Eye-Gaze Steered Beamforming for Hearing Aids

Master Thesis

Simone Birk Bols Thomsen

Aalborg University
Mathematical Engineering

This project has been written in LaTeX with figures produced in Ti*k*Z and Python 3.6. Scripts have been made using Python 3.6.

AALBORG UNIVERSITY
STUDENT REPORT

Mathematical Engineering
Aalborg University
http://www.aau.dk

Title:
Eye-gaze Steered Beamforming for Hearing Aids

Theme:
Master Thesis

Project Period:
Fall semester 2021 and spring semester 2022

Participants:
Simone Birk Bols Thomsen

Supervisors:
Poul Hoang
Jesper Jensen
Morten Nielsen

Copies: 1

Number of Pages: 120

Date of Completion:
June 3, 2022

Abstract:

If multiple microphones are available in a hearing aid (HA) device, beamformers can be applied to enhance target speech signals in noisy environments. Many common beamformers require knowledge of the target sound source location relative to the HA user. Traditional beamforming methods are equipped with techniques that try to localize the target source acoustically, i.e., only using microphone signals. However, localizing the target source in presence of competing speakers remains an unsolvable problem. In this thesis, we study the use of an additional modality, apart from sound, to help enhancing the target signal. Specifically, we aim to use the HA user's eye-gaze as an asset to efficiently identity the target direction. Initially, we examine the potential performance benefits of using eye-gaze steered beamforming under ideal conditions. Subsequently, we propose two eye-gaze based beamforming systems, namely a Bayesian beamformer with the posterior probability on the target direction estimated based on a prior probability derived from the user's eye-gaze, and a Bayesian beamformer with the posterior jointly estimated from the HA microphone signals and the HA user's eye-gaze signal. The performance of the proposed methods are compared with current audio-only methods. The main conclusion is that, under certain conditions, the proposed eye-gaze based beamformers are able to outperform audio-only methods in terms of estimated speech intelligibility and quality.

Abstract

De fleste moderne beamformers anvendt i høreapparater kræver adgang til information om placeringen på den ønskede taler relativt til høreapparatbrugeren. Placeringen af den ønskede taler er sjældent kendt på forhånd og skal derfor estimeres online fra de støjfyldte mikrofonsignaler. Traditionelle algoritmer, der kan bruges til at estimere retningen på den ønskede taler, bruger kun adgang til de støjfyldte mikrofonsignaler. Disse algoritmer er kendt for at performe dårligt i komplekse akustiske miljøer, hvor der er flere samtidige talere tilstede samt i meget støjfydldte omgivelser.

I denne afhandling undersøges idéen om at styre høreapparat-beamformers ved hjælp af brugeres øjne, som ofte hviler på den ønskede taler for blandt andet mundaflæsning. Idéen om at bruge brugerens øjeretnign er motiveret af den kendsgerning, at når vi kommunikerer med hinanden igennem tale, så involverer vores opførsel både auditiv og visuel opmærksomhed. Da visuel opmærksomhed kræver, at vi retter blikket mod den ønskede taler, formoder vi, at information om brugerens øjeretning - som er uafhængig af støjen i det akustiske miljø - må kunne bidrage til at forbedre støjreduktionen i fremtidige høreapparater. Målet med denne afhandling er derfor at udvikle og foreslå et beamforming-system til høreapparater, som inkorporerer brugeres øjeretning i kombination med de støjfyldte mikrofonsignaler. I den forbindelse foreslåes to beamforming-systemer til høreapparater, nemlig 1) en Bayesian beamformer, hvor *a posteriori* sandsynlighedsfordelingen på den ønskede talers retning er estimeret baseret på en *a priori* sandsynlighedsfordeling udledt af brugeres øjeretning, samt 2) en Bayesian beamformer, hvor *a posteriori* sandsynlighedsfordelingen er estimeret simultant fra de støjfyldte mikrofonsignaler og burgeres øjeretning.

I denne afhandling udføres først et eksplorativt eksperiment, hvor den potentielle performance gevinst ved at bruge øjestyret beamformers undersøges. Dernæst studeres der en Bayesiansk tilgang til at kombinere de støjfyldte mikofonsignaler og brugerens øjeretning, hvorunder de foreslåede metoder præsenteres. Den resterende del af afhandlingen beskæftiger sig med simuleringsaspekter, herunder en gennemgang af det anvendte data samt evaluering af de undersøgte algoritmer.

På baggrund af resultaterne fundet i denne afhandling, kan det konkluderes, at de foreslåede metoder, der inkluderer brugerens øjeretning - under nogle forhold - er bedre til at ekstrahere det rene talesignal fra de støjfyldte mikrofonsignaler, sammenlignet med de undersøgte eksisterende metoder, der ikke inkluderer brugerens øjeretning.

Preface

This Master's Thesis (60 ECTS) is written by Simone Birk Bols Thomsen of the Master program Mathematical Engineering at Aalborg University, Department of Mathematical Sciences in the period from 01/09/2021 to 03/06/2022.

The topic of interest in this thesis is *Eye-Gaze Steered Beamforming for Hearing Aids*.

For references throughout the thesis, the IEEE-method is used with specification of pages, sections or chapters. Additional information about the sources can be seen in the bibliography.

All figures and tables throughout the thesis have been created by the author and are generated with Python 3.8.8 and the Ti*k*z-package in LaTeX. In addition, Python 3.8.8 is used to develop software to perform the numerical calculations related to the thesis. The signal processing has been performed using the Python libraries `NumPy`, `SciPy`, `Matplotlib`, `glob`, `os`, and `Numba`, while performance scores are computed using MATLAB®.

In this thesis, mathematical quantities are specified as elements of a relevant mathematical space where necessary. However, in general it should appear from the context which space a particular variable is contained in. Furthermore, in regards to notation, we refer to functions as well as their function values interchangeably, i.e., depending on the context, $f(x)$ can both be taken to mean the function f as well as the function value $f(x)$. Strictly speaking, from a mathematical point of view, functions should be denoted without arguments, as the central concern is the function itself. However, as this thesis is mostly aimed at an engineering audience, we choose to adopt the convention of referring to functions with arguments. Moreover, boldface lowercase and boldface capital letters are used to indicate vectors and matrices, respectively, while scalars are portrayed as non-boldface letters. All vectors are considered as column vectors unless otherwise specified. Types of spaces and mathematical quantities used in this thesis appear from the *nomenclature*.

The author would like to thank the supervisors Jesper Jensen (Demant A/S, Department of Electronic Systems), Morten Nielsen (Department of Mathematical Sciences), and Poul Hoang (Demant A/S) for their supervision throughout the development of this thesis. Furthermore, the author would like to thank Eriksholm Research Centre for providing data as well as guidance throughout the development of the thesis.

Nomenclature

List of Abbreviations

AIR	Acoustic impulse response.
ATF	Acoustis transfer function.
CPSD	Cross power spectral density.
DFT	Discrete Fourier transform.
DOA	Direction-of-arrival.
ERH	Eriksholm Research Centre.
ESTOI	Extended short-time objective intelligibility.
HA	Hearing Aid.
HAD	HA device.
MMSE	Minimum mean square error.
MSE	Mean square error.
MVDR	Minimum variance distortionless response.
PDF	Probability density function.
PMF	Probability mass function.
PSD	Power spectral density.
RTF	Relative transfer function.
segSNR	Segmental SNR.
SNR	Signal-to-noise ratio.
STFT	Short-time Fourier transform.

VAD	Voice activity detector.

List of Symbols

$(\cdot)^*$	Complex conjugate.
$(\cdot)^H$	Conjugate transposition.
$(\cdot)^{-1}$	Matrix inversion.
$\boldsymbol{\Gamma_v}(k,l)$	Normalized noise CPSD matrix.
$\boldsymbol{a}(n,\theta_s)$	Vector containing the AIRs from the target speaker to each of the M microphone.
$\boldsymbol{C_x}(k,l), \boldsymbol{C}_s(k,l), \boldsymbol{C_v}(k,l)$	Noisy, target, and noise CPSD matrix.
$\boldsymbol{d}(k,l,\theta)$	RTF vector of the target sound source to the M microphones.
$\boldsymbol{I}_{M\times M}$	$M \times M$ identity matrix.
$\boldsymbol{v}(n)$	Vector containing the noise signals for each microphone.
$\boldsymbol{w}_{\mathrm{B}}(k,l)$	Bayesian beamformer weight vector.
$\boldsymbol{w}_{\mathrm{MVDR}}(k,l,\theta_s)$	MVDR beamformer weight vector.
$\boldsymbol{X}(k,l)$	Sequence of L consecutive frames of noisy microphone observations.
$\boldsymbol{x}(n)$	Vector containing the noisy signals for each microphone.
$\hat{\boldsymbol{d}}(k,l,\hat{\theta}_s)$	Estimated RTF vector.
$\hat{\boldsymbol{R}}(k,l)$	Sample estimate of the noisy CPSD matrix.
$\hat{\lambda}_{s,\mathrm{ML}}(k,l,\theta_i), \hat{\lambda}_{v,\mathrm{ML}}(k,l,\theta_i)$	Maximum likelihood estimates of $\lambda_s(k,l)$ and $\lambda_s(k,l)$, respectively.
$\hat{\theta}_{s,\mathrm{ML}}$	Maximum likelihood estimate of the target direction θ_s.
$\hat{\tilde{s}}(k,l)$	Estimated target speech signal in the time-frequency domain.
$\hat{s}(n)$	Estimated target speech signal in the time domain.

$\lambda_s(k,l), \lambda_v(k,l)$	PSD of the target and noise signals at the reference microphone, respectively.
$\mathbb{C}$	The set of complex numbers.
$\mathbb{N}_0$	The set of non-negative integers.
$\mathbb{R}$	The set of real numbers.
$\mathcal{D}$	RTF dictionary.
$\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$	VAD states for speech-absense, single-talk for target talker 1 and 2, and double-talk.
$\phi_j(l)$	Discretized eye-gaze measurement.
$\mathrm{Re}\{\cdot\}, \mathrm{Im}\{\cdot\}$	Real and imaginary part operators, repsectively.
$\exp(\cdot)$	Exponential function.
$\mathrm{tr}(\cdot)$	Trace operator.
Θ	Discrete candidate set of directions from which the target signal can arrive.
θ_i	DOA parameter.
θ_s	Target DOA.
$\tilde{\boldsymbol{a}}(k,l,\theta_s)$	Vector of ATFs.
$\tilde{\boldsymbol{v}}(k,l)$	Vector of noise signals.
$\tilde{\boldsymbol{x}}(k,l)$	Vector of noisy microphone signals in the time-frequency domain.
$\tilde{a}_m(k,l,\theta_s)$	ATF from the target sound source to the m'th microphone.
$\tilde{s}(k,l)$	Time-frequency domain representation of $s(n)$.
$\tilde{v}_m(k,l)$	Time-frequency domain representation of $v_m(n)$.
$\tilde{x}_m(k,l)$	Noisy microphone signal in time-frequency domain.
$\lvert\cdot\rvert$	Matrix determinant.
$a_m(n,\theta_s)$	AIR from the target sound source to the m'th microphone.

$d_m(k,l,\theta)$	RTF of the target sound source to the m'th microphone.
$E[\cdot]$	Expectation.
$f(\boldsymbol{X}(k,l)\|\theta_i)$	Likelihood function for the noisy microphone signals $\boldsymbol{X}(k,l)$ given θ_i.
$f\left(\boldsymbol{X}(k,l)\|\theta_i,\phi_j(l)\right)$	Likelihood function for the noisy microphone signals $\boldsymbol{X}(k,l)$ given θ_i and $\phi_j(l)$.
$f\left(\boldsymbol{X}(k,l)\|\phi_j(l)\right)$	Likelihood function for the noisy microphone signals $\boldsymbol{X}(k,l)$ given $\phi_j(l)$.
j	Imaginary unit.
k	Frequency bin index.
l	Time frame index.
M	Number of microphones.
n	Discrete-time index.
$p(\theta_i)$	Prior probability of the target DOA.
$p(\theta_i\|\boldsymbol{X}(k,l))$	Posterior probability density function of θ_i, given $\boldsymbol{X}(k,l)$.
$p\left(\theta_i\|\boldsymbol{X}(k,l),\phi_j(l)\right)$	Posterior probability density function of θ_i given $\boldsymbol{X}(k,l)$ and $\phi_j(l)$.
$p\left(\theta_i\|\phi_j(l)\right)$	Conditional probability of θ_i given $\phi_j(l)$.
$s(n)$	Time domain target speech signal.
$s_p(n)$	Time domain speech signal for target p.
$s_p^l(n)$	l'th windowed segment of $s_p(n)$.
$v_m(n)$	Additive noise signal at m'th microphone.
$x_m(n)$	Noisy signal at m'th microphone.

Contents

1. Introduction

1.1 Haring Loss

The human hearing is one of our most important senses which enables us, among other things, to communicate through spoken language. Speech communication plays an important role in our everyday life as it e.g., is important for social interaction and for the opportunities for effective learning in school and in higher education for children and young people. [1] Unfortunately, people with hearing impairment have reduced ability to understand speech meaning that hearing impairment may have a negative impact on quality of life. World Health Organization (WHO) [2] estimates that approximately 430 million people of the world's population have disabling hearing loss, making hearing loss one of the most common sensory processing disorders. In fact, it is estimated that over 700 million people will have disabling hearing loss by 2050.

1.2 Hearing Aids

To accommodate individuals with hearing impairments, hearing aids (HAs) can be used, which are devices that are capable of applying advanced digital signal processing on sampled sound signals with the main goal of increasing speech intelligibility and quality in order to deliver the intended sound to the HA user [3, p. 4].

Today, there exists various types and designs of HAs [1]. Among these, is the behind-the-ear (BTE) HA which is a device where most of the components are placed behind the pinna, the microphones are positioned just above the pinna, and where the sound is led to the ear canal through a small plastic tube [1, p. 57]. A typical modern BTE HA is shown in Fig. 1.1 [4]. The microphones of a modern HA convert the sound pressure measured on the microphone array into electrical signals which are then sampled into discrete time signals and thereby processed on the HA device (HAD) [1].

Figure 1.1: An Oticon Opn behind-the-ear with receiver-in-canal hearing aid [4].

1.3 Directional Microphones and Beamforming

One of the most challenging situations for hearing-impaired people is to understand speech in acoustically noisy environments [3, p. 110]. As a consequence, most hearing-impaired people find it extremely difficult to communicate through speech in e.g. larger groups of people. One example is a cocktail party, in which the hearing impaired is involved in a listening environment involving multiple speakers and noise sources. In such scenarios, hearing-impaired people are in general heavily challenged, specifically when the signal-to-noise ratio (SNR) is low [1, p. 41]. Due to the fact that difficulty when listening in noisy environments is a common speech intelligibility complaint, this aspect has had, and still has, a lot of effort in the design of HAs [3, p. 169].

To help the hearing impaired to understand speech in noisy environments, noise reduction technologies are often implemented in HAs [5, p. 269]. The overall goal of those noise reduction algorithms is to reduce the noise and increase speech intelligibility and quality of the desired speech in the acoustic environment [3, pp. 9, 111].

A large portion of acoustic noise reduction algorithms are implemented as linear filters. Usually, the acoustic signal is picked up by $M \geq 2$ microphones and passed through a linear filter which suppresses the noise. Ideally, the filter creates an acoustic beam towards the desired speaker by the use of directional microphones. This can be extremely useful as the beam can enhance the sound from the location of the desired speaker, i.e., from the target speaker direction, while attenuating other sounds from locations of non-interest [3, p. 111], [6, p. 3]. The method of processing multi-channel signals in order to enhance signals from a particular spatial direction is also called beamforming, and is a type of signal processing methods implemented in hearing aids, that has proven effective at increasing speech intelligibility [3, p. 9][6, p. 3].

1.4 Direction-of-Arrival Estimation

Many acoustic beamformers used for noise reduction in HAs require knowledge of the location of the desired speaker with respect to the microphones of the HA in order to steer the beam. The location of the desired speaker is in many realistic situations, however, not know in advance and has to be estimated online from the observable noisy microphone signals, or assumed to be known, e.g., assumed to be directly in the front of the user [7], [8, p. 265]. The relative location of a sound source with respect to the microphones of the HA is generally given in terms of the direction-of-arrival (DOA) of the sound wave impinging from that direction [7], [9]. Hence, a class of algorithms often used for estimating the location of the desired speaker is DOA estimation algorithms [7], [9]. Traditional DOA estimation algorithms typically try to localize the target sound source acoustically, i.e., only using the microphone signals. However, these methods are known to perform poorly in complex situations, particularly in acoustic scenes with loud competing speakers and when the SNR is low [10, p. 243]. For example, current DOA estimators such as maximum likelihood [11]–[13], and deep learning-based DOA estimators [9], are known to suffer from not being able to robustly handle a conversational partner in a multi-speaker environment, without additional a priori information on the conversational partner's location [14]. This is due to the fact that competing speakers share similar signal characteristics to the desired speaker, and hence, most conventional audio-only DOA estimation algorithms struggle at determining if a competing speaker is desired or not. In worst cases, the DOA estimation algorithm may erroneously classify a competing speaker as being desired and instead the beamformer might enhance the noise and suppress the desired speech, i.e., working against the overall objective of helping the hearing impaired understand speech in noisy environments. Hence, in general, the task of estimating the DOA is not simple, and the consequences of DOA estimation errors can be severe [7], [8]. In other words, accurate target sound source localization and target DOA estimation are crucial for beamformers to steer the acoustic beam towards the target target [14].

Problems similar to the ones mentioned above may occur when using beamformers that assume the target talker to be directly in the front of the user. This is due to the fact that such frontal steered beamformers do not take into account where the HA user is looking [15]. For instance, it may be the case that the HA user is turning his or her head towards person A, but in fact be listening to and gazing at person B. Another typical example one could imagine would be that in a multi-talker scenario, it is likely that the HA user is turning his or her head to be in between person A and B while the eyes are jumping back and forth between person A and B. In this case, it is possible that none of the target signals would be enhanced by the frontal steered beamforming system, and in worst case, the beamforming system would determine the target speakers as being noise. Furthermore, studies have shown that in the case a target talker is positioned on the side of the HA user, the user do typically not turn

his or her all the way to directly face the target talker, even though the HA user may be looking at the target talker [16].

As mentioned, what is shared about the traditional methods mentioned above is that they only consider acoustic signals, i.e., they are described as so-called audio-only systems. However, the behaviour of a person during a conversation typically involves both auditory and visual attention [17], where visual attention implies that the person direct his or her eye-gaze toward the target sound source. Furthermore, visual information is essentially not affected by the acoustic noise and competing speakers in a listing environment, which makes vision a reliable cue to exploit in difficult acoustics conditions [18].

1.5 How to Improve Hearing Aids

In future HA systems, additional information apart from sound signals captured by microphones may be available. For instance, one could envision future HAs which could measure the eye-gaze direction of the user, e.g., via cameras pointing towards the eyes of the user (e.g., mounted on glasses), or using electrodes (e.g., in-ear electrodes), which may reveal the direction as a function of time of the user's eye. In many situations, this additional information can provide very strong evidence of the direction of an active target talker, and hence, help identify the target direction. For example, it is often the case that a HA user looks at the target sound source of interest, at least now and then, e.g. for lip reading in acoustically difficult situations [19]. Based on the fact that eye-gaze is described as an excellent predictor of conversational attention [19], and that eye contact furthermore is a natural human response in a social environment [19], it seems reasonable to suppose that the use of information about the HA users' eye position to help derive the target location, to some extent can have a beneficial contribution in noise reduction technology for future hearing aids.

1.6 Eye-Gaze Steering

The idea of using the HA user's eye-gaze to steer a HA has already been explored in several studies [16], [17], [20]–[22]. The eye-gaze steering described in previous studies works theoretically by enhancing the sound in the direction of the HA user's eye-gaze to any given time. Even though these previous studies confirms that such beamformers, which are steered toward the direction the user's eye-gaze, can be a good way to improve future HAs, such "hard" eye-gaze steered beamformers may still have some limitations. For instance, the user's eye-gaze may not always be directed towards the desired speaker although the user's attention is at the desired speaker, e.g., the eye-gaze might be slightly offset or the user might be gazing at something else than the desired speaker briefly. In these cases, a hard eye-gaze steered beamformer will likely classify the target talker as being noise, since it enhance in the direction

of the user's eye-gaze. However, sine natural listening behaviours involves that the user may not always direct his of he eyes toward the target speaker at any moment in time, it would be preferable to develop systems that allow for such uncertainty about the target direction, such that we are able to develop beamforming algorithms that better reflect and allow for real-life communication abilities. Specifically, instead of only relying on the acoustic microphone signals or on the user's eye-gaze to steer the beamformer, it could be interesting to examine how the eye-gaze and microphone signals from the HA can be combined to, for example, jointly estimate the direction of the desired speaker.

When the target DOA is uncertain, a so-called Bayesian approach to beamforming can be taken. In [23] and [24], methods have been proposed where a probability distribution on the target DOA is used to consider the Bayesian beamforming approach for noise reduction. These proposed methods rely on acoustic information only to estimate the target signal, but may be extended to be useful in a situation where an additional signal is available, e.g., a signal representing the HA user's eye-gaze.

1.7 Problem Statement

To summarize, even though several DOA estimation methods exists in the literature, achieving effective suppression of loud competing speakers remains extremely challenging and a remarkably difficult problem to solve even with the most state-of-the-art speech enhancement systems [14]. This fact makes the foundation for the motivation behind this thesis. In contrast to current audio-only beamforming methods, which try to localize the target source acoustically, i.e., using the microphones only, in this thesis, we study an alternative means for target sound source localization that exploits an additional modality, apart from sound, to help localizing and enhancing the target source. Specifically, based on the fact that the HA user's tend to look at the target sound source, e.g., for lip reading, it appears feasible to use information about the HA users' eye position to help derive the target location. Therefore, the motivation of this thesis is to investigate the possibility of incorporating the HA user's' eye-gaze into a HA beamformer, such the proposed method is able to perform better or at least on par with current state-of-the-art beamforming methods that uses microphone signals only. In this thesis, we therefore seek to answer the following main question.

Main Question:

How can information provided by the HA user's eye-gaze and by the HA microphone signals be combined and used to construct a beamformer for HA applications, and can such a beamformer potentially outperform current audio-only beamforming methods in terms of predicted speech intelligibility and predicted speech quality in noisy acoustic scenes?

In the process of developing and proposing a beamforming system for HAs which incorporates the HA user's eye-gaze, a natural and simple starting point, would be to examine an upper performance bound of using eye-gaze steered beamforming. Beside aiming in laying the foundation for study and propose more complex systems that are not developed under such ideal conditions, from a technical point of view, such a feasibility study may in addition provide valuable insights in determining the value of the concept for future HAs. To this end, we formulate the following sub question:

Sub Questions:

i) Under ideal conditions, what is the potential performance benefit of using eye-gaze steered beamforming?

The idealistic scenario is achieved by considering a synthetic situation where the user's eye-gaze is assumed precisely pointing towards the desired speaker at any moment in time.

1.7.1 Delimitations

In this thesis, we consider a Bayesian approach to fuse information provided by the HA user's eye-gaze and by the HA microphone signals. Parts of the theory presented throughout this thesis are general, but due to time constraints and in order to focus on technical aspect and application in real-world scenarios, we have needed to restrict our work in several ways. First of all, we have chosen to keep the mathematics regarding probability theory as simple as possible. In that regard, it should be noted that we present the results from probability theory using the notations that are traditionally used conventions in the engineering literature, and so, we do not treat the concepts in its most general measure-theoretic setting. In such a more general, measure-theoretic treatment of probability, every functions would be considered as densities with respect to different base measures. However, such a measure-theoretic treatment is beyond the scope of this thesis, hence, the reader is referred to [25, Sec. II] for a more rigorous measure-theoretic treatment of probability and statistics.

In this thesis, a dataset containing real-world measurements of HA users eye-gaze recorded in synchronization with presented audio-visual stimuli is available, which allows the study of beamforming systems for HAs which incorporates the user's eye-gaze. This dataset is provided by Eriksholm Research Centre (ERH) which is a part of Oticon. The dataset will be used to construct the proposed beamformers and to compare and determine the performance of different beamforming systems. Some information regarding the applied dataset will be not be provided in this thesis, due to the fact that the full report, in which the dataset is described is confidential, and therefore not public available.

Furthermore, in the simulations carried out in this thesis, we have chosen to remain as close as possible to the acoustic conditions of the experiments conducted

at ERH. Even though acoustic situations including competing speakers are where we expect the inclusion of eye-gaze in a beamforming system to be especially beneficial, we do not consider acoustic scenes with competing speakers, as data supporting this is not included in the dataset. In other words, we limit our simulations to the specific audio-visual stimuli of the experimental setup conducted at ERH, and hence, limit the extend to how explorative our simulations are.

1.8 Outline

The remainder of this thesis is structured as follows: Chapter 2 presents the acoustic signal model considered in this thesis and discusses the employed statistical assumptions. Given the acoustic signal model, the thesis moves on to cover acoustic beamforming, specifically, the minimum variance distortionless response beamformer, in Chapter 3. Chapter 4 will give an introduction to a well-known audio-only model-based DOA estimation algorithm which will be used as a competing method throughout this thesis. In Chapter 5, a detailed feasibility test is performed and the upper bound performance of an eye-gaze steered beamformer is compared to state-of-the-art audio-only beamforming methods. This chapter aim at answering the sub question formulated for this thesis. After accessing the upper performance bound of eye-gaze steered beamforming under ideal conditions, the thesis moves on to cover the study on how eye-gaze information can be incorporated in a beamforming system, with the aim of answering the main question of this thesis. In Chapter 6, a Bayesian approach is taken to the fusion of eye-gaze signals and acoustic signals, in which the necessary mathematical concepts of Bayesian beamforming, is introduced. Following this, the chapter provides a detailed presentation of our two proposed Bayesian beamforming methods. Chapter 7 introduces to the employed audio-visual dataset which contains real-world eye-gaze measurements, and in Chapter 8, the performance of the proposed beamforming methods are evaluated through numerical simulations using the eye-gaze data and sound signals from the dataset. Following this, Chapter 9 presents a discussion of the results obtained from the simulation experiments as well as of the work presented in this thesis as a whole. Finally, in Chapter 10, we conclude on the study presented in this thesis as well as present thoughts on aspects for interesting further development.

2. Acoustic Signal Model and Assumptions

The microphones of a modern HA convert the sound pressure measured on the microphone array into electrical signals which are then sampled into discrete time signals and subsequently processed on the HAD [6, p. 6]. As mentioned, noise reduction systems, such as beamforming systems, are often used in state-of-the-art HADs to process the microphone signals in order to enhance a target speech signal while suppressing the noise signals from the surroundings with the aim of improving speech intelligibility and quality. These noise reduction algorithms are often based on some optimality criteria and a statistical model of the microphone signals [26, p. 17],[8]. When constructing a mathematical model of acoustic environments where sound waves impinge on an array of microphones on an HA, we have to make assumptions about the sound field in which the sound waves propagate from the sound source to the array of microphones. Therefore, the focus of this chapter is on modeling the acoustic environments and the assumptions made in relation to the sound field and the sound sources, with the purpose of deriving a signal model that sufficiently describes the noisy microphone signals as they are picked up by the microphones on the HA. The signals will be considered as discrete time signals, as we process the signals after being received by the microphones, and thereby omitting the analog-to-digital conversion.

