periods

As only time instants of speech activity are employed in the comparison, a voice activity detection system is implemented in MATLAB according to [28]. After dividing the signal into frames $f[k]$ of a length of $F = 0.01 f_s \sim 10\,\text{ms}$ (f_s denotes the sampling frequency), some feature are extracted at each frame, namely, the energy

$$E = \sum_{\kappa=0}^{F-1} \|f[\kappa]\|_2^2 \,, \tag{4.1}$$

the frequency exhibiting the maximum value of the spectrum magnitude $\|F[\mu]\|_2$ and the *spectral flatness measure* (SFM)

$$SFM = \frac{\exp\left(\frac{1}{F}\sum_{\mu=0}^{F-1} \ln \|F[\mu]\|_2\right)}{\frac{1}{F}\sum_{\mu=0}^{F-1} \|F[\mu]\|_2}. \tag{4.2}$$

If at least one of the features exceeds the corresponding threshold, the frame currently processed is marked as speech segment otherwise as noise block.

The thresholds are set to the ones suggested in [28], an energy threshold of 40, an SFM value of 5 and the threshold of the frequency denoting, where the maximum value of the spectrum magnitude is located, is defined as 185 Hz. Furthermore, speech segments lasting less than 50 ms are ignored as well as non-speech sections of less than 100 ms.

The speech activity detection is applied to the known emitted signals. Afterward, the location of the maximum of the cross-correlation between the recorded signal and the emitted one is computed. This knowledge is used to shift the results of the voice activity detection in time so the sections marked as speech match the speech segments of the recorded signals. As a last step, the results are converted from the frame domain back to the sample domain for reusability purposes.

Chapter 5

Results

After discussing the different TDOA calculation methods and sound source localization algorithms, the results and the comparison of these methods are presented in the following. As the localization procedures require the conversion from DOAs to TDOAs for their computations, the comparison starts with the different TDOA models.

5.1 Time Difference Of Arrival

As remarked in Section 2.2 the approximation of the mapping from DOAs to TDOAs is one of the limiting factors when it comes to sound source localization. Thus, the different models introduced in Section 2.2 will be adapted to the physical properties of the robot NAO and the results will be compared in this section.

5.1.1 Estimation from Dataset

Before the models can be compared, the actual TDOAs have to be estimated from known data. In this thesis this is done the same way as explained in Section 3.1. But instead of evaluating the cross-covariances of windows of length $1.0 f_s$ (f_s denotes the sampling frequency; i.e. of 1.0 s length) only at specific TDOAs, the cross-covariances for all possible delays are computed. The index of the maximum of the cross-covariance

then denotes the TDOA in samples at this time instant. After clipping impossible TDOAs (> 100 samples)[1], the estimated TDOA for each azimuth angle is obtained as the mean of the remaining results. The white noise recordings in the absorbing room configuration of the dataset are provided as input to this procedure.
In the following, it will turn out that if channel 1 is chosen as reference, the approximation of the TDOAs using the different models exhibits the smallest errors relative to the TDOAs estimated in this section. Thus, only the TDOAs between the channels 1 and 2, 1 and 3 and 1 and 4 are depicted in Figure 5.1.

5.1.2 Comparison

As already discovered during the project HUMAVIPS providing the free-field TDOA models with the documented data leads to rather poor results [4]. Thus, all formulae are adapted to the physical properties of the robot NAO utilizing a reflective trust region algorithm. For completeness reasons the free-field model based on the data from the documentation of the robot NAO ('Free-Field - doc') is also included in the comparison.
The reflective trust region method approximates a nonlinear function using a (quadratic) model and searches for the minimum in the trust region with respect to the given bounds. If this step is successful, the found minimum is set as the starting point to the next iteration. The size of the trust region is then updated depending on the success of the previous step (see [29, 30] for more details).
The parameters of all models (see Equations 2.6, 2.7, 2.9 and 2.22) can be learnt from the TDOAs estimated in the previous section of a source at a height of 93 cm, 143 cm or at both elevations. In this section the results of all three variants are presented.
The start parameters for the adaptation of the linear model utilized during the project HUMAVIPS are chosen to be $(p_1, p_2, p_3) = (1, 1, 0)$ for every channel combination. The 'Free-Field' model is initialized to the same data as used by the 'Free-Field (doc)' for-

[1] 100 samples correspond to a distance of 0.71 m at a sampling frequency of 48 kHz and an assumed velocity of sound of 343 $\frac{\text{m}}{\text{s}}$

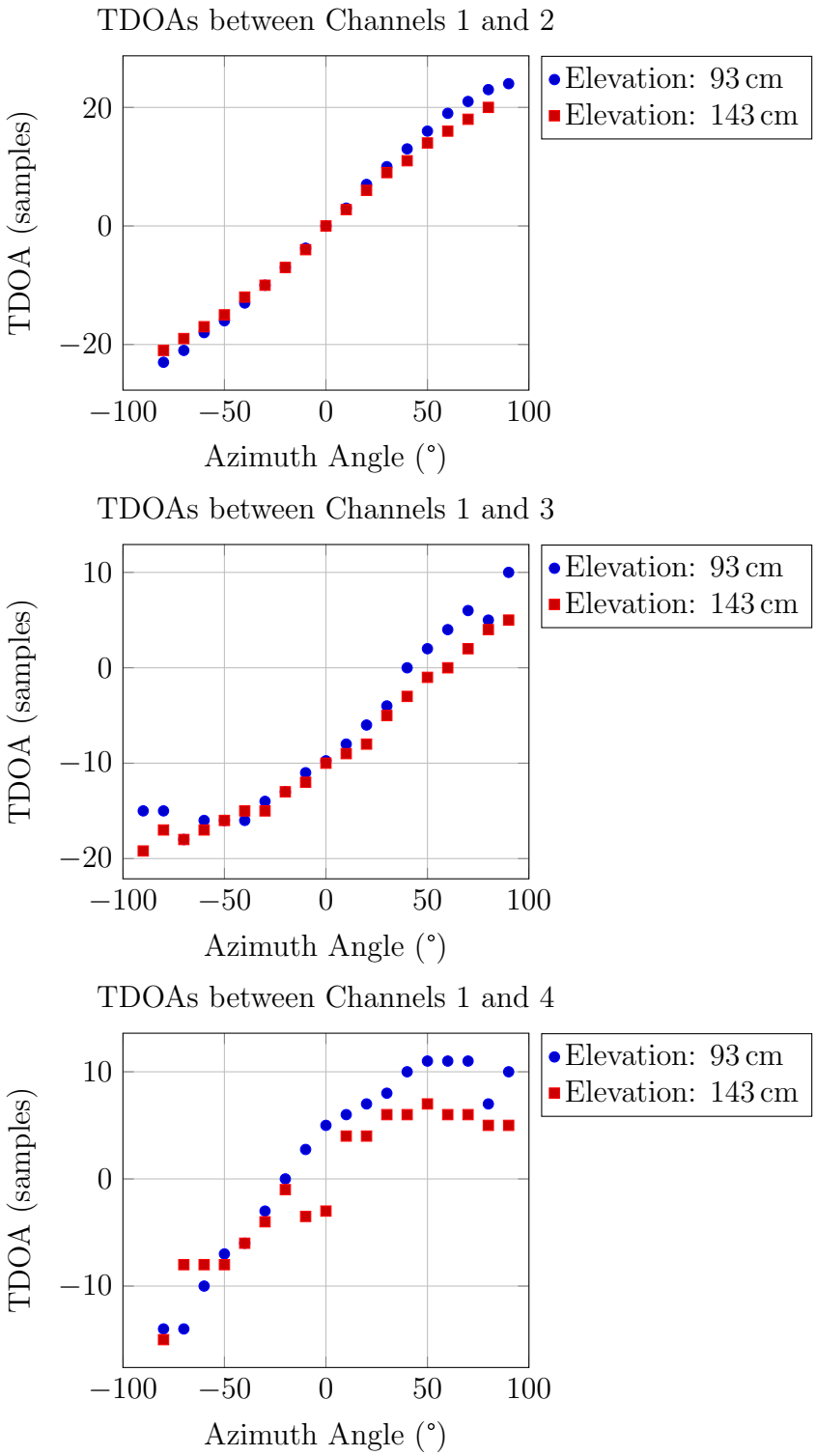


Figure 5.1: Estimated TDOAs between the channels 1 and 2, 1 and 3 and 1 and 4 with channel 1 as reference

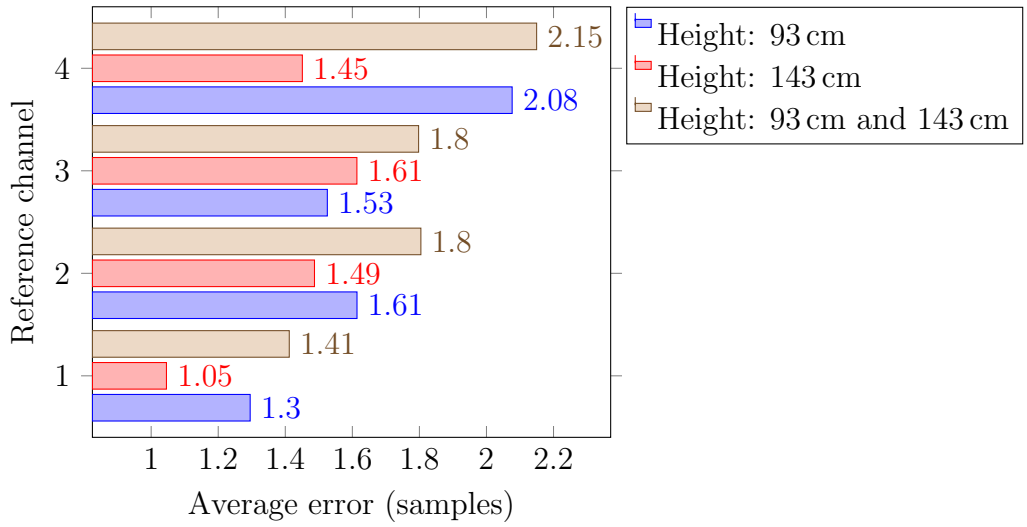


Figure 5.2: Average errors in samples depending on a chosen reference channel and the data used for training

mula. The initial parameters of the spherical head models are set to the corresponding values derived from the microphone positions given at Chapter 4 and [18, Location]

In the following, the errors of the presented models are discussed; the results of the regression can be reviewed in the Appendix A.1.

Four microphones are embedded into the head of the robot NAO. Thus, three TDOAs can be measured and utilized for the estimation of the DOA of the sound source. In a first step, one of the microphones has to be chosen as the reference point according to the signal model (cf. Section 2.1). Figure 5.2 depicts the average error over all 19 sound source positions and all models for each possible reference channel depending on the height of the white-noise source which is used for the training of the models. It shows that the best results with an average error of 1.26 samples only are archived if sensor 1 is chosen as the reference microphone. This observation is also confirmed by Tables 5.1, 5.2 and 5.3 which show the average errors between the different models and the estimated TDOAs for different input data using each microphone as reference as the errors of all models seem to behave the same way. Hence the further comparison is focused on this configuration.

Furthermore, the different models mapping DOAs to TDOAs can be compared with

	Reference Channel			
	1	2	3	4
Free-Field (doc)	2.615	2.480	2.080	2.683
Free-Field	0.711	1.358	1.098	1.695
HUMAVIPS	1.797	1.832	2.373	2.774
Woodworth	0.653	1.359	1.288	1.844
Extended Woodworth	0.701	1.040	0.788	1.387

Table 5.1: Average errors in samples depending on the utilized TDOA model learnt from the estimated TDOAs of a source at a height of 93 cm and the reference channel chosen

	Reference Channel			
	1	2	3	4
Free-Field (doc)	1.775	2.090	2.652	2.381
Free-Field	0.740	1.072	1.219	0.968
HUMAVIPS	1.283	2.206	1.744	2.013
Woodworth	0.716	0.999	1.166	0.996
Extended Woodworth	0.713	1.069	1.287	0.896

Table 5.2: Average errors in samples depending on the utilized TDOA model learnt from the estimated TDOAs of a source at a height of 143 cm and the reference channel chosen

	Reference Channel			
	1	2	3	4
Free-Field (doc)	2.195	2.285	2.366	2.532
Free-Field	0.884	1.370	1.609	1.824
HUMAVIPS	1.668	2.131	2.082	2.478
Woodworth	1.142	1.801	1.475	2.100
Extended Woodworth	1.171	1.435	1.456	1.814

Table 5.3: Average errors in samples depending on the utilized TDOA model learnt from the estimated TDOAs of a source at a height of 93 cm and 143 cm and the reference channel chosen

the help of the Tables 5.1, 5.2 and 5.3. As expected from the previous results [4], the error of the model 'Free-Field - doc' is one of the largest in all configurations, followed by the simple model used during the project HUMAVIPS. Surprisingly, the learnt free-field model performs much better than the one derived from the documentation. The learnt model decreases the error by about 1.4 samples on average. Overall the error difference between the two spherical head models is very small and at least one of both formulas performs better than the other ones if the models are learnt and tested on data of sources located at only one height. If the two available heights are employed for training and comparing the models, the learnt 'Free-Field' model offers the smallest errors (cf. Figure 5.3).

According to these results, the 'Woodworth' and the 'Extended Woodworth' model seem to offer the best approximation of the physical properties if the models are learnt and tested on data of a source which is placed at one height. This is not supprising as the TDOAs are estimated on the entire front hemicycle and not on the narrow visual field of view like during the project HUMAVIPS. The microphones are embedded into the head of the robot NAO, which the 'Woodworth' and the 'Extended Woodworth' model honors in its design, in contrast to the 'Free-Field' models.