2.1 Overview of the Acoustic Scene

An HA user may experience a huge variety of different acoustic environments including everything from cocktail party like environments in which the hearing impaired is situated in a listening environment involving, e.g., a target speaker, multiple undesirable speakers, and other additive noise sources, where the sound waves approximately impinge from all directions, to listening environments where only a single target speaker and a single point-source interferer, e.g., a competing speaker, is present [6]. Furthermore, speech generated by, e.g., the target speaker does not only reach the microphones on the HA via the direct propagation from the source, but also via reflections off the objects and surfaces in the room. In most realistic situations, the microphone signals are additionally disrupted by noise generated by the microphones themselves, referred to as microphone self-noise [3, pp. 110-114],[26,

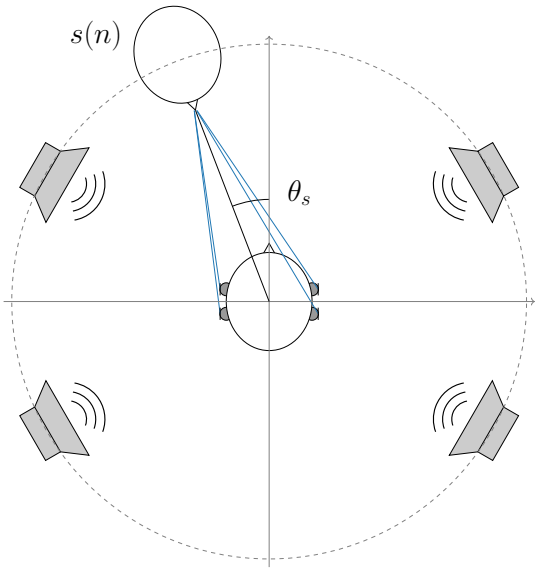


Figure 2.1: An example of an acoustic environment with an HA user wearing two behind-the-ear HAs with two microphones on each, a target source which emitting a signal $s(n)$ which impinges on the microphone array from direction θ_s, and arbitrary noise sources represented by the loudspeakers. It is assumed that the target sound source can be modeled as a point source and that the only objects affecting any sound source before reaching the microphones are the head, torso and pinnae of the HA user. Blue lines depict the sound transmission from a point sound source to the two ears [27, Fig. 1.14].

p. 71]. For simplicity, we choose in this thesis to limit the scope to only include acoustic environments where the sound waves propagating from a source are affected by the head, torso, and pinnae of the HA user, but otherwise propagates in free field.

A typical example of an acoustic environment is illustrated in Fig. 2.1, where an HA user, wearing BTE HAs with two microphones on each ear, is located in a noisy environment with multiple sound sources. One of these sound sources is the target speaker which is the person whom the HA user wants to listen to. The speech signal emitting from the target speaker is referred to as the target speech signal and is denoted $s(n)$, with $n \in \mathbb{N}_0$ indicating a discrete-time index. The other sound sources, which are depicted as loudspeakers in Fig. 2.1, are noise sources which may be either diffuse noise, interfering point sources, or any other additive source of noise which may mask the target speech signal. However, since any of these undesired sources are disadvantageous for speech quality and intelligibility, they may be regarded as a single entity.

Based on the acoustic setup examplified in Fig. 2.1, we will in the following section derive a signal model for the noisy microphone signals as they impinge on the microphone array of the HA.

2.2 Signal Model in the Time Domain

According to the acoustic scenario in Fig. 2.1, each microphone receives a sound signal which is assumed to consist of two components, namely a desired speech signal and an undesired noise signal covering any signal that is not originating from the target speech source. We assume that the target source can be modeled as a point source, meaning that the sound source produces spherical waves which propagate omnidirectionally in space [28]. This assumption implies that the propagation of the sound signal from the target source to a microphone can be expressed by a linear convolution between the target signal and an acoustic impulse response (AIR) [29, p. 694]. Each microphone is associated with different AIRs since the propagation of a sound signal from a particular source location to each microphone is different [29, p. 694]. This issue is evident by the blue lines in Fig. 2.1 illustrating the sound transmission from a point sound source to each of the four microphones [27, p. 20]. Generally, the AIRs are functions of both the azimuth and elevation angle of the sound source in relation to the microphone array. This is due to the fact that signals originating from different azimuth and elevation angles propagate from the sources to the m'th microphone differently. However, for simplicity, we will in this thesis limit ourselves to the horizontal plane, i.e., we will only consider the dependency of the azimuth angle. Furthermore, we assume that the position of a sound source in the acoustic environment as well as the position of the HA user is static, meaning that the AIRs are assumed to be time-invariant [29, p. 694]. These two assumptions together implies that the sound transmission from a point source to the microphones can be expressed by a linear time-invariant system. [29, p. 694]. Hence, for a microphone array with M microphones, where each microphone picks up the sound from the noisy acoustic environment, the noisy signal $x_m(n)$, for $n \in \mathbb{N}_0$, at the m'th microphone can be modeled as

$$x_m(n) = (s * a_m(\bullet, \theta_s))(n) + v_m(n), \quad m = 1, \dots M, \tag{2.1}$$

where $*$ denotes the linear convolution operator, which is defined in Appendix A.1, $a_m(n, \theta_s)$ denotes the AIR from the target to the m'th microphone, $s(n)$ is the target signal measured at the target source and impinges on the microphone array from direction θ_s, and $v_m(n)$ denotes an overall additive noise component containing a sum of all undesired signals received at the m'th microphone, e.g., interfering point sources, diffuse background noise, and microphone self-noise.

2.3 Signal Model in the Time-Frequency Domain

Due to the wide-band and non-stationary nature of speech, speech processing such as beamforming is conveniently performed in the time-frequency domain [26, p. 72],[8]. Typically, the time-frequency domain representation of the noisy microphone signals is obtained by making use of the short-time Fourier transform (STFT) which is defined in Appendix A.2 [30, p. 230].

Let k and l be the frequency bin index and time frame index, respectively, for $k = 0, \ldots, N-1$ where N is the window length used in the STFT. Then, by applying the STFT to the noisy microphone signal $x_m(n)$ in (2.1), we obtain a time-frequency domain representation, which for a given frequency bin index k and time frame index l is denoted $\tilde{x}_m(k,l) \in \mathbb{C}$, for $m = 1, \ldots, M$, and is given as

$$\begin{aligned} \tilde{x}_m(k,l) &= STFT\{x_m(n)\}(k,l) && (2.2)\\ &= STFT\{(s * a_m(\bullet, \theta_s))(n)\}(k,l) + STFT\{v_m(n)\}(k,l) && (2.3)\\ &= \tilde{s}(k,l)\tilde{a}_m(k,l,\theta_s) + \tilde{v}_m(k,l), && (2.4) \end{aligned}$$

where $\tilde{a}_m(k,l,\theta_s) \in \mathbb{C}$ is the acoustic transfer function (ATF) from the target source to the m'th microphone, (2.3) follows by linearity of the STFT, and (2.4), which is known as the narrowband approximation of the noisy microphone signal [29, p. 696], is due the fact that convolution in the time domain can be approximated as a multiplication in the short-time frequency domain, provided the window length is appropriately large [31]. In this thesis, we adopt the standard made assumption in speech processing that $\tilde{x}_m(k,l) \in \mathbb{C}$, for $m = 1, \ldots, M$, are approximately independent across time l and frequency k, which allows us to treat each STFT coefficient independently. It should be noted that this independency assumption is valid when the correlation time of the signal is short compared to the window length N and when successive frames are spaced sufficiently far apart [13], [32]. For a given frequency bin k and time frame l, we stack, for notational conciseness, the complex-valued STFT coefficients of all M noisy microphone signals in an $M \times 1$ vector $\tilde{\boldsymbol{x}}(k,l) \in \mathbb{C}^M$, by defining

$$\begin{aligned} \tilde{\boldsymbol{x}}(k,l) &= \begin{bmatrix} \tilde{x}_1(k,l) & \ldots & \tilde{x}_M(k,l) \end{bmatrix}^T,\\ \tilde{\boldsymbol{a}}(k,l,\theta_s) &= \begin{bmatrix} \tilde{a}_1(k,l,\theta_s) & \ldots & \tilde{a}_M(k,l,\theta_s) \end{bmatrix}^T, \qquad (2.5)\\ \tilde{\boldsymbol{v}}(k,l) &= \begin{bmatrix} \tilde{v}_1(k,l) & \ldots & \tilde{v}_M(k,l) \end{bmatrix}^T, \end{aligned}$$

which allow us to express the complex-valued STFT coefficients of all M noisy microphone signals in vector notation as

$$\tilde{\boldsymbol{x}}(k,l) = s(k,l)\tilde{\boldsymbol{a}}(k,l,\theta_s) + \tilde{\boldsymbol{v}}(k,l), \qquad (2.6)$$

where it follows that $\tilde{s}(k,l)$ is the scalar STFT coefficient of the target speech signal, $\tilde{\boldsymbol{a}}(k,l,\theta_s)$ is the vector of ATFs from the target signal to all microphones.

When performing beamforming in HA applications, we are typically interested in extracting the target speech signal measured at a pre-selected reference microphone as opposed to the target speech signal as measured at the target source. This is due to the fact that the signal at the reference microphone already has propagated from the source to the HA, which means it contains information about reflections and influences made by the HA users head, pinnae and torso. This situation is more natural, compared to the situation where the target signal is measured at the target source, as it corresponds to how we naturally perceive sound, that is, at our ears

and not at the source of the sound. Therefore, it is often seen in the literature, e.g., [29, p. 694], that the ATFs are normalized with respect to a pre-selected reference microphone. After normalization, we refer to the ATF vector as the relative acoustic transfer function (RTF) vector [29, p. 696]. More specifically, let

$$\boldsymbol{d}(k,l,\theta_s) = \frac{\tilde{\boldsymbol{a}}(k,l,\theta_s)}{\tilde{a}_{m^\star}(k,l,\theta_s)}, \tag{2.7}$$

denotes a vector whose elements m'th element $d_m(k,l,\theta_s)$, for $m = 1\ldots,M$, represents the RTF from the target source to the m'th microphone, and where $m^\star$ is the reference microphone index. As a result, the $m^\star$'th element in $\boldsymbol{d}(k,l,\theta_s)$ equals one, while the other elements define the RTFs of the target signal from the reference microphone to all of the microphones, i.e.,

$$\boldsymbol{d}(k,l,\theta_s) = \begin{bmatrix} 1 & d_2(k,l,\theta_s) & \ldots & d_M(k,l,\theta_s) \end{bmatrix}^T, \tag{2.8}$$

where we, without loss of generality, have defined $m^* = 1$ to be the reference microphone. By substituting the ATFs $\tilde{\boldsymbol{a}}(k,l,\theta_s)$ with the RTFs $\boldsymbol{d}(k,l,\theta_s)$ in (2.6), a modified signal model is obtained as

$$\tilde{\boldsymbol{x}}(k,l) = \tilde{s}_{ref}(k,l)\boldsymbol{d}(k,l,\theta_s) + \tilde{\boldsymbol{v}}(k,l), \tag{2.9}$$

in which the target speech signal at the microphones is described in terms of the RTFs $\boldsymbol{d}(k,l,\theta_s)$ and the target speech signal measured at the reference microphone given as

$$\tilde{s}_{ref}(k,l) = \tilde{s}(k,l)\tilde{a}_1(k,l,\theta_s). \tag{2.10}$$

The aim of beamforming, which will be covered in the next chapter, is then to obtain an estimate of $\tilde{s}_{ref}(k,l)$, which is free from noise [33, p. 2974].

Beside the assumptions of a target point source and an additive noise component in the signal model, we assume, for mathematical convenience, that the noisy microphone signals $\tilde{\boldsymbol{x}}(k,l)$ are realizations of random processes, and adopt the often-made statistical assumptions that the noise component is uncorrelated to the speech component [29]. This allows us to model the $M \times M$ inter-microphone cross power spectral density (CPSD) matrix of $\tilde{\boldsymbol{x}}(k,l)$, which is defined as $\boldsymbol{C}_x(k,l) = E\left[\tilde{\boldsymbol{x}}(k,l)\tilde{\boldsymbol{x}}^H(k,l)\right]$, where $E[\cdot]$ and $(\cdot)^H$ denote the expectation and conjugate transpose operators, respectively, as a sum of the CPSD matrices of the two individual signal components [26], i.e.,

$$\boldsymbol{C}_{\{}\boldsymbol{x}(\boldsymbol{k},\boldsymbol{l}) = \boldsymbol{C}_s(k,l) + \boldsymbol{C}_{\boldsymbol{v}}(k,l). \tag{2.11}$$

Under the assumption that the target RTF vector $\boldsymbol{d}(k,l)$ is deterministic with respect to expectation, the target CPSD matrix $\boldsymbol{C}_s(k,l) \in \mathbb{C}^{M\times M}$ can be expressed as

$$\boldsymbol{C}_s(k,l) = \lambda_s(k,l)\boldsymbol{d}(k,l,\theta_s)\boldsymbol{d}^H(k,l,\theta_s), \tag{2.12}$$

where $\lambda_s(k,l) = E[\tilde{s}_{ref}(k,l)\tilde{s}^*_{ref}(k,l)]$, with $(\cdot)^*$ denoting the complex conjugate, is defined as the power spectral density (PSD) of the target signal at the reference

microphone. Furthermore, it is often assumed that the noise CPSD matrix $\boldsymbol{C_v}(k,l) \in \mathbb{C}^{M\times M}$ can be modeled as [12], [13]

$$\boldsymbol{C_v}(k,l) = \lambda_v(k,l)\boldsymbol{\Gamma_v}(k,l_0), \quad l > l_0, \tag{2.13}$$

where $\lambda_v(k,l) = E[\tilde{v}_{ref}(k,l)\tilde{v}^*_{ref}(k,l)]$ is the PSD of the noise signal at the reference microphone, $\boldsymbol{\Gamma_v}(k,l_0)$ is the normalized noise CPSD matrix which contains a value of 1 at the diagonal element corresponding to the reference microphone index [34], and l_0 denotes the most recent frame index with speech absence.Substituting (2.12) and (2.13) into (2.11), the noisy CPSD matrix then becomes

$$\boldsymbol{C}_x(k,l) = \lambda_s(k,l)\boldsymbol{d}(k,l,\theta_s)\boldsymbol{d}^H(k,l,\theta_s) + \lambda_v(k,l)\boldsymbol{\Gamma_v}(k,l_0). \tag{2.14}$$

For notational convenience, we will from this point omit the indexing of ref in the notation of the target speech signal measured at the reference microphone. Hence, we use $\tilde{s}(k,l)$ to denote both the target speech signal measured at the source and the target speech signal measured at the reference microphone. The representations of the target speech signal are distinguished by whether the target speech component is written in terms of the ATF vector $\tilde{\boldsymbol{a}}(k,l,\theta_s)$ or the RTF vector $\boldsymbol{d}(k,l,\theta_s)$.

In this chapter, the employed signal model has been derived in both the time domain and in the time-frequency domain, and assumptions with regard to the employed signal model have been presented. The time domain signal model in (2.1) will be used to simulate the noisy microphone signals received at the HA microphones in the different acoustic scenes studied in this thesis. The next chapter will cover an introduction to some basic concepts of acoustic beamforming.

3. Acoustic Beamforming

In acoustic environments consisting of an HA user wearing HADs with multiple microphones, a target speaker and multiple noise sources originating from the surroundings, the microphones on the HAD do not only pick up the clean target speech signal but the noisy signal. In such an environment, a hearing impaired may struggle to understand the desired speech compared to persons with normal hearing. In order to make challenging listening environments more accessible to the HA user, the overall goal of the HADs is to enhance the speech signal impinging from the direction of the target source from the observed noisy microphone signals. As mentioned, beamforming methods are typically implemented in modern HAs for such multi-channel noise reduction purposes. The basic concept of acoustic beamforming is to create a directional microphone that is steered towards e.g., the direction of the target speaker, by optimally combining the noisy signals picked up by the microphones into one signal to be presented to the user [8]. In this chapter, we introduce the concept of acoustic beamforming. Specifically, we focus on the minimum variance distortionless response (MVDR) beamformer [10], [29], [35], as this beamformer will be widely used throughout the thesis. Later, in Chapter 6, we will move on to consider so-called Bayesian beamformers, and as we will se, Bayesian beamfomers can be considered as a weighted sum of MVDR bemformers.

It should be noted that often the structure of a noise reduction system as employed in a modern HA comprises a beamformer coupled with a post-filter, resulting in a two-step algorithm [8]. However, for simplicity, we do not consider post-filtering methods in this thesis.

An overview block diagram of a simple beamforming system is illustrated in Fig. 3.1. As evident from the block diagram, the overall beamforming system can be split up into three stages: An analysis stage where the received signals are converted into a time-frequency representation using the STFT, processing stage where the beamformer parameters are estimated and where the beamformer is applied to the noisy signals to obtain an estimate of the target speech signal, and a synthesis stage where the estimated target speech signal is transformed back into the time domain using the ISTFT. In this chapter, we focus solely on the processing stage. The necessary concepts of analysis and synthesis of the spatio-temporal signals will be discussed in Chapter 5 where the practical implementation of the proposed beamforming noise

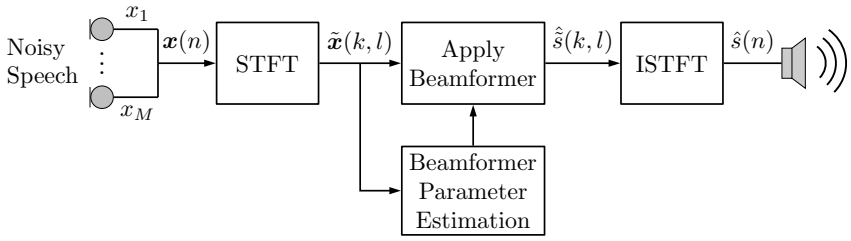


Figure 3.1: Overview of the elements used in a beamforming noise reduction system in a HA.

reduction system is covered.

3.1 Linear Beamformers

Referring back to the signal model for the noisy observations in the time-frequency domain given as

$$\tilde{\boldsymbol{x}}(k,l) = \boldsymbol{d}(k,l,\theta_s)\tilde{s}(k,l) + \tilde{\boldsymbol{v}}(k,l), \tag{3.1}$$

where $\tilde{s}(k,l) \in \mathbb{C}$ is the target signal measured at the reference microphone, $\boldsymbol{d}(k,l,\theta_s) \in \mathbb{C}^M$ is the RTF vector, and $\tilde{\boldsymbol{v}}(k,l) \in \mathbb{C}^M$ is the overall additive noise component which is assumed to be uncorrelated with the target signal.

A beamformer is a linear spatial filter defined by a vector of M complex weights $\boldsymbol{w}(k,l) \in \mathbb{C}^M$ consisting of one weight per microphone in the microphone array. The beamformer is then applied as an inner product between the beamformer weights and the noisy microphone signals such that the output of the beamformer for the k'th frequency bin and the l'th time frame is [29, p. 698]

$$y(k,l) = \boldsymbol{w}^H(k,l)\tilde{\boldsymbol{x}}(k,l). \tag{3.2}$$

Substituting $\tilde{\boldsymbol{x}}(k,l)$ in (3.1) into (3.2), the processed signal $y(k,l) \in \mathbb{C}$ can be represented as

$$\begin{aligned} y(k,l) &= \boldsymbol{w}^H(k,l)\left(\boldsymbol{d}(k,l,\theta_s)\tilde{s}(k,l) + \tilde{\boldsymbol{v}}(k,l)\right) \\ &= \tilde{s}(k,l)\boldsymbol{w}^H(k,l)\boldsymbol{d}(k,l,\theta_s) + \boldsymbol{w}^H(k,l)\tilde{\boldsymbol{v}}(k,l). \end{aligned} \tag{3.3}$$

Using the assumption that the noisy microphone signals $\tilde{\boldsymbol{x}}(k,l)$ are realizations of a random process, and assuming that the the weight vector $\boldsymbol{w}(k,l)$ is deterministic with respect to expectation, the output power of a beamformer is defined as

$$\begin{aligned} E[|y(k,l)|^2] &= E[y(k,l)y^*(k,l)] \\ &= E[\boldsymbol{w}^H(k,l)\boldsymbol{x}(k,l)\boldsymbol{x}^H(k,l)\boldsymbol{w}(k,l)] \\ &= \boldsymbol{w}^H(k,l)\boldsymbol{C}_{\boldsymbol{x}}(k,l)\boldsymbol{w}(k,l), \end{aligned} \tag{3.4}$$

where $\boldsymbol{C}_{\boldsymbol{x}}(k,l) \in \mathbb{C}^{M\times M}$ is the noisy CPSD matrix.

The weight vectors of the individual beamformers can be computed considering some optimization criterion, such as minimum mean square error (MMSE), minimum variance distortionless response (MVDR), linearly constrained minimum variance (LCMV), etc. [7]. As mentioned, consider in this chapter the MVDR criterion to obtain the weight vectors.

3.1.1 Minimum Variance Distortionless Response Beamformer

In this section, the MVDR beamformer will be introduced. The MVDR beamformer is of interest, at it can be shown that the MVDR beamformer is the beamformer which maximizes the SNR of the output of the beamformer [29, p. 702],[36, p. 1367].

The MVDR beamformer collects statistics about the listening environment to derive beamformer weights that 1) attenuate the noise energy as much as possible, i.e., achieve minimum variance, while 2) ensuring that the sounds from the target direction are not attenuated or amplified, i.e., achieve a distortionless response towards the target [8]. Hence, for the MVDR beamformer, the weight vector is be obtained as the solution to the constrained optimization problem [10], [29], [35]

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \boldsymbol{w}^H(k,l)\boldsymbol{C}_{\boldsymbol{v}}(k,l)\boldsymbol{w}(k,l) \\ \text{s.t.} \quad & \boldsymbol{w}^H(k,l)\boldsymbol{d}(k,l,\theta_s) = 1. \end{aligned} \tag{3.5}$$

The solution to the optimization problem for the MVDR beamformer is given by Theorem 3.1.

Theorem 3.1 (Optimal MVDR beamformer coefficients)
Let M be the number of microphones and assume $\boldsymbol{C}_{\boldsymbol{v}}(k,l) \in \mathbb{C}^{M\times M}$ is positive definite and Hermitian. The solution to the optimization problem

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \boldsymbol{w}^H(k,l)\boldsymbol{C}_{\boldsymbol{v}}(k,l)\boldsymbol{w}(k,l) \\ \text{s.t.} \quad & \boldsymbol{w}^H(k,l)\boldsymbol{d}(k,l,\theta_s) = 1, \end{aligned} \tag{3.6}$$

is the optimal beamformer coefficients given as

$$\boldsymbol{w}_{\text{MVDR}}(k,l,\theta_s) = \frac{\boldsymbol{C}_{\boldsymbol{v}}^{-1}(k,l)\boldsymbol{d}(k,l,\theta_s)}{\boldsymbol{d}^H(k,l,\theta_s)\boldsymbol{C}_{\boldsymbol{v}}^{-1}(k,l)\boldsymbol{d}(k,l,\theta_s)}. \tag{3.7}$$

Proof.
The optimization problem in (3.6) is a constrained minimization problem where the real-valued and non-negative function of the complex weight vector $\boldsymbol{w}$ to be minimized is

$$f(\boldsymbol{w}) = \boldsymbol{w}(k,l)^H\boldsymbol{C}_{\boldsymbol{v}}(k,l)\boldsymbol{w}(k,l), \tag{3.8}$$

subject to the constraint that

$$\boldsymbol{w}^H(k,l)\boldsymbol{d}(k,l,\theta_s) = 1. \tag{3.9}$$

In order to obtain the solution to this optimization problem, the well-known approach is to rewrite the constrained minimization problem in (3.6) using the method of complex Lagrange multipliers [35, p. 442], [10, p. 25].

Let the complex Lagrange multiplier, denoted $\lambda \in \mathbb{C}$, be given as [37, p. 793]

$$\lambda = \lambda_r + j\lambda_i, \tag{3.10}$$

where $\lambda_r = \mathrm{Re}\{\lambda\}$ and $\lambda_i = \mathrm{Im}\{\lambda\}$ is the real and imaginary part of λ, respectively, and j is the imaginary unit. We then convert the constrained minimization problem in (3.6) into an unconstrained minimization by forming the Lagrangian function as [37, pp. 793-794], [38]

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}) &= f(\boldsymbol{w}) + \lambda_r \mathrm{Re}\{\boldsymbol{w}^H(k,l)\boldsymbol{d}(k,l,\theta_s) - 1\} \\ &\qquad + \lambda_i \mathrm{Im}\{\boldsymbol{w}^H(k,l)\tilde{\boldsymbol{d}}(k,l,\theta_s) - 1\} \qquad (3.11)\\ &= f(\boldsymbol{w}) + \mathrm{Re}\{\lambda^*(\boldsymbol{w}^H(k,l)\boldsymbol{d}(k,l,\theta_s) - 1)\} \qquad (3.12)\\ &= f(\boldsymbol{w}) + \frac{1}{2}\lambda^*\left(\boldsymbol{w}^H(k,l)\boldsymbol{d}(k,l,\theta_s) - 1\right) \qquad (3.13)\\ &\qquad + \frac{1}{2}\lambda\left(\boldsymbol{d}^H(k,l,\theta_s)\boldsymbol{w}(k,l) - 1\right),\end{aligned}$$

where we have used the relations

$$\mathrm{Re}\{z\} = \frac{z+z^*}{2} \quad \text{and} \quad \mathrm{Im}\{z\} = \frac{z-z^*}{2j}, \quad z \in \mathbb{C}. \tag{3.14}$$

As it appears from (3.13), we may consider $\mathcal{L}$ as a real function of the two complex variables $\boldsymbol{w}$ and $\boldsymbol{w}^H$, so we denote it by $\mathcal{L}(\boldsymbol{w}, \boldsymbol{w}^H)$, [39, pp. 518-519]. Note that, the notations in this proof is slightly different from the ones in the rest of the thesis, as we sometimes consider $\boldsymbol{w}$ without arguments, to accord with the literature.

Invoking the rules of partial differentiation under Wirtinger calculus [37, pp .785-794], we assume that $\mathcal{L}(\boldsymbol{w}, \boldsymbol{w}^H)$ is analytic in $\boldsymbol{w}$ and $\boldsymbol{w}^H$ independently, in the sense of partial differentiation. Then, it can be shown [38] that a necessary and sufficient condition for $\mathcal{L}(\boldsymbol{w}, \boldsymbol{w}^H)$ to be minimized can be obtained by taking the complex gradient of $\mathcal{L}(\boldsymbol{w}, \boldsymbol{w}^H)$ with respect to $\boldsymbol{w}^H$ and equating the result to zero.

Differentiating $\mathcal{L}(\boldsymbol{w}, \boldsymbol{w}^H)$ with respect to $\boldsymbol{w}^H$ and formally treating $\boldsymbol{w}$ as a constant [37, pp .785-794], we obtain

$$\frac{\partial \mathcal{L}(\boldsymbol{w}, \boldsymbol{w}^H)}{\partial \boldsymbol{w}^H} = \boldsymbol{C}_{\boldsymbol{v}}(k,l)\boldsymbol{w}(k,l) + \frac{1}{2}\lambda^*\boldsymbol{d}(k,l,\theta_s). \tag{3.15}$$

Setting (3.15) equal to zero and solve for $\boldsymbol{w}$, yields the structure of the optimum weight vector, i.e.,

$$\boldsymbol{w}_{\mathrm{opt}}(k,l) = -\frac{1}{2}\lambda^*\boldsymbol{C}_{\boldsymbol{v}}^{-1}(k,l)\boldsymbol{d}(k,l,\theta_s), \tag{3.16}$$

where λ^* remains to be determined. Note that $\boldsymbol{C}_{\boldsymbol{v}}^{-1}(k,l)$ is valid by the assumption that $\boldsymbol{C}_{\boldsymbol{v}}(k,l)$ is positive definite.