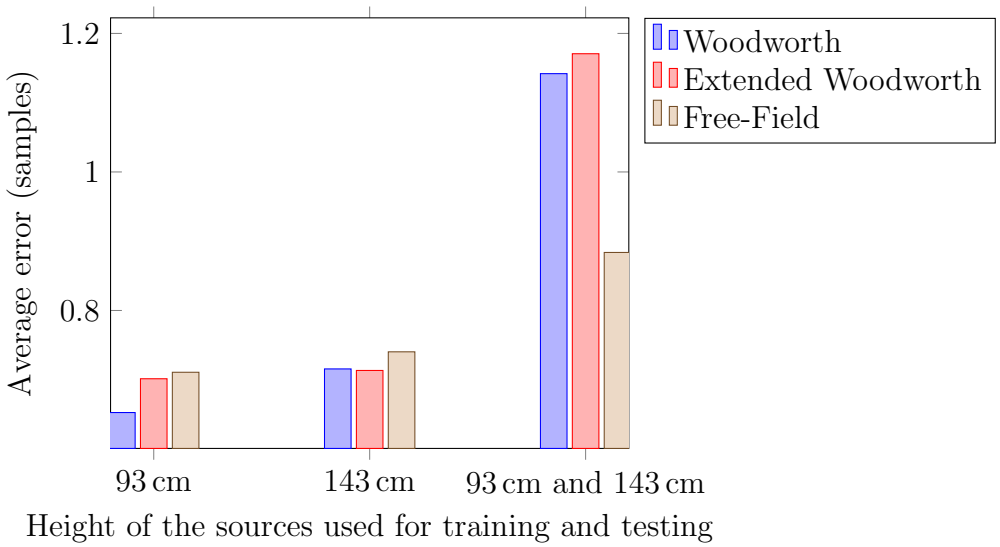


Figure 5.3: Average errors in samples of chosen models depending on the height of the source(s) whose recordings are utilized for training and evaluation of these models

But these models also exhibit a big disadvantage: Although the pure 'Woodworth' model should be able to localize sources at every height, this model shows significant errors if multiple elevations are utilized for learning and testing. The 'Extended Woodworth' model also performs worse in this case as it is limited to the azimuthal plane only at the moment. In the case of multiple heights, models like the learnt free-field formula that also employ the elevation of the sound source in their calculations, show better results.

As the dataset only offers recordings under reflecting conditions with a source at a height of 143 cm besides the ones in the absorbing room configuration, most of the following comparisons will be done on data recorded from loudspeakers located at that height. Thus, the 'Extended Woodworth' model trained on recordings of a white-noise source placed at a height of 143 cm is employed as the default formula mapping DOAs to TDOAs. Besides that, the 'Extended Woodworth' model should offer results close to its actual TDOA estimation quality at this height despite its limitation to the azimuthal plane as the left and the right microphone are located at an elevation angle of $\approx -20°$ relative to center of the head of the robot in spherical coordinates at the back of the robot. The elevation angle of a loudspeaker at a distance of 200 cm and

at a height of 143 cm amounts nearly the same, namely $\approx 25°$. Therefore, the source should be located at least near the azimuth plane of the left and the right microphone.

5.2 Methods

As the best TDOA models for the robot NAO have been found, all requirements for the comparison of the algorithms introduced in Chapter 3 are fulfilled.

The evaluation of the sound source localization methods implemented in MATLAB is done with all four channels at discrete time instants from 100 ms to the end of the signal with a step size of 100 ms on all single source recordings of the dataset presented in Chapter 4. The objective is the estimation of the source azimuth angle knowing the source elevation angle and the number of sources. All computations are repeated with and without Wiener prefiltering utilizing the approach described in Section 2.3. The transformation of the noise and the recorded signal into the frequency domain is done by a STFT with a frame length of 1024 and 50% overlap. The regularization constant is chosen as 10^{-10} and the weighting factor ν as 1. The correlation matrix of the speech signal $\mathbf{R_{MM}}[b, \mu]$ is initialized with an identity matrix and computed using the recursive averaging formula (Equation 2.27) with a forgetting factor of 0.999. The noise correlation matrix $\mathbf{R}_{\mathbf{N}_{ref}\mathbf{N}_{ref}}[b, \mu]$ is estimated using the averaging approach (see Equation 2.26) across all available blocks. In order to compute the latest, the recordings without active sources from the dataset are used as input instead of the actual speech signals. The elements of the matrix are set to zeros as start values.

All algorithms are only compared at time instants with speech activity. The results of the computations are separated into inliers and outliers (OUT) depending on the distance between the estimated azimuth angle and the known one of the sound source. Afterward, the average (AVG) and the standard deviation (SD) of the absolute error of the inliers are computed. As threshold dividing the results into both categories an absolute error of 20° is chosen. Furthermore, the time elapsed during the computation of the results of the different algorithms is measured (CT) and normalized to the signal

length (CTC), i.e.

$$\text{CTC} = \frac{\text{Computation Time (s)}}{\text{Signal Length (s)}}. \tag{5.1}$$

The computation of the DOA is restarted for each azimuth angle of the source and each type of sound (white-noise, speech) played by the source. For all results of the evaluation the interested reader is referred to the Appendix A.2.

The comparison focuses mainly on the amount of outliers as a small bias and/or a wider spread of the estimated source position within a specific bound should be preferred over a bigger amount of calculated DOAs, which are located far away from the actual position of the emitter.

HUMAVIPS

As window length of the correlation-based algorithm HUMAVIPS the same length $L = 100\,\text{ms} = 0.1 f_s$ is chosen as used during the project itself [4]. The parameter κ_{max} is set to 30.

As it can be obtained by the results in Tables 5.4, 5.5, 5.6, 5.7 and 5.8 and Appendix A.2, the method performs better on speech signals if the SDW-MWF is applied to the input signal before processing. This decreases the amount of noise in the signal and also the amount of false positives. In the case of white-noise emission the same behavior is observed in most of the configurations.

MUSIC

The transformation from the time into the frequency domain needed by the MUSIC algorithm is done by a STFT using a von Hann window, a frame length of $F = 1024$

and 50% overlap. The mode vectors are approximated as

$$\mathbf{A}\left[\mu,\theta,\phi\right] = \frac{e^{-\jmath\frac{2\pi\mu}{F}\mathcal{K}(\theta,\phi)}}{\left\|e^{-\jmath\frac{2\pi\mu}{F}\mathcal{K}(\theta,\phi)}\right\|_2} \tag{5.2}$$

$$= \frac{\left(e^{-\jmath\frac{2\pi\mu}{F}\mathcal{K}_1(\theta,\phi)}, e^{-\jmath\frac{2\pi\mu}{F}\mathcal{K}_2(\theta,\phi)}, ..., e^{-\jmath\frac{2\pi\mu}{F}\mathcal{K}_M(\theta,\phi)}\right)^T}{\left\|\left(e^{-\jmath\frac{2\pi\mu}{F}\mathcal{K}_1(\theta,\phi)}, e^{-\jmath\frac{2\pi\mu}{F}\mathcal{K}_2(\theta,\phi)}, ..., e^{-\jmath\frac{2\pi\mu}{F}\mathcal{K}_M(\theta,\phi)}\right)^T\right\|_2}, \tag{5.3}$$

where F denotes the frequency resolution, i.e. the block length of the STFT. For the computation of the pseudospectra the frequencies from 0 Hz to $\frac{f_s}{2}$ are utilized.

The MUSIC algorithm is tested using almost all possible combinations of the variants introduced in Section 3.2. The correlation matrix is always estimated using the recursive averaging approach with a forgetting factor γ of 0.999. If the correlation matrix is calculated limited to a window, the same window length is employed as during the evaluation of the cross-correlation method, i.e. $L = 100$ ms, and the correlation matrix is reset to zeros after each DOA estimation. The factor 0.999 is also used for the recursive averaged estimation of the noise correlation matrix in the case of the GEVD-MUSIC variants.

During the tests (see Tables 5.4, 5.5, 5.6, 5.7 and 5.8 and Appendix A.2) it turns out that most of the time the variants using the window-limited approximation of the correlation matrix perform worse than the ones which are able to remember the past correlation matrices. Furthermore, it seems to be inconsequential, whether the mode vectors are projected onto noise or the signal subspace. The only exceptions to this observation are the GEVD-MUSIC variants. Here, the performance of the versions which are projecting the mode vectors onto the signal subspace is worse than the others. The sharpening of the results as mentioned in [8, Section 2] due the use of DOA normalization is not reproducible with this setup in general; in fact these variants often lead to more outliers (for more details see Appendix A.2). Thus, only the SEVD-MUSIC and the GEVD-MUSIC variants projecting the mode vectors onto the noise subspace and using the recursive averaged estimation of the correlation matrix which is not limited to a window are discussed in the following.

Like in the case of the correlation-based algorithm, the results of the SEVD-MUSIC approach get better if the input data is preprocessed with the SDW-MWF for the same reasons as above. The GEVD-MUSIC algorithm on the other hand returns better results if the input data is not modified. This is explainable as the noise correlation matrix is learnt from the no source recordings of the dataset. An additional preprocessing with a SDW-MWF changes the properties of the noise in the input signal of the GEVD-MUSIC algorithm. This decreases the effect of the GEVD.

BSS-ADP

The transformation into the frequency domain for the BSS-ADP algorithm is done by a filterbank of 256 channels, a decimation ratio of 64 and a prototype length of 512. The prototype is designed by an iterative least squares method. Afterward, an ideal highpass with a cutoff frequency of 160 Hz is applied to the signal.
The diagonal coefficients of the demixing system are initialized to ones and the offdiagonal coefficients to zeros. For every frame returned by the filterbank the demixing system is updated 100 times at a step size of 0.001. The weighting factor ρ is chosen as 0.8 and the regularization constant δ_i as 10^{-9}.
For the ADPs the mode vectors are calculated as

$$\mathbf{A}\left[\mu, \theta, \phi\right] = e^{-\jmath \frac{2\pi\mu}{F} \mathcal{K}(\theta,\phi)} \tag{5.4}$$

$$= \left(e^{-\jmath \frac{2\pi\mu}{F} \mathcal{K}_1(\theta,\phi)}, e^{-\jmath \frac{2\pi\mu}{F} \mathcal{K}_2(\theta,\phi)}, ..., e^{-\jmath \frac{2\pi\mu}{F} \mathcal{K}_M(\theta,\phi)}\right)^T, \tag{5.5}$$

where F denotes the frequency resolution, here 256. In contrast to the mode vectors utilized in the computations of the MUSIC variants, no normalization is performed here. The directivity patterns are averaged over the frequency bins from 0 Hz to $\frac{f_s}{2}$. The nonlinear transform $g(x)$ is chosen to be

$$g(x) = \tanh(2x). \tag{5.6}$$

Both variants, the full-channel and the pairwise full-channel, of the BSS-ADP algorithm are evaluated using the dataset. In the absorbing room configuration the BSS-

ADP method often returns better results if the preprocessing with the SDW-MWF is applied. Under reflective conditions the outcomes show contrary behavior in most configurations. The results of the pairwise BSS-ADP variant get better with SDW-MWF prefiltering in most cases. Overall the pairwise variant often performs worse than BSS-ADP without the application of the SDW-MWF, but exhibits better results than the full-channel approach in many cases with SDW-MWF prefiltering.

Comparison

After the choice of the parameters of the different algorithms the methods presented in Chapter 3 are compared.

As it can be obtained by Figure 5.4 which shows the averaged absolute of all azimuth errors over time of chosen algorithms, the MUSIC and the BSS-ADP variants profit from the ability to remember the their previous results leading to a convergence toward an absolute error of about 5.5° relative to the actual sound source DOA over time. But the convergence behavior of both algorithms seems to be unstable. The methods utilized during the project HUMAVIPS lacks this facility. Thus, the absolute error does not converge, but oscillates around 24°.

The Tables 5.4, 5.5, 5.6, 5.7 and 5.8 show the most important results of the evaluation. In the case of the learnt elevation of the 'Extended Woodworth' model under absorbing conditions the GEVD-MUSIC algorithm which projects the mode vectors onto the noise subspace performs best on both signal types, i.e. the speech signal from the TIMIT database [27] and the white-noise signal. The speech source is located in 94.8% of the cases with a bias of the absolute errors of 3.1 ° and a standard deviation of the absolute errors of 3.0°. The pairwise BSS-ADP method with previous Wiener filtering of the input signal localizes the emitter with similar properties in 92.5% of the cases (see Table 5.4). If white-noise is emitted by the loudspeakers, the amount of outliers decreases to about 2.5% utilizing the GEVD-MUSIC approach for localization, whereas the correlation- and the BSS-based methods seem to offer worse results in this case (see Table 5.5). If the position of a speech source should be estimated under

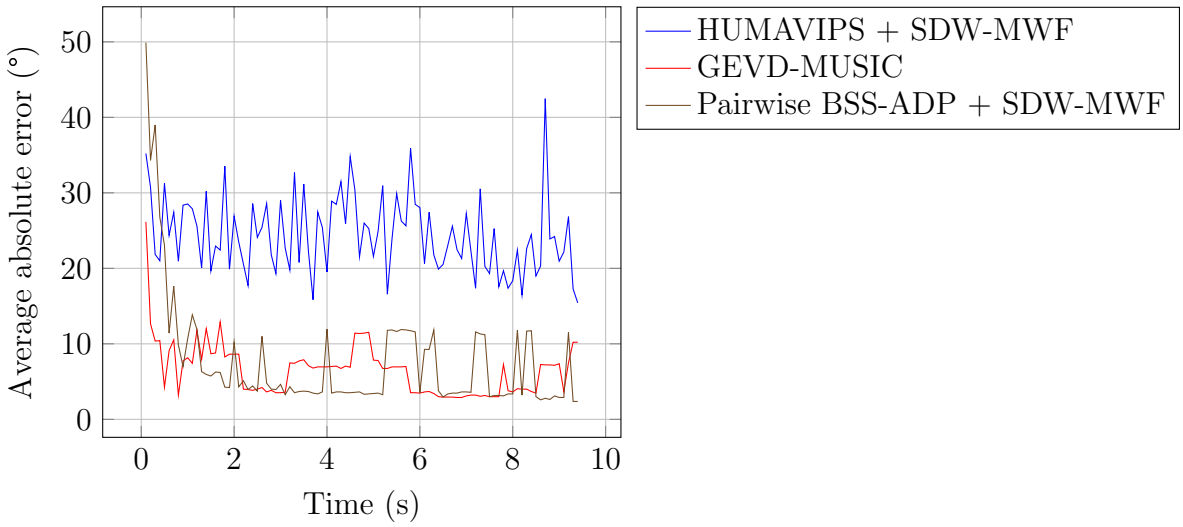


Figure 5.4: Averaged absolute of all azimuth errors over time obtained by the recordings done in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with one source at a height of 143 cm emitting sound from the TIMIT database [27]

reflecting conditions, the correlation-based method exhibits the smallest amount of outliers (27.2%) if a SDW-MWF is applied to the recorded signals. But the performance of all algorithms tested here is rather poor in this case (see Table 5.6). The averaged azimuth errors over time in the reflecting room configuration is depicted in Figure 5.5, which shows besides the non-convergence of the GEVD-MUSIC variant a divergence of the pairwise BSS-ADP algorithm. If the algorithms are evaluated on a signal from a speech source at a height of 93 cm using an 'Extended Woodworth' model, which is only trained on white-noise recordings of sources at a height of 143 cm, the BSS-ADP approach offers the best results with only 22.6% of outliers. The GEVD-MUSIC variant shows 3.8% more outliers in this case (cf. Table 5.7).

Employing the 'Free-Field' model trained on recordings of both available elevations from the dataset (93 cm and 143 cm) for the conversion from DOAs to TDOAs, the overall picture does not change much (cf. Tables A.13, A.14, A.15, A.16, A.17, A.18, A.19, A.20, A.21, A.22, A.23 and A.24 in Appendix A.2). The GEVD-MUSIC and the pairwise BSS-ADP algorithms estimate the position of the source very well with exception of the recordings done in the reflecting room configuration. Here, all algorithms

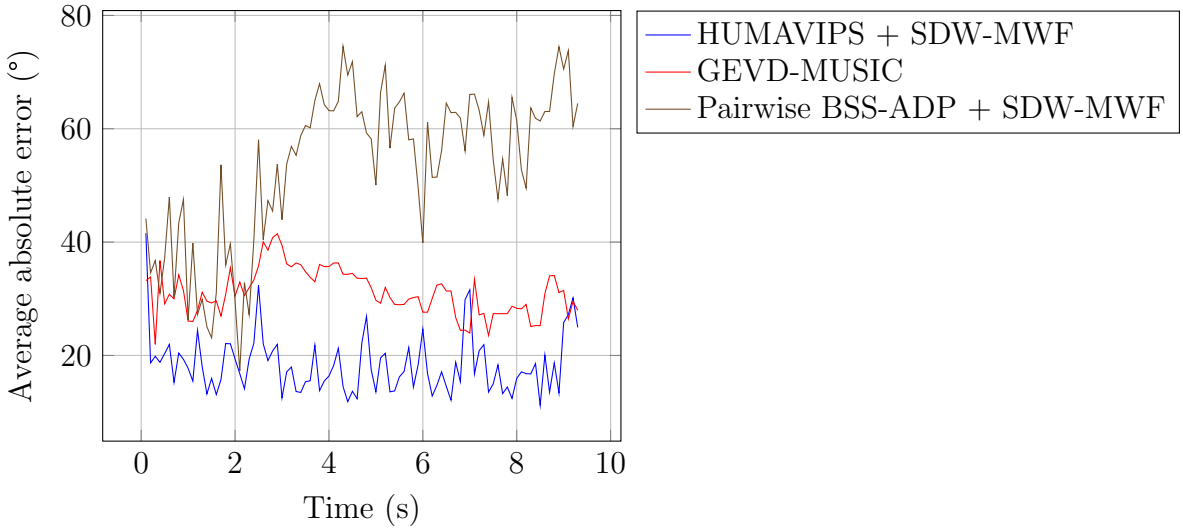


Figure 5.5: Averaged absolute of all azimuth errors over time obtained by the recordings done in the reflecting room configuration ($T_{60} = 510\,\text{ms}$) with one source at a height of 143 cm emitting sound from the TIMIT database [27]

				Wiener prefiltered			
	AVG (°)	SD (°)	OUT (%)	AVG (°)	SD (°)	OUT (%)	CTC
HUMAVIPS	6.1	5.6	47.2	5.8	5.2	31.5	1.4
SEVD-MUSIC (NS)	3.3	3.1	23.7	4.4	4.6	12.1	2.0
GEVD-MUSIC (NS)	**3.1**	**3.0**	**5.2**	3.9	4.3	9.1	2.2
BSS-ADP	2.8	3.2	32.1	4.7	4.5	12.8	199.9
Pairwise BSS-ADP	2.5	3.2	53.7	2.6	3.7	7.5	355.3

Table 5.4: Evaluation results of chosen algorithms using the 'Extended Woodworth' model obtained in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with one source at a height of 143 cm uttering speech signal from the TIMIT database [27]

				Wiener prefiltered			
	AVG (°)	SD (°)	OUT (%)	AVG (°)	SD (°)	OUT (%)	CTC
HUMAVIPS	2.1	1.8	42.1	2.4	2.6	44.9	1.3
SEVD-MUSIC (NS)	2.9	3.8	26.7	1.8	2.7	11.1	2.2
GEVD-MUSIC (NS)	**3.1**	**3.9**	**2.5**	1.6	1.8	13.5	2.3
BSS-ADP	2.7	3.1	17.8	2.3	3.1	20.3	201.1
Pairwise BSS-ADP	2.0	2.5	11.3	1.6	1.7	13.1	357.4

Table 5.5: Evaluation results of chosen algorithms using the 'Extended Woodworth' model obtained in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with one source at a height of 143 cm emitting a white-noise signal

				Wiener prefiltered			
	AVG (°)	SD (°)	OUT (%)	AVG (°)	SD (°)	OUT (%)	CTC
HUMAVIPS	8.3	6.2	48.4	**8.8**	**6.1**	**27.2**	1.3
SEVD-MUSIC (NS)	6.3	5.5	64.9	7.4	5.8	63.6	2.1
GEVD-MUSIC (NS)	4.8	4.8	42.9	6.0	5.1	40.0	2.2
BSS-ADP	4.2	4.3	52.7	5.7	5.3	78.6	199.9
Pairwise BSS-ADP	5.0	5.9	68.5	6.0	5.0	56.2	355.5

Table 5.6: Evaluation results of chosen algorithms using the 'Extended Woodworth' model obtained in the reflecting room configuration ($T_{60} = 510\,\text{ms}$) with one source at a height of 143 cm uttering speech signal from the TIMIT database [27]

				Wiener prefiltered			
	AVG (°)	SD (°)	OUT (%)	AVG (°)	SD (°)	OUT (%)	CTC
HUMAVIPS	6.5	5.0	50.2	7.6	5.8	33.7	1.3
SEVD-MUSIC (NS)	5.3	5.3	38.7	6.6	4.1	35.4	2.1
GEVD-MUSIC (NS)	6.0	3.8	26.4	6.1	4.4	27.8	2.2
BSS-ADP	4.8	4.2	37.0	6.9	4.8	32.9	199.9
Pairwise BSS-ADP	3.7	3.7	61.8	**4.3**	**4.1**	**22.6**	354.5

Table 5.7: Evaluation results of chosen algorithms, which are using the 'Extended Woodworth' model trained on recordings of a white-noise source at a height of 143 cm as a mapping function from TDOA to DOA, obtained in the absorbing room configuration ($T_{60} = 190$ ms) with one source at a height of 93 cm uttering speech signal from the TIMIT database [27]

exhibit a large amount of outliers. Re-evaluating all algorithms on the recordings done with sources at a height of 93 cm (see Table 5.8), it turns out that the lower percentage of inliers in the previous setup (cf. Table 5.7) can be explained by the use of a model not configured for that elevation.

According to these results, the GEVD-MUSIC and the pairwise BSS-ADP algorithms are the best methods for sound source localization in terms of accuracy excluding the evaluation results of the algorithms in the reflecting room configuration, where the performance of all algorithms is very weak. But elapsing about 355 s $\approx$ 6 min per 1 s of input data for computation, the BSS-based approach does not seem to be capable for real-time use; especially if the method should run on a robot with limited processing power like the robot NAO. As the GEVD-MUSIC variant which projects the mode vectors onto the noise subspace only consumes about 2.2 s per 1 s of input signal and additionally offers a silghtly smaller amount of outliers, this method should be the preferred algorithm for sound source localization with the robot NAO.

				Wiener prefiltered			
	AVG (°)	SD (°)	OUT (%)	AVG (°)	SD (°)	OUT (%)	CTC
HUMAVIPS	6.5	4.9	48.7	7.4	5.4	40.7	1.3
SEVD-MUSIC (NS)	4.9	5.7	47.4	2.3	3.3	5.5	2.0
GEVD-MUSIC (NS)	2.1	2.7	4.2	**2.6**	**3.5**	**2.4**	2.2
BSS-ADP	2.8	3.1	11.5	3.5	3.7	16.5	199.5
Pairwise BSS-ADP	3.4	4.2	44.9	2.9	3.4	3.6	354.9

Table 5.8: Evaluation results of chosen algorithms using the 'Free-Field' model obtained in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with one source at a height of 93 cm uttering speech signal from the TIMIT database [27]

Chapter 6

Conclusion

6.1 Summary

In this thesis different models which can be used to convert DOAs into TDOAs and different sound source localization algorithms are presented and compared on a new dataset recorded by the robot NAO.

As already stated in [4] the free-field model using the microphone positions denoted in the documentation [18, Location] offers a rather bad estimate of the TDOA. But the simple model utilized during the project HUMAVIPS also exhibits significant errors. The learnt free-field, the pure Woodworth and the extended Woodworth model feature more accuracy than those. Overall the pure Woodworth and the extended Woodworth model by [21] seem to be the best choice for the robot NAO if the sources which should be localized are placed at one known height. Otherwise the learnt free-field model offers a good DOAs-TDOAs conversion.

The correlation-based sound source localization algorithm exhibits the smallest runtime and a stable amount of outliers. This method is best in the case of the recordings in the reverberant room configuration. In the case of the recordings in the absorbing room configuration, the pairwise BSS-ADP variant and the GEVD-MUSIC algorithm which projects the mode vectors onto the noise subspace feature the smallest amount

of outliers. Having a look at the runtime, the BSS-ADP methods do not seem to be capable for real-time implementations in contrast to the correlation-based algorithm and the MUSIC variants. Furthermore, the advantage of remembering the previous results is shown.

6.2 Future Work

Utilizing the extended Woodworth model presented in Subsection 2.2.3, the computation of the TDOA depending on the elevation θ_s of the sound source should also be possible. The deviation of the corresponding formula is beyond the scope of this thesis, but might be an interesting topic for further research.

The sound source algorithms do not seem to be robust to reverberation. This shows a large room for improvements. Additionally, the MUSIC variants might benefit from the knowledge of the head related transfer functions [7] in terms of accuracy and in terms of speed from coherent signal subspace processing [31, 32]. Furthermore, the methods are only tested with one active source. In the future, evaluation with multiple sound sources can be envisioned.

Appendix A

Results

A.1 Approximation of the different Time Difference Of Arrival Models

In the following, the results of the different learnt TDOA models are shown. Each figure contains the estimated TDOAs from the dataset as explained in Subsection 5.1.1 and the according model with learnt parameters sampled from $-90°$ to $90°$ with a $1°$ resolution.

In addition to the regression results obtained from data captured from a source at a height of 143 cm a 'Free-Field' model learnt from data from both elevations is shown in Figure A.6.

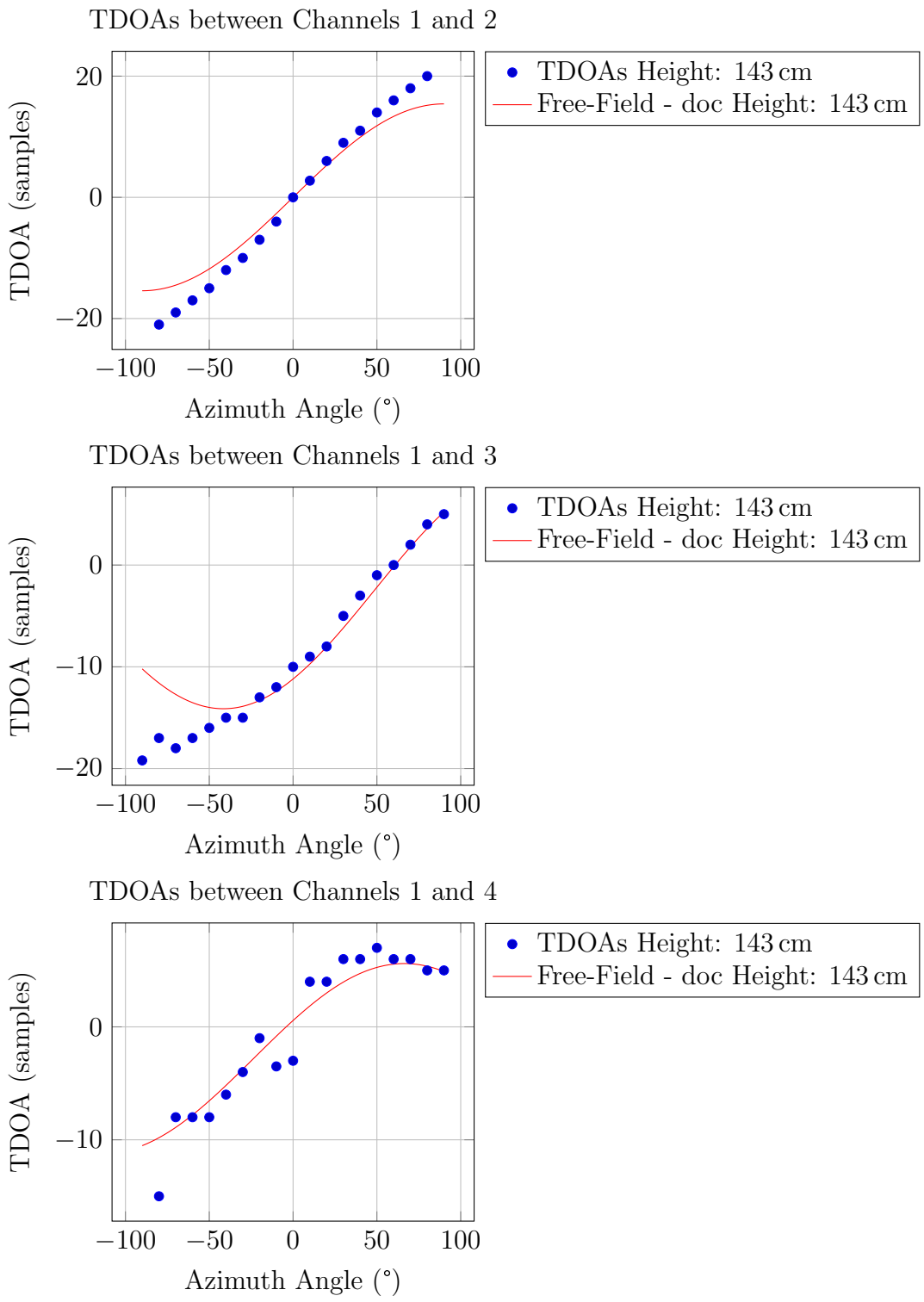


Figure A.1: Estimated TDOAs from the dataset and its approximation using the free-field model with parameters from the NAO documentation [18, Location] between the channels 1 and 2, 1 and 3 and 1 and 4 with channel 1 as reference