In order to determine λ^*, we first rewrite the constraint equation in (3.6) as

$$\boldsymbol{d}^H(k,l,\theta_s)\boldsymbol{w}(k,l) = 1. \tag{3.17}$$

Then, we impose this constraint by substituting (3.16) into (3.17), which leads to

$$\boldsymbol{d}^H(k,l,\theta_s)\boldsymbol{w}_{\text{opt}}(k,l) = -\frac{1}{2}\lambda^*\boldsymbol{d}^H(k,l,\theta_s)\boldsymbol{C}_{\boldsymbol{v}}^{-1}(k,l)\boldsymbol{d}(k,l,\theta_s) = 1. \tag{3.18}$$

Finally, solving for λ^* in (3.18), yields

$$\lambda^* = \frac{-2}{\boldsymbol{d}^H(k,l,\theta_s)\boldsymbol{C}_{\boldsymbol{v}}^{-1}(k,l)\boldsymbol{d}(k,l,\theta_s)}, \tag{3.19}$$

where the existence of $\left(\boldsymbol{d}^H(k,l,\theta_s)\boldsymbol{C}_{\boldsymbol{v}}^{-1}(k,l)\boldsymbol{d}(k,l,\theta_s)\right)^{-1}$ is guaranteed by the fact that $\boldsymbol{C}_{\boldsymbol{v}}(k,l)$ is assumed to be positive definite. Substituting (3.19) into (3.16), we arrive at the optimum beamforming coefficients, i.e.,

$$\boldsymbol{w}_{\text{opt}}(k,l,\theta_s) = \frac{\boldsymbol{C}_{\boldsymbol{v}}^{-1}(k,l)\boldsymbol{d}(k,l,\theta_s)}{\boldsymbol{d}^H(k,l,\theta_s)\boldsymbol{C}_{\boldsymbol{v}}^{-1}(k,l)\boldsymbol{d}(k,l,\theta_s)}, \tag{3.20}$$

denoted $\boldsymbol{w}_{\text{MVDR}}(k,l,\theta_s)$ in (3.7). ■

Theorem 3.1 gives a closed form expression for the MVDR beamformer weights $\boldsymbol{w}_{\text{MVDR}}(k,l,\theta_s)$ as a function of the RTF vector $\boldsymbol{d}(k,l,\theta_s)$ and the inverse noise CPSD matrix $\boldsymbol{C}_{\boldsymbol{v}}^{-1}(k,l)$. Hence, in order to compute $\boldsymbol{w}_{\text{MVDR}}(k,l,\theta_s)$ for a given time-frequency tile, access to $\boldsymbol{d}(k,l,\theta_s)$ and $\boldsymbol{C}_{\boldsymbol{v}}^{-1}(k,l)$ at that same time-frequency tile is required. In practice, the RTF vector is typically not known in advance and neither is the noise CPSD matrix $\boldsymbol{C}_{\boldsymbol{v}}(k,l) = E\left[\tilde{\boldsymbol{v}}(k,l)\tilde{\boldsymbol{v}}^H(k,l)\right]$ in expected sense, and thus, these have to be estimated from the microphone signals. The next chapter covers estimation of the RTF vector, while we briefly discuss the estimation of the noise CPSD matrix in the following.

Using a voice activity detector (VAD) to identify time-frequency tiles of noise-only, realizations of the noise process $\tilde{\boldsymbol{v}}(k,l)$ can be observed in isolation and used to estimate the noise CPSD matrix [40]. Specifically, for the l'th time frame and the k'th frequency bin, the noise CPSD matrix can be estimated as a moving average over L time frames during speech absence [40], i.e.,

$$\hat{\boldsymbol{C}}_{\boldsymbol{v}}(k,l) = \frac{1}{L}\sum_{j=l-L+1}^{l}\boldsymbol{v}(k,j)\boldsymbol{v}^H(k,j). \tag{3.21}$$

Choosing the optimum number of time frames L to estimate the CPSD over is non-trivial. Generally, the more frames used for estimating the noise CPSD matrix, the

lower the variance of the estimator is. However, there is a trade-off when considering acoustic environments, which is due to how non-stationary the noise sources are. The change in spatial position will affect the estimate of the noise CPSD matrix, and thus, it will influence the performance of MVDR beamformer. Hence, when choosing the number of frames over which the moving average is taken, we have to make a choice between better noise reduction and being reactive to spatial changes in the acoustic environment [6, p. 18]. In Chapter 5, we expand upon the estimation of $\boldsymbol{C_v}(k, l)$ in more detail.

4. Dictionary-Based Maximum Likelihood DOA Estimation

As we saw in the previous chapter, the implementation of the MVDR beamformer requires the knowledge of the RTF from the target speaker to the HA microphones. Optimal noise reduction can therefore only be achieved if the MVDR beamformer is provided the true target RTF vector, or equivalently, the true direction of the target sound source [6, p. 25]. However, as mentioned, the RTF-vectors are typically not known in advance, and thus, have to be estimated. Plenty of RTF estimation methods exists in the literature, and among these are methods that treat the RTF estimation problem as a DOA estimation problem. Specifically, the idea behind the DOA-based RTF estimation methods is to obtain an estimate of the RTF vector by mapping the estimated DOA into a RTF vector from a predefined dictionary of RTF vectors [6]. In practice, these RTF dictionaries are often constructed in a simple and straightforward manner where each element of the RTF dictionary is associated with one particular candidate target direction.

Today, a state-of-the-art approach for model-based DOA estimation used in the context of beamforming for HA applications, relies on the maximum likelihood principle. Specifically, the method we will use in this thesis is a dictionary-based maximum likelihood DOA estimation method, which is described in e.g., [11], [24], [41]. Closed-form expressions for the employed maximum likelihood estimates were derived in [11] for a similar signal model to the one considered in this thesis, although in a non-acoustic context. However, equivalent expressions for the maximum likelihood estimates are derived in e.g., [12], [13], [41] and used in e.g., [24], in an acoustic context, in particular, in the study of algorithms for HA applications. The following theory is primarily based on [11]–[13], [24], [41], and we therefore also refer the reader to these sources for a more in-depth theory behind the method. In this chapter, we present the method from a practical standpoint, where we, based on some model assumptions, derive the likelihood function, but simply present the closed-form maximum likelihood estimation solutions with a reference to the original sources. Our objective in this chapter is to ensure understanding of the method in a sufficient way such that we are able to understand its implementation and use the method for comparison in the following chapter. Furthermore, the likelihood function of the noisy microphone signals will play a central role in the Bayesian framework that will be

studied in details in Chapter 6 in the context of our proposed beamforming methods.

This chapter is organized by first recalling the signal model for the noisy microphone observations as well as the statistical assumptions made on the signal model. Following this, we introduce some additional key assumptions that will be made in order to employ the dictionary-based maximum likelihood DOA estimation method, and lastly, we introduce the theory behind and results of the method.

4.1 Signal Model and Assumptions

Recall the time-frequency domain signal model for the noisy observations $\tilde{\boldsymbol{x}}(k,l) \in \mathbb{C}^M$ given as

$$\tilde{\boldsymbol{x}}(k,l) = \tilde{s}(k,l)\boldsymbol{d}(k,l,\theta_s) + \tilde{\boldsymbol{v}}(k,l), \tag{4.1}$$

where $\tilde{s}(k,l)$ is the scalar STFT coefficient of the target signal measured at a pre-selected reference microphone which impinges on the microphone array from direction θ_s, $\boldsymbol{d}(k,l,\theta_s)$ is the target RTF-vector, and $\tilde{\boldsymbol{v}}(k,l)$ is the overall additive noise component which is assumed to be uncorrelated with the target signal. As we saw in Chapter 2, the assumption of $\tilde{s}(k,l)$ and $\tilde{\boldsymbol{v}}(k,l)$ to be mutually uncorrelated random processes implies that that CPSD matrix of $\tilde{\boldsymbol{x}}(k,l)$ can be modeled as

$$\begin{aligned} \boldsymbol{C}_{\boldsymbol{x}}(k,l) &= \boldsymbol{C}_s(k,l) + \boldsymbol{C}_{\boldsymbol{v}}(k,l) \\ &= \lambda_s(k,l)\boldsymbol{d}(k,l,\theta_s)\boldsymbol{d}^H(k,l,\theta_s) + \lambda_v(k,l)\boldsymbol{\Gamma}_{\boldsymbol{v}}(k,l_0), \quad l > l_0, \end{aligned} \tag{4.2}$$

where $\lambda_s(k,l)$ and $\lambda_v(k,l)$ are PSDs of the target and noise at the reference microphone, respectively, and the matrix $\boldsymbol{\Gamma}_{\boldsymbol{v}}(k,l_0)$ is the normalized noise CPSD matrix at the most recent time frame index l_0 where speech was absent [34]. The target RTF vector $\boldsymbol{d}(k,l,\theta_s)$ and the time-varying PSDs $\lambda_s(k,l)$ and $\lambda_v(k,l)$ are all unknown, whereas the matrix $\boldsymbol{\Gamma}_{\boldsymbol{v}}(k,l_0)$ may be estimated in speech absent regions, identified by the use of an ideal VAD algorithm, and is therefore assumed known [13].

To consider the maximum likelihood method to estimate the target DOA θ_s, it is assumed that the target sound source can arrive from one out of I pre-selected source directions θ_i, for $i = 1,\dots,I$, such that each possible source direction can be represented by an associated RTF vector $\boldsymbol{d}(k,l,\theta_i)$, for $i = 1,\dots,I$. To represent the discrete set of possible target source directions, we use assume a predefined dictionary of RTF vectors $\mathcal{D} = \{\boldsymbol{d}(k,l,\theta_1),\dots,\boldsymbol{d}(k,l,\theta_I)\}$ to be available such that each dictionary element, i.e., each RTF vector $\boldsymbol{d}(k,l,\theta_i)$, for $i = 1,\dots,I$, is associated with one particular candidate target direction. In Fig. 4.1, a graphical example of such dictionary is depicted, where candidate target directions are confined to a circle around the user, and where each green cross on the circle represents an RTF vector in the dictionary. With this notation of an RTF dictionary, the signal model for the noisy microphone signals in (4.1) may be written as

$$\tilde{\boldsymbol{x}}(k,l) = \tilde{s}(k,l)\boldsymbol{d}(k,l,\theta_i) + \tilde{\boldsymbol{v}}(k,l), \quad i = 1\dots,I, \tag{4.3}$$

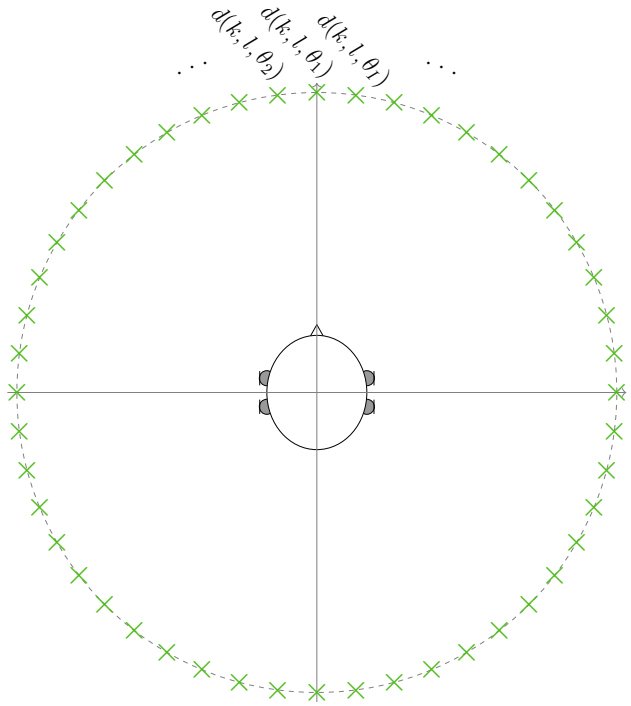


Figure 4.1: Potential target sound source locations and their associated relative acoustic transfer functions (RTFs) $\boldsymbol{d}(k,l,\theta_i)$ for a particular frequency index k and time index l.

where $\boldsymbol{d}(k,l,\theta_i)$ is one particular RTF vector from the dictionary $\mathcal{D}$. Similarly, the model for the noisy CPSD matrix in (4.2) may be written as

$$\boldsymbol{C}_{\boldsymbol{x}}(k,l,\theta_i) = \lambda_s(k,l)\boldsymbol{d}(k,l,\theta_i)\boldsymbol{d}^H(k,l,\theta_i) + \lambda_v(k,l)\boldsymbol{\Gamma}_{\boldsymbol{v}}(k,l_0), \quad l > l_0, \tag{4.4}$$

where we have now indicated that the noisy CPSD matrix explicitly depends on the target DOA θ_i, for $i = 1,\ldots,I$.

To employ the maximum likelihood method to estimate the target DOA θ_s, the probability density function (PDF) of the noisy microphone signals is required. To this end, we use the signal model for the noisy microphone observations in (4.3), and, unless otherwise is stated, the assumptions made on the signal model in Chapter 2 are assumed to be the same in the proceeding. Additionally, we make the assumption that the noisy microphone signals $\tilde{\boldsymbol{x}}(n)$ are realizations of zero-mean Gaussian random processes, hence, their complex STFT coefficients $\tilde{\boldsymbol{x}}(k,l)$ are circularly-symmetric complex Gaussian distributed with CPSD matrix given in (4.2), i.e., [24]

$$\tilde{\boldsymbol{x}}(k,l) \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{x}}(k,l,\theta_i)). \tag{4.5}$$

Furthermore, in the following, $\tilde{\boldsymbol{x}}(k,l)$ are assumed to be independent across time l and frequency k [13].

4.2 Maximum Likelihood DOA Estimation

Based on the Gaussian assumption, the likelihood function of the noisy observations $\tilde{\boldsymbol{x}}(k,l)$, which is obtained by considering the complex Gaussian PDF of $\tilde{\boldsymbol{x}}(k,l)$ as a function of the target DOA parameter, is given by

$$f(\tilde{\boldsymbol{x}}(k,l)|\theta_i) = \frac{1}{\pi^M \left|\boldsymbol{C}_{\boldsymbol{x}}(k,l,\theta_i)\right|} \exp\left(\tilde{\boldsymbol{x}}^H(k,l)\boldsymbol{C}_{\boldsymbol{x}}{}^{-1}(k,l,\theta_i)\tilde{\boldsymbol{x}}(k,l)\right), \tag{4.6}$$

where $\theta_i \in \Theta$, with $\Theta = \{\theta_1, \ldots, \theta_I\}$ being a discrete candidate set of directions from which the target signal can arrive, $|\cdot|$ denotes the matrix determinant, and $\boldsymbol{C}_{\boldsymbol{x}}(k,l,\theta_i)$ is given in (4.4) and is a function of the scalar PSDs $\lambda_s(k,l)$ and $\lambda_v(k,l)$ as well as the target DOA θ_i, which dependency is represented through the target RTF vector $\boldsymbol{d}(k,l,\theta_i)$. Note that from (4.6), it is seen that $\boldsymbol{C}_{\boldsymbol{x}}(k,l,\theta_i)$ is required to be invertible. In practice, this is not a problem as e.g., microphone self-noise will ensure that $\boldsymbol{\Gamma}_{\boldsymbol{v}}(k,l_0)$ and hence $\boldsymbol{C}_{\boldsymbol{x}}(k,l,\theta_i)$ has full rank [12].

To facilitate the derivation of the maximum likelihood estimate of the target DOA, we assume that $\lambda_s(k,l)$, $\lambda_v(k,l)$, and $\boldsymbol{d}(k,l,\theta_i)$ can be considered approximately constant across a certain number of L consecutive time frames of the STFT coefficients $\tilde{\boldsymbol{x}}(k,l)$. Note that this assumption is also known as the short-time stationarity assumption [13]. Following this assumption, let $\boldsymbol{X}(k,l) \in \mathbb{C}^{M\times L}$ denote a matrix with L observed vectors $\tilde{\boldsymbol{x}}(k,j)$, for $j = l-L+1, \ldots, l$, as columns such that

$$\boldsymbol{X}(k,l) = [\tilde{\boldsymbol{x}}(k,l-L+1), \ldots, \tilde{\boldsymbol{x}}(k,l)]. \tag{4.7}$$

Then, based on the assumption that the noisy microphone observations $\tilde{\boldsymbol{x}}(k,l)$ are independent across time l, it follows that the joint likelihood function of $\boldsymbol{X}(k,l)$, i.e., the joint likelihood function of successive observations, is given by the product of the likelihood functions of successive observations, i.e.,

$$f(\boldsymbol{X}(k,l)|\theta_i) = \prod_{j=l-L+1}^{l} f(\tilde{\boldsymbol{x}}(k,j)|\theta_i), \quad i = 1\ldots, I. \tag{4.8}$$

Let $\hat{\boldsymbol{R}}(k,l) \in \mathbb{C}^{M\times M}$ denotes the sample estimate of the noisy CPSD matrix which we defined as

$$\hat{\boldsymbol{R}}(k,l) = \frac{1}{L}\boldsymbol{X}(k,l)\boldsymbol{X}^H(k,l). \tag{4.9}$$

Then, under the aforementioned assumptions, it can be shown that the joint likelihood of $\boldsymbol{X}(k,l)$ can be expressed as

$$f(\boldsymbol{X}(k,l)|\theta_i) = \frac{\exp\left(-L\mathrm{tr}\left(\hat{\boldsymbol{R}}(k,l)\boldsymbol{C}_{\boldsymbol{x}}{}^{-1}(k,l,\theta_i)\right)\right)}{\pi^{LM}\left|\boldsymbol{C}_{\boldsymbol{x}}(k,l,\theta_i)\right|^L}, \tag{4.10}$$

where $\mathrm{tr}\,(\cdot)$ denotes the trace operator defined as the sum of the main diagonal elements of a square matrix. Based on the likelihood function (4.10), the maximum

likelihood DOA estimation approach first derives the maximum likelihood estimation of $\lambda_s(k,l)$ and $\lambda_v(k,l)$ conditioned on the DOA parameter θ_i, for $i = 1,\ldots,I$. Then, the estimated parameters are substituted into (4.15), in which a concentrated likelihood function is obtained, and finally, this concentrated likelihood function is maximized with respect to θ_i to obtain the maximum likelihood estimate of the target DOA.

Maximum Likelihood Estimates of Target and Noise PSDs

In [11], closed-form expressions for the maximum likelihood estimates of $\lambda_s(k,l)$ and $\lambda_v(k,l)$ are derived, and equivalent expressions are derived in e.g., [12], [13] for similar signal models to the one employed in this thesis. It can be shown that the maximum likelihood estimate of the noise PSD $\lambda_v(k,l)$, conditioned on the DOA parameter θ_i, is [13]

$$\hat{\lambda}_{v,\mathrm{ML}}(k,l,\theta_i) = \frac{1}{M-1} \operatorname{tr}\left(\boldsymbol{B}(k,l,\theta_i)^H \hat{\boldsymbol{R}}(k,l)\boldsymbol{B}(k,l,\theta_i)\left(\boldsymbol{B}^H(k,l,\theta_i)\boldsymbol{\Gamma}_{\boldsymbol{v}}(k,l_0)\boldsymbol{B}(k,l,\theta_i)\right)^{-1}\right), \tag{4.11}$$

where the dependency on θ_i in $\hat{\lambda}_{v,\mathrm{ML}}(k,l,\theta_i)$ is introduced to indicate the that the maximum likelihood estimate of the noise PSD $\lambda_v(k,l)$ depends on the choice of θ_i, and where $\boldsymbol{B}(k,l,\theta_i) \in \mathbb{C}^{M\times M-1}$ denotes a so-called blocking-matrix given as [12], [13]

$$\boldsymbol{B}(k,l,\theta_i) = \left(\boldsymbol{I}_{M\times M} - \frac{\boldsymbol{d}(k,l,\theta_i)\boldsymbol{d}^H(k,l,\theta_i)}{\boldsymbol{d}^H(k,l,\theta_i)\boldsymbol{d}(k,l,\theta_i)}\right)\boldsymbol{I}_{(M\times M-1)}, \quad i = 1,\ldots,I. \tag{4.12}$$

The interpretation of the blocking matrix is that it is used to project $\tilde{\boldsymbol{x}}(k,l)$ into the null-space of $\boldsymbol{d}(k,l,\theta_i)\boldsymbol{d}^H(k,l,\theta_i)$, i.e., we block the speech component

$$\lambda_s(k,l)\boldsymbol{d}(k,l,\theta_i)\boldsymbol{d}^H(k,l,\theta_i)$$

of $\tilde{\boldsymbol{x}}(k,l)$ [12], [13], [42].

Given the maximum likelihood estimate of the noise PSD in (4.11), it can be shown [11], [13] that the maximum likelihood estimate of the target PSD, conditioned on θ_i, is

$$\hat{\lambda}_{s,\mathrm{ML}}(k,l,\theta_i) = \boldsymbol{w}_{\mathrm{MVDR}}^H(k,l,\theta_i)\left(\hat{\boldsymbol{R}}(k,l) - \hat{\lambda}_{v,\mathrm{ML}}(k,l,\theta_i)\boldsymbol{\Gamma}_{\boldsymbol{v}}(k,l_0)\right)\boldsymbol{w}_{\mathrm{MVDR}}(k,l,\theta_i), \tag{4.13}$$

where $\boldsymbol{w}_{\mathrm{MVDR}}(k,l,\theta_i) \in \mathbb{C}^M$, for $i = 1,\ldots,I$, are MVDR beamformers steered towards each direction θ_i defined in the discrete set Θ of possible target DOAs, i.e., [12], [13]

$$\boldsymbol{w}_{\mathrm{MVDR}}(k,l,\theta_i) = \frac{\boldsymbol{\Gamma}_{\boldsymbol{v}}^{-1}(k,l_0)\boldsymbol{d}(k,l,\theta_i)}{\boldsymbol{d}^H(k,l,\theta_i)\boldsymbol{\Gamma}_{\boldsymbol{v}}^{-1}(k,l_0)\boldsymbol{d}(k,l,\theta_i)}, \quad i = 1,\ldots,I. \tag{4.14}$$

Concentrated Log-Likelihood Function

Inserting the maximum likelihood estimates of $\lambda_s(k,l)$ and $\lambda_v(k,l)$ into the likelihood function in (4.10), an expression for the concentrated likelihood function $f(\boldsymbol{X}(k,l)|\theta_i, \hat{\lambda}_{s,\mathrm{ML}}(k,l,\theta_i), \hat{\lambda}_{v,\mathrm{ML}}(k,l,\theta_i))$, which we may denote $\bar{f}(\boldsymbol{X}(k,l)|\theta_i)$, is obtained, i.e.,

$$\bar{f}(\boldsymbol{X}(k,l)|\theta_i) = \frac{\exp\left(-L\mathrm{tr}\left(\hat{\boldsymbol{R}}(k,l)\hat{\boldsymbol{C}}_{\boldsymbol{x}}^{-1}(k,l,\theta_i)\right)\right)}{\pi^{LM}\left|\hat{\boldsymbol{C}}_{\boldsymbol{x}}(k,l,\theta_i)\right|^{L}}, \tag{4.15}$$

where we have defined

$$\hat{\boldsymbol{C}}_{\boldsymbol{x}}(k,l,\theta_i) = \hat{\lambda}_{s,\mathrm{ML}}(k,l,\theta_i)\boldsymbol{d}(k,l,\theta_i)\boldsymbol{d}^H(k,l,\theta_i) + \hat{\lambda}_{v,\mathrm{ML}}(k,l,\theta_i)\boldsymbol{\Gamma}_{\boldsymbol{v}}(k,l_0). \tag{4.16}$$

For practical convenience, often the logarithm of the likelihood function is maximized instead of the likelihood function. First of all, taking the logarithm of the likelihood function simplifies the subsequent mathematical analysis, and secondly, and even more importantly for our present purpose, it also provides numerical stability because the product of a large number of small probabilities can easily make the computations numerical unstable, which is resolved by instead computing the sum of the log-probabilities [43, p.26]. As the logarithm is a monotonically increasing function of its argument, maximization of the log of a function is equivalent to maximization of the function itself. Taking the natural logarithm of (4.15), and using the fact that [11]

$$\mathrm{tr}\left(\hat{\boldsymbol{R}}(k,l)\hat{\boldsymbol{C}}_{\boldsymbol{x}}^{-1}(k,l,\theta_i)\right) = M, \tag{4.17}$$

the concentrated log-likelihood function can be written as

$$\begin{aligned}\log\big(\bar{f}(\boldsymbol{X}(k,l)|\theta_i)\big) = &- LM\log(\pi)\\ &- L\log\Big(\Big|\hat{\boldsymbol{C}}_{\boldsymbol{x}}(k,l,\theta_i)\Big|\Big)\\ &- LM,\end{aligned} \tag{4.18}$$

where $\log(\cdot)$ is taken to mean the natural logarithm. As we want to maximize the log-likelihood function with respect to the DOA parameter θ_i, we can ignore terms in (4.18) not involving this parameter. By doing so, we arrive at the reduced expression of the concentrated log-likelihood function given as [34]

$$\log\big(\bar{f}(\boldsymbol{X}(k,l)|\theta_i)\big) = -\log\Big(\Big|\hat{\boldsymbol{C}}_{\boldsymbol{x}}(k,l,\theta_i)\Big|\Big). \tag{4.19}$$

Wideband Maximum Likelihood Estimate of the Target DOA

From the concentrated log-likelihood function in (4.19), we are finally able to obtain the maximum likelihood estimate of the target direction θ_s for a given time-frequency tile. However, as an acoustic target sound sources plausibly occupy multiple frequency

bins at once, due to the wide-band nature of speech, we may utilize this information to obtain a more robust estimate of the target DOA [42, Sec. III-D]. We do so by utilizing the assumption of $\tilde{\boldsymbol{x}}(k,l)$ to be independent across frequency k, which allows us to jointly estimate the target DOA across frequencies by summation of the concentrated log-likelihood functions for all k. In other words, we choose to maximize the wideband concentrated log-likelihood function with respect to θ_i [24], [41], i.e.,

$$\hat{\theta}_{s,\mathrm{ML}} = \arg\max_{\theta_i \in \Theta} \sum_{k=1}^{K} \log\left(\bar{f}(\boldsymbol{X}(k,l)|\theta_i)\right), \tag{4.20}$$

where K is the total number of frequency bins of the one-sided spectrum. As θ_i belongs to a relatively small discrete set of directions, the wideband maximum likelihood estimate of θ_s is obtained through an exhaustive search over θ_i. Given the maximum likelihood estimate of θ_s, an estimate of the target RTF vector is obtained by mapping $\hat{\theta}_{s,\mathrm{ML}}$ into an associated target RTF from a predefined dictionary of target RTF vectors [41]. The dictionary-based maximum likelihood DOA estimation method is summarized in Algorithm 1 as pseudo-code.

Algorithm 1 Dictionary-Based Maximum Likelihood RTF Estimation

Input:

$\hat{\boldsymbol{R}}(k,l) \in \mathbb{C}^{M\times M},\ \forall\ k$: Sample estimate of noisy CPSD matrix.
$\boldsymbol{\Gamma}_v(k,l_0) \in \mathbb{C}^{M\times M},\ \forall\ k$: Normalized noise CPSD matrix.
$\boldsymbol{\Gamma}_v^{-1}(k,l_0) \in \mathbb{C}^{M\times M},\ \forall\ k$: Inverse normalized noise CPSD matrix.
$\mathcal{D} = \{\boldsymbol{d}(k,l,\theta_1),\ldots,\boldsymbol{d}(k,l,\theta_I)\} \in \mathbb{C}^{M\times K\times I}$: Dictionary of RTF vectors.