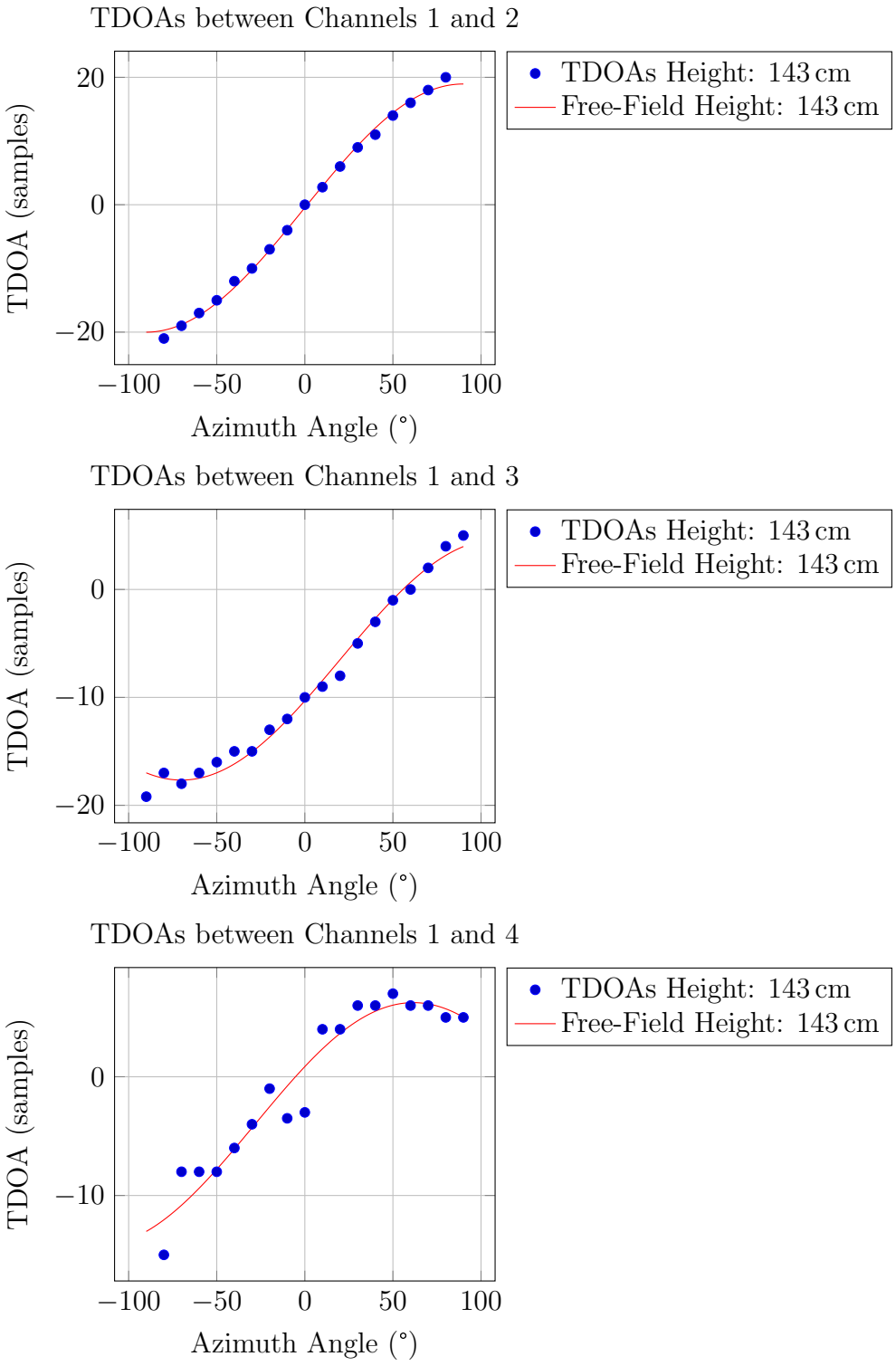


Figure A.2: Estimated TDOAs from the dataset and its approximation using the free-field model with learnt parameters between the channels 1 and 2, 1 and 3 and 1 and 4 with channel 1 as reference

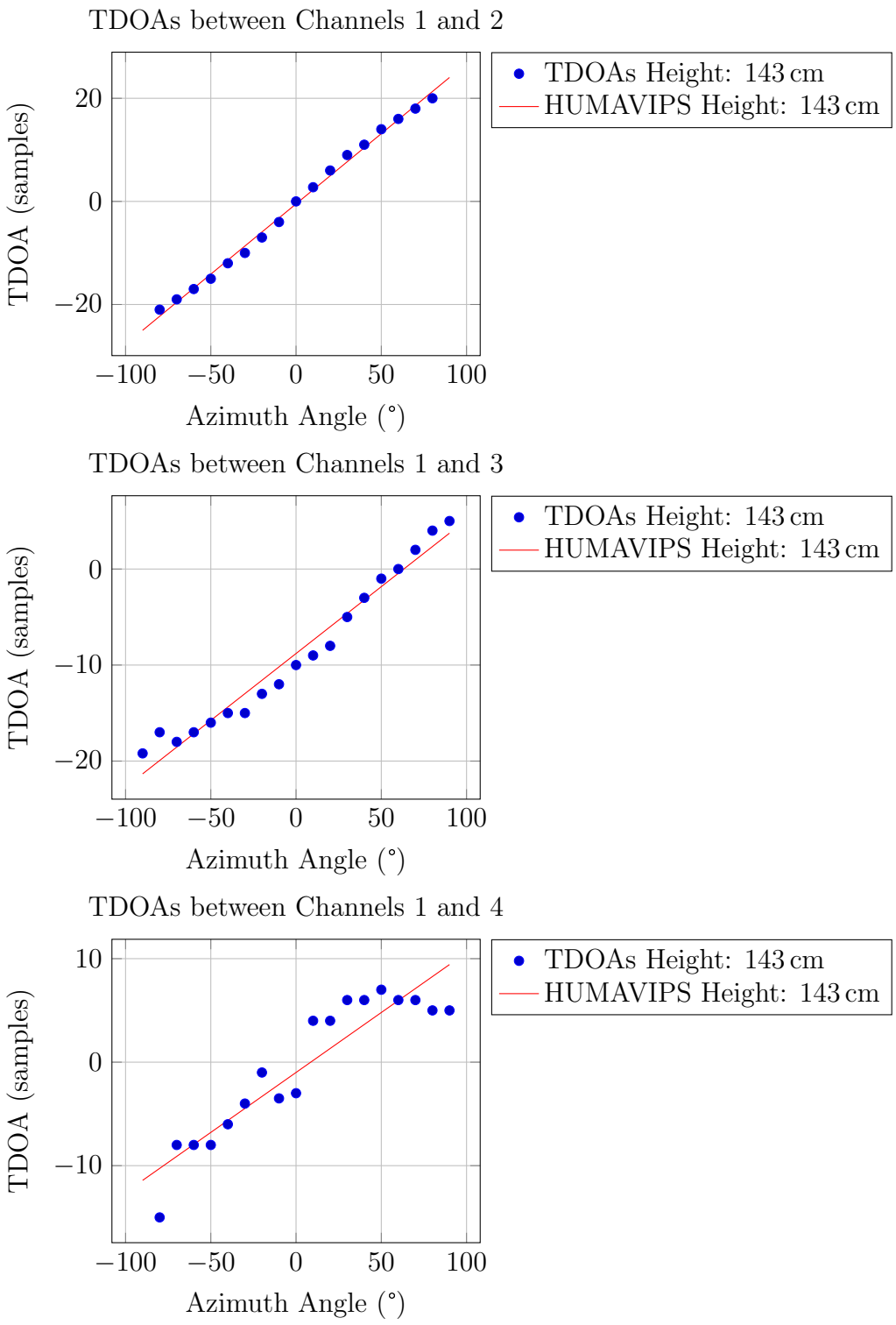


Figure A.3: Estimated TDOAs from the dataset and its approximation using the model developed during the project HUMAVIPS between the channels 1 and 2, 1 and 3 and 1 and 4 with channel 1 as reference

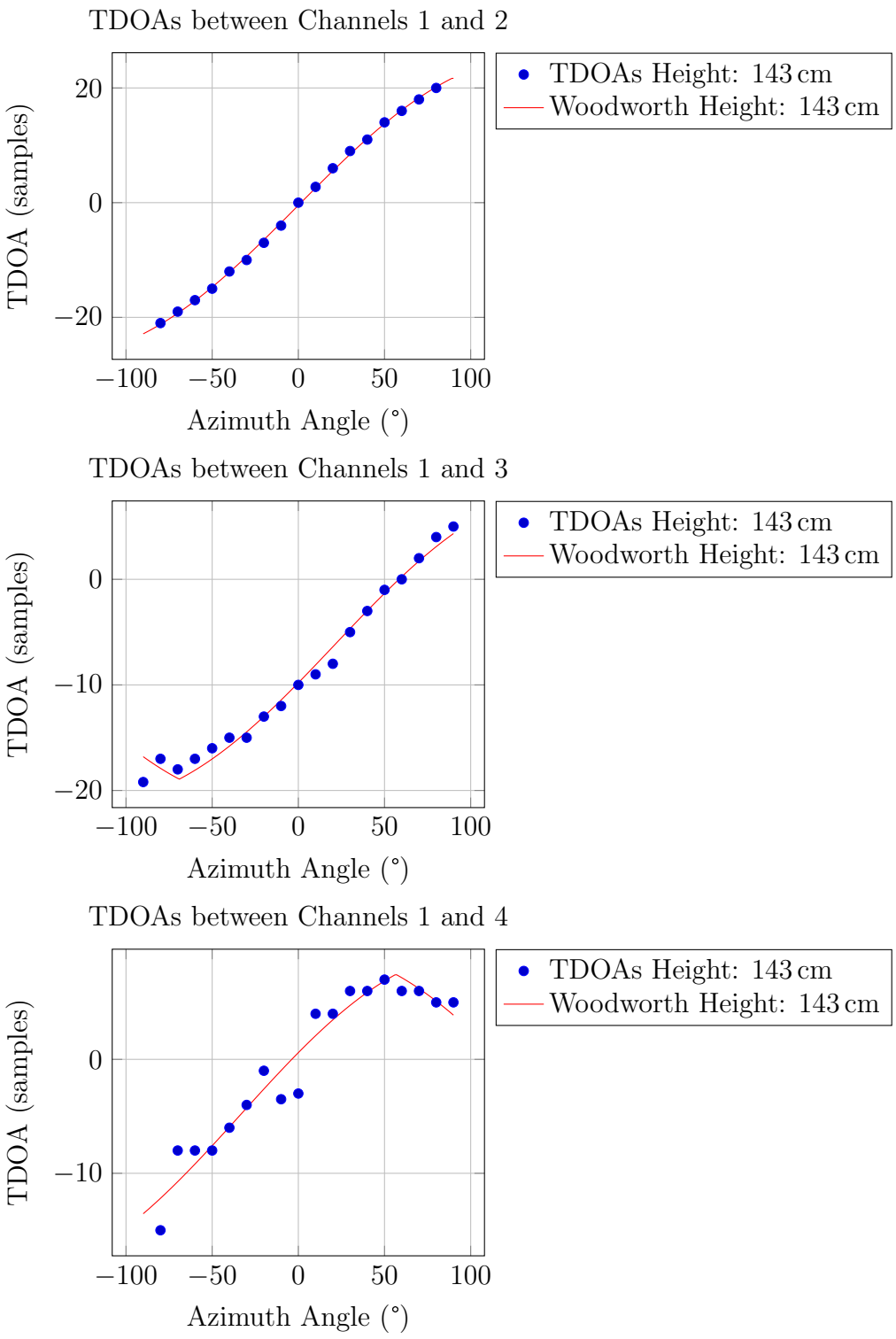


Figure A.4: Estimated TDOAs from the dataset and its approximation using the Woodworth model between the channels 1 and 2, 1 and 3 and 1 and 4 with channel 1 as reference

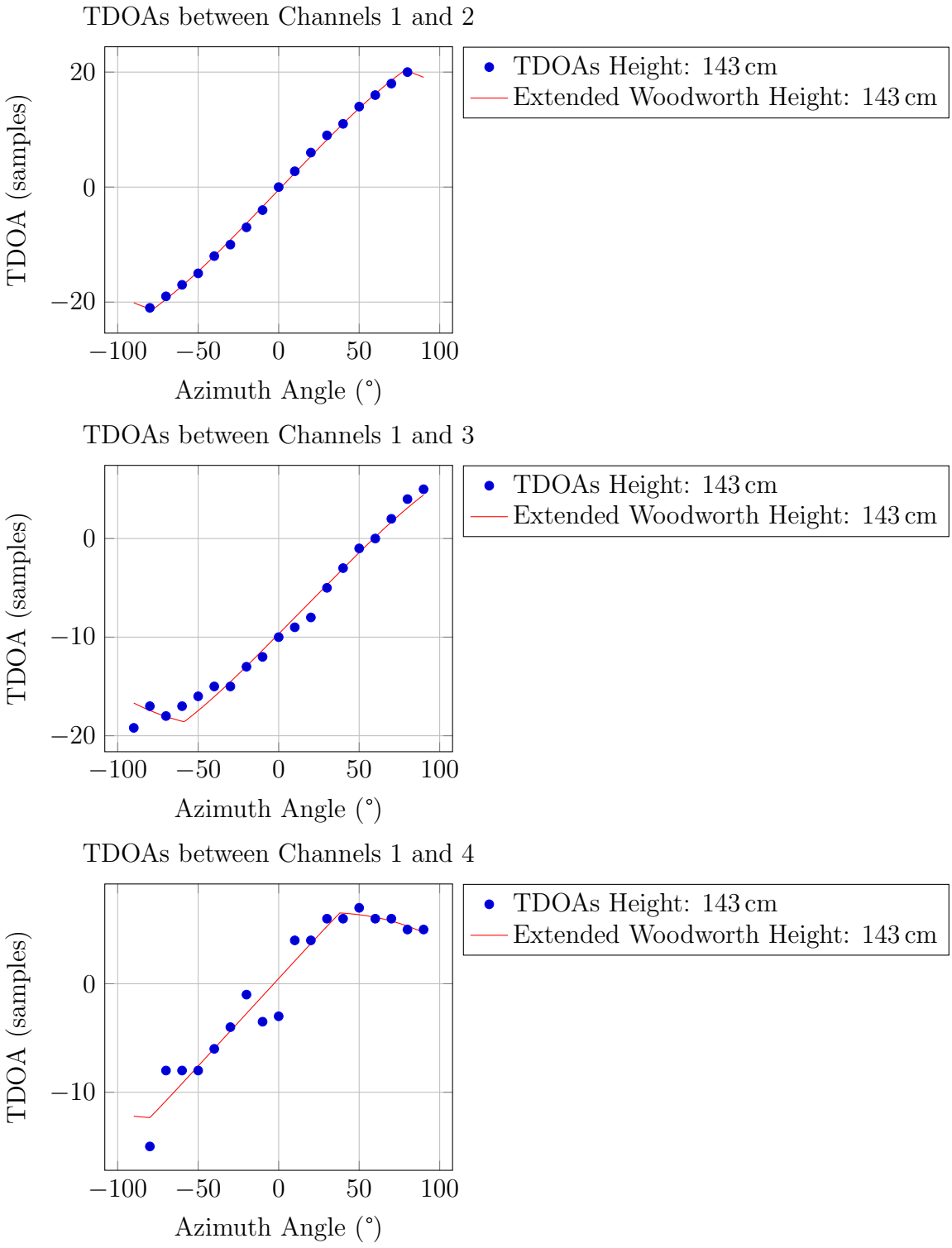


Figure A.5: Estimated TDOAs from the dataset and its approximation using the extended Woodworth model between the channels 1 and 2, 1 and 3 and 1 and 4 with channel 1 as reference