Output:

$\hat{\boldsymbol{d}}(k,l,\hat{\theta}_{s,\mathrm{ML}}) \in \mathbb{C}^{M},\ \forall\ k$: Maximum likelihood estimate of RTF vector.

for $k = 0$ **to** $K-1$ **do**
  for $i = 1$ **to** I **do**
    Obtain $\boldsymbol{d}(k,l,\theta_i)$ from $\mathcal{D}$.
    $\boldsymbol{B}(k,l,\theta_i) = \left(\boldsymbol{I}_{M\times M} - \dfrac{\boldsymbol{d}(k,l,\theta_i)\boldsymbol{d}^H(k,l,\theta_i)}{\boldsymbol{d}^H(k,l,\theta_i)\boldsymbol{d}(k,l,\theta_i)}\right)\boldsymbol{I}_{(M\times M-1)}$.
    $\hat{\lambda}_{v,\mathrm{ML}}(k,l,\theta_i) = \dfrac{1}{M-1}\mathrm{tr}\left(\boldsymbol{B}(k,l,\theta_i)^H\hat{\boldsymbol{R}}(k,l)\boldsymbol{B}(k,l,\theta_i)\left(\boldsymbol{B}^H(k,l,\theta_i)\boldsymbol{\Gamma}_{\boldsymbol{v}}(k,l_0)\boldsymbol{B}(k,l,\theta_i)\right)^{-1}\right)$.
    $\boldsymbol{C}_{\boldsymbol{v},\mathrm{ML}}(k,l,\theta_i) = \hat{\lambda}_{v,\mathrm{ML}}(k,l,\theta_i)\boldsymbol{\Gamma}_{\boldsymbol{v}}(k,l_0)$.
    $\boldsymbol{w}_{\mathrm{MVDR}}(k,l,\theta_i) = \dfrac{\boldsymbol{\Gamma}_{\boldsymbol{v}}^{-1}(k,l_0)\boldsymbol{d}(k,l,\theta_i)}{\boldsymbol{d}^H(k,l,\theta_i)\boldsymbol{\Gamma}_{\boldsymbol{v}}^{-1}(k,l_0)\boldsymbol{d}(k,l,\theta_i)}$.
    $\hat{\lambda}_{s,\mathrm{ML}}(k,l,\theta_i) = \boldsymbol{w}_{\mathrm{MVDR}}^H(k,l,\theta_i)\left(\hat{\boldsymbol{R}}(k,l) - \boldsymbol{C}_{v,\mathrm{ML}}(k,l,\theta_i)\right)\boldsymbol{w}_{\mathrm{MVDR}}(k,l,\theta_i)$.
    $\boldsymbol{C}_{\boldsymbol{x},\mathrm{ML}}(k,l,\theta_i) = \hat{\lambda}_{s,\mathrm{ML}}(k,l,\theta_i)\boldsymbol{d}(k,l,\theta_i)\boldsymbol{d}^H(k,l,\theta_i) + \boldsymbol{C}_{\boldsymbol{v},\mathrm{ML}}(k,l,\theta_i)$.
    $\log\left(\bar{f}(\boldsymbol{X}(k,l)|\theta_i)\right) = -\log\left(|\boldsymbol{C}_{\boldsymbol{x},\mathrm{ML}}(k,l,\theta_i)|\right)$.
  end for
end for
$\hat{\theta}_{s,\mathrm{ML}} = \arg\max_{\theta_i\in\Theta}\sum_{k=0}^{K-1}\log\left(\bar{f}(\boldsymbol{X}(k,l)|\theta_i)\right)$,
Obtain RTF vector from $\mathcal{D}$ corresponding to $\hat{\theta}_{s,\mathrm{ML}}$ for all k.

5. Feasibility Test and Upper Bound Performance of Eye-Gaze Steered Beamformers

The first task of this thesis has been to examine the potential performance benefits of using eye-gaze steered beamformers. In order to access whether eye-gaze steered beamforming is worth the extra cost of having additional sensors, e.g., electrodes or cameras, in future hearing aids to measure the user's eye-gaze, a feasibility study has been carried out. The purpose of this study is to establish fundamental knowledge on the potential performance benefits of using eye-gaze steered beamforming. To determine an upper performance bound of an eye-gaze steered beamforming system, the feasibility study is performed under ideal conditions where the user's eye-gaze is assumed precisely pointing towards the desired speaker. The conclusions are drawn based on empirical experiments. Specifically, to examine the potential performance benefits of using eye-gaze steered beamformers, an evaluation and comparison are carried out between an oracle eye-gaze steered MVDR beamformer and a fixed MVDR beamformer, that is always steered towards the front of the user, as well as an MVDR beamformer steered using the microphone-only DOA estimator based on maximum likelihood estimation described in the previous chapter.

The chapter begins with a presentation of the implementation of the overall simulation framework that has been constructed in order to perform the simulation experiments related to the feasibility study. This includes a description of how the acoustic environments, in which the beamformers are tested, are created as well as a description of the practical implementation of the beamforming algorithms. Afterwards, the performance measures used to evaluate the beamformers are presented, and finally, the performance of the beamforming methods are evaluated and compared through simulation experiments.

5.1 Implementation

In order to apply the MVDR beamformers, we need access to the noisy microphone signals $\boldsymbol{x}(n)$ as they are picked up by the microphones on the HA devices and the

MVDR beamformer weights $\boldsymbol{w}_{\text{MVDR}}(k,l)$. In the following sections, we describe how these are obtained through simulations.

5.1.1 Simulating HA Microphone Signals $\boldsymbol{x}(n)$

In order to simulate noisy microphone signals, as they are picked up by the microphones on the HA devices with $M \geq 2$ microphones, we have made a general framework which can be used to simulate from the signal model presented in (2.1), i.e.,

$$\boldsymbol{x}(n) = (s * \boldsymbol{a}(\bullet, \theta_s))(n) + \boldsymbol{v}(n), \tag{5.1}$$

where $s(n)$ is the target signal measured at the source location, $\boldsymbol{a}(n, \theta_s)$ is a vector containing the AIRs from the target speaker to each of the M microphones on the HA, and $\boldsymbol{v}(n)$ is an additive noise term. Hence, in order to simulate from the signal model in (5.1), we need access to the clean target signal $s(n)$, the AIRs $\boldsymbol{a}(n, \theta_s)$, and some noise signals $\boldsymbol{v}(n)$. In the following, we describe how these signals, as well as the parameters that will be used to simulate the acoustic scenes, are obtained, and in addition, present the general framework for using the signal model in (5.1) to simulate the sound received at the HA microphones in the studied acoustic scenes.

Acoustic Impulse Response and Sound Databases: The AIRs used to simulate the wave propagation from a sound source to the microphones on the HAs are obtained from a database provided by Oticon. The AIRs are obtained from empirical measurements with HAs placed on a real human head [33]. The measurement setup consists of a spherical loudspeaker array with a HA user, seated in the center of the array, wearing a BTE HA on each ear. We refer the reader to Fig. 1.1 for an example of a BTE HA. Each BTE HA has three microphones where two are placed in a front/rear configuration on the HA and the third is placed in the ear canal. For the purpose of the simulation experiments carried out in this chapter, different subsets of the microphones on the HAs are used in both a monaural and in binaural HA configurations. For the binaural HA configurations, we assume wireless, simultaneous, and error-free signal exchange between the left and the right HA. The AIRs only incorporates the influence of the head, torso and pinnae, as well as the position of the sound source relative to the HA user, while the small amount of reverberation present in the original AIRs has been removed by truncating the AIRs. The AIRs are sampled at a sampling frequency of 44.1 kHz and were measured 360° around the head with an azimuth resolution of 7.5°, with 0° defined as the frontal direction from the HA user's point of view, and the azimuth is counterclockwise rotating. The AIRs were also measured at different elevations, but in this thesis, we will focus on simulating the target and noise sources only from the azimuth angle, as, in many real acoustic scenarios, the target may primarily be located at about the same elevation as the HA user. Hence, when simulating acoustic scenes, the sound waves can arrive from a discrete set of $\frac{360^\circ}{7.5^\circ} = 48$ possible locations which are placed uniformly around

a circle, with a radius of 1.5 meters, in a horizontal plane approximately at the height of the subjects ears [14], [33], [44]

Clean speech signals used for the target are speech obtained from the TIMIT Corpus [45]. The TIMIT Corpus contains a total of 6300 recorded sentences, 10 sentences spoken by each of 630 speakers from eight major dialect regions of the United States. For the purpose of this thesis, for each speaker, the recordings are concatenated into a single recording. The noise types used in the simulations are synthetic babble noise as well as speech shaped noise. The babble noise is created using speech signals from the TIMIT corpus. The speech shaped noise (SSN) is obtained from [46] where the SSN sequence is constructed by filtering a 50 minute Gaussian white noise sequence through a 12th-order all-pole filter with coefficients found from linear predictive coding analysis of 100 randomly chosen sentences from a Danish speech corpus. The clean speech signals from TIMIT and noise signals from [46] are sampled at 16 kHz.

Preprocessing: The first part of the simulation of acoustic scenes constitutes a preprocessing stage, where the audio signals used to generate the acoustic scene are resampled such that the sampling frequency of the audio signals are in agreement with the sampling frequency of the AIRs. As the clean speech signals and noise signals are sampled at 16 kHz, while the AIRs are sampled at 44.1 kHz, the clean speech signals and noise signals are upsampled to 44.1 kHz. The resampling is done using a polyphase *up/down* method [47] where the signal is upsampled by a factor P, a zero-phase low-pass finite impulse response filter is applied, and then the signal is downsampled by a factor Q. The specific values of P and Q are chosen based on the sampling frequency of the considered audio signals. The duration of an acoustic scene is chosen to be 5 seconds, and for each scene realization, a new target speech signal as well as noise signals are randomly chosen and kept fixed during the acoustic scene.

Specification of Acoustic Scene: In the simulation of the acoustic scenes, we choose to fix the noise fields to be approximately isotropic, as this may be the case in many realistic acoustic scenes, such as interfering speakers at a cocktail party [6, p. 43]. We construct the approximately isotropic babble or SSN noise fields by modeling noise received at each microphone as a superposition of mutually temporal uncorrelated speech signals or SSN sequences impinging uniformly from all of the 48 possible directions.

For the possible target directions, we use a subset of the AIR database, namely $\theta_s \in \Theta_s = \{-90°, -75°, \ldots, 75°, 90°\}$, corresponding to the frontal-horizontal plane. Hence, the target is located in the frontal half-plane. We choose this range of target directions as in many realistic situations, the target speaker is usually located at the front of the user.

To control the SNR in the acoustic scenes, we scale the target speech signal by a gain factor g. Specifically, we define the input SNR as the ratio between the average

target power and the average noise power. The target speech power and noise power are computed prior to convolution between the signals and the AIRs. Hence, we have decided to define the SNR using signal and noise powers computed at the source locations. In this way, the SNR is not a function of direction and therefore, the input SNR measured at the reference microphone, which is placed on the left ear of the HA user, is biased, due to the head-shadow effect. To be more precise, when the target signal impinges on the microphone array from the user's right hand side, the SNR measured at the reference microphone is reduced due to head shadow effects, but in turn, the SNR measured on the reference microphone is almost unaltered by the head and the torso of the user, when the target signal impinges on the microphone array from the user's left hand side. Hence, if SNR was measured at the reference microphone, the SNR would be higher for target signals arriving from the left, than for target signals arriving from the right side.

The average power of a target speech signal, $s(n)$, for $n = 0, \ldots, N_s - 1$, is computed as

$$\hat{\varsigma}_s^2 = \frac{1}{N_{s,\text{active}}} \sum_{n=0}^{N_s-1} (s(n))^2, \tag{5.2}$$

where $N_{s,\text{active}}$ is the number of samples in the target speech signal where speech is present. When we compute the average power of the target speech signal, we remove speech absent segments in the computation by applying a VAD. We do this since the different speech signals from the TIMIT corpus might contain different amount of silent segments. The procedure for using the VAD in the computation of the average power is summarized in Algorithm 2 as pseudo-code.

In order to obtain the average power of the noise signal $v(n) = \sum_{q=1}^{Q} v_q(n)$, where $v_q(n)$ is the q'th noise signal measured the q'th noise source, the average power of each of the Q noise signals $v_q(n)$ is computed as

$$\hat{\varsigma}_{q,v}^2 = \frac{1}{N_v} \sum_{n=0}^{N_v-1} (v_q(n))^2, \quad q = 1, ..., Q,$$

where N_v is the number of samples $v_q(n)$. Next, since the noise sources are assumed uncorrelated, the average power of the noise signal $v(n)$ in the acoustic scene is determined as the sum of the average noise power of the Q noise signals, i.e.,

$$\hat{\varsigma}_v^2 = \sum_{q=1}^{Q} \hat{\varsigma}_{q,v}^2.$$

When the noise type is babble, an identical VAD approach, as the one in Algorithm 2, is used for $v(n)$ in order to identify rare events of simultaneous speech absence from all Q noise sources.

The SNR will be expressed in decibel (dB), thus, the SNR is given by [48, p. 229]

$$\text{SNR} = 10 \log_{10} \left(\frac{\hat{\varsigma}_s^2}{\hat{\varsigma}_v^2} \right) \text{dB}. \tag{5.3}$$

Algorithm 2 VAD used in computation of average power of target speech signal and babble noise

Input:

$s(n)$, for $n = 0, \ldots, N_s - 1$: Target speech signal.

Output:

$N_{s,\text{active}}$: Number of samples in the target signal with speech presences.

1: Set threshold value, $\delta_{\text{th}} = 40$ dB.
2: Set $N_{s,\text{active}} = 0$.
3: Segment target signal into non-overlapping frames of $N = 256$ samples as $s^l(n) = s(n + lN)$.
4: **for all** l **do**
5: Compute frame energy $E(l) = 10 \log_{10} \left(\sum_{n=0}^{N-1} (s^l(n))^2 + \epsilon \right)$.
6: **end for**
7: Determine maximum frame energy as $E_{\max} = \max_l (E(l))$.
8: **for all** l **do**
9: **if** $E_{\max} - E(l) \leq \delta_{\text{th}}$ **then**
10: Add N to $N_{s,\text{active}}$.
11: **end if**
12: **end for**

In order to control the SNR in an acoustic scene, we then scale the target speech signal with a gain factor g, such that the SNR is determined as

$$\text{SNR} = 10 \log_{10} \left(\frac{\frac{1}{N_{s,\text{active}}} \sum_{n=0}^{N_s-1} (g s(n))^2}{\sum_{q=1}^{Q} \frac{1}{N_v} \sum_{n=0}^{N_v-1} (v_q(n))^2} \right) = 10 \log_{10} \left(\frac{g^2 \hat{\varsigma}_s^2}{\hat{\varsigma}_v^2} \right) \text{dB}. \tag{5.4}$$

Isolating g in (5.4), yields

$$g = \left(\frac{\hat{\varsigma}_s^2}{\hat{\varsigma}_v^2} \cdot 10^{-\frac{\text{SNR}}{10}} \right)^{-\frac{1}{2}}. \tag{5.5}$$

Multiplying the target speech signal $s(n)$ with g ensures that the SNR in the acoustic scene has the desired value.

Filtering and Generation of Acoustic Scene: After preprocessing and specification of the acoustic scene, the structure of a spatial setup, where the audio signals can arrive from 48 equidistant angles, is generated. This simulation of the acoustic path for each 48 angles is done by convolving the audio signals with an associated AIR. Afterwards, the resulting signals are downsampled to 16 kHz. The noisy microphone signals are then simulated by adding the convolved target signal with the convolved noise signals such that the resulting noisy microphone signals are in accordance to (5.1) with the target signal scaled by g, i.e.,

$$\boldsymbol{x}(n) = ((g s(\bullet)) * \boldsymbol{a}(\bullet, \theta_s))(n) + \boldsymbol{v}(n), \tag{5.6}$$

with

$$\boldsymbol{v}(n) = \sum_{q=1}^{Q} (v_q * \boldsymbol{a}(\bullet, \theta_{v_q}))(n), \tag{5.7}$$

where θ_{v_q} is the direction associated with the q'th noise signal.

In order to generate a number of different acoustic scenes, the code is executed by choosing a specific acoustic scene according to target direction, noise type, and input SNR and by specifying the number of microphones M as well as the microphone array geometry used to construct $\boldsymbol{x}(n)$. For each scene realization, the target direction, noise type, and the input SNR are chosen and kept fixed during the acoustic scene.

5.1.2 Implementation of MVDR Beamforming System

As we saw in Chapter 3, some parameters need to be estimated in order to implement the MVDR beamformer efficiently, as these are unknown in practice. Specifically, this involves knowledge of the RTF vector of the target speech signal as well as the noise CPSD matrix. In this section, we will describe how we obtain these beamformer parameters in our implementation of the MVDR beamformer. Before presenting the beamformer parameter estimation, the necessary specifications related to the transformation of the time-domain noisy microphone signals into the their corresponding time-frequency domain representations are briefly covered.

Analysis and Synthesis

Whenever a beamformer is implemented in this thesis, the signal processing is performed in the time-frequency domain. To obtain the time-frequency domain representation of the noisy microphone signals, we use the STFT, as mentioned in Chapter 2. In this thesis, we apply the STFT with a window length of $N = 256$ samples, which with a sampling frequency of 16 kHz corresponds to 16 ms. We choose this window length since a commonly made assumption in speech processing applications is that speech, which has a time varying spectrum, can be considered wide sense stationary in time intervals around $20 - 30$ ms [48, pp. 866-867]. With a sampling frequency of 16 kHz, 20 ms will correspond to a window length of 320 samples, and so, for convenience, the window length is rounded to 256 samples. For the window, we use the square root Hanning window which is constructed by taking the square root of the Hanning window. The Hanning window is defined by [48, p. 858]

$$w(n) = \begin{cases} \sqrt{\frac{\left(1-\cos\left(\frac{2\pi n}{N}\right)\right)}{2}}, & 0 \leq n \leq N, \\ 0, & \text{otherwise}, \end{cases} \qquad n \in \mathbb{N}_0. \tag{5.8}$$

We choose this window, as it is widely used in the context of beamforming when the goal is to enhance a target speech signal by noise reduction [6, p. 16], [49, p. 50]. The square root Hanning window is implemented with a hop size of $D = 128$, which

corresponds to an overlap of 50%. We choose this hop size as it guarantees that the window satisfies the so-called overlap-add property [30, p. 232]

$$\sum_{l=-\infty}^{\infty} w(n - lD) = 1. \tag{5.9}$$

The overlap-add property is desirable, as it ensures perfect reconstruction of the signal as long as an equivalent square root Hanning window is used for synthesis [30, p. 232] and as long as $D \leq N \leq K$, where K is the number of frequency bins [48, pp. 856-858]. For convenience, we let $N = K$. Our choice of settings for the STFT are summarized in Table 5.1.

Window	Square root Hanning
Window length	$N = 256$
Number of frequency bins	$K = 256$
Hop size	$D = 128$
Overlap	50%

Table 5.1: Settings of the STFT.

After application of the beamformer, the estimated target speech signal, which is obtained as the output of the beamformer, is transformed back into the time domain using the ISTFT [48, pp. 850-581], with the same settings as for the STFT.

Beamformer Parameter Estimation

After transforming the noisy microphone signal used in a given simulation using the STFT, the beamforming algorithms are applied to each frequency subband. Due to the fact that $x_m(n)$ is real-valued, the spectrum is symmetric, and therefore, only the first $K = N/2 + 1$ frequency bins of $\tilde{x}_m(k, l)$, corresponding to the number of frequency bins of the one-sided spectrum, are processed. Application of the MVDR beamformer results in an estimate of the target speech signal received at the reference microphone given as

$$\hat{\tilde{s}}(k, l) = \hat{\boldsymbol{w}}_{\text{MVDR}}^{H}(k, l, \hat{\theta}_s)\tilde{\boldsymbol{x}}(k, l), \tag{5.10}$$

where $\hat{\boldsymbol{w}}_{\text{MVDR}}^{H}$ is an estimate of the MVDR beamformer weights, which for each time-frequency tile is implemented as

$$\hat{\boldsymbol{w}}_{\text{MVDR}}(k, l, \hat{\theta}_s) = \frac{\hat{\boldsymbol{C}}_{\boldsymbol{v}}^{-1}(k, l)\hat{\boldsymbol{d}}(k, l, \hat{\theta}_s)}{\hat{\boldsymbol{d}}^{H}(k, l, \hat{\theta}_s)\hat{\boldsymbol{C}}_{\boldsymbol{v}}^{-1}(k, l)\hat{\boldsymbol{d}}(k, l, \hat{\theta}_s)}, \tag{5.11}$$

where $\hat{\boldsymbol{C}}_{\boldsymbol{v}}(k, l)$ and $\hat{\boldsymbol{d}}(k, l, \hat{\theta}_s)$ are the estimated noise CPSD matrix and RTF vector, respectively. In the following, we describe how these estimates are obtained and used to implement the MVDR beamformers. The overall procedure for implementing the MVDR beamforming system is summarized in Algorithm 3 as pseudo-code.

Noise CPSD matrix estimation, $\hat{\boldsymbol{C}}_{\boldsymbol{v}}(k,l)$: In this thesis, we estimate the noise CPSD matrix by using noise dominant time-frequency tiles to update the noise CPSD matrix and use the resulting estimate during speech presence. In practice, detecting noise dominant time-frequency tiles requires e.g., the use of a VAD. For the implementation of the MVDR beamformer, we have decided to let the first second of each acoustic scene realization consisting of noise-only samples such that we can use these samples to obtain an estimate of the noise CPSD matrix using

$$\hat{\boldsymbol{C}}_{\boldsymbol{v}}(k,l_0) = \frac{1}{L}\sum_{j=l_0-L+1}^{l_0} \tilde{\boldsymbol{v}}(k,j)\tilde{\boldsymbol{v}}^H(k,j), \quad k=0,\ldots,K-1, \tag{5.12}$$

over $L = 125$ time frames corresponding to the first second of an acoustic scene realization, where l_0 denotes the last time frame index in the first second with speech absence. By obtaining the noise CPSD matrix this way, the underlying assumption is that the structure of the noise CPSD matrix found during speech absence remains identical during target speech presence.

Implementation of RTF vector estimation, $\hat{\boldsymbol{d}}(k,l,\hat{\theta}_s)$: In order to implement the closed form expression for the MVDR beamformer for a given time-frequency tile as given in (5.11), access to the target RTF vector at this time-frequency tile is required. The target RTF vectors used to implement the beamformers are obtained from a predefined RTF dictionary where each of the possible target RTF vectors are associated with a corresponding target DOA. In the following, we first describe how the RTF dictionary is constructed, and afterwards, we present the implementation of the baseline maximum likelihood estimation of the RTF vectors which consists of a maximum likelihood DOA estimation followed by a look-up of the RTF in the predefined RTF dictionary.

A. Construction of RTF dictionary: The RTF dictionary used in the determination of the beamformer coefficients is constructed such that $\mathcal{D} = \{\boldsymbol{d}(k,l,\theta_1),\ldots,\boldsymbol{d}(k,l,\theta_I)\}$ is an ordered tuple where the elements $\boldsymbol{d}(k,l,\theta_i)$, for $i = 1,\ldots,I$ are RTF vectors associated with the sound sources impinging from direction $\theta_i = (i-1)7.5^\circ$ in the horizontal plane where $\theta_1 = 0^\circ$ is the frontal direction with respect to the HA user. This RTF dictionary is graphically depicted in Fig. 5.1, where each RTF vector in the constructed dictionary $\mathcal{D}$ for a particular frequency index k and time index l is represented by the green crosses on the circle around the HA user.

The predefined database of RTFs are made by first downsampling the AIRs in the AIR database described in Section 5.1.1 from 44.1 kHz to 16 kHz using the polyphase up/down method [47]. In this way, the sampling frequency of the AIRs are in agreement with the desired sampling frequency of the simulated received microphone signals.

In order to transform the AIRs to ATFs, we apply the discrete Fourier transform (DFT) [48, p. 654], which allows us to use a window length equal to the length of the

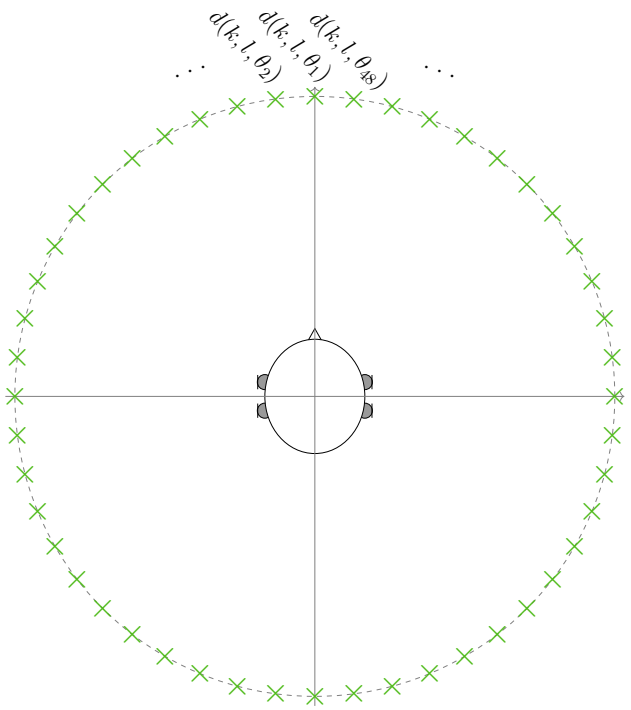


Figure 5.1: Potential target sound source locations and their associated relative acoustic transfer functions (RTFs) $\boldsymbol{d}(k, l, \theta_i)$ for a particular frequency index k and time index l.

downsampled AIRs and thereby prevent any additional information loss. However, the window length of the DFT should not only be in accordance to the length of the downsampled AIRs, but also to the window length of the STFT applied to the noisy microphone signals, i.e., $N = 256$. The downsampled AIRs have a length of 372 samples, of which the last 223 samples are observed to be zeros. Hence, we can simply truncate the AIRs to 256 samples without any information loss, as samples with a value of zero do not contribute to the DFT. Thus, the ATFs are obtained by applying the DFT on the truncated AIRs. After obtaining the ATFs, we obtain the RTF for the m-th microphone by using

$$d_m(k, \theta) = \frac{1}{\tilde{a}_1(k, \theta)} \tilde{a}_m(k, \theta), \quad m = 1, \ldots, M, \tag{5.13}$$

where the frontal left microphone with reference microphone index 1 is chosen as the reference microphone. Due to the fact that we have applied the DFT, the RTF vectors only depend on the frequency bins and the estimated target DOA and not the time frame. However, the RTF vectors are still implicitly dependent on the time frame, as we for each time frame estimate a target DOA which is not guaranteed to be constant over all time frames. When the estimate of the target DOA $\hat{\theta}_s$ is mapped to an RTF vector, we will explicitly denote the dependence on the time frame of the RTF vector by $\hat{\boldsymbol{d}}(k, l, \hat{\theta}_s)$.

B. Implementation of dictionary-based maximum likelihood DOA estimation: In the following, we describe how we have implemented the maximum likelihood based DOA estimation scheme employed in this thesis. The implementation of this method is made by translating an existing MATLAB function which return a wide-band maximum likelihood estimate of the target RTF vector given a dictionary of RTF vectors for K frequency bins, into a Python function. Hence, in order to use the maximum likelihood method as a competing DOA estimation method used to steer an MVDR beamformer, it is of crucial importance to ensure that the translation of the MATLAB function is correct. In order to ensure that our Python implementation is a correct translation of the MATLAB function, we have evaluated and compared the two implementations in different typically occurring acoustic scenes that a HA user may experience. The comparisons were done by providing the two functions with the same input variables and then see if identical results for both implementations are obtained. Specifically, for each time frame, we investigated whether the implementations provide identical estimates of the element from the RTF dictionary. From the results of the comparison, it was found that the two implementations provide identical results in the different studied acoustic scenes, and therefore, we conclude that our Python implementation can be considered a valid implementation of the dictionary based maximum likelihood DOA estimation method. An example of the comparison is shown in Appendix B.1.

5.2 Performance Measures

To quantify the performance of the beamformers, the beamforming performance will be reported in terms of extended short-time objective intelligibility (ESTOI) [50] and segmental SNR (segSNR) [51, p. 9], which yield an estimate of predicted speech intelligibility and quality, respectively. Specifically, ESTOI and segSNR can be used to evaluate changes in predicted speech intelligibility and quality as a result of applying a beamformer by comparing the estimated time-domain target speech signal $\hat{s}(n)$ from the beamformer with the clean target speech signal $s(n)$ received at the reference microphone [50], [51, p. 9].

Extended Short-time Objective Intelligibility

ESTOI is a quantity which predicts speech intelligibility. The details of ESTOI will not be covered in this thesis, but can be found in [50]. The lower and upper bound in the predicted speech intelligibility score is -1 and 1, respectively, where a higher value reflects higher speech intelligibility.

Segmental Signal-to-Noise Ratio

SegSNR is a simple objective measure to evaluate speech enhancement algorithms. The segSNR measure takes both noise reduction and speech distortion into account

[51, p. 9]. In this thesis, segSNR will be used to measure the degree of noise reduction, as a measure for speech quality. However, it should be mentioned that, although segSNR is widely used in the context of evaluation of speech enhancement algorithms, it has been shown to correlate poorly with speech quality [51, sec. 2.2].