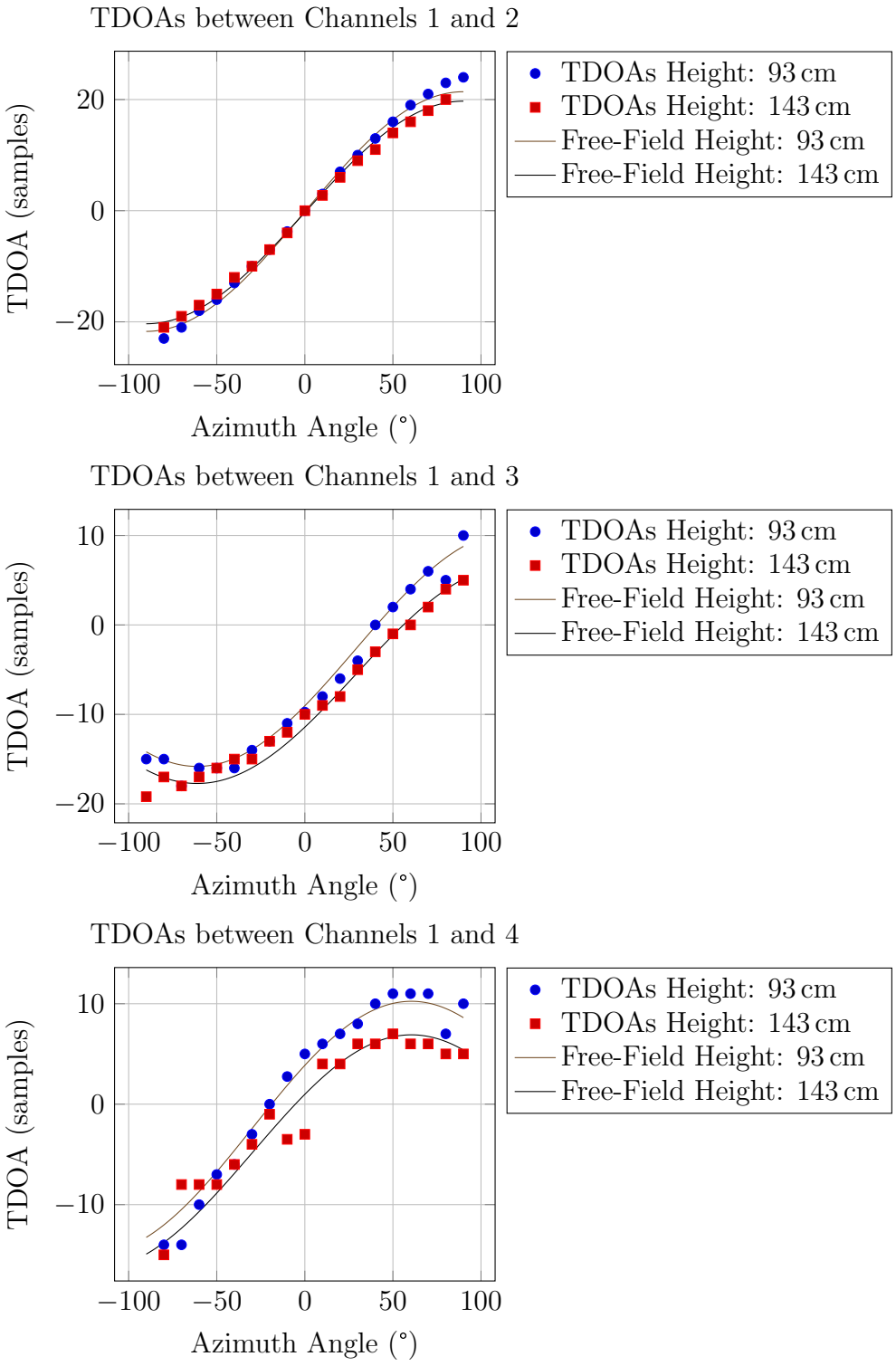


Figure A.6: Estimated TDOAs from the dataset and its approximation using the free-field model with learnt parameters between the channels 1 and 2, 1 and 3 and 1 and 4 with channel 1 as reference

A.2 Methods

In the following, the results of the evaluation of the different algorithms and its variants are detailed. Each method is labeled according to Chapter 3. Additional information is added to the different MUSIC variants. A 'NS' in brackets denotes projection of the mode vectors onto the noise subspace and a 'SS' onto the signal subspace. A preceding 'W' indicates the limitation to a window with the same length as the one used for the HUMAVIPS method for estimation of the correlation matrix. If the 'W' is omitted the recursive averaging technique which is not limited to a window is applied.
In the header of the tables the same abbreviations as in Chapter 5 are used, i.e.

AVG	AVeraGe of the absolute error
CT	Computation Time
CTC	Computation Time Coefficient
OUT	OUTliers
SD	Standard Deviation of the absolute error

The conversion from DOAs to TDOAs in Tables A.1, A.2, A.3, A.4, A.5, A.6, A.7, A.8, A.9, A.10, A.11 and A.12 is done using the 'Extended Woodworth' model trained on the recordings of a white-noise source at a height of 143 cm, in Tables A.13, A.14, A.15, A.16, A.17, A.18, A.19, A.20, A.21, A.22, A.23 and A.24 the 'Free-Field' model learnt from white-noise data of both elevations is utilized.

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	6.5	5.0	50.2	21.6	1.3
SEVD-MUSIC (NS)	5.3	5.3	38.7	34.1	2.1
GEVD-MUSIC (NS)	6.0	3.8	26.4	36.1	2.2
DN-SEVD-MUSIC (NS)	5.6	5.7	41.2	34.0	2.1
DN-GEVD-MUSIC (NS)	5.7	4.5	45.1	36.1	2.2
SEVD-MUSIC (SS)	5.3	5.3	38.7	32.3	2.0
GEVD-MUSIC (SS)	8.2	6.0	73.0	34.3	2.1
DN-SEVD-MUSIC (SS)	5.6	5.7	41.2	32.2	2.0
DN-GEVD-MUSIC (SS)	8.2	6.0	73.1	34.3	2.1
SEVD-MUSIC (W,NS)	7.4	5.2	67.9	33.6	2.0
GEVD-MUSIC (W,NS)	5.9	4.2	40.4	35.6	2.2
DN-SEVD-MUSIC (W,NS)	7.6	5.4	71.8	33.6	2.0
DN-GEVD-MUSIC (W,NS)	5.9	4.3	49.2	36.0	2.2
SEVD-MUSIC (W,SS)	7.4	5.2	67.9	31.9	1.9
GEVD-MUSIC (W,SS)	9.0	5.9	72.2	34.2	2.1
DN-SEVD-MUSIC (W,SS)	7.6	5.4	71.8	32.2	2.0
DN-GEVD-MUSIC (W,SS)	8.9	6.0	72.3	34.4	2.1
BSS-ADP	4.8	4.2	37.0	3295.4	200.0
Pairwise BSS-ADP	3.7	3.7	61.8	5832.4	353.9

Table A.1: Evaluation of the results of the algorithms for the recordings done using the TIMIT database [27] in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with a single source at a height of 93 cm and utilizing the 'Extended Woodworth' model trained on white-noise data of a source at a height of 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	7.6	5.8	33.7	21.5	1.3
SEVD-MUSIC (NS)	6.6	4.1	35.4	33.8	2.1
GEVD-MUSIC (NS)	6.1	4.4	27.8	35.7	2.2
DN-SEVD-MUSIC (NS)	6.5	4.1	41.9	33.9	2.1
DN-GEVD-MUSIC (NS)	6.4	4.4	39.2	35.8	2.2
SEVD-MUSIC (SS)	6.6	4.1	35.4	32.2	2.0
GEVD-MUSIC (SS)	10.6	5.9	71.6	34.0	2.1
DN-SEVD-MUSIC (SS)	6.5	4.1	41.9	32.2	2.0
DN-GEVD-MUSIC (SS)	10.6	5.9	71.7	34.1	2.1
SEVD-MUSIC (W,NS)	6.9	4.8	35.5	33.9	2.1
GEVD-MUSIC (W,NS)	6.4	4.3	32.4	36.1	2.2
DN-SEVD-MUSIC (W,NS)	6.9	4.8	42.9	34.0	2.1
DN-GEVD-MUSIC (W,NS)	6.6	4.5	42.5	36.2	2.2
SEVD-MUSIC (W,SS)	6.9	4.8	35.5	32.2	2.0
GEVD-MUSIC (W,SS)	9.2	6.3	71.9	34.4	2.1
DN-SEVD-MUSIC (W,SS)	6.9	4.8	42.9	32.4	2.0
DN-GEVD-MUSIC (W,SS)	9.1	6.3	71.8	34.5	2.1
BSS-ADP	6.9	4.8	32.9	3293.0	199.8
Pairwise BSS-ADP	4.3	4.1	22.6	5852.3	355.1

Table A.2: Evaluation of the results of the algorithms for the recordings done using the TIMIT database [27] in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with a single source at a height of 93 cm, Wiener prefiltering and utilizing the 'Extended Woodworth' model trained on white-noise data of a source at a height of 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	6.1	5.6	47.2	22.6	1.4
SEVD-MUSIC (NS)	3.3	3.1	23.7	33.8	2.0
GEVD-MUSIC (NS)	3.1	3.0	5.2	36.0	2.2
DN-SEVD-MUSIC (NS)	3.3	3.1	29.7	34.1	2.1
DN-GEVD-MUSIC (NS)	3.3	3.4	18.9	36.2	2.2
SEVD-MUSIC (SS)	3.3	3.1	23.7	32.3	1.9
GEVD-MUSIC (SS)	7.1	5.8	64.0	34.4	2.1
DN-SEVD-MUSIC (SS)	3.3	3.1	29.7	32.4	2.0
DN-GEVD-MUSIC (SS)	7.0	5.7	64.3	34.5	2.1
SEVD-MUSIC (W,NS)	6.3	5.2	60.7	34.0	2.0
GEVD-MUSIC (W,NS)	4.0	3.9	31.3	36.2	2.2
DN-SEVD-MUSIC (W,NS)	6.4	5.2	65.5	34.2	2.1
DN-GEVD-MUSIC (W,NS)	4.0	4.0	43.2	36.3	2.2
SEVD-MUSIC (W,SS)	6.3	5.2	60.7	32.3	1.9
GEVD-MUSIC (W,SS)	7.0	5.4	74.2	34.6	2.1
DN-SEVD-MUSIC (W,SS)	6.4	5.2	65.5	32.4	2.0
DN-GEVD-MUSIC (W,SS)	7.0	5.5	74.5	34.6	2.1
BSS-ADP	2.8	3.2	32.1	3313.8	199.7
Pairwise BSS-ADP	2.5	3.2	53.7	5888.8	354.9

Table A.3: Evaluation of the results of the algorithms for the recordings done using the TIMIT database [27] in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with a single source at a height of 143 cm and utilizing the 'Extended Woodworth' model trained on white-noise data of a source at a height of 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	5.8	5.2	31.5	22.8	1.4
SEVD-MUSIC (NS)	4.4	4.6	12.1	33.9	2.0
GEVD-MUSIC (NS)	3.9	4.3	9.1	35.8	2.2
DN-SEVD-MUSIC (NS)	4.5	4.7	17.6	34.0	2.1
DN-GEVD-MUSIC (NS)	4.2	4.5	18.3	35.9	2.2
SEVD-MUSIC (SS)	4.4	4.6	12.1	32.3	1.9
GEVD-MUSIC (SS)	8.8	5.4	71.9	34.1	2.1
DN-SEVD-MUSIC (SS)	4.5	4.7	17.6	32.3	1.9
DN-GEVD-MUSIC (SS)	8.9	5.4	72.3	34.3	2.1
SEVD-MUSIC (W,NS)	4.5	4.2	21.1	34.0	2.1
GEVD-MUSIC (W,NS)	4.3	4.1	18.3	36.1	2.2
DN-SEVD-MUSIC (W,NS)	4.4	4.3	28.1	34.2	2.1
DN-GEVD-MUSIC (W,NS)	4.4	4.3	30.3	36.3	2.2
SEVD-MUSIC (W,SS)	4.5	4.2	21.1	32.4	2.0
GEVD-MUSIC (W,SS)	8.4	6.1	73.6	34.5	2.1
DN-SEVD-MUSIC (W,SS)	4.4	4.3	28.1	32.5	2.0
DN-GEVD-MUSIC (W,SS)	8.4	6.1	73.7	34.6	2.1
BSS-ADP	4.7	4.5	12.8	3318.3	200.0
Pairwise BSS-ADP	2.6	3.7	7.5	5901.9	355.7