The segSNR and is computed by averaging frame level SNR estimates as [52, p. 480], [51, p. 9]

$$\text{segSNR} = \frac{1}{N_y} \sum_{i=1}^{N_y} 10 \log_{10} \left(\frac{\sum_{n=Ni}^{Ni+N-1} y(n)^2}{\sum_{n=Ni}^{Ni+L-1} [(y(n) - \hat{y}(n))^2]} \right) \text{ dB}, \tag{5.14}$$

where $y(n)$ is the unprocessed time-domain signal, $\hat{y}(n)$ is the enhanced time-domain signal, N is the frame length, i.e., the number of samples in each frame, and N_y is the number of frames in the signal. When computing the segSNR scores, we exclude silent frames from the sum in (5.14) in order to avoid large negative segSNR values which will bias the overall measure [51, p. 9].

In this tesis, we use the implementation from [53] for segSNR and the one from [50] for ESTOI.

5.3 Beamformer Evaluation

In this section, we examine an upper bound performance of eye-gaze steered beamformers. To this end, the experiments are performed under ideal conditions where the user's eye-gaze is assumed precisely pointing towards the desired speaker. To examine the potential performance benefits of using eye-gaze steered beamformers, an evaluation and comparison are carried out between the following beamforming methods:

- **MVDR-Eye-Gaze**: The beamforming system MVDR-Eye-Gaze is used as an eye-gaze controlled reference system to indicate the upper bound performance of the eye-gaze steered MVDR beamformer if the eye-gaze of the HA user is precisely pointing towards the desired speaker. This is relevant to examine, since, if no significant gain is obtained under ideal conditions, then eye-gaze steered beamforming may not we worthwhile for future HAs. For this method, $\hat{\boldsymbol{d}}(k, l, \hat{\theta}_s)$ is therefore the RTF vector associated with the true target direction, i.e., $\hat{\theta}_s = \theta_s$, and is used to implement the MVDR beamformer.

- **MVDR-Fixed**: In the context of HA systems, the target speaker is often assumed to be frontal with respect to the HA user [8]. The beamforming system MVDR-Fixed is used to emulate such system. For this method, $\hat{\boldsymbol{d}}(k, l, \hat{\theta}_s)$ is the RTF vector associated with the frontal direction, i.e., $\hat{\theta}_s = 0°$ and is used to implement the MVDR beamformer.

- **MVDR-ML:** The beamforming system MVDR-ML is used as an example of an MVDR beamformer steered using a microphone-only DOA estimator. For this method, the target RTF vector $\hat{\boldsymbol{d}}(k,l,\hat{\theta}_s)$ is estimated using the directional-based maximum likelihood DOA estimation method in Algorithm 1, and is used to implement the MVDR beamformer. I.e., for this method $\hat{\boldsymbol{d}}(k,l,\hat{\theta}_s)$ is the target RTF vector associated with the dictionary maximum likelihood DOA estimate of θ_s, i.e., $\hat{\theta}_s = \theta_{s,\mathrm{ML}}$.

5.3.1 Experimental Setup

The performance of the beamforming systems is evaluated on noisy microphone signals which are simulated in accordance to the description in Section 5.1. The settings for generating the acoustic scenes are summarized in Table 5.2. As presented in the

Noise Field	Noise Type	SNR	Target Direction
Isotropic	Babble, SSN	$-10, 0, 10$ dB	$\theta_s \in \Theta_s = \{-90°, -75°, \ldots, 75°, 90°\}$

Table 5.2: Settings for parameters in simulation of acoustic environments used for evaluation of the beamformer performance.

table, we are simulating noisy acoustic microphone signals for varying input SNRs, specifically -10 dB, 0 dB, and 10 dB. Examining the influence the input SNR has on the performance of the beamforming systems is motivated by the fact that the state-of-the-art dictionary-based maximum likelihood DOA estimation method is known to perform poorly in acoustic scenes where the SNR is low. In acoustic scenes where the SNR is low, we therefore expect the MVDR-Eye-Gaze to perform much better than the MVDR-ML, while we on the other hand expect the MVDR-ML to perform almost on par with the MVDR-Eye-Gaze at a high SNR, given that a good estimate of $\boldsymbol{\Gamma}_{\boldsymbol{v}}(k,l_0)$ can be obtained. Furthermore, we evaluate the performance for three different microphone array configurations; a 2-microphone configuration, a 4-microphone configuration, and a 6-microphone configuration. Exploring the influence of the number of microphones used in a HA system on the performance of the MVDR beamformers, is motivated by the fact that the beamwidth of a 2-microphone frontal MVDR beamformer may allow for considerable larger deviations of the target location from the assumed frontal DOA than might be the case for $M = 4$ and $M = 6$ microphone beamformers. For $M = 2$ microphones, we therefore expect the MVDR-Fixed to preform on par with the performance of the MVDR-Eye-Gaze.

The settings for analysis and synthesis, for the implementation of the MVDR beamformers, and for the maximum likelihood estimation of the RTF vectors are all specified in accordance with the descriptions in Section 5.1.2.

5.3.2 Evaluation and Simulation Results

In this section, we compare the performance of the MVDR-Eye-Gaze, MVDR-Fixed and MVDR-ML in terms of average ESTOI and segSNR scores, in order to examine the potential performance improvements of the oracle eye-gaze steered beamformer MVDR-Eye-Gaze in different conditions where the baseline methods potentially may fail. Furthermore, to get an indication of an upper bound on the noise reduction performance of the MVDR beamformer, an oracle MVDR beamformer, denoted MVDR-Ideal is implemented. This beamformer is implemented using the true target DOA as well as an estimate of the noise CPSD matrix which for each time-frequency tile is updated using the most recent 125 samples of the isolated noise signal $\tilde{\boldsymbol{v}}(k,l)$. The ESTOI and segSNR scores for the MVDR-Ideal as well as for the unprocessed noisy microphone signal at the reference microphone, denoted "Noisy", are included as references for comparison.

The ESTOI and segSNR scores will be reported as functions of the direction of the target speaker relative to the user, i.e., as a function of target DOA. The performance as a function of DOA will be reported for the three different input SNRs and for the two different noise types specified in Table 5.2.

For each simulated acoustic scene, ESTOI and segSNR scores have been computed for the unprocessed noisy signal at the reference microphone as well as for the estimated target speech signal obtained by the MVDR-Eye-Gaze, MVDR-Fixed, MVDR-ML, and the MVDR-Ideal. After obtaining ESTOI and segSNR scores associated to each acoustic scene, we determine the average ESTOI and segSNR scores with respect to DOA by averaging the scores over 10 scene realizations with same DOA. Furthermore, in order to visualize the improvement in the ESTOI and segSNR scores after application of the beamformers, we subtract the ESTOI and segSNR scores of the unprocessed noisy signal from the ESTOI and segSNR scores of the MVDR-Eye-Gaze, MVDR-Fixed, MVDR-ML, and the MVDR-Ideal.

In the following, for the sake of readability, we only show results using acoustic scenes with babble noise. The simulation experiments with the noise type SSN lead to similar conclusions, although for SSN, the performance difference between the MVDR-ML and the MVDR-Eye-Gaze is much less significant. The better performance of the MVDR-ML in SSN compared to babble noise may be explained by the fact that the maximum likelihood DOA estimation method used in MVDR-ML to obtain an estimate of the RTF vector search for a point sound source for the target speaker. The babble noise is generated as a superposition of interfering speech signals which share similar signal characteristics to the target speech signal, and although, the babble noise is composed of 48 equiangular located point sources, these are not necessarily always active at the same time. For that reason, it may be the case that, at some point in time, one of the interfering speakers may be more prominent in the noisy signal. This may result in the maximum likelihood DOA estimation method used in MVDR-ML erroneously estimating the direction of an interfering speaker as the target direction and instead suppress the true target speaker, which would decrease the performance of the MVDR-ML. Therefore, it is expected that the

MVDR-ML may struggle to determining the direction of the target speaker when the simulated noise type is babble and when the SNR is low, but that the performance of the MVDR-ML increases in acoustic scenes with SSN where the noise field is more diffuse. The results for SNN can be found in Appendix B.

Estimated Speech Intelligibility With Respect to DOA

In Fig. 5.2, the average ESTOI scores (left figures) and the improvement of ESTOI score relative to the unprocessed noisy signal at the reference microphone (right figures) are plotted as functions of target direction. The figure shows the results from three experiments with different microphone configurations. In the first experiment, we evaluate the performance for a 2-microphone configuration, using only the front and rear microphone, respectively, on the left HA (Fig. 5.2a). In the second experiment, we evaluate the peformance for a 4-microphone configuration, using the front and rear microphones on both the left and right HA (Fig. 5.2b). Finally, in the third experiment, we evaluate the performance for a 6-microphone configuration, using all three microphones on both HAs (Fig. 5.2c). In these experiments, the input SNR is fixed to -10 dB. For each experiment, we sweep over the different target directions in the discrete DOA range, $\theta_s \in \Theta_s$, and then, for each DOA, we compute the average ESTOI score as well as the average improvement in ESTOI score. Fig. 5.3 and Fig. 5.4 show similar experiments, but where we evaluate the performance under different input SNRs. Specifically, in Fig. 5.3, the input SNR is fixed to 0 dB while the input SNR is fixed to 10 dB in the experiments for which the results are illustrated in Fig. 5.4.

From Figs. 5.2 to 5.4, it is seen that, in general, the average ESTOI scores improve as the SNR increases, as expected. Furthermore, for the individual SNR levels, we observe that the ESTOI scores obtained for all the beamforming methods, expect for the MVDR-Fixed, improve as the number of microphone increases. For the MVDR-Fixed this is only the case when $\theta_s = 0°$, as expected, since this beamforming method assumes a frontal target direction. Also as expected, the MVDR-Eye-Gaze performs on par with the the upper bound performance of the MVDR beamformer (MVDR-Ideal) in all of the considered conditions. However, it is seen that the performance difference between the MVDR-Ideal and MVDR-Eye-Gaze becomes more significant as the number of microphones increases. This may be explained by the fact that the variance of the noise CPSD matrix estimate used to implement the beamformers increases when M, and thereby the dimension of $\boldsymbol{C}_{\boldsymbol{v}}(k, l_0)$, increases while the amount of data used to calculate the estimates remains unchanged. Furthermore, from the results for $M = 2$ microphones (Figs. 5.2a, 5.3a, and 5.4a), we see a general tendency that the beamforming methods obtain higher scores for positive target DOAs, i.e., for targets arriving from the left of the HA user, than for negative target DOAs, i.e, for targets arriving from the right. This may be explained by the fact that the SNR is defined at the source locations and not at the reference microphone, and so, the input SNR measured at the reference microphone is biased due to the head-shadow, as mentioned in Section 5.1.1. Hence, this behaviour is in line with our expectations.

In relation to the influence of the number of microphones on the breamforming performance, we consider again the results for $M = 2$ microphones. Comparing the scores obtained for MVDR-Eye-Gaze and MVDR-Fixed, we see that the difference in ESTOI scores obtained from these two methods are very small when $\theta_s \in \{-60°, \cdots, 60°\}$, and in fact, they obtain almost identical performance scores when $\theta_s \in \{-30°, \cdots, 30°\}$. Following this, based on the results obtained using $M = 4$ microphones (Figs. 5.2b, 5.3b, and 5.4b) and $M = 6$ microphones (Figs. 5.2c, 5.3c, and 5.4c), we see that the difference in performance scores of the MVDR-Eye-Gaze and the MVDR-Fixed is much more significant when $\theta_s \neq 0°$, as we obtain a loss in performance scores for the MVDR-Fixed as we move away from the assumed frontal target direction. These observations can be explained by the fact that for 2-microphone MVDR beamformers, the width of the beam is wider, and so, the MVDR-Fixed is insensitive to mismatches between the true target direction and its assumed frontal target direction, i.e, these mismatches do not deteriorate the MVDR-Fixed performance. On the other hand, when using $M = 4$ and $M = 6$ microphones, these mismatches deteriorate the MVDR-Fixed performance significantly, whereas the MVDR-Eye-Gaze performs well for all target directions. These results suggest that the 2-microphone array configuration of the MVDR-Eye-Gaze appear to be superior to MVDR-Fixed only for relatively large deviations of the assumed frontal DOA of the MVDR-Fixed (approximatly $\pm 60°$), which is an important result for the practical use of the eye-gaze steered beamformer in HAs. Specifically, the range of DOAs where the $2-$microphone MVDR-Eye-Gaze offers any advantage over the simpler 2-microphone MVDR-Fixed, may be too far from the angle range in which the eye-gaze lies in realistic communicational situations to justify the computational and implementational cost of having additional sensors in future HAs to measure the user's eye-gaze. In summary, the use of $M > 2$ microphone array configurations of the MVDR beamformer seems necessary in order to achieve the full potential of using eye-gaze steered beamformers in HAs.

By comparing Figs. 5.2 to 5.4 in terms of SNR, it is seen that the MVDR-ML has a very degraded performance at an SNR of -10 dB compared to all the other beamforming methods. In fact, for targets arriving from the sides, we see that the application of the MVDR-ML deteriorates the performance, as we obtain ESTOI scores that are lower than that of the unprocessed noisy signal in those angle ranges. However, as the SNR increases to 10 dB, the performance of the MVDR-ML approaches the performance of the MVDR-Eye-Gaze and MVDR-Ideal. This result is in line with results reported in e.g., [24], and is expected sine the maximum likelihood estimates of the target DOAs, or equivalently, of the RTFs, are known to perform well when used in a beamforming context when the SNR is sufficiently high [12], [34].

Estimated Speech Quality With Respect to DOA

In Figs. 5.5 to 5.7, the segSNR scores as well as the improvement in segSNR score relative to the unprocessed noisy signal as functions of DOA, are illustrated. In general, the segSNR scores are observed to be very small. One important aspect to

emphasize in this context is that we measure the performance with segSNR but this is not what we are optimizing for with the MVDR beamformers. In other words, minimizing the objective function of the MVDR beamformers, does not necessarily translate directly to optimum segSNR performance [6, p. 77].

As expected, segSNR for the noisy signal is always maximal for a target direction of ≈ 90 degrees, i.e., at the left ear where the reference microphone is placed. Interestingly, when $M = 2$ (Figs. 5.5a, 5.6a, and 5.7a), the MVDR-Eye-Gaze performs better for target directions $\theta_s \approx 30^\circ$ and not at $\theta_s \approx 90$ degrees, where the input segSNR is largest. This optimum angle may be understood as a trade-off between the frontal direction, where the 2-microphone MVDR beamformer is most efficient and 90 degrees, where the input SNR is largest. [13] Finally, as for the ESTOI results, performance is relatively lower for target angles around $\theta_s \in \{-90^\circ, \cdots, -15^\circ\}$, because the SNR at the reference microphone is reduced due to the head shadow effects. [13]

From Figs. 5.5 to 5.7, it is observed that, generally, MVDR-Eye-Gaze performs better than the baseline methods MVDR-ML and MVDR-Fixed in terms of segSNR. However, from Figs. 5.5a and 5.6a, it appears that, when using $M = 2$ microphones and the SNR is low (-10 dB and 0 dB), MVDR-Fixed is able to outperform the upper boound performance of the MVDR beamformer (MVDR-Ideal) as well as the MVDR-Eye-Gaze at all target directions, except when $\theta_s \in \{-30^\circ, \cdots, 15^\circ\}$, where they perform on par. The exact reason remains unknown, but a possible explanation may be as follows. If we consider the situation where the true target direction is $\theta_s = 60^\circ$, then MVDR-Fixed, which points towards $\theta_s = 0^\circ$, is able to obtain a higher noise suppression at the expense of target distortion, compared to MVDR-Ideal and MVDR-Eye-Gaze, which are both steered towards the true target direction ($\theta_s = 60^\circ$) to maintain the distortionless constraint in the target direction. In addition, 2-microphone MVDR beamformers are known to be more efficient for targets located parallel to the microphone axis, i.e., $\theta_s \approx 0^\circ$ rather then perpendicular to the microphone axis, i.e., $\theta_s \approx \pm 90^\circ$[13], and so, the 2-microphone configuration of MVDR-Fixed has an advantage over the correpsonding 2-microphone configurations of MVDR-Ideal and MVDR-Eye-Gaze when targets arrive from the sids of the HA user. Although, the above explanations rely on theory for 2-microphone MVDR beamformers, it might be that it is the same mechanism that applies for $M = 4$ and $M = 6$ microphones for low SNRs, where we see a similar behaviour (Figs. 5.5b, 5.5c, and 5.6b), and for the MVDR-ML which is also observed to be able to obtain higher segSNR scores than the MVDR-Ideal and MVDR-Eye-Gaze in some DOA ranges (Figs. 5.5, 5.6a, and 5.6b).

5.4 Summary

In summary, the results obtained from the evaluation carried out in this chapter, suggest that, under ideal conditions where the user's eye-gaze is assumed pointing precisely towards the target speaker, the MVDR-Eye-Gaze has an advantage over

the state-of-the-art MVDR-Fixed in situations where $M > 2$ microphone is used and when the target is placed away from frontal direction, i.e., when $\theta_s \neq 0°$, and that it has an advantage over the state-of-the-art MVDR-ML in situations where the SNR is low. Following this, we can conclude that using $M = 2$ microphones, the performance gain of using an oracle eye-gaze steered MVDR beamformer, compared to a frontal fixed MVDR beamfomer, is marginal. This is an important result, as many HAs, but not all, are equipped with only $M = 2$ local microphones. However, using $M = 4$ and $M = 6$ microphones, we can not reject the possibility that there is something to gain by using eye-gaze steered beamforming in future HAs. Furthermore, we can conclude that at low SNR levels, there may be a potential of obtaining a performance improvement using information provided by the user's eye-gaze to help steer a beamformer for HA applications, compared to using an audio-only method to estimate the direction of the target speaker.

Having examined the upper performance bound for eye-gaze steered beamforming, we will in the rest of this thesis focus on how eye-gaze information can be incorporated in a beamforming system, with the aim of proposing a beamforming method for HA applications which potentially may outperform current audio-only methods which soley rely on the noisy microphone signals to solve the problem of enhancing the target speech signal. Specifically, we will study the use of an additional modality, apart from sound, in our specific case the user's eye-gaze, to help localizing and enhancing the target sound source. To this end, we will in the following chapter present our proposed solution to this approach, i.e., we will describe how we propose to design a signal processing algorithm for target localization using information provided by both acoustic and eye direction information, and how to use this information for estimating the underlying clean target signal by the means of a beamforming algorithm.

Algorithm 3 MVDR Beamforming System

Input:

$\boldsymbol{x}(n) = [x_1(n), \ldots, x_M(n)]^T$: Noisy microphone signal in time-domain,

$\mathcal{D} = \{\boldsymbol{d}(k, l, \theta_1), \ldots, \boldsymbol{d}(k, l, \theta_I)\} \in \mathbb{C}^{M \times K \times I}$: Dictionary of RTF vectors.

Output:

$\hat{s}(n)$: Estimated target speech signal at the reference microphone in the time-domain.

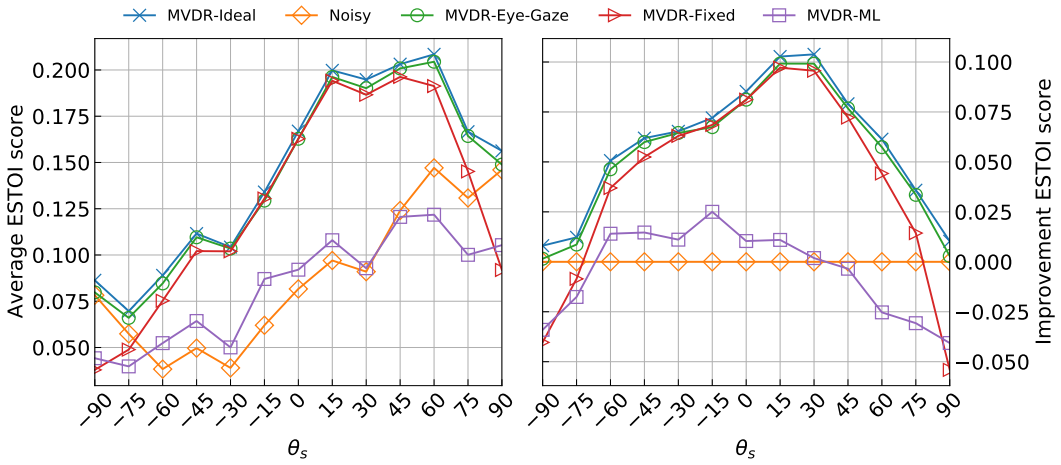
1: Apply STFT to $\boldsymbol{x}(n)$ to obtain $\tilde{\boldsymbol{x}}(k, l)$ for all k and l.
2: **for all** l **do**
3: **for** $k = 0$ **to** $K - 1$ **do**
4: Get $\hat{\boldsymbol{d}}(k, l, \hat{\theta}_s) \in \mathbb{C}^M$ from $\mathcal{D}$ determined by the beamforming method in question.
5: **end for**
6: **end for**
7: **for** $k = 0$ **to** $K - 1$ **do**
8: Estimate noise CPSD matrix from the first second of noise-only as

$$\hat{\boldsymbol{C}}_{\tilde{\boldsymbol{v}}}(k, l_0) = \frac{1}{L} \sum_{l=0}^{L-1} \tilde{\boldsymbol{v}}(k, l) \tilde{\boldsymbol{v}}^H(k, l).$$

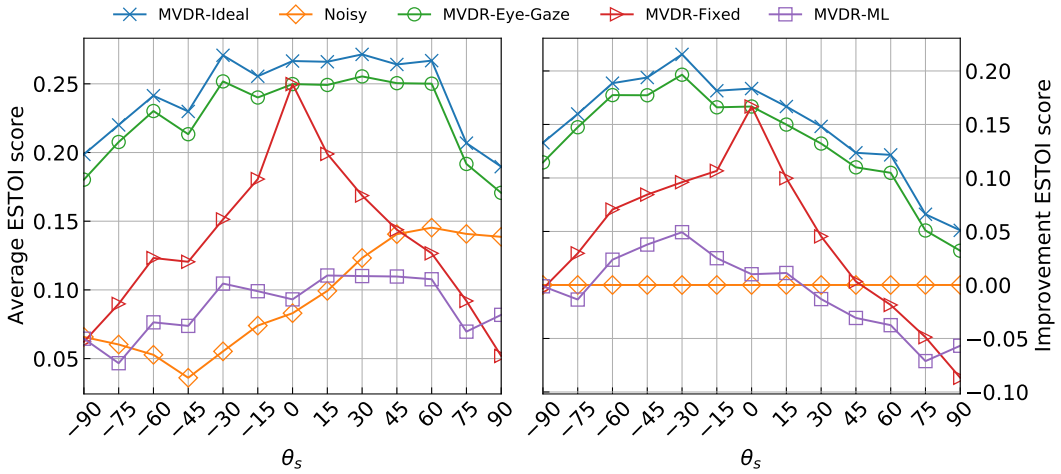
9: **end for**
10: **for all** l **do**
11: **for** $k = 0$ **to** $K - 1$ **do**
12: Compute MVDR beamformer weight vector as

$$\hat{\boldsymbol{w}}_{\text{MVDR}}(k, l) = \frac{\hat{\boldsymbol{C}}_{\boldsymbol{v}}^{-1}(k, l_0) \hat{\boldsymbol{d}}(k, l, \hat{\theta}_s)}{\hat{\boldsymbol{d}}^H(k, l, \hat{\theta}_s) \hat{\boldsymbol{C}}_{\boldsymbol{v}}^{-1}(k, l_0) \hat{\boldsymbol{d}}(k, l, \hat{\theta}_s)}.$$

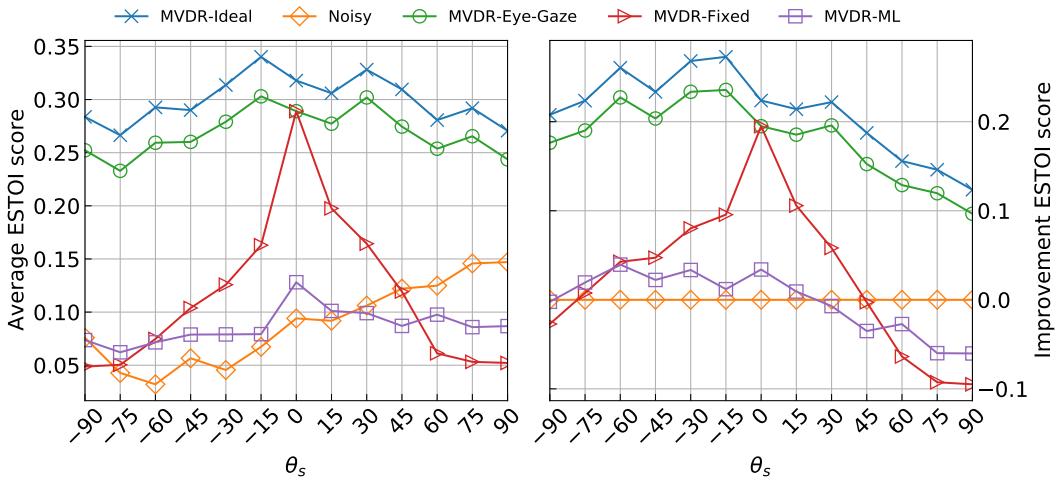
13: Apply beamformer weights as $\tilde{\boldsymbol{s}}(k, l) = \hat{\boldsymbol{w}}_{\text{MVDR}}^H(k, l) \tilde{\boldsymbol{x}}(k, l)$
14: **end for**
15: **end for**
16: Apply ISTFT to $\tilde{\boldsymbol{s}}(k, l)$ to obtain $\hat{s}(n)$ for all n.



(a) A 2-microphone monaural configuration, using the front and rear microphone on the left HA.

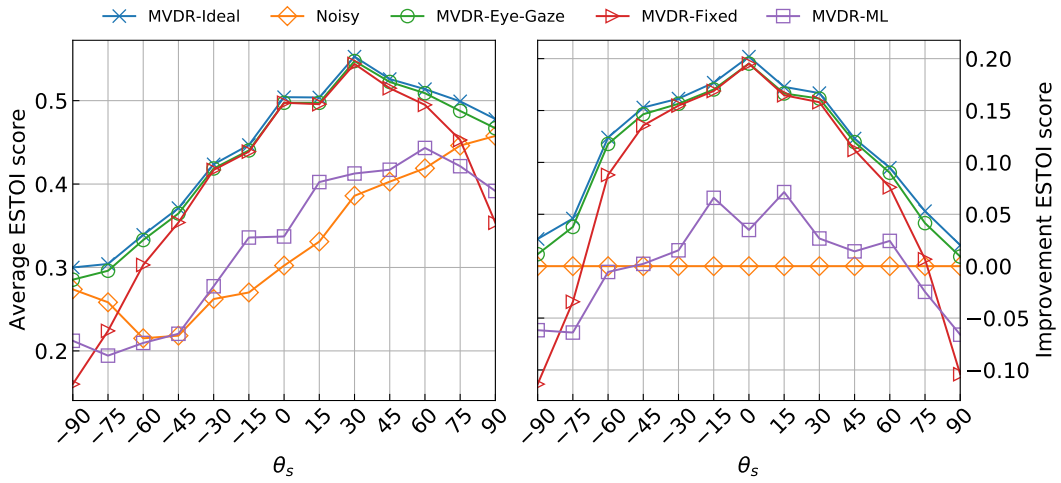


(b) A 4-microphone binaural configuration, using the front and rear microphones on both the left and right HA.

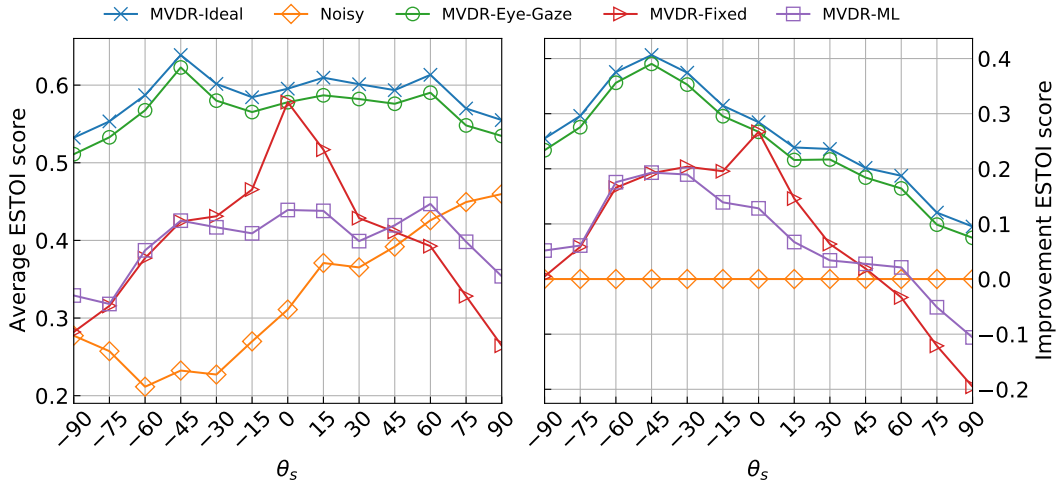


(c) A 6-microphone binaural configuration, using all three microphones on both HAs.

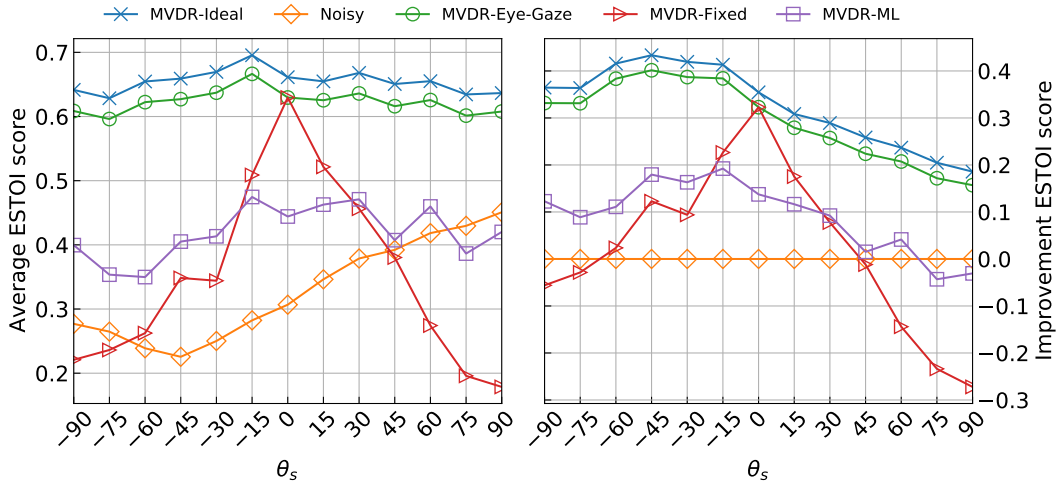
Figure 5.2: Average ESTOI scores and average improvement ESTOI scores as a function of target DOA. The noise type is babble in an approximately isotropic noise field with an input SNR of −10 dB.



(a) A 2-microphone monaural configuration, using the front and rear microphone on the left HA.

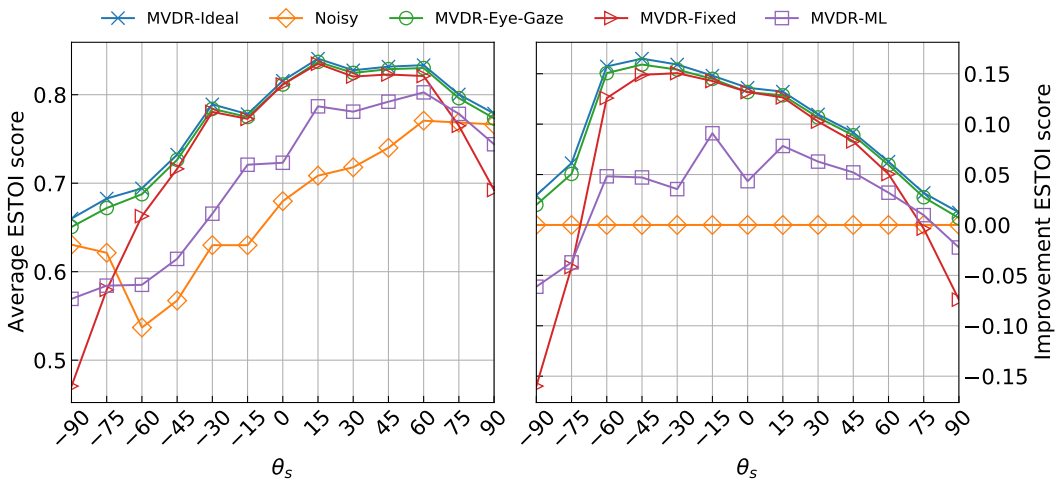


(b) A 4-microphone binaural configuration, using the front and rear microphones on both the left and right HA.

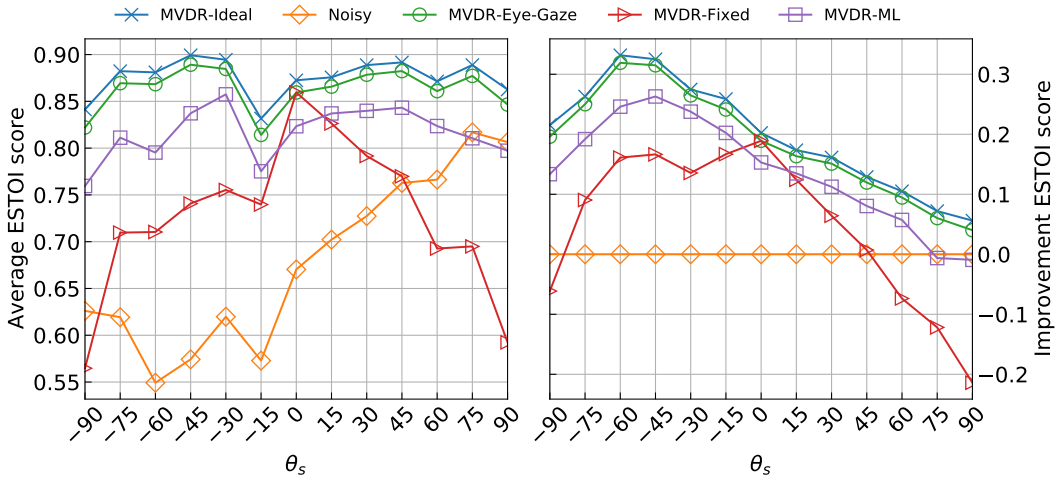


(c) A 6-microphone binaural configuration, using all three microphones on both HAs.

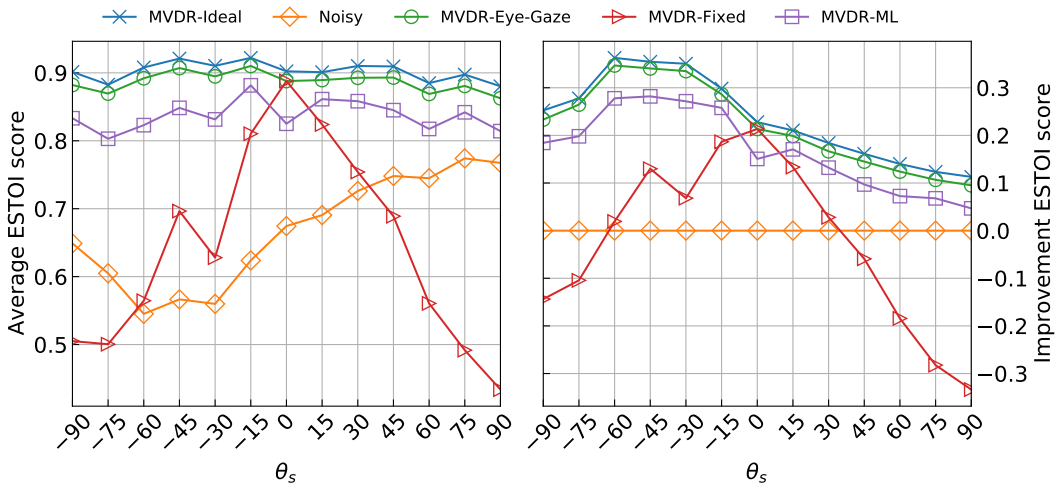
Figure 5.3: Average ESTOI scores and average improvement ESTOI scores as a function of target DOA. The noise type is babble in an approximately isotropic noise field with an input SNR of 0 dB.



(a) A 2-microphone monaural configuration, using the front and rear microphone on the left HA.

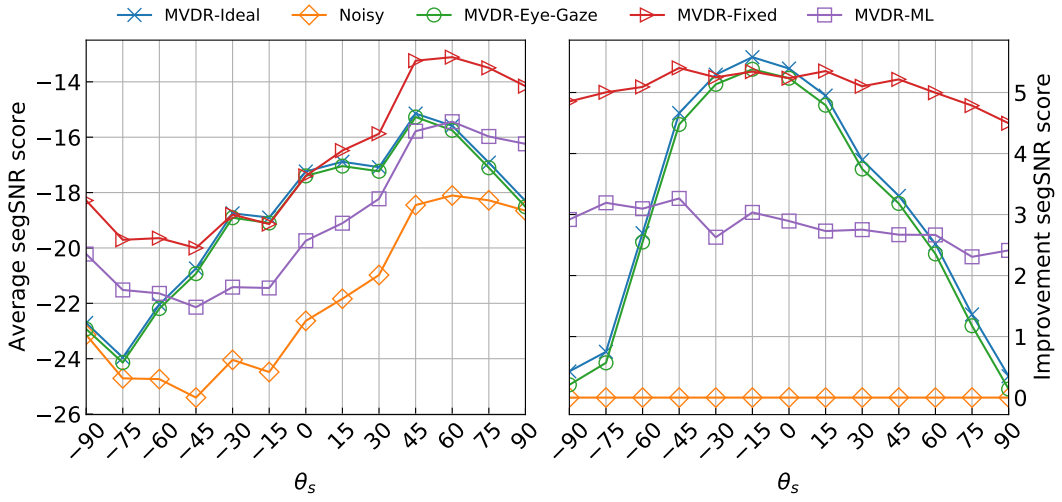


(b) A 4-microphone binaural configuration, using the front and rear microphones on both the left and right HA.

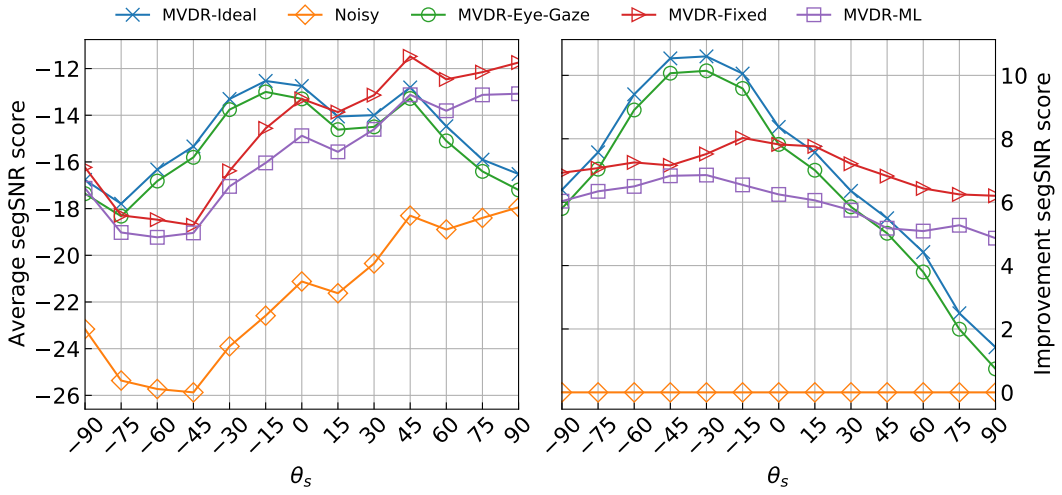


(c) A 6-microphone binaural configuration, using all three microphones on both HAs.

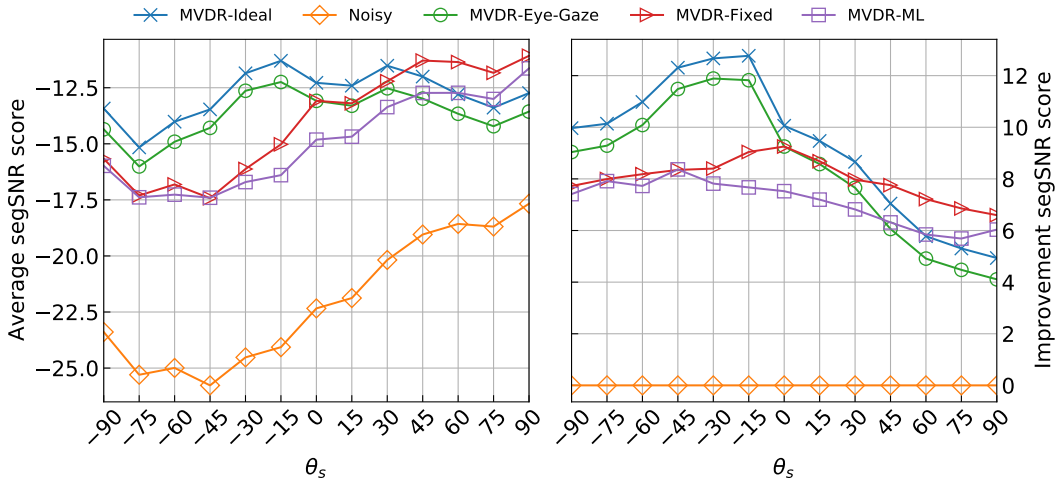
Figure 5.4: Average ESTOI scores and average improvement ESTOI scores as a function of target DOA. The noise type is babble in an approximately isotropic noise field with an input SNR of 10 dB.



(a) A 2-microphone monaural configuration, using the front and rear microphone on the left HA.

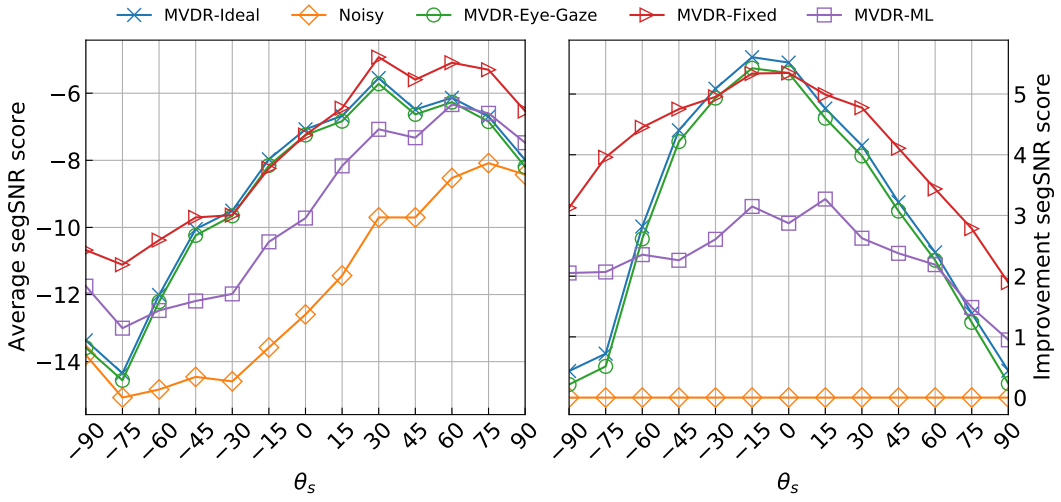


(b) A 4-microphone binaural configuration, using the front and rear microphones on both the left and right HA.

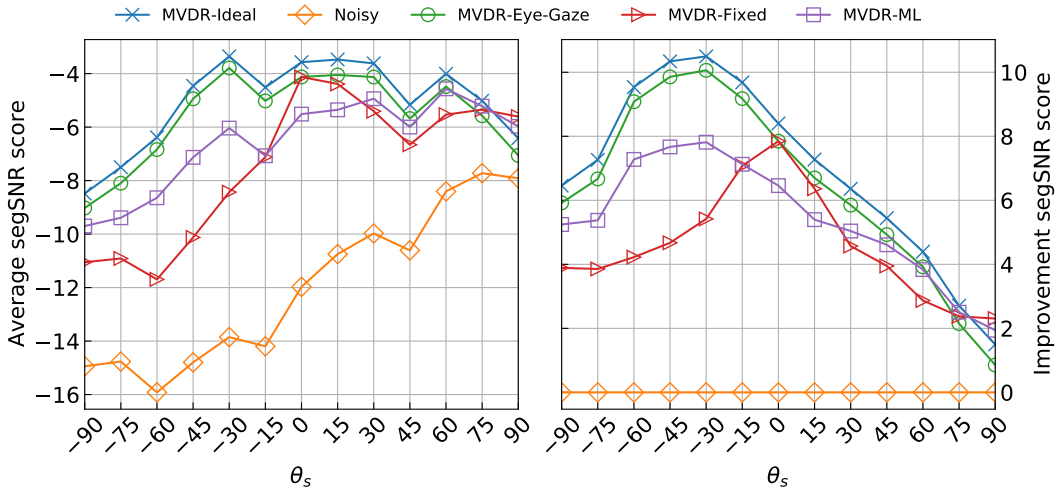


(c) A 6-microphone binaural configuration, using all three microphones on both HAs.

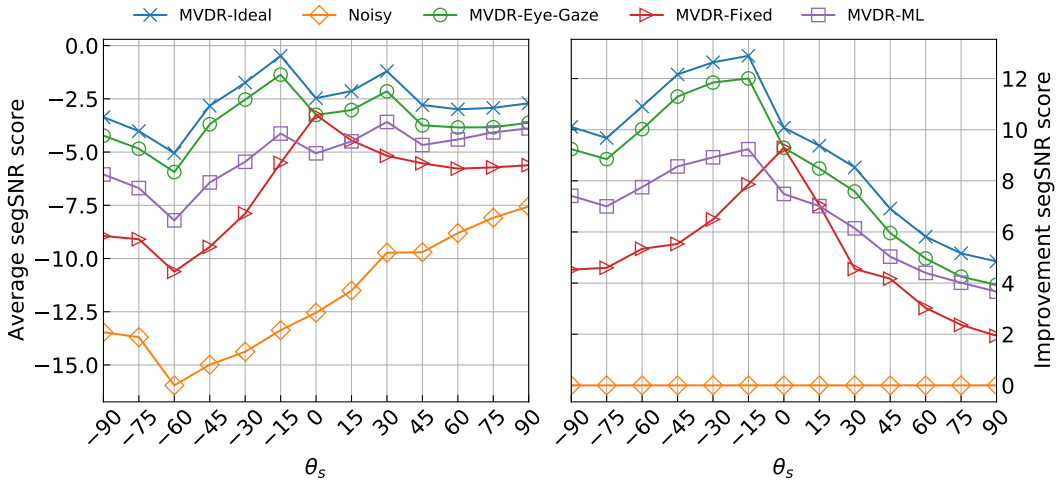
Figure 5.5: Average segSNR scores and average improvement segSNR scores as a function of target DOA. The noise type is babble in an approximately isotropic noise field with an input SNR of −10 dB.



(a) A 2-microphone monaural configuration, using the front and rear microphone on the left HA.

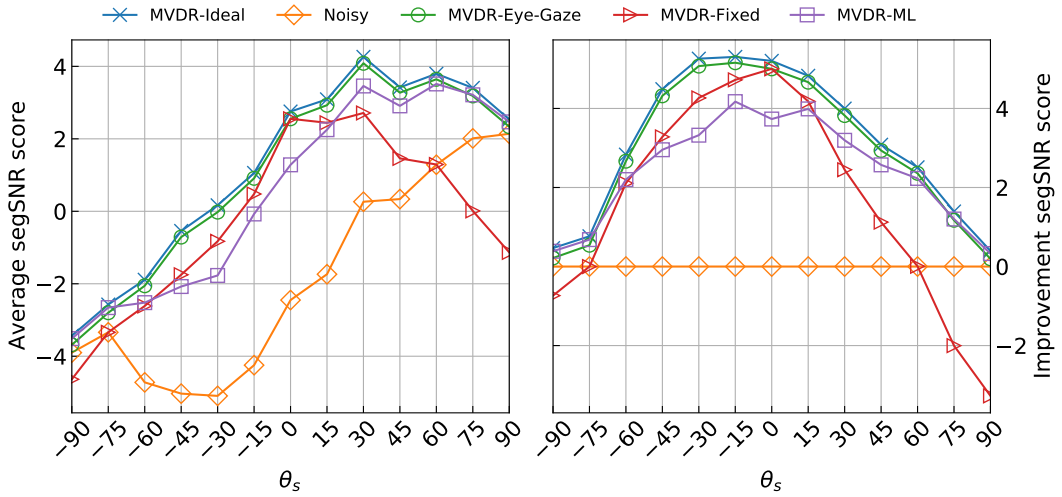


(b) A 4-microphone binaural configuration, using the front and rear microphones on both the left and right HA.

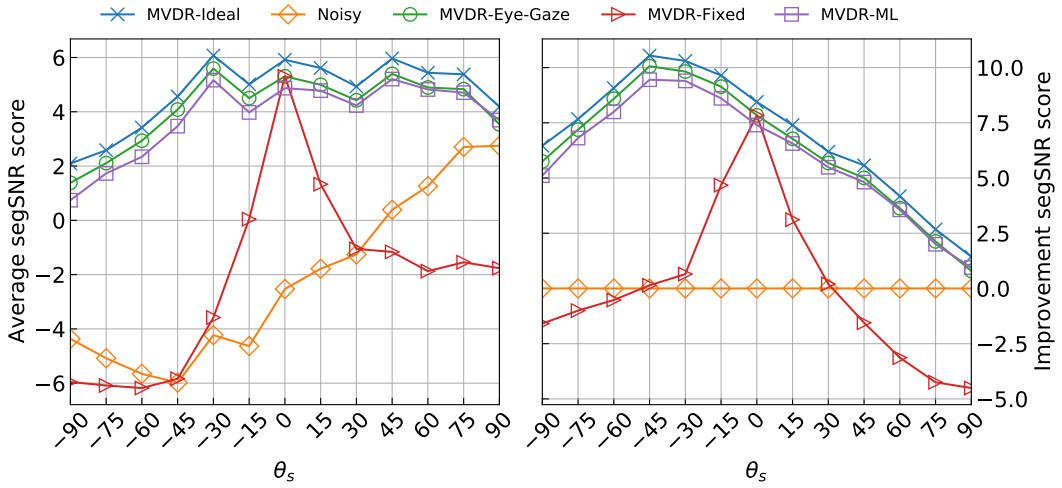


(c) A 6-microphone binaural configuration, using all three microphones on both HAs.

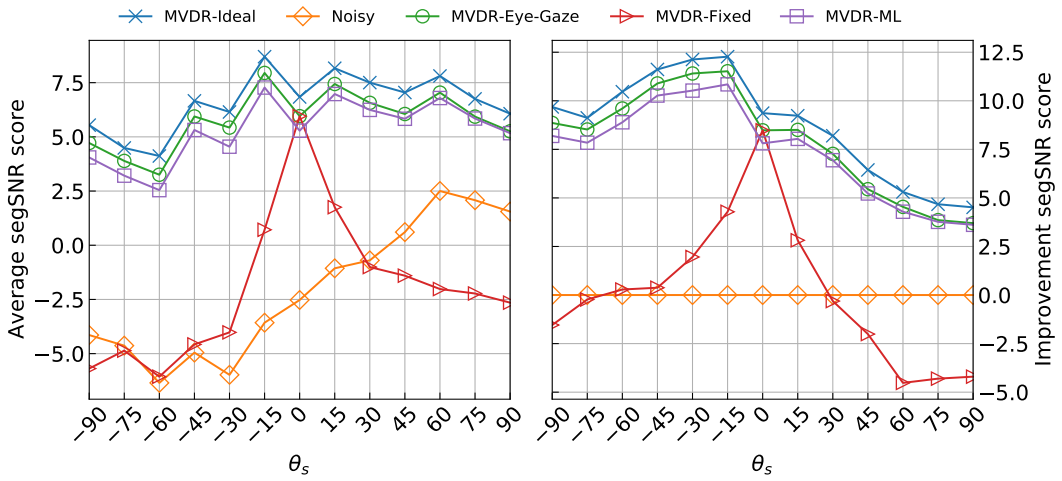
Figure 5.6: Average segSNR scores and average improvement segSNR scores as a function of target DOA. The noise type is babble in an approximately isotropic noise field with an input SNR of 0 dB.



(a) A 2-microphone monaural configuration, using the front and rear microphone on the left HA.



(b) A 4-microphone binaural configuration, using the front and rear microphones on both the left and right HA.



(c) A 6-microphone binaural configuration, using all three microphones on both HAs.

Figure 5.7: Average segSNR scores and average improvement segSNR scores as a function of target DOA. The noise type is babble in an approximately isotropic noise field with an input SNR of 10 dB.

6. Proposed Audio-Gaze Beamforming Methods

The overall purpose of this thesis is to investigate the use of an additional modality apart from sound to help localizing and enhancing the target signal in HA applications. Specifically, in this thesis, we aim to use a signal related to the HA user's eye-gaze as this additional source of information. In this context, our goal is to develop and propose a beamforming system for HAs which incorporates the HA user's eye-gaze to explore the potential advantages of using this additional modality. One of the milestones in this process is to study and design statistical signal processing algorithms for target localization using information provided by both acoustic and eye direction information. This chapter aims to provide a detailed presentation of our proposed HA beamforming methods which incorporates the user's eye gaze.

In general, there may be several mathematical methods which can be employed to propose algorithms that make use of an additional modality to extract information related to the direction of the target sound source. In this thesis, we have decided to employ a Bayesian approach. One of the motivations behind employing a Bayesian approach is that when prior information about the target sound source direction is available, Bayesian beamforming offers a framework for incorporating this prior information about the target sound source DOA to form optimal beamformers under DOA uncertainty [23], [24], [54]. In general, estimation within a Bayesian framework involves the choice of a loss function, e.g., the mean-square-error (MSE), as well as minimization of the expected value of this loss function. In the special case of parameter estimation, the most direct Bayesian approach is to assume that the uncertain DOA is a random variable with a prior distribution that reflects its uncertainty, and then optimize relative to the corresponding posterior distribution of the DOA given the observed data [54], [55]. Such a random formulation of the target DOA is preferred over a deterministic formulation because it considers the average effect of the DOA error instead of a particular perturbed value that may not be representative enough to describe the uncertain scenario [54].

In [23] and [24], methods have been proposed where a posterior DOA probability is used to consider the Bayesian beamforming approach for noise reduction. However,

these proposed methods rely solely on acoustic information to estimate the target signal. The overall idea of our proposed methods is to extend these audio-only Bayesian methods to use access to the user's eye-gaze information too. Specifically, we propose to use information provided by both the microphone signals and the HA user's eye-gaze to estimate a posterior probability distribution of the direction of the target talker, i.e., for each possible target direction assigning a probability. This estimated posterior probability distribution is then used in the design of Bayesian beamformers, which we can apply to the noisy microphone signals to obtain estimates of the target speech signal measured at the reference microphone.

Due to the fact that our proposed methods can be seen as an extension of the Bayesian methods in [23], [24] which rely on acoustic information only, we begin this chapter with a derivation of the traditional Bayesian beamformer which uses acoustic information only. Afterwards, this framework is extended to be useful in situations where an additional information, apart from sound, is available.