Table A.4: Evaluation of the results of the algorithms for the recordings done using the TIMIT database [27] in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with a single source at a height of 143 cm, Wiener prefiltering and utilizing the 'Extended Woodworth' model trained on white-noise data of a source at a height of 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	8.3	6.2	48.4	21.6	1.3
SEVD-MUSIC (NS)	6.3	5.5	64.9	33.8	2.0
GEVD-MUSIC (NS)	4.8	4.8	42.9	35.9	2.2
DN-SEVD-MUSIC (NS)	5.1	4.7	82.9	34.0	2.1
DN-GEVD-MUSIC (NS)	4.9	5.0	53.7	36.1	2.2
SEVD-MUSIC (SS)	6.3	5.5	64.9	32.2	1.9
GEVD-MUSIC (SS)	9.2	5.9	77.9	34.2	2.1
DN-SEVD-MUSIC (SS)	5.1	4.7	82.9	32.3	2.0
DN-GEVD-MUSIC (SS)	9.2	5.9	78.2	34.4	2.1
SEVD-MUSIC (W,NS)	8.0	5.8	72.6	34.0	2.1
GEVD-MUSIC (W,NS)	6.0	5.2	54.5	36.0	2.2
DN-SEVD-MUSIC (W,NS)	8.5	5.8	76.6	34.0	2.1
DN-GEVD-MUSIC (W,NS)	6.3	5.6	63.6	36.2	2.2
SEVD-MUSIC (W,SS)	8.0	5.8	72.6	32.3	2.0
GEVD-MUSIC (W,SS)	8.6	6.1	76.5	34.4	2.1
DN-SEVD-MUSIC (W,SS)	8.5	5.8	76.6	32.4	2.0
DN-GEVD-MUSIC (W,SS)	8.6	6.1	76.6	34.5	2.1
BSS-ADP	4.2	4.3	52.7	3302.3	199.7
Pairwise BSS-ADP	5.0	5.9	68.5	5871.8	355.2

Table A.5: Evaluation of the results of the algorithms for the recordings done using the TIMIT database [27] in the reflecting room configuration ($T_{60} = 510\,\text{ms}$) with a single source at a height of 143 cm and utilizing the 'Extended Woodworth' model trained on white-noise data of a source at a height of 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	8.8	6.1	27.2	21.7	1.3
SEVD-MUSIC (NS)	7.4	5.8	63.6	33.9	2.1
GEVD-MUSIC (NS)	6.0	5.1	40.0	35.9	2.2
DN-SEVD-MUSIC (NS)	7.6	5.9	72.4	34.1	2.1
DN-GEVD-MUSIC (NS)	5.2	4.7	56.4	36.0	2.2
SEVD-MUSIC (SS)	7.4	5.8	63.6	32.3	2.0
GEVD-MUSIC (SS)	9.0	5.7	73.5	34.2	2.1
DN-SEVD-MUSIC (SS)	7.6	5.9	72.4	32.3	2.0
DN-GEVD-MUSIC (SS)	9.2	5.7	74.5	34.3	2.1
SEVD-MUSIC (W,NS)	7.0	5.5	57.2	34.0	2.1
GEVD-MUSIC (W,NS)	6.6	5.4	51.6	36.1	2.2
DN-SEVD-MUSIC (W,NS)	7.2	5.5	64.5	34.0	2.1
DN-GEVD-MUSIC (W,NS)	6.9	5.6	63.0	36.3	2.2
SEVD-MUSIC (W,SS)	7.0	5.5	57.2	32.3	2.0
GEVD-MUSIC (W,SS)	8.6	6.0	75.1	34.5	2.1
DN-SEVD-MUSIC (W,SS)	7.2	5.5	64.5	32.3	2.0
DN-GEVD-MUSIC (W,SS)	8.6	6.0	75.5	34.5	2.1
BSS-ADP	5.7	5.3	78.6	3308.5	200.1
Pairwise BSS-ADP	6.0	5.0	56.2	5881.1	355.7

Table A.6: Evaluation of the results of the algorithms for the recordings done using the TIMIT database [27] in the reflecting room configuration ($T_{60} = 510\,\text{ms}$) with a single source at a height of 143 cm, Wiener prefiltering and utilizing the 'Extended Woodworth' model trained on white-noise data of a source at a height of 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	7.3	5.1	62.9	7.7	1.4
SEVD-MUSIC (NS)	7.2	5.2	29.3	11.9	2.2
GEVD-MUSIC (NS)	6.6	4.8	20.2	12.5	2.3
DN-SEVD-MUSIC (NS)	7.6	5.5	28.7	11.9	2.2
DN-GEVD-MUSIC (NS)	7.0	5.3	21.4	12.6	2.3
SEVD-MUSIC (SS)	7.2	5.2	29.3	11.3	2.1
GEVD-MUSIC (SS)	5.6	4.4	44.0	12.0	2.2
DN-SEVD-MUSIC (SS)	7.6	5.5	28.7	11.4	2.1
DN-GEVD-MUSIC (SS)	5.5	4.4	44.0	12.0	2.2
SEVD-MUSIC (W,NS)	7.1	5.4	29.3	11.9	2.2
GEVD-MUSIC (W,NS)	6.1	4.7	26.7	12.7	2.3
DN-SEVD-MUSIC (W,NS)	7.4	5.6	30.3	12.0	2.2
DN-GEVD-MUSIC (W,NS)	6.9	5.1	29.8	12.7	2.3
SEVD-MUSIC (W,SS)	7.1	5.4	29.3	11.4	2.1
GEVD-MUSIC (W,SS)	5.8	4.9	44.0	12.0	2.2
DN-SEVD-MUSIC (W,SS)	7.4	5.6	30.3	11.4	2.1
DN-GEVD-MUSIC (W,SS)	5.8	4.8	43.8	12.0	2.2
BSS-ADP	7.2	6.4	24.6	1108.1	201.5
Pairwise BSS-ADP	6.5	5.5	22.4	1961.5	356.6

Table A.7: Evaluation of the results of the algorithms for the recordings done using a white-noise signal in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with a single source at a height of 93 cm and utilizing the 'Extended Woodworth' model trained on white-noise data of a source at a height of 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	7.7	4.8	58.4	7.5	1.4
SEVD-MUSIC (NS)	6.2	4.6	32.9	11.9	2.2
GEVD-MUSIC (NS)	7.0	5.3	29.2	12.5	2.3
DN-SEVD-MUSIC (NS)	6.1	4.6	34.5	11.9	2.2
DN-GEVD-MUSIC (NS)	7.3	5.5	31.8	12.5	2.3
SEVD-MUSIC (SS)	6.2	4.6	32.9	11.3	2.1
GEVD-MUSIC (SS)	3.3	2.8	52.5	11.9	2.2
DN-SEVD-MUSIC (SS)	6.1	4.6	34.5	11.4	2.1
DN-GEVD-MUSIC (SS)	3.3	2.8	52.5	12.0	2.2
SEVD-MUSIC (W,NS)	6.6	4.9	29.5	11.9	2.2
GEVD-MUSIC (W,NS)	6.9	5.0	34.1	12.6	2.3
DN-SEVD-MUSIC (W,NS)	6.5	4.9	30.9	11.9	2.2
DN-GEVD-MUSIC (W,NS)	6.9	5.1	36.4	12.7	2.3
SEVD-MUSIC (W,SS)	6.6	4.9	29.5	11.4	2.1
GEVD-MUSIC (W,SS)	4.0	3.7	48.2	12.1	2.2
DN-SEVD-MUSIC (W,SS)	6.5	4.9	30.9	11.4	2.1
DN-GEVD-MUSIC (W,SS)	4.0	3.7	48.1	12.1	2.2
BSS-ADP	7.2	5.2	36.1	1109.9	201.8
Pairwise BSS-ADP	4.9	5.1	14.7	1968.5	357.9

Table A.8: Evaluation of the results of the algorithms for the recordings done using white-noise signal in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with a single source at a height of 93 cm, Wiener prefiltering and utilizing the 'Extended Woodworth' model trained on white-noise data of a source at a height of 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	2.1	1.8	42.1	7.3	1.3
SEVD-MUSIC (NS)	2.9	3.8	26.7	11.8	2.1
GEVD-MUSIC (NS)	3.1	3.9	2.5	12.5	2.3
DN-SEVD-MUSIC (NS)	3.2	4.0	30.1	11.9	2.2
DN-GEVD-MUSIC (NS)	3.0	4.1	6.3	12.6	2.3
SEVD-MUSIC (SS)	2.9	3.8	26.7	11.3	2.1
GEVD-MUSIC (SS)	5.9	6.4	14.8	12.0	2.2
DN-SEVD-MUSIC (SS)	3.2	4.0	30.1	11.3	2.1
DN-GEVD-MUSIC (SS)	5.9	6.4	14.4	12.0	2.2
SEVD-MUSIC (W,NS)	3.0	3.7	23.7	11.9	2.2
GEVD-MUSIC (W,NS)	2.8	3.9	13.8	12.6	2.3
DN-SEVD-MUSIC (W,NS)	3.4	4.1	24.9	11.9	2.2
DN-GEVD-MUSIC (W,NS)	2.8	4.0	16.7	12.6	2.3
SEVD-MUSIC (W,SS)	3.0	3.7	23.7	11.3	2.1
GEVD-MUSIC (W,SS)	5.4	6.3	20.0	12.0	2.2
DN-SEVD-MUSIC (W,SS)	3.4	4.1	24.9	11.3	2.1
DN-GEVD-MUSIC (W,SS)	5.3	6.2	20.8	12.0	2.2
BSS-ADP	2.7	3.1	17.8	1105.5	201.0
Pairwise BSS-ADP	2.0	2.5	11.3	1963.3	357.0

Table A.9: Evaluation of the results of the algorithms for the recordings done using a white-noise signal in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with a single source at a height of 143 cm and utilizing the 'Extended Woodworth' model trained on white-noise data of a source at a height of 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	2.4	2.6	44.9	7.2	1.3
SEVD-MUSIC (NS)	1.8	2.7	11.1	11.8	2.2
GEVD-MUSIC (NS)	1.6	1.8	13.5	12.4	2.3
DN-SEVD-MUSIC (NS)	1.9	3.2	10.1	11.9	2.2
DN-GEVD-MUSIC (NS)	1.6	1.8	13.4	12.5	2.3
SEVD-MUSIC (SS)	1.8	2.7	11.1	11.3	2.0
GEVD-MUSIC (SS)	4.8	6.8	49.9	11.9	2.2
DN-SEVD-MUSIC (SS)	1.9	3.2	10.1	11.3	2.1
DN-GEVD-MUSIC (SS)	4.4	6.5	51.2	11.9	2.2
SEVD-MUSIC (W,NS)	1.9	2.9	11.5	11.9	2.2
GEVD-MUSIC (W,NS)	1.8	2.2	15.7	12.5	2.3
DN-SEVD-MUSIC (W,NS)	2.0	3.1	10.1	12.0	2.2
DN-GEVD-MUSIC (W,NS)	1.8	2.1	15.9	12.6	2.3
SEVD-MUSIC (W,SS)	1.9	2.9	11.5	11.3	2.1
GEVD-MUSIC (W,SS)	4.0	6.1	45.4	11.9	2.2
DN-SEVD-MUSIC (W,SS)	2.0	3.1	10.1	11.3	2.1
DN-GEVD-MUSIC (W,SS)	4.0	6.1	45.7	12.0	2.2
BSS-ADP	2.3	3.1	20.3	1106.8	201.2
Pairwise BSS-ADP	1.6	1.7	13.1	1967.8	357.8

Table A.10: Evaluation of the results of the algorithms for the recordings done using a white-noise signal in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with a single source at a height of 143 cm, Wiener prefiltering and utilizing the 'Extended Woodworth' model trained on white-noise data of a source at a height of 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	4.4	4.7	52.0	7.2	1.3
SEVD-MUSIC (NS)	5.5	5.7	40.2	11.8	2.1
GEVD-MUSIC (NS)	4.4	4.6	3.7	12.5	2.3
DN-SEVD-MUSIC (NS)	5.9	6.0	42.6	11.9	2.2
DN-GEVD-MUSIC (NS)	4.5	4.8	5.6	12.6	2.3
SEVD-MUSIC (SS)	5.5	5.7	40.2	11.3	2.1
GEVD-MUSIC (SS)	4.4	5.5	22.8	12.0	2.2
DN-SEVD-MUSIC (SS)	5.9	6.0	42.6	11.3	2.1
DN-GEVD-MUSIC (SS)	4.2	5.3	23.6	12.0	2.2
SEVD-MUSIC (W,NS)	5.3	5.2	39.5	11.9	2.2
GEVD-MUSIC (W,NS)	4.2	4.6	17.8	12.6	2.3
DN-SEVD-MUSIC (W,NS)	5.6	5.5	40.4	11.9	2.2
DN-GEVD-MUSIC (W,NS)	4.2	4.5	20.4	12.6	2.3
SEVD-MUSIC (W,SS)	5.3	5.2	39.5	11.3	2.1
GEVD-MUSIC (W,SS)	4.5	5.5	33.0	12.0	2.2
DN-SEVD-MUSIC (W,SS)	5.6	5.5	40.4	11.4	2.1
DN-GEVD-MUSIC (W,SS)	4.3	5.2	34.7	12.0	2.2
BSS-ADP	3.1	4.0	18.4	1105.8	201.1
Pairwise BSS-ADP	2.8	3.8	21.7	1967.6	357.8

Table A.11: Evaluation of the results of the algorithms for the recordings done using a white-noise signal in the reflecting room configuration ($T_{60} = 510\,\text{ms}$) with a single source at a height of 143 cm and utilizing the 'Extended Woodworth' model trained on white-noise data of a source at a height of 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	3.5	4.3	40.8	7.2	1.3
SEVD-MUSIC (NS)	2.9	2.8	17.3	11.8	2.2
GEVD-MUSIC (NS)	2.7	2.5	16.8	12.5	2.3
DN-SEVD-MUSIC (NS)	2.7	2.5	18.9	11.9	2.2
DN-GEVD-MUSIC (NS)	2.5	2.1	25.1	12.5	2.3
SEVD-MUSIC (SS)	2.9	2.8	17.3	11.3	2.1
GEVD-MUSIC (SS)	3.0	3.8	52.4	11.9	2.2
DN-SEVD-MUSIC (SS)	2.7	2.5	18.9	11.3	2.1
DN-GEVD-MUSIC (SS)	3.2	4.0	53.7	11.9	2.2
SEVD-MUSIC (W,NS)	3.3	3.6	15.1	11.9	2.2
GEVD-MUSIC (W,NS)	2.9	3.0	21.3	12.5	2.3
DN-SEVD-MUSIC (W,NS)	3.0	3.0	16.9	11.9	2.2
DN-GEVD-MUSIC (W,NS)	2.6	2.3	26.7	12.6	2.3
SEVD-MUSIC (W,SS)	3.3	3.6	15.1	11.3	2.1
GEVD-MUSIC (W,SS)	3.5	4.2	47.6	11.9	2.2
DN-SEVD-MUSIC (W,SS)	3.0	3.0	16.9	11.3	2.1
DN-GEVD-MUSIC (W,SS)	3.5	4.2	48.2	12.0	2.2
BSS-ADP	2.6	3.8	37.4	1106.1	201.1
Pairwise BSS-ADP	3.2	5.2	15.2	1962.3	356.8

Table A.12: Evaluation of the results of the algorithms for the recordings done using a white-noise signal in the reflecting room configuration ($T_{60} = 510\,\text{ms}$) with a single source at a height of 143 cm, Wiener prefiltering and utilizing the 'Extended Woodworth' model trained on white-noise data of a source at a height of 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	6.5	4.9	48.7	21.7	1.3
SEVD-MUSIC (NS)	4.9	5.7	47.4	33.5	2.0
GEVD-MUSIC (NS)	2.1	2.7	4.2	35.6	2.2
DN-SEVD-MUSIC (NS)	5.4	6.0	54.4	33.7	2.0
DN-GEVD-MUSIC (NS)	1.9	2.6	16.3	35.8	2.2
SEVD-MUSIC (SS)	4.9	5.7	47.4	31.9	1.9
GEVD-MUSIC (SS)	6.4	7.0	57.7	33.9	2.1
DN-SEVD-MUSIC (SS)	5.4	6.0	54.4	32.0	1.9
DN-GEVD-MUSIC (SS)	6.6	7.1	58.4	34.0	2.1
SEVD-MUSIC (W,NS)	6.0	5.6	64.5	33.6	2.0
GEVD-MUSIC (W,NS)	3.2	3.9	17.8	35.7	2.2
DN-SEVD-MUSIC (W,NS)	6.4	5.8	69.1	33.8	2.0
DN-GEVD-MUSIC (W,NS)	3.2	4.1	27.1	35.8	2.2
SEVD-MUSIC (W,SS)	6.0	5.6	64.5	32.0	1.9
GEVD-MUSIC (W,SS)	6.8	6.5	67.9	34.1	2.1
DN-SEVD-MUSIC (W,SS)	6.4	5.8	69.1	32.1	1.9
DN-GEVD-MUSIC (W,SS)	6.8	6.6	68.2	34.2	2.1
BSS-ADP	2.8	3.1	11.5	3287.9	199.5
Pairwise BSS-ADP	3.4	4.2	44.9	5847.5	354.8

Table A.13: Evaluation of the results of the algorithms for the recordings done using the TIMIT database [27] in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with a single source at a height of 93 cm and utilizing the 'Free-Field' model trained on white-noise data of a source at heights of 93 cm and 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	7.4	5.4	40.7	22.0	1.3
SEVD-MUSIC (NS)	2.3	3.3	5.5	33.6	2.0
GEVD-MUSIC (NS)	2.6	3.5	2.4	35.5	2.2
DN-SEVD-MUSIC (NS)	2.3	3.2	12.1	33.8	2.0
DN-GEVD-MUSIC (NS)	2.7	3.3	6.8	35.7	2.2
SEVD-MUSIC (SS)	2.3	3.3	5.5	32.0	1.9
GEVD-MUSIC (SS)	8.8	7.6	62.0	33.8	2.1
DN-SEVD-MUSIC (SS)	2.3	3.2	12.1	32.1	1.9
DN-GEVD-MUSIC (SS)	8.9	7.6	62.5	34.1	2.1
SEVD-MUSIC (W,NS)	2.9	3.5	10.6	33.8	2.1
GEVD-MUSIC (W,NS)	3.0	3.8	9.9	36.0	2.2
DN-SEVD-MUSIC (W,NS)	2.8	3.5	17.7	33.9	2.1
DN-GEVD-MUSIC (W,NS)	3.0	3.9	18.5	36.2	2.2
SEVD-MUSIC (W,SS)	2.9	3.5	10.6	32.1	1.9
GEVD-MUSIC (W,SS)	8.1	7.3	66.3	34.4	2.1
DN-SEVD-MUSIC (W,SS)	2.8	3.5	17.7	32.3	2.0
DN-GEVD-MUSIC (W,SS)	8.2	7.4	67.1	34.6	2.1
BSS-ADP	3.5	3.7	16.5	3286.2	199.4
Pairwise BSS-ADP	2.9	3.4	3.6	5850.2	355.0

Table A.14: Evaluation of the results of the algorithms for the recordings done using the TIMIT database [27] in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with a single source at a height of 93 cm, Wiener prefiltering and utilizing the 'Free-Field' model trained on white-noise data of a source at heights of 93 cm and 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	7.4	5.1	57.7	22.9	1.4
SEVD-MUSIC (NS)	2.8	3.1	12.7	34.0	2.0
GEVD-MUSIC (NS)	3.8	3.9	3.5	36.1	2.2
DN-SEVD-MUSIC (NS)	3.0	3.4	18.1	34.1	2.1
DN-GEVD-MUSIC (NS)	4.0	4.1	20.3	36.2	2.2
SEVD-MUSIC (SS)	2.8	3.1	12.7	32.4	2.0
GEVD-MUSIC (SS)	7.6	5.9	59.2	34.4	2.1
DN-SEVD-MUSIC (SS)	3.0	3.4	18.1	32.5	2.0
DN-GEVD-MUSIC (SS)	7.6	6.0	61.3	34.5	2.1
SEVD-MUSIC (W,NS)	5.9	5.1	55.6	34.1	2.1
GEVD-MUSIC (W,NS)	4.1	4.2	24.2	36.2	2.2
DN-SEVD-MUSIC (W,NS)	6.1	5.2	60.2	34.2	2.1
DN-GEVD-MUSIC (W,NS)	4.1	4.2	37.2	36.3	2.2
SEVD-MUSIC (W,SS)	5.9	5.1	55.6	32.4	2.0
GEVD-MUSIC (W,SS)	7.3	5.7	69.6	34.6	2.1
DN-SEVD-MUSIC (W,SS)	6.1	5.2	60.2	32.5	2.0
DN-GEVD-MUSIC (W,SS)	7.2	5.7	70.3	34.7	2.1
BSS-ADP	3.8	3.4	19.2	3316.3	199.8
Pairwise BSS-ADP	4.0	3.9	46.1	5902.2	355.7

Table A.15: Evaluation of the results of the algorithms for the recordings done using the TIMIT database [27] in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with a single source at a height of 143 cm and utilizing the 'Free-Field' model trained on white-noise data of a source at heights of 93 cm and 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	5.8	4.9	36.6	22.9	1.4
SEVD-MUSIC (NS)	3.2	3.5	12.3	33.9	2.0
GEVD-MUSIC (NS)	3.5	3.8	5.6	35.8	2.2
DN-SEVD-MUSIC (NS)	3.5	3.6	20.6	34.1	2.1
DN-GEVD-MUSIC (NS)	3.6	4.0	20.2	36.0	2.2
SEVD-MUSIC (SS)	3.2	3.5	12.3	32.3	1.9
GEVD-MUSIC (SS)	8.4	6.5	73.6	34.2	2.1
DN-SEVD-MUSIC (SS)	3.5	3.6	20.6	32.4	2.0
DN-GEVD-MUSIC (SS)	8.6	6.5	74.5	34.3	2.1
SEVD-MUSIC (W,NS)	3.6	3.9	17.6	34.0	2.1
GEVD-MUSIC (W,NS)	3.7	3.8	15.0	36.1	2.2
DN-SEVD-MUSIC (W,NS)	3.7	4.0	26.9	34.1	2.1
DN-GEVD-MUSIC (W,NS)	3.7	3.9	26.8	36.3	2.2
SEVD-MUSIC (W,SS)	3.6	3.9	17.6	32.3	1.9
GEVD-MUSIC (W,SS)	8.1	6.1	71.1	34.4	2.1
DN-SEVD-MUSIC (W,SS)	3.7	4.0	26.9	32.4	2.0
DN-GEVD-MUSIC (W,SS)	8.2	6.1	71.5	34.6	2.1
BSS-ADP	3.9	3.8	15.2	3320.8	200.1
Pairwise BSS-ADP	2.7	3.0	6.3	5908.0	356.0

Table A.16: Evaluation of the results of the algorithms for the recordings done using the TIMIT database [27] in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with a single source at a height of 143 cm, Wiener prefiltering and utilizing the 'Free-Field' model trained on white-noise data of a source at heights of 93 cm and 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	7.6	5.4	55.6	21.6	1.3
SEVD-MUSIC (NS)	6.0	5.3	64.8	33.8	2.0
GEVD-MUSIC (NS)	4.1	3.8	40.7	35.9	2.2
DN-SEVD-MUSIC (NS)	6.0	5.1	71.8	33.9	2.1
DN-GEVD-MUSIC (NS)	4.1	3.9	55.2	36.0	2.2
SEVD-MUSIC (SS)	6.0	5.3	64.8	32.2	1.9
GEVD-MUSIC (SS)	7.3	6.1	73.4	34.2	2.1
DN-SEVD-MUSIC (SS)	6.0	5.1	71.8	32.3	2.0
DN-GEVD-MUSIC (SS)	7.5	6.2	74.3	34.3	2.1
SEVD-MUSIC (W,NS)	7.6	5.7	71.9	33.9	2.0
GEVD-MUSIC (W,NS)	5.2	5.0	49.4	36.0	2.2
DN-SEVD-MUSIC (W,NS)	7.9	5.8	74.9	34.0	2.1
DN-GEVD-MUSIC (W,NS)	5.4	5.3	61.6	36.1	2.2
SEVD-MUSIC (W,SS)	7.6	5.7	71.9	32.2	1.9
GEVD-MUSIC (W,SS)	8.6	6.3	76.2	34.3	2.1
DN-SEVD-MUSIC (W,SS)	7.9	5.8	74.9	32.3	2.0
DN-GEVD-MUSIC (W,SS)	8.6	6.3	76.3	34.4	2.1
BSS-ADP	5.1	4.6	43.8	3318.1	200.7
Pairwise BSS-ADP	5.9	5.3	65.4	5892.3	356.4

Table A.17: Evaluation of the results of the algorithms for the recordings done using the TIMIT database [27] in the reflecting room configuration ($T_{60} = 510\,\text{ms}$) with a single source at a height of 143 cm and utilizing the 'Free-Field' model trained on white-noise data of a source at heights of 93 cm and 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	8.0	5.7	38.4	21.6	1.3
SEVD-MUSIC (NS)	5.0	4.7	61.7	33.8	2.0
GEVD-MUSIC (NS)	6.1	4.7	36.7	35.8	2.2
DN-SEVD-MUSIC (NS)	5.4	4.7	71.2	34.0	2.1
DN-GEVD-MUSIC (NS)	5.5	5.0	51.6	35.9	2.2
SEVD-MUSIC (SS)	5.0	4.7	61.7	32.2	1.9
GEVD-MUSIC (SS)	7.6	5.9	69.8	34.1	2.1
DN-SEVD-MUSIC (SS)	5.4	4.7	71.2	32.3	2.0
DN-GEVD-MUSIC (SS)	7.7	5.9	71.2	34.2	2.1
SEVD-MUSIC (W,NS)	6.3	5.5	52.3	34.0	2.1
GEVD-MUSIC (W,NS)	5.8	5.3	46.3	36.1	2.2
DN-SEVD-MUSIC (W,NS)	6.3	5.5	60.9	34.1	2.1
DN-GEVD-MUSIC (W,NS)	5.9	5.4	60.1	36.3	2.2
SEVD-MUSIC (W,SS)	6.3	5.5	52.3	32.2	1.9
GEVD-MUSIC (W,SS)	8.0	6.0	73.5	34.5	2.1
DN-SEVD-MUSIC (W,SS)	6.3	5.5	60.9	32.3	2.0
DN-GEVD-MUSIC (W,SS)	8.1	6.0	74.3	34.5	2.1
BSS-ADP	5.5	5.3	78.4	3314.5	200.5
Pairwise BSS-ADP	5.6	4.7	56.1	5892.6	356.4

Table A.18: Evaluation of the results of the algorithms for the recordings done using the TIMIT database [27] in the reflecting room configuration ($T_{60} = 510\,\text{ms}$) with a single source at a height of 143 cm, Wiener prefiltering and utilizing the 'Free-Field' model trained on white-noise data of a source at heights of 93 cm and 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	3.2	2.6	61.9	7.7	1.4
SEVD-MUSIC (NS)	2.4	1.9	23.1	11.8	2.1
GEVD-MUSIC (NS)	2.6	2.7	6.4	12.5	2.3
DN-SEVD-MUSIC (NS)	2.5	2.0	23.2	11.9	2.2
DN-GEVD-MUSIC (NS)	2.6	2.7	8.9	12.5	2.3
SEVD-MUSIC (SS)	2.4	1.9	23.1	11.3	2.1
GEVD-MUSIC (SS)	2.4	1.3	26.6	11.9	2.2
DN-SEVD-MUSIC (SS)	2.5	2.0	23.2	11.3	2.1
DN-GEVD-MUSIC (SS)	2.4	1.3	27.1	12.0	2.2
SEVD-MUSIC (W,NS)	2.4	2.2	23.7	11.9	2.2
GEVD-MUSIC (W,NS)	2.5	2.4	12.1	12.5	2.3
DN-SEVD-MUSIC (W,NS)	2.5	2.4	24.6	11.9	2.2
DN-GEVD-MUSIC (W,NS)	2.6	2.5	15.2	12.6	2.3
SEVD-MUSIC (W,SS)	2.4	2.2	23.7	11.3	2.1
GEVD-MUSIC (W,SS)	2.4	1.5	29.9	12.0	2.2
DN-SEVD-MUSIC (W,SS)	2.5	2.4	24.6	11.3	2.1
DN-GEVD-MUSIC (W,SS)	2.4	1.5	30.9	12.0	2.2
BSS-ADP	3.8	3.5	10.2	1111.4	202.1
Pairwise BSS-ADP	2.7	3.0	17.7	1970.8	358.3

Table A.19: Evaluation of the results of the algorithms for the recordings done using a white-noise signal in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with a single source at a height of 93 cm and utilizing the 'Free-Field' model trained on white-noise data of a source at heights of 93 cm and 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	2.8	1.9	66.4	7.5	1.4
SEVD-MUSIC (NS)	2.4	2.3	15.7	11.8	2.1
GEVD-MUSIC (NS)	2.5	2.3	16.3	12.5	2.3
DN-SEVD-MUSIC (NS)	2.4	2.3	15.7	11.9	2.2
DN-GEVD-MUSIC (NS)	2.5	2.2	16.3	12.5	2.3
SEVD-MUSIC (SS)	2.4	2.3	15.7	11.3	2.0
GEVD-MUSIC (SS)	3.4	3.6	27.4	11.9	2.2
DN-SEVD-MUSIC (SS)	2.4	2.3	15.7	11.3	2.1
DN-GEVD-MUSIC (SS)	3.4	3.6	27.7	11.9	2.2
SEVD-MUSIC (W,NS)	2.5	2.4	13.9	11.9	2.2
GEVD-MUSIC (W,NS)	2.4	2.2	17.2	12.7	2.3
DN-SEVD-MUSIC (W,NS)	2.5	2.4	14.3	12.0	2.2
DN-GEVD-MUSIC (W,NS)	2.4	2.2	16.5	12.7	2.3
SEVD-MUSIC (W,SS)	2.5	2.4	13.9	11.3	2.1
GEVD-MUSIC (W,SS)	3.1	3.2	28.0	12.1	2.2
DN-SEVD-MUSIC (W,SS)	2.5	2.4	14.3	11.4	2.1
DN-GEVD-MUSIC (W,SS)	3.1	3.2	28.3	12.1	2.2
BSS-ADP	3.1	3.4	16.5	1110.2	201.9
Pairwise BSS-ADP	2.1	1.5	12.5	1969.6	358.1