6.1 Bayesian Beamforming

In this section, we introduce the necessary background of the Bayesian approach to adaptive beamforming using acoustic information only, aiming to lay the foundation for us to be able to extend this approach to incorporate eye-gaze information in addition to the traditional acoustic information. First, we recall the signal model for the noisy microphone observations as well as the statistical assumptions made on this signal model. Following this, we derive the Bayesian beamformer, and lastly, we discuss the building blocks of the Bayesian beamfromer.

6.1.1 Acoustic Information - HA Microphone Signals

Recall the signal model for the noisy observations $\tilde{\boldsymbol{x}}(k,l) \in \mathbb{C}^M$ in the time-frequency domain which for a particular time-frequency tile is given as

$$\tilde{\boldsymbol{x}}(k,l) = \tilde{s}(k,l)\boldsymbol{d}(k,l,\theta_s) + \tilde{\boldsymbol{v}}(k,l), \tag{6.1}$$

where $\tilde{s}(k,l) \in \mathbb{C}$ is the scalar STFT coefficient of the target signal measured at a pre-selected reference microphone which arrives from direction θ_s, $\boldsymbol{d}(k,l,\theta_s) \in \mathbb{C}^M$ is the RTF vector of the target signal from the chosen reference microphone to all of the microphones, and $\tilde{\boldsymbol{v}}(k,l) \in \mathbb{C}^M$ is the overall additive noise component which is assumed to be uncorrelated with the target signal.

To employ a Bayesian approach, we use the signal model for the noisy microphone observations in (6.1), and, unless otherwise is stated, the assumptions made on the individual signal components and their interrelationships in Chapter 2 are assumed to be the same in the subsequent presentation. Furthermore, as in Chapter 4, we assume in the following that the target sound source can arrive from one out of I pre-selected source directions θ_i, for $i = 1, \ldots, I$, where each source direction is

represented by an RTF vector $\boldsymbol{d}(k,l,\theta_i)$, for $i = 1,\ldots,I$ from a predefined RTF dictionary $\mathcal{D} = \{\boldsymbol{d}(k,l,\theta_1),\ldots,\boldsymbol{d}(k,l,\theta_I)\}$. Therefore, we write the signal model for the noisy microphone signals observed at the HA microphones as

$$\tilde{\boldsymbol{x}}(k,l) = \tilde{s}(k,l)\boldsymbol{d}(k,l,\theta_i) + \tilde{\boldsymbol{v}}(k,l), \tag{6.2}$$

with $\boldsymbol{d}(k,l,\theta_i)$ being one particular RTF vector from the dictionary $\mathcal{D}$.

6.1.2 Derivation of the Bayesian Beamformer

In this section, the Bayesian beamformer is derived. This derivation is primarily based on [24], [23], and [54]. As mentioned, in a Bayesian approach, the target DOA is modeled as a random variable with a prior probability distribution defined over a candidate set of target DOAs. Since it is assumed that the target sound source can arrive from one out of I possible source directions θ_i, for $i = 1,\ldots,I$, the candidate set of target DOAs is assumed to be discrete. Specifically, in the following description, we assume the target DOA to be a discrete random variable with possible outcomes θ_i, for $i = 1,\ldots,I$, and with prior PMF, denoted $p(\theta_i)$, defined on the discrete candidate set of DOAs $\Theta = \{\theta_i,\ldots,\theta_I\}$.

As mentioned, the Bayesian beamformer refers to an optimal beamformer that minimizes the MSE between the target signal $\tilde{s}(k,l)$ and the estimated target signal $\hat{\tilde{s}}(k,l)$, under DOA uncertainty. In order to derive the Bayesian beamformer, we therefore consider the minimum mean-square error (MMSE) estimator of the target signal based on observations of the noisy microphone signals. To this end, let $\boldsymbol{X}(k,l) = [\tilde{\boldsymbol{x}}(k,l-L+1),\ldots,\tilde{\boldsymbol{x}}(k,l)] \in \mathbb{C}^{M\times L}$ denote a dataset of L consecutive time frames of noisy observations. Given the observation set $\boldsymbol{X}(k,l)$, the goal is to retrieve the target signal $\tilde{s}(k,l)$ at the reference microphone. From a Bayesian approach, this means that we are interested in the MMSE estimate of $\tilde{s}(k,l)$. Per definition, the MMSE estimate of the target signal $\tilde{s}(k,l)$ is the conditional expectation of $\hat{\tilde{s}}(k,l)$, given the observed data $\boldsymbol{X}(k,l)$ [56, p. 156], [52, pp. 208-210], i.e., for a given time-frequency tile, we can write the MMSE estimate of $\tilde{s}(k,l)$ as [54]

$$\hat{\tilde{s}}(k,l) = \mathrm{E}\left[\tilde{s}(k,l)|\boldsymbol{X}(k,l)\right]. \tag{6.3}$$

Using the definition of the conditional expectation of a continuous random variable, we can expand the conditional expectation of $\tilde{s}(k,l)$ given $\boldsymbol{X}(k,l)$ in (6.3) as

$$\begin{aligned}\hat{\tilde{s}}(k,l) &= \mathrm{E}\left[\tilde{s}(k,l)|\boldsymbol{X}(k,l)\right] &(6.4)\\ &= \int_{\mathbb{C}} \tilde{s}(k,l) f(\tilde{s}(k,l)|\boldsymbol{X}(k,l))\, d\tilde{s}(k,l), &(6.5)\end{aligned}$$

where $f(\tilde{s}(k,l)|\boldsymbol{X}(k,l))$ denotes the conditional PDF of the target signal $\tilde{s}(k,l)$, given the noisy observations $\boldsymbol{X}(k,l)$. Due to the fact that the target DOA is modeled as a discrete random variable with a prior PMF $p(\theta_i)$ defined over the candidate prior set

$\Theta = \{\theta_1, \ldots, \theta_I\}$, we can use the law of total probability [57, Th. 1.1] to expand the expression for the conditional PDF $f(\tilde{s}(k,l)|\boldsymbol{X}(k,l))$ as [23], [24]

$$f(\tilde{s}(k,l)|\boldsymbol{X}(k,l)) = \sum_{i=1}^{I} f(\tilde{s}(k,l)|\boldsymbol{X}(k,l),\theta_i)p(\theta_i|\boldsymbol{X}(k,l)), \tag{6.6}$$

where $p(\theta_i|\boldsymbol{X}(k,l))$, for $i = 1,\ldots,I$, denotes the posterior probability of θ_i given $\boldsymbol{X}(k,l)$. Substituting (6.6) into (6.5), the Bayesian MMSE estimate [52, p. 209] of the target signal at a given time-frequency tile becomes

$$\hat{\tilde{s}}(k,l) = \int_{\mathbb{C}} \tilde{s}(k,l) \sum_{i=1}^{I} f(\tilde{s}(k,l)|\boldsymbol{X}(k,l),\theta_i)p(\theta_i|\boldsymbol{X}(k,l))\, d\tilde{s}(k,l) \tag{6.7}$$

$$= \sum_{i=1}^{I} p(\theta_i|\boldsymbol{X}(k,l)) \int_{\mathbb{C}} \tilde{s}(k,l) f(\tilde{s}(k,l)|\boldsymbol{X}(k,l),\theta_i)\, d\tilde{s}(k,l) \tag{6.8}$$

$$= \sum_{i=1}^{I} p(\theta_i|\boldsymbol{X}(k,l)) \mathrm{E}\left[\tilde{s}(k,l)|\boldsymbol{X}(k,l),\theta_i\right], \tag{6.9}$$

where the second equality holds due to the fact that the integral is a linear operator [58, p. 70], and the last equality follows from the definition of conditional expectation. From (6.9), it is seen that the MMSE estimate of $\tilde{s}(k,l)$ can be viewed as a linear mixture of directional MMSE estimates $\mathrm{E}\left[\tilde{s}(k,l)|\boldsymbol{X}(k,l),\theta_i\right]$ of $\tilde{s}(k,l)$ combined according to the posterior distributions $p(\theta_i|\boldsymbol{X}(k,l))$. For Gaussian signals, it can be shown, see e.g., [54], that these directional MMSE estimates are the output of so-called multichannel Wiener filters (MWFs) steered towards each direction θ_i, for $i = 1,\ldots,I$, and the corresponding MMSE estimator is what is known as the Bayesian beamformer [54], and is given in terms of its beamformer coefficients as

$$\boldsymbol{w}_{\mathrm{B}}(k,l) = \sum_{i=1}^{I} p(\theta_i|\boldsymbol{X}(k,l))\boldsymbol{w}_{\mathrm{MWF}}(k,l,\theta_i), \tag{6.10}$$

where $\boldsymbol{w}_{\mathrm{MWF}}(k,l,\theta_i)$ is the MWF beamformer weights which are given as the solution to the optimization problem [10, pp. 41-46]

$$\boldsymbol{w}_{\mathrm{MWF}}(k,l,\theta_i) = \arg\min_{\boldsymbol{w}} \mathrm{E}\left[|\tilde{s}(k,l) - \boldsymbol{w}^H(k,l)\tilde{\boldsymbol{x}}(k,l)|^2\right]. \tag{6.11}$$

It is known that the MWF beamformers used in (6.10) sometimes lead to audible distortions when implemented in practice [59]. To avoid this, it has been proposed in the literature [23], [24] to use a more heuristically motivated Bayesian MVDR beamformer instead where the MWF beamformers $\boldsymbol{w}_{\mathrm{MWF}}(k,l,\theta_i)$, for $i = 1,\ldots,I$, in (6.10) are substituted by MVDR beamformers steered towards each direction θ_i in the discrete set Θ of candidate target DOAs, i.e., [24]

$$\check{\boldsymbol{w}}_{\mathrm{B}}(k,l) = \sum_{i=1}^{I} p(\theta_i|\boldsymbol{X}(k,l))\boldsymbol{w}_{\mathrm{MVDR}}(k,l,\theta_i), \tag{6.12}$$

where $\boldsymbol{w}_{\text{MVDR}}(k,l,\theta_i)$, for $i = 1\ldots,I$, are the MVDR beamformer coefficients derived in Chapter 3 which for each $\theta_i \in \Theta$ is given as [23], [24],

$$\boldsymbol{w}_{\text{MVDR}}(k,l,\theta_i) = \frac{\boldsymbol{C}_{\boldsymbol{v}}^{-1}(k,l)\boldsymbol{d}(k,l,\theta_i)}{\boldsymbol{d}^H(k,l,\theta_i)\boldsymbol{C}_{\boldsymbol{v}}^{-1}(k,l)\boldsymbol{d}(k,l,\theta_i)}. \tag{6.13}$$

Note that, we have denoted $\check{\boldsymbol{w}}_{\text{B}}(k,l)$ in (6.12) with a check sign to highlight that there is not equality between the left-hand side in (6.12) and the left-hand side in (6.10). Due to the fact that we in this thesis have decided to consider MVDR beamformers and not MWF beamformers, we will use the approximated Bayesian beamformer in (6.12) when implementing the Bayesin beamformers. For brevity, we will in the rest of this thesis refer to the beamformer in (6.12) as the Bayesian beamformer as well as ignore the check sign, however, still keeping in mind the deviation from the proper theoretical Bayesian beamformer in (6.10).

From (6.12), it is seen that in order to implement the Bayesian beamformer, knowledge about the posterior probability $p(\theta_i|\boldsymbol{X}(k,l))$, for $i = 1,\ldots,I$, is required. Therefore, in the following we will describe how this probability distribution is obtained.

Computing a Posterior DOA Probabilities

To compute the posterior DOA probability $p(\theta_i|\boldsymbol{X}(k,l))$, for $i = 1,\ldots,I$, in (6.12), we can use Bayes' theorem [56, p. 151, eq. (2.251)] to expand it as

$$p(\theta_i|\boldsymbol{X}(k,l)) = \frac{f(\boldsymbol{X}(k,l)|\theta_i)p(\theta_i)}{f\left(\boldsymbol{X}(k,l)\right)}, \quad i = 1,\ldots I, \tag{6.14}$$

where $f(\boldsymbol{X}(k,l)|\theta_i)$ is the likelihood function for the noisy microphone signals, $\boldsymbol{X}(k,l)$ conditioned on θ_i, $p(\theta_i)$ is the prior probability of the target DOA, which describes the underlying probability that a target signal arrives from a particular direction θ_i, for $i = 1,\ldots,I$, and $f(\boldsymbol{X}(k,l))$ is the marginal PDF of $\boldsymbol{X}(k,l)$, which can be written in terms of the likelihood function and the prior as

$$f(\boldsymbol{X}(k,l)) = \sum_{i=1}^{I} f(\boldsymbol{X}(k,l),\theta_i) \tag{6.15}$$

$$= \sum_{i=1}^{I} f(\boldsymbol{X}(k,l)|\theta_i)p(\theta_i), \tag{6.16}$$

where the first equality is the so-called marginalization by summing out the θ_i's, and where the second equality follows from the fact that a joint density can be written as the product of a conditional density and a marginal probability [25, pp. 225+226]. Note that the denominator in (6.14) acts as a normalizing constant to make $p(\theta_i|\boldsymbol{X}(k,l))$ on the left-hand side sum to a value of 1 over the DOA range Θ,

i.e., $\sum_{i=1}^{I} p(\theta_i|\boldsymbol{X}(k,l)) = 1$. Substituting (6.16) into (6.14), the posterior probability is expressed as

$$p(\theta_i|\boldsymbol{X}(k,l)) = \frac{f(\boldsymbol{X}(k,l)|\theta_i)p(\theta_i)}{\sum_{i=1}^{I} f(\boldsymbol{X}(k,l)|\theta_i)p(\theta_i)}, \quad i = 1, \dots I. \tag{6.17}$$

In the following, we briefly describe how to evaluate the likelihood function $f(\boldsymbol{X}(k,l)|\theta_i)$ and the prior probability $p(\theta_i)$ on the right-hand side of (6.17) in order to be able to compute the value of the posterior DOA probability $p(\theta_i|\boldsymbol{X}(k,l))$ on the left-hand side.

The likelihood function $f(\boldsymbol{X}(k,l)|\theta_i)$: The likelihood $f(\boldsymbol{X}(k,l)|\theta_i)$ for the noisy microphone observations can be computed following the approach outlined in Chapter 4. Specifically, we recall that to compute $f(\boldsymbol{X}(k,l)|\theta_i)$, it is assumed that the noisy observations $\boldsymbol{X}(k,l)$ are realizations of circularly-symmetric complex Gaussian distributed random processes with CPSD matrix given as

$$\boldsymbol{C_x}(k,l,\theta_i) = \lambda_s(k,l,\theta_i)\boldsymbol{d}(k,l,\theta_i)\boldsymbol{d}^H(k,l,\theta_i) + \lambda_v(k,l,\theta_i)\boldsymbol{\Gamma_v}(k,l_0), \quad l > l_0, \tag{6.18}$$

where $\lambda_s(k,l)$ and $\lambda_v(k,l)$ are the scalar PSDs of the target and noise signal at the reference microphone, respectively, $\boldsymbol{\Gamma_v}(k,l_0)$ is the normalized noise CPSD matrix which contain a value of 1 at the diagonal element corresponding to the reference microphone index [34], and where $l_0 < l$ denotes the last time frame of a speech absence period. As we saw in Chapter 4, this Gaussian assumption implies that the joint likelihood function $f(\boldsymbol{X}(k,l)|\theta_i)$ can be written as

$$f(\boldsymbol{X}(k,l)|\theta_i) = \frac{\exp\left(-L\text{tr}\left(\hat{\boldsymbol{R}}(k,l)\boldsymbol{C_x}^{-1}(k,l,\theta_i)\right)\right)}{\pi^{LM}\left|\boldsymbol{C_x}(k,l,\theta_i)\right|^L}, \quad i = 1, \dots, I. \tag{6.19}$$

To evaluate $\boldsymbol{C_x}(k,l,\theta_i)$ for a particular frequency bin index k and time frame l, i.e., for a particular time-frequency tile, we can follow the procedure outlined in Chapter 4, where $\boldsymbol{\Gamma_v}(k,l_0)$ is estimated from speech absence time-frequency tiles, and where $\lambda_s(k,l)$ and $\lambda_v(k,l)$ are estimated using the closed-form maximum likelihood estimate solutions in (4.13) and (4.11), respectively [11], [13]

The prior probability $p(\theta_i)$: In order to find the prior probability $p(\theta_i)$, for $i = 1, \dots, I$, any prior information about the target direction θ_i may be utilized. For instance, if we do not believe that the target signal tends to arrive from a particular direction over another, a uniform prior can be used, i.e., $p(\theta_i) = 1/I$, for $i = 1, \dots, I$, where I is the cardinality of the discrete set of direction from which the target signal can arrive. Similarly, if we expect a priori that the target signal primarily arrives from e.g. the frontal plane with respect to the HA user, this prior information may be reflected in the prior by increasing the probabilities corresponding to the frontal directions. As we will see in the next chapter, in this thesis, we propose

to use information provided by the HA user's eye-gaze to derive a prior probability distribution which will be used to evaluate the posterior DOA probabilities.

The description in this section considers the situation where the estimation of the posterior DOA probabilities are based on microphone signals only. As our goal is to use eye-gaze and microphone signals to estimate the target signal $\tilde{s}(k,l)$ impinging on the reference microphone, we will in the following sections extend the audio-only Bayesian framework to take into account additional information provided by the user's eye-gaze. Specifically, we propose to compute a posterior DOA probability distribution based on microphone signals and eye-gaze data in two different ways. In the first method, we utilize prior information of the target sound source DOA obtained from the user's eye-gaze direction, and derive a posterior DOA probability distribution of the target DOAs which only access the eye-gaze through the prior probability distribution. In the second approach, we compute a posterior probability of DOAs conditioned on both the microphone signals and the eye-gaze direction, meaning that a joint description of the posterior DOA probabilities based on microphone and eye-gaze data is obtained. Finally, these posterior DOA probabilities will then be used to form Bayesian beamformers.

6.2 Proposed Gaze-Prior Bayesain Beamforming Method

As we saw in Section 6.1.2, computing the posterior DOA probabilities

$$p(\theta_i|\boldsymbol{X}(k,l)) = \frac{f(\boldsymbol{X}(k,l)|\theta_i)p(\theta_i)}{\sum_{i=1}^{I} f(\boldsymbol{X}(k,l)|\theta_i)p(\theta_i)}, \quad i = 1,\dots I, \tag{6.20}$$

requires a prior probability on the target DOAs θ_i, for $i = 1,\dots,I$, which describes the intrinsic probability that a target signal arrives from a particular direction θ_i. As mentioned, information about the HA user's eye-gaze direction may in many situations provide very strong evidence of the direction of an active target talker, and, hence, help identify the target direction. For example, it is often the case that a HA user looks at the target talker, at least now and then, e.g. for lip reading in acoustically difficult situations. A simple heuristic extension of the Bayesian framework outlined in Section 6.1.2 that take advantages of the additional eye-gaze information, may therefore be to incorporate the user's eye-gaze in the derivation of the prior $p(\theta_i)$ in (6.17) over the I discrete target DOAs. This eye-gaze based prior can then subsequently be combined with the acoustic information via the acoustic likelihood function $f(\boldsymbol{X}(k,l)|\theta_i)$ to obtain an estimate of the posterior DOA probabilities $p(\theta_i|\boldsymbol{X}(k,l))$ which then relies on both acoustic and eye-gaze information. In the following, we describe how we propose to obtain the estimate of the prior probabilities $p(\theta_i)$, for $i = 1,\dots,I$.

6.2.1 Computing a Prior Probability of Target DOAs From Eye-Gaze Data

In this method, we propose to use the eye-gaze data to estimate the prior probabilities $p(\theta_i)$, for $i = 1, \ldots, I$ via the non-parametric density estimation approach of histograms. To this end, let $\phi(l)$ denote a time-varying signal which provides information about the user's eye direction, where l is a time variable which, for convenience, is assumed to be synchronized with the time variable l related to the acoustic information used in the indexing of e.g., $\tilde{\boldsymbol{x}}(k, l)$. Furthermore, assume for simplicity that the information represented by $\phi(l)$ is the user's eye-gaze angle in the horizontal plane with respect to the frontal direction from the HA user's point-of-view, as a function of time, then, $\phi(l)$ is a scalar. Based on $\phi(l)$, we propose to compute a histogram over the HA user's eye-gaze direction across a past period of time. Assuming that the HA user looks at the target source, at least now and then, e.g., for lip-reading, we hypothesize that the histogram may show higher occurrences of that particular target direction than others. By normalizing the histogram into a PMF, we obtain an estimate of the prior probabilities $p(\theta_i)$, for $i = 1, \ldots, I$, which we can then use to evaluate the posterior DOA probabilities in (6.17). These posterior DOA probabilities can then subsequently be used to implement a Bayesian beamformer. Note that an important consideration is how to choose the duration of previous eye-gaze measurements used in the histogram, i.e., how long the time period in which we estimate our prior over should be. For small periods of time, the prior probability may adapt quickly to new data, however, a potential drawback of having too few data points in the density estimation, is that the histogram can become overfitted to the variance in the data. For large periods of time, the histogram may be more slowly varying, and thereby, not as responsive to quick changes. In the next chapter, we will expand upon the computations of the prior probabilities in more practical details.

Before proceeding, an important consideration has to be emphasized. We abuse the term prior probability distribution, as in a rigorous use of the terminology, a prior probability distribution refers to a prior knowledge, or expectation, of a distribution of a parameter before data is observed. In this thesis, we use the observed eye-gaze angles to adaptively form the prior of the target DOA parameter, and hence, violates with the rigorous definition of being a prior probability.

6.2.2 Implementation of the Proposed Gaze-Prior Beamformer

To implement the proposed gaze-prior Bayesian beamforming system, the discrete prior probability distribution $p(\theta_i)$, for $i = 1, \ldots, I$, is estimated from real-world eye-gaze data and used to evaluate the posterior probability distribution $p(\theta_i|\boldsymbol{X}(k, l))$, for $i = 1, \ldots, I$. Due to numerical tractability, we choose to first compute the posterior probability $p(\theta_i|\boldsymbol{X}(k, l))$ in the logarithmic domain and afterwards transform the density back into the linear domain. Taking the natural logarithm of (6.20), yields

$$\ln(p(\theta_i|\boldsymbol{X}(k, l))) = \ln(f(\boldsymbol{X}(k, l)|\theta_i)) + \ln(p(\theta_i)) - \ln(c), \quad i = 1, \ldots, I, \tag{6.21}$$

where $c = \sum_{i=1}^{I} f(\boldsymbol{X}(k,l)|\theta_i)p(\theta_i)$. The log-likelihood function $\ln(f(\boldsymbol{X}(k,l)|\theta_i))$ can be computed in accordance to the description in Chapter 4, while the log-prior probability $\ln(p(\theta_i))$ is found by computing a histogram over the the HA user's eye-gaze direction $\phi(l)$ across a past period of time T. Specifically, at the l'th time frame, we choose to compute the log-prior probability $\ln(p(\theta_i))$ based on eye-gaze measurement from time frame l and the measurements from the previous $T-1$ time frames. As mentioned in the previous section, when choosing the number of time frames over which the histogram is computed, we have to make a choice between robustness against rapid changes in the data and being quickly adaptive to new data.

In creating a histogram, a few parameters must be considered to obtain a sensible result. This includes the number of bins that $\phi(l)$ is partitioned into, the width of these bins, and lastly, the placement of the bin edges. Since the AIR database available for this thesis has an angular resolution of 7.5°, we have chosen to associate a bin with each discrete candidate target DOA, hence we have $I = 48$ bins in our histogram. Note that one could have grouped multiple angles into one bin, resulting in fewer bins, an option often done if the histogram captures too much noise in the data, giving a poor estimate of the density. Since we have chosen $I = 48$ bins and the continuous eye-gaze signal $\phi(l)$ can take values in the continuous range $[-180°, 180°)$, the width of the bins becomes $360°/48 = 7.5°$. Lastly, consider the placement of the bin edges, which we choose such that a given bin is centered around the associated angle, e.g., for the bin associated with a target direction of $\theta_i = 15°$, the bin edges are 11.25° and 18.75°, respectively. Finally, the histogram is computed by counting the fraction of measured eye-gaze angles falling in each of the $I = 48$ bins, and normalizing the histogram into a PMF, we obtain an estimate of the prior probabilities $p(\theta_i)$, for $i = 1, \ldots, I$. At last, taking the natural logarithm of the histogram, with a very small number added to all values to avoid taking the logarithm of zero, we arrive at the log-prior probabilities $\ln(p(\theta_i))$, for $I = 1, \ldots, I$.

Finally, in order to compute the log-posterior probability, the normalization constant c is needed. In the linear domain c is given as the product of the likelihood and the prior, summed over all I elements in the discrete set of candidate target directions. Inserting the evaluated log-likelihood function $\ln(f(\boldsymbol{X}(k,l)|\theta_i))$ and the estimated log-prior probability distribution $\ln(p(\theta_i))$ into (6.21), and applying the exponential function $\exp(\cdot)$ to (6.21), the posterior DOA probabilities $p(\theta_i|\boldsymbol{X}(k,l))$ is found.

Having computed the posterior DOA probabilities $p(\theta_i|\boldsymbol{X}(k,l))$, for $i = 1, \ldots, I$, these are used to implement the Bayesian beamformer in (6.12). Finally, to obtain the estimated target signal $\hat{\tilde{s}}(k,l)$ for a given time-frequency tile, the Bayesian beamformer is applied to the noisy microphone signals by taking the inner product between the beamformer weights and the noisy microphone signals $\boldsymbol{x}(k,l)$ such that the estimated target signal $\hat{\tilde{s}}$, which is given as the output of the beamformer for the k'th

frequency bin and the l'th time frame, is

$$\hat{\tilde{s}}(k,l) = \hat{\boldsymbol{w}}_{\mathrm{B}}^{H}(k,l)\boldsymbol{x}(k,l). \tag{6.22}$$

The implementation of this proposed Bayesian beamforming method is summarized in Algorithm 4 as pseudo-code.

6.3 Proposed Joint Audio-Gaze Beamforming Method

Instead of extending the framework in Section 6.1.2 to take into account the eye-gaze information via a prior, one may consider to incorporate the eye-gaze information in a way such that a joint description of the posterior DOA probabilities based on microphone and eye-gaze data is obtained. To this end, assume that the HA system has access to the eye-gaze signal $\phi(l)$, in addition to access to the traditional noisy microphone signals $\tilde{\boldsymbol{x}}(k,l)$. Furthermore, we assume for simplicity that this eye-gaze signal can be discretized such that $\phi(l) \in \{\phi_1, \ldots, \phi_J\}$, meaning that the eye-gaze angle at a particular moment in time l is one out of J possible eye-gaze angles ϕ_j, for $j = 1, \ldots, J$. In other words, we define a discrete set of candidate eye-gaze angles. An example of such a discrete set of candidate eye-gaze directions is depicted in Fig. 6.1, where each blue cross indicates in which directions the eye-gaze may be.

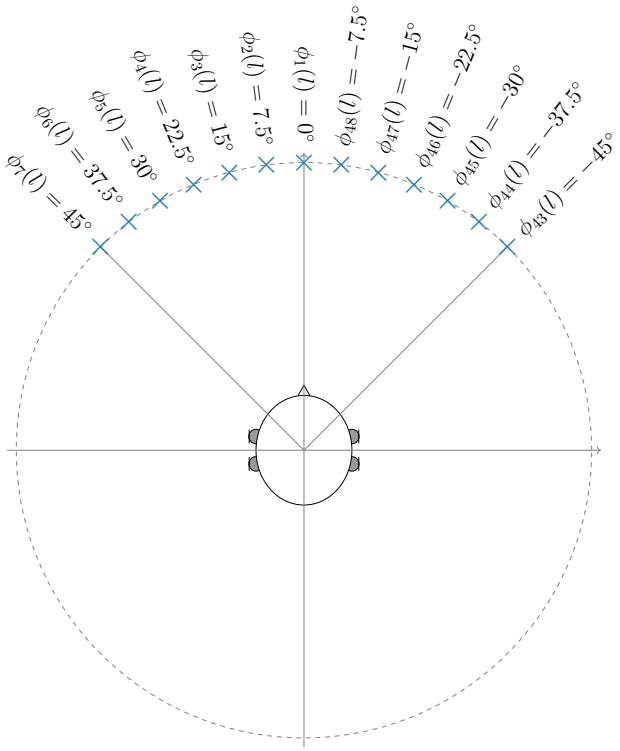


Figure 6.1: Example of user's candidate eye-gaze direction $\phi_j(l)$ at time l.