Table A.20: Evaluation of the results of the algorithms for the recordings done using white-noise signal in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with a single source at a height of 93 cm, Wiener prefiltering and utilizing the 'Free-Field' model trained on white-noise data of a source at heights of 93 cm and 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	6.2	6.3	40.6	7.3	1.3
SEVD-MUSIC (NS)	2.9	2.9	13.7	11.8	2.2
GEVD-MUSIC (NS)	1.7	1.2	2.5	12.5	2.3
DN-SEVD-MUSIC (NS)	3.4	3.5	13.2	11.9	2.2
DN-GEVD-MUSIC (NS)	1.7	1.2	10.5	12.5	2.3
SEVD-MUSIC (SS)	2.9	2.9	13.7	11.3	2.0
GEVD-MUSIC (SS)	2.5	1.6	21.5	11.9	2.2
DN-SEVD-MUSIC (SS)	3.4	3.5	13.2	11.3	2.1
DN-GEVD-MUSIC (SS)	2.6	1.7	21.6	12.0	2.2
SEVD-MUSIC (W,NS)	3.1	3.1	14.8	11.9	2.2
GEVD-MUSIC (W,NS)	2.0	1.9	8.0	12.5	2.3
DN-SEVD-MUSIC (W,NS)	3.4	3.5	14.1	11.9	2.2
DN-GEVD-MUSIC (W,NS)	2.0	1.7	12.6	12.6	2.3
SEVD-MUSIC (W,SS)	3.1	3.1	14.8	11.3	2.1
GEVD-MUSIC (W,SS)	2.8	2.2	24.3	12.0	2.2
DN-SEVD-MUSIC (W,SS)	3.4	3.5	14.1	11.3	2.1
DN-GEVD-MUSIC (W,SS)	2.9	2.2	24.4	12.0	2.2
BSS-ADP	2.1	1.9	11.1	1106.9	201.3
Pairwise BSS-ADP	2.2	1.5	11.1	1969.4	358.1

Table A.21: Evaluation of the results of the algorithms for the recordings done using a white-noise signal in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with a single source at a height of 143 cm and utilizing the 'Free-Field' model trained on white-noise data of a source at heights of 93 cm and 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	5.8	5.5	33.7	7.4	1.3
SEVD-MUSIC (NS)	2.1	2.1	3.2	11.8	2.1
GEVD-MUSIC (NS)	2.0	1.4	6.7	12.4	2.3
DN-SEVD-MUSIC (NS)	2.1	2.1	2.9	11.9	2.2
DN-GEVD-MUSIC (NS)	2.0	1.4	7.9	12.5	2.3
SEVD-MUSIC (SS)	2.1	2.1	3.2	11.3	2.1
GEVD-MUSIC (SS)	4.0	6.3	63.6	11.9	2.2
DN-SEVD-MUSIC (SS)	2.1	2.1	2.9	11.3	2.1
DN-GEVD-MUSIC (SS)	4.0	6.4	64.9	11.9	2.2
SEVD-MUSIC (W,NS)	2.3	2.2	4.9	11.9	2.2
GEVD-MUSIC (W,NS)	2.1	1.5	7.4	12.5	2.3
DN-SEVD-MUSIC (W,NS)	2.3	2.2	4.2	11.9	2.2
DN-GEVD-MUSIC (W,NS)	2.1	1.6	8.6	12.5	2.3
SEVD-MUSIC (W,SS)	2.3	2.2	4.9	11.3	2.1
GEVD-MUSIC (W,SS)	3.1	4.0	49.9	11.9	2.2
DN-SEVD-MUSIC (W,SS)	2.3	2.2	4.2	11.4	2.1
DN-GEVD-MUSIC (W,SS)	3.2	4.3	51.8	12.0	2.2
BSS-ADP	2.9	2.2	9.8	1108.3	201.5
Pairwise BSS-ADP	2.2	1.6	2.2	1969.5	358.1

Table A.22: Evaluation of the results of the algorithms for the recordings done using a white-noise signal in the absorbing room configuration ($T_{60} = 190\,\text{ms}$) with a single source at a height of 143 cm, Wiener prefiltering and utilizing the 'Free-Field' model trained on white-noise data of a source at heights of 93 cm and 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	2.8	2.4	48.1	7.2	1.3
SEVD-MUSIC (NS)	2.9	2.7	20.4	11.8	2.1
GEVD-MUSIC (NS)	3.0	3.1	1.4	12.4	2.3
DN-SEVD-MUSIC (NS)	3.0	2.7	20.9	11.8	2.1
DN-GEVD-MUSIC (NS)	3.0	3.1	5.5	12.5	2.3
SEVD-MUSIC (SS)	2.9	2.7	20.4	11.2	2.0
GEVD-MUSIC (SS)	3.4	3.3	8.0	11.9	2.2
DN-SEVD-MUSIC (SS)	3.0	2.7	20.9	11.3	2.0
DN-GEVD-MUSIC (SS)	3.3	3.3	8.1	11.9	2.2
SEVD-MUSIC (W,NS)	2.9	2.7	18.6	11.8	2.2
GEVD-MUSIC (W,NS)	3.2	3.2	8.3	12.5	2.3
DN-SEVD-MUSIC (W,NS)	3.0	2.8	17.3	11.8	2.2
DN-GEVD-MUSIC (W,NS)	3.3	3.2	7.5	12.5	2.3
SEVD-MUSIC (W,SS)	2.9	2.7	18.6	11.3	2.0
GEVD-MUSIC (W,SS)	3.9	4.1	19.9	11.9	2.2
DN-SEVD-MUSIC (W,SS)	3.0	2.8	17.3	11.3	2.1
DN-GEVD-MUSIC (W,SS)	3.8	4.1	20.6	11.9	2.2
BSS-ADP	2.6	2.5	11.1	1102.8	200.5
Pairwise BSS-ADP	2.8	2.7	8.8	1945.5	353.7

Table A.23: Evaluation of the results of the algorithms for the recordings done using a white-noise signal in the reflecting room configuration ($T_{60} = 510\,\text{ms}$) with a single source at a height of 143 cm and utilizing the 'Free-Field' model trained on white-noise data of a source at heights of 93 cm and 143 cm for the conversion from DOAs to TDOAs

	AVG (°)	SD (°)	OUT (%)	CT (s)	CTC
HUMAVIPS	4.7	4.7	41.0	7.0	1.3
SEVD-MUSIC (NS)	3.0	3.2	5.3	11.6	2.1
GEVD-MUSIC (NS)	2.9	3.0	7.5	12.2	2.2
DN-SEVD-MUSIC (NS)	2.9	3.0	5.9	11.7	2.1
DN-GEVD-MUSIC (NS)	2.9	3.0	7.8	12.2	2.2
SEVD-MUSIC (SS)	3.0	3.2	5.3	11.1	2.0
GEVD-MUSIC (SS)	3.0	3.2	36.7	11.6	2.1
DN-SEVD-MUSIC (SS)	2.9	3.0	5.9	11.1	2.0
DN-GEVD-MUSIC (SS)	3.0	3.2	37.7	11.7	2.1
SEVD-MUSIC (W,NS)	3.0	3.1	4.7	11.6	2.1
GEVD-MUSIC (W,NS)	2.9	3.0	8.6	12.2	2.2
DN-SEVD-MUSIC (W,NS)	3.0	3.1	5.8	11.7	2.1
DN-GEVD-MUSIC (W,NS)	2.9	2.8	13.4	12.3	2.2
SEVD-MUSIC (W,SS)	3.0	3.1	4.7	11.1	2.0
GEVD-MUSIC (W,SS)	3.5	3.8	34.2	11.7	2.1
DN-SEVD-MUSIC (W,SS)	3.0	3.1	5.8	11.2	2.0
DN-GEVD-MUSIC (W,SS)	3.4	3.9	35.1	11.7	2.1
BSS-ADP	2.7	3.2	25.8	1085.2	197.3
Pairwise BSS-ADP	2.5	2.6	11.3	1933.4	351.5

Table A.24: Evaluation of the results of the algorithms for the recordings done using a white-noise signal in the reflecting room configuration ($T_{60} = 510\,\text{ms}$) with a single source at a height of 143 cm, Wiener prefiltering and utilizing the 'Free-Field' model trained on white-noise data of a source at heights of 93 cm and 143 cm for the conversion from DOAs to TDOAs

List of Figures

List of Tables

References

[1] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]," *IEEE Robotics Automation Magazine*, vol. 19, no. 2, pp. 98–100, June 2012.

[2] D. Hanson, A. Olney, I. A. Pereira, and M. Zielke, "Upending the uncanny valley," in *Proceedings of the twentieth national conference on artificial intelligence*, July 2005.

[3] J. Goetz, S. Kiesler, and A. Powers, "Matching robot appearance and behavior to tasks to improve human-robot cooperation," in *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003*, October 2003, pp. 55–60.

[4] J. Cech, R. Mittal, A. Deleforge, J. Sanchez-Riera, X. Alameda-Pineda, and R. Horaud, "Active-speaker detection and localization with microphones and cameras embedded into a robotic head," in *2013 IEEE International Conference on Humanoid Robots*, September 2013.

[5] G. Athanasopoulos, H. Brouckxon, and W. Verhelst, "Sound Source Localization for Real-World Humanoid Robots," in *Proceedings of the 11th International Conference on Signal Processing (SIP 2012)*. WSEAS, 2012, pp. 131–136.

[6] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, March 1986.

[7] D. S. Talagala, W. Zhang, and T. D. Abhayapala, "Broadband doa estimation using sensor arrays on complex-shaped rigid bodies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1573–1585, August 2013.

[8] F. S. Avarvand, A. Ziehe, and G. Nolte, "Self-consistent music algorithm to localize multiple sources in acoustic imaging," in *Proceedings on CD of the 4th Berlin Beamforming Conference, 22-23 February 2012*. GFaI, Gesellschaft zu Förderung angewandter Informatik e.V., Berlin, February 2012.

[9] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 7, pp. 984–995, July 1989.

[10] S. Argentieri and P. Danès, "Broadband variations of the music high-resolution method for sound source localization in robotics," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, October 2007, pp. 2009–2014.

[11] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environments," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, October 2009, pp. 664–669.

[12] K. Furukawa, K. Okutani, K. Nagira, T. Otsuka, K. Itoyama, K. Nakadai, and H. G. Okuno, "Noise correlation matrix estimation for improving sound source localization by multirotor uav," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, November 2013, pp. 3943–3948.

[13] H. Buchner, R. Aichner, and W. Kellermann, "Trinicon: a versatile framework for multichannel blind signal processing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings.*, vol. 3, May 2004, pp. III–889–III–892.

[14] C. A. Anderson, S. Meier, W. Kellermann, P. D. Teal, and M. A. Poletti, "A gpu-accelerated real-time implementation of trinicon-bss for multiple separation units," in *2014 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, May 2014, pp. 102–106.

[15] A. Lombard, T. Rosenkranz, H. Buchner, and W. Kellermann, "Exploiting the self-steering capability of blind source separation to localize two or more sound sources in adverse environments," in *2008 ITG Conference on Voice Communication (SprachKommunikation)*, October 2008, pp. 1–4.

[16] ——, "Multidimensional localization of multiple sound sources using averaged directivity patterns of blind source separation systems," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 233–236.

[17] A. Lombard, Y. Zheng, H. Buchner, and W. Kellermann, "Tdoa estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1490–1503, August 2011.

[18] Nao documentation - microphones. Aldebaran Robotics. [Online]. Available: https://community.aldebaran.com/doc/1-14/family/robots/microphone_robot.html

[19] R. S. Woodworth and H. Schlosberg, *Experimental Psychology*, third (revised and reset) ed. Methuen & Co. LTD, 1954, ch. Audition, pp. 349–361.

[20] H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," in *Proceedings of the 6th International Conference on Digital Audio Effects*, September 2003, pp. 209–213.

[21] N. L. Aaronson and W. M. Hartmann, "Testing, correcting, and extending the woodworth model for interaural time difference," *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 817–823, February 2014.

[22] Nao documentation - nao technical overview. Aldebaran Robotics. [Online]. Available: https://community.aldebaran.com/doc/1-14/family/robots/index_robots.html

[23] A. Spriet, M. S. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, 2004.

[24] E. A. P. Habets, J. Benesty, S. Gannot, and I. Cohen, "The mvdr beamformer for speech enhancement," in *Speech Processing in Modern Communication: Challenges and Perspectives*, ser. Springer Topics in Signal Processing, I. Cohen, J. Benesty, and S. Gannot, Eds. Springer Berlin Heidelberg, 2010, vol. 3, ch. 9, pp. 225–254.

[25] Nao documentation - links. Aldebaran Robotics. [Online]. Available: https://community.aldebaran.com/doc/1-14/family/robots/links_robot.html

[26] (2003) Genelec 1029a datasheet. Genelec. [Online]. Available: http://www.genelec.com/pdf/DS1029a.pdf

[27] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue. (1993) Timit acoustic-phonetic continuous speech corpus. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S1

[28] M. H. Moattar and M. M. Homayounpour, "A simple but efficient real-time voice activity detection algorithm," in *17th European Signal Processing Conference*, August 2009, pp. 2549–2553.

[29] T. F. Coleman and Y. Li, "On the convergence of interior-reflective newton methods for nonlinear minimization subject to bounds," *Mathematical Programming*, vol. 67, no. 1–3, pp. 189–224, October 1994.

[30] ——, "An interior trust region approach for nonlinear minimization subject to bounds," *SIAM Journal on Optimization*, vol. 6, no. 2, pp. 418–445, May 1996.

[31] H. Wang and M. M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 4, pp. 823–831, August 1985.

[32] T. D. Abhayapala and H. Bhatta, "Coherent broadband source localization by modal space processing," in *2003 International Conference on Telecommunications*, vol. 2, February 2003, pp. 1617–1623.