Note that in the proceeding description, we will denote the possible realizations ϕ_j, for $j = 1, \ldots, J$, of the random variable $\phi(l)$ as $\phi_j(l)$, for $j = 1, \ldots, J$. The variable $\phi_j(l)$ is in fact not a time-varying signal, however, we still choose to denote it with the index l to coincide with the indexing of the noisy microphone signals $\tilde{\boldsymbol{x}}(k,l)$. Furthermore, it should be noted that in practice, the eye-gaze angle is definitely a continuous signal, and hence, the assumption that the eye-gaze angle at a particular moment in time is discrete may not be a very realistic assumption. However, due to the fact that we in this thesis only have access to a discrete database of AIRs, we can only evaluate our systems for specific discrete candidate target directions, and hence, the data we use to extract information about the target directions may, for practical reasons, indeed be discrete as well.

6.3.1 Computing a Probability of Target DOAs Conditioned on Eye-gaze Data

The idea of this proposed framework is that instead of estimating the target signal relying on $p(\theta_i|\boldsymbol{X}(k,l))$, i.e., using access to acoustic information alone, we propose to extend the method to use the user's eye-gaze information too. In other words, rather than using $p(\theta_i|\boldsymbol{X}(k,l))$, for $i = 1 \ldots, I$, to estimate the target signal, we propose to use a joint posterior DOA probability distribution $p(\theta_i|\boldsymbol{X}(k,l), \phi_j(l))$, for $i = 1 \ldots, I$, linking observations $\boldsymbol{X}(k,l)$ and $\phi_j(l)$ to the target DOA θ_i. In this way, we propose to build a joint audio-gaze Bayesian beamformer to estimate the probability of a target direction.

In order to express the posterior probability $p(\theta_i|\boldsymbol{X}(k,l), \phi_j(l))$ for each $i = 1, \ldots, I$, we first expand it using Bayes' theorem such that

$$p\left(\theta_i|\boldsymbol{X}(k,l), \phi_j(l)\right) = \frac{f\left(\boldsymbol{X}(k,l)|\theta_i, \phi_j(l)\right) p\left(\theta_i|\phi_j(l)\right)}{f\left(\boldsymbol{X}(k,l)|\phi_j(l)\right)}, \tag{6.23}$$

where $f(\boldsymbol{X}(k,l)|\theta_i, \phi_j(l))$ is the likelihood of $\boldsymbol{X}(k,l)$ given θ_i and $\phi_j(l)$, $p(\theta_i|\phi_j(l))$ is a conditional PMF of θ_i given $\phi_j(l)$, and $f(\boldsymbol{X}(k,l)|\phi_j(l))$ acts as a normalizing constant to make $p\left(\theta_i|\boldsymbol{X}(k,l), \phi_j(l)\right)$ on the right-hand side a proper density.

From (6.23), it is seen that in order to implement the Bayesian beamformer, using the joint posterior DOA probability on the left-hand side of (6.23), we need to be able to evaluate the factors on the right-hand side. The joint conditional likelihood function $f(\boldsymbol{X}(k,l)|\theta_i, \phi_j(l))$ describes the likelihood of observing $\tilde{\boldsymbol{x}}(k,l)$ given that both the target DOA θ_i and the eye-gaze direction $\phi_j(l)$ is known. However, it may be reasonable to assume that, having knowledge of the user's eye-gaze direction $\phi_j(l)$, may not provide us with any additional information as already provided by the knowledge provided by the given target DOA θ_i. Hence, it seems reasonable to consider $\tilde{\boldsymbol{x}}(k,l)$ as being conditional independent of $\phi_j(l)$, given that the target DOA θ_i is known. We assume therefore that $\boldsymbol{X}(k,l)$ is conditionally independent of $\phi_j(l)$ given that θ_i is known. With this assumption, the joint conditional likelihood function $f\left(\boldsymbol{X}(k,l)|\theta_i, \phi_j(l)\right)$ simplifies to the acoustic likelihood $f\left(\boldsymbol{X}(k,l)|\theta_i\right)$ [43, pp. 46, 372],

and hence, the posterior probability $p(\theta_i|\boldsymbol{X}(k,l),\phi_j(l))$ in (6.23) becomes

$$p\left(\theta_i|\boldsymbol{X}(k,l),\phi_j(l)\right)=\frac{f\left(\boldsymbol{X}(k,l)|\theta_i\right)p\left(\theta_i|\phi_j(l)\right)}{\sum_{i=1}^{I}f\left(\boldsymbol{X}(k,l)|\theta_i\right)p\left(\theta_i|\phi_j(l)\right)}, \tag{6.24}$$

where we, to obtain the expression in the denominator, have used the fact that $p\left(\theta_i|\boldsymbol{X}(k,l),\phi_j(l)\right)$ should sum to a value of 1 over the DOA range $\Theta=\{\theta_1,\ldots,\theta_I\}$, i.e., $\sum_{i=1}^{I}p\left(\theta_i|\boldsymbol{X}(k,l),\phi_j(l)\right)=1$. It is seen that the only term on the right-hand side of (6.24) that depends on the eye-gaze signal $\phi_j(l)$ is now the conditional PMF $p\left(\theta_i|\phi_j(l)\right)$, and hence, it is this quantity that describes how knowledge of the HA users' horizontal eye-gaze direction provides information about the target direction. In the following, we describe how we choose to express the conditional probability $p(\theta_i|\phi_j(l))$ in order to be able to compute the posterior probability on the left-hand side of (6.24).

Probability of θ_i given $\phi_j(l)$: In order to derive the conditional PMF $p(\theta_i|\phi_j(l))$ of the discrete variable θ_i, given the realization of the discrete random variable $\phi_j(l)$, we need to know their joint PMF [57, Def. 3.6]. However, using Bayes' theorem, the conditional PMF can be derived from the joint PMF, and also the other way around, and hence, to compute the conditional PMF $p(\theta_i|\phi_j(l))$, for $i=1,\ldots,I$ and $j=1,\ldots,J$, we may expand it using Bayes' theorem such that

$$p(\theta_i|\phi_j(l))=\frac{p(\phi_j(l)|\theta_i)p(\theta_i)}{p(\phi_j(l))}, \tag{6.25}$$

in which we express the conditional PMF $p(\theta_i|\phi_j(l))$ in terms of the conditional PMF $p(\phi_j(l)|\theta_i)$. Hence, to compute $p(\theta_i|\phi_j(l))$, we are in need of $p(\phi_j(l)|\theta_i)$, which is the probability that the eye-gaze is in direction $\phi_j(l)$ given that the target direction is θ_i, and we are in the need of the prior probability $p(\theta_i)$ of the target DOAs θ_i, for $i=1,\ldots,I$, which describes the underlying probability that a target signal arrives from a particular direction θ_i.

For mathematical tractability and simplicity, the idea of the proposed method is to build a "look-up" table with histograms over $p(\phi_j(l)|\theta_i)$, for $i=1,\ldots,I$, such that we can use this table of histograms to evaluate the conditional probabilities $p(\theta_i|\phi_j(l))$, which we subsequently can use to estimate the joint posterior DOA probabilities $p(\theta_i|\boldsymbol{X}(k,l),\phi_j(l))$ to be used in the proposed Bayesian beamformer. Hence, we propose to compute the conditional probabilities $p(\phi_j(l)|\theta_i)$, for $i=1,\ldots,I$, "offline", i.e., before application of our proposed beamforming methods. In the following, will describe how computing the conditional probabilities $p(\theta_i|\phi_j(l))$ for $i=1,\ldots,I$ and $j=1,\ldots,J$, can be considered as constructing an $(I\times J)$ matrix of estimated conditional probabilities, as illustrated in Table 6.1. Specifically, in creating the matrix of the conditional probabilities $p(\theta_i|\phi_j(l))$, we first estimate the conditional probabilities $p(\phi_j(l)|\theta_i)$, which can also be viewed as filling in an $(I\times J)$ matrix of histograms, where we for each of the I rows fill in a histogram to estimate $p(\phi_j(l)|\theta_i)$,

	$\phi_1(l)$	$\phi_2(l)$	$\cdots$	$\phi_J(l)$
θ_1	$p\left(\theta_1\|\phi_1(l)\right)$	$p\left(\theta_1\|\phi_2(l)\right)$	$\cdots$	$p\left(\theta_1\|\phi_J(l)\right)$
θ_2	$p\left(\theta_2\|\phi_1(l)\right)$	$p\left(\theta_2\|\phi_2(l)\right)$	$\cdots$	$p\left(\theta_2\|\phi_J(l)\right)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
θ_I	$p\left(\theta_I\|\phi_1(l)\right)$	$p\left(\theta_I\|\phi_2(l)\right)$	$\cdots$	$p\left(\theta_I\|\phi_J(l)\right)$

Table 6.1: Pre-computed look-up table with conditional probabilities $p(\theta_i|\phi_j(l))$.

	$\phi_1(l)$	$\phi_2(l)$	$\cdots$	$\phi_J(l)$
θ_1	$p\left(\phi_1(l)\|\theta_1\right)$	$p\left(\phi_2(l)\|\theta_1\right)$	$\cdots$	$p\left(\phi_J(l)\|\theta_1\right)$
θ_2	$p\left(\phi_1(l)\|\theta_2\right)$	$p\left(\phi_2(l)\|\theta_2\right)$	$\cdots$	$p\left(\phi_J(l)\|\theta_2\right)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
θ_I	$p\left(\phi_1(l)\|\theta_I\right)$	$p\left(\phi_2(l)\|\theta_I\right)$	$\cdots$	$p\left(\phi_J(l)\|\theta_I\right)$

Table 6.2: Pre-computed look-up table with conditional probabilities $p(\phi_j(l)|\theta_i)$.

for $j = 1, \ldots, J$. This matrix is illustrated in Table 6.2. Note that since each row of the matrix of the probabilities $p(\phi_j(l)|\theta_i)$ is a density over $\phi_j(l)$, the rows will sum to one. Retaining the mindset of considering the probabilities as matrices, we can construct a new $(I \times J)$ joint PMF matrix $p(\phi_j(l), \theta_i) = p(\phi_j(l)|\theta_i)p(\theta_i)$, by multiplying each column of the matrix containing the probabilities $p(\phi_j(l)|\theta_i)$ with the associated entry of the prior probability $p(\theta_i)$. Next, we compute the probabilities $p(\phi_j(l))$, for $j = 1, \ldots, J$, by summing the columns of the matrix containing the probabilities $p(\phi_j(l)|\theta_i)p(\theta_i)$, for $i = 1, \ldots, I$. At last, we can construct the desired matrix of the conditional probabilities $p(\theta_i|\phi_j(l))$ by taking each column of $p(\phi_j(l)|\theta_i)p(\theta_i)$ and dividing by the associated entry of $p(\phi_j(l))$. Now the columns of $p(\theta_i|\phi_j(l))$ are probabilities over θ_i, for $i = 1, \ldots, I$, and therefore, sum to one. In the next chapter, we will present the specific estimation of each row in Table 6.2 using real-world eye-gaze data from experiments where the HA user's eye-gaze is measured while the user is following a conversation between two persons in a noisy environment.

In the following, we outline how the posterior DOA probabilities $p(\theta_i|\boldsymbol{X}(k,l), \phi_j(l))$, for $i = 1, \ldots, I$, after being computed, are used to implement the proposed Bayesian beamforming system to estimate the target signal $s(k,l)$ impinging on the reference

microphone.

6.3.2 Implementation of the Proposed Joint Audio-Gaze Beamformer

The implementation of the proposed joint audio-gaze beamforming method can we considered as a two-step procedure. First, the conditional probabilities $p(\theta_i|\phi_j(l))$, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, are computed using real-world eye-gaze data and stored in a look-up table, as illustrated in Table 6.1. This is done offline, i.e., before application of the proposed beamforming system. Secondly, the posterior probabilities $p(\theta_i|\boldsymbol{X}(k,l), \phi_j(l))$ are computed and used in a Bayesian beamformer to estimate the target signal impinging on the reference microphone. To do so, we choose to first compute the posterior probability $p(\theta_i|\boldsymbol{X}(k,l), \phi_j(l))$ in the logarithmic domain for numerical stability, and afterwards transform the probabilities back into the linear domain. Taking the natural logarithm of (6.24), yields

$$\ln(p(\theta_i|\boldsymbol{X}(k,l)), \phi_j(l)) = \ln(f(\boldsymbol{X}(k,l)|\theta_i)) + \ln(p(\theta_i|\phi_j(l))) - \ln(c), \quad i = 1, \ldots, I, \tag{6.26}$$

where $c = \sum_{i=1}^{I} f(\boldsymbol{X}(k,l)|\theta_i)p(\theta_i|\phi_j(l))$. Computing the acoustic log-likelihood function $\ln(f(\boldsymbol{X}(k,l)|\theta_i))$ in accordance to the description in Chapter 4, the posterior probability is found by substituting the evaluated log-likelihood function and the pre-computed look-up table with conditional PMFs $p(\theta_i|\phi_j(l))$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, into (6.26) and applying the exponential function $\exp(\cdot)$ to (6.26).

Having computed the posterior DOA probabilities $p(\theta_i|\boldsymbol{X}(k,l), \phi_j(l))$, for $i = 1, \ldots, I$, these are used to implement the Bayesian beamformer in (6.12). Finally, to obtain the estimated target signal $\hat{\tilde{s}}(k,l)$ for a given time-frequency tile, the Bayesian beamformer is applied to the noisy microphone signals by taking the inner product between the beamformer weights and the noisy microphone signals $\tilde{\boldsymbol{x}}(k,l)$ such that the estimated target signal $\hat{\tilde{s}}(k,l)$, which is given as the output of the beamformer for the k'th frequency bin and the l'th time frame, is

$$\hat{\tilde{s}}(k,l) = \hat{\boldsymbol{w}}_{\mathrm{B}}^{H}(k,l)\boldsymbol{x}(k,l). \tag{6.27}$$

The implementation of this proposed Bayesian beamforming system is summarized in Algorithm 5 as pseudo-code.

Having presented our two proposed beamforming methods, we will in the next chapter study the eye-gaze data and audio-visual data used in this thesis to design and evaluate our proposed methods.

Algorithm 4 Proposed Beamforming System: Bayesian-Gaze-Prior

Input:

$\boldsymbol{x}(n) = [x_1(n), \ldots, x_M(n)]^T$: Noisy microphone signals in time-domain.

$\mathcal{D} = \{\boldsymbol{d}(k,l,\theta_1), \ldots, \boldsymbol{d}(k,l,\theta_I)\} \in \mathbb{C}^{K\times M\times I}$: Dictionary of RTF vectors.

$\boldsymbol{\phi}(l) = [\phi_j(l-T+1), \ldots, \phi_j(l))]$: Vector of T time frames of eye-gaze measurements.

Output:

$\hat{s}(n)$: Estimated target speech signal at the reference microphone in the time-domain.

1: Apply STFT to $\boldsymbol{x}(n)$ to obtain $\tilde{\boldsymbol{x}}(k,l)$ for all k and l.
2: **for all** l **do**
3: **for** $k = 0$ **to** $K-1$ **do**
4: Let $\boldsymbol{X}(k,l) = [\tilde{\boldsymbol{x}}(k,l-L+1), \ldots, \tilde{\boldsymbol{x}}(k,l)]$ (negative indexing refers to the 1 second prepended noise).
5: Compute log-likelihood function $\ln(f(\boldsymbol{X}(k,l)|\theta_i))$.
6: **end for**
7: Compute log-prior $\ln(p(\theta_i))$ as a histogram of eye-gaze measurements in $\boldsymbol{\phi}(l)$. If T prior time frames are not available, all available time frames are used.
8: Estimate noisy CPSD matrix, $\hat{\boldsymbol{C}}_{\boldsymbol{v}}(k,l)$, for all k, according to (5.12).
9: Compute the log-normalization constant as

$$\ln(c) = \ln\left(\sum_{i=1}^{I}\left(\sum_{k=0}^{K-1}(\ln(f(\boldsymbol{X}(k,l)|\theta_i))) + \ln(p(\theta_i))\right)\right).$$

10: Obtain the posterior probability $p(\theta_i|\boldsymbol{X}(k,l))$ as

$$p(\theta_i|\boldsymbol{X}(k,l)) = \exp\left(\sum_{k=0}^{K-1}(\ln(f(\boldsymbol{X}(k,l)|\theta_i))) + \ln(p(\theta_i)) - \ln(c)\right), \quad \forall i.$$

11: Compute the Bayesian beamformer weights as

$$\hat{\boldsymbol{w}}_{\mathrm{B}}(k,l) = \sum_{i=1}^{I} p(\theta_i|\boldsymbol{X}(k,l))\hat{\boldsymbol{w}}_{\mathrm{MVDR}}(k,l,\theta_i), \quad \forall k.$$

12: Apply proposed Bayesian beamformer as $\tilde{\boldsymbol{s}}(k,l) = \hat{\boldsymbol{w}}_{\mathrm{B}}(k,l)^H\tilde{\boldsymbol{x}}(k,l), \quad \forall k.$
13: **end for**
14: Apply ISTFT to $\tilde{\boldsymbol{s}}(k,l)$ to obtain $\hat{s}(n)$ for all n.

Algorithm 5 Proposed Beamforming System, Audio-Gaze

Input:

$\boldsymbol{x}(n) = [x_1(n), \ldots, x_M(n)]^T$: Noisy microphone signals in time-domain.
$\mathcal{D} = \{\boldsymbol{d}(k,l,\theta_1), \ldots, \boldsymbol{d}(k,l,\theta_I)\} \in \mathbb{C}^{K \times M \times I}$: Dictionary of RTF vectors.
$\phi_j(l)$: Eye-gaze angle at time frame l.
$\boldsymbol{Z} \in \mathbb{R}^{I \times J}$: Conditional probability matrix $p(\theta_i|\phi_j(l))$.

Output:

$\hat{s}(n)$: Estimated time-domain target speech signal at the reference microphone.

Apply STFT to $\boldsymbol{x}(n)$ to obtain $\tilde{\boldsymbol{x}}(k,l)$ for all k and l.
for all l **do**
  for $k = 0$ **to** $K-1$ **do**
    Let $\boldsymbol{X}(k,l) = [\tilde{\boldsymbol{x}}(k,l-L+1), \ldots, \tilde{\boldsymbol{x}}(k,l)]$ (negative indexing refers to the 1 second prepended noise).
    Compute log-likelihood $\ln(f(\boldsymbol{X}(k,l)|\theta_i))$.
  end for
  Set $\ln(p(\theta_i|\phi_j(l)))$ equal to the logarithm of the j'th column of $\boldsymbol{Z}$ corresponding to $\phi_j(l)$.
  Estimate noisy CPSD matrix, $\hat{\boldsymbol{C}}_{\boldsymbol{v}}(k,l)$ for all k, according to (5.12).
  Compute the log-normalization constant as

$$\ln(c) = \ln\left(\sum_{i=1}^{I}\left(\sum_{k=0}^{K-1}(\ln(f(\boldsymbol{X}(k,l)|\theta_i))) + \ln(p(\theta_i))\right)\right).$$

  Obtain the posterior probability $p(\theta_i|\boldsymbol{X}(k,l), \phi_j(l))$ as

$$p(\theta_i|\boldsymbol{X}(k,l)) = \exp\left(\sum_{k=0}^{K-1}(\ln(f(\boldsymbol{X}(k,l)|\theta_i))) + \ln(p(\theta_i)) - \ln(c)\right), \quad \forall i.$$

  Compute the Bayesian beamformer weights as

$$\hat{\boldsymbol{w}}_{\mathrm{B}}(k,l) = \sum_{i=1}^{I} p(\theta_i|\boldsymbol{X}(k,l), \phi_j(l))\hat{\boldsymbol{w}}_{\mathrm{MVDR}}(k,l,\theta_i), \quad \forall k$$

  Apply Bayesian beamformer weights as $\tilde{\boldsymbol{s}}(k,l) = \boldsymbol{w}(k,l)^H\tilde{\boldsymbol{x}}(k,l)$, for all k.
end for
Apply ISTFT to $\tilde{\boldsymbol{s}}(k,l)$ to obtain $\hat{s}(n)$ for all n.

7. Eye-Gaze and Audio-Visual Data Study

A key aspect that allow us to study the construction and performance evaluation of beamforming systems for HAs which incorporates the user's eye-gaze, is the availability of real-world measurements of a HA user's eye-gaze recorded in synchronization with presented audio-visual stimuli. As mentioned in Chapter 1, the dataset used in this thesis is provided by Eriksholm Research Centre which is a part of Oticon. The dataset will be used to construct the proposed beamformers and to compare and determine the performance of different beamforming systems.

In this chapter, we introduce to the dataset provided by ERH. This include an overview of the relevant backgrounds behind the dataset as well as the test setup and procedure. The purpose of this overview is to provide the necessary background for understanding the data collection process and to get insight into the possible limitations involved in using the dataset in the design and simulation of the studied beamforming systems. Following this introduction to the dataset, we describe how we use the acoustic stimuli from the experiments, i.e., the clean sound signals from several loudspeakers, to create realistic synthetic acoustic scenes with target sound sources and background noise. Furthermore, we discuss how we process the raw eye-gaze data included in the dataset to obtain a set of eye-gaze measurements which we can use in our specific simulation framework. Finally, the chapter moves on to a description of how the preprocessed eye-gaze measurements are used to compute the look-up table in Table 6.1 with conditional probabilities before application of our proposed beamforming systems. The application and evaluation of the proposed bamformers will be studies in Chapter 8.

7.1 Audio-Visual Material and Methodology

The dataset provided by ERH contains eye-gaze measurements from 24 hearing impaired test participants and was measured in the Sound Wave Laboratory at ERH in June 2021. A panoramic view capture of the physical test setup used in the Soundwave laboratory at ERH is depicted in Fig. 7.2, while a schematic overview of the test setup can be seen in Fig. 7.2.

The test setup aims to measure the HA user's eye-gaze while they are following

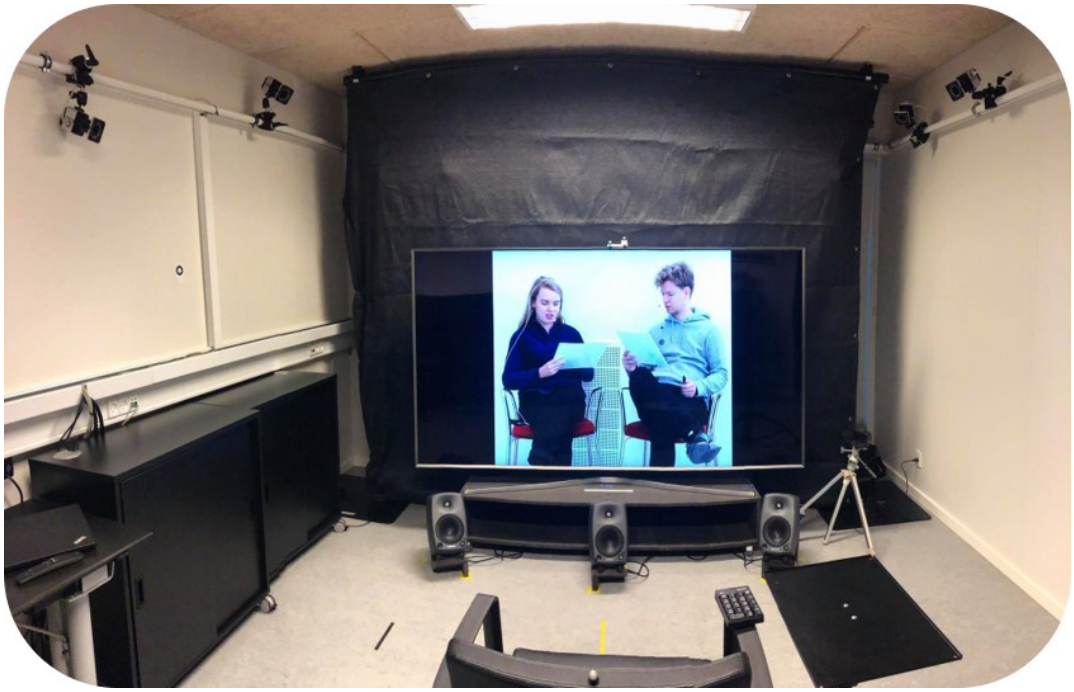


Figure 7.1: Panoramic view capture of the experimental setup used in the Soundwave laboratory at Eriksholm Research Center (ERH). This picture is provided by ERH.

a conversation between two people in a noisy environment. In short, the test setup consists of:

1. A test participant with a head-mounted Tobii Pro 2 eye-tracker.
2. An 88″ curved TV-screen on which audio-visual stimuli is presented. The test participant was seated 1.8 meters away from the TV. The angle between the two talkers presented on the TV is approximately 20 degrees. Note that this angle can vary a bit as the two talkers can move in their sitting position.
3. A total of 10 loudspeakers, where three of the loudspeakers are located in front of the TV screen and seven are located behind the TV in a height so that the TV screen does not block the direct sound from the loudspeakers.

Referring to Fig. 7.2, the two green loudspeakers are used to play the sound from the two conversing talkers presented on the screen, while the red loudspeakers are used to play background noise. The background noise was created as one synthetic babble noise realization played throughout all 8 speakers. The angles between the two loudspeakers used to play the speech signals was measured in the Sound Wave Laboratory to be approximately 27°, with the left loudspeaker placed approximately at 14.6° and the right loudspeaker placed approximately at $-12°$ with 0° defined as the frontal direction from the HA user's point of view, and the azimuth is counter-clockwise rotating. Note that the visual position of the talkers on the TV screen

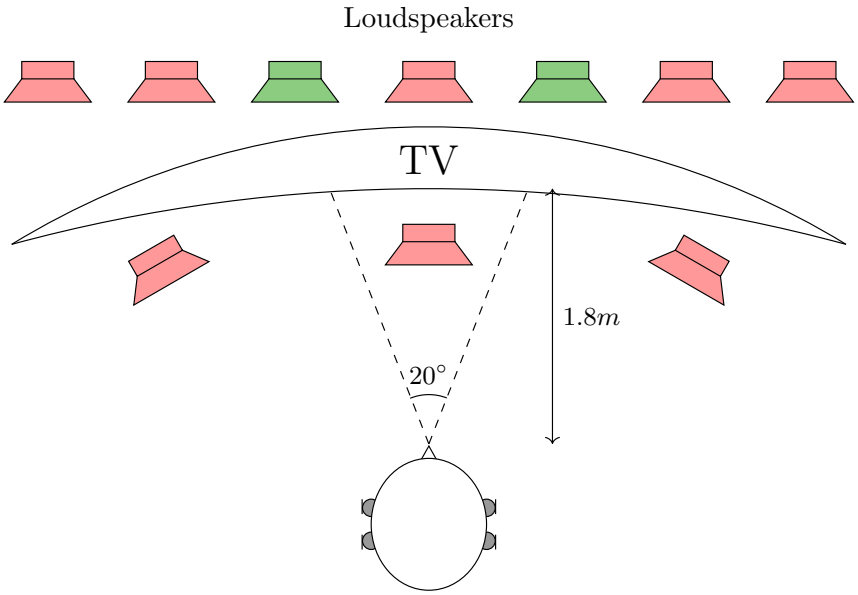


Figure 7.2: A schematic diagram depicting the test setup, showing the location of the speakers and noise sources with respect to the test subject. The test setup aims to measure eye-gaze of the HA user following a conversation between two people in babble noise.

did not precisely coincide with the loudspeaker locations, introducing a mismatch between the visual and acoustic stimuli presented for the test participants. During the recording of the eye-gaze, no fixation of the participants heads were enforced, yet the participants were instructed to keep their head as fixated as possible.

The audio-visual stimuli contains 26 recorded two-talker scenario videos. The talkers were four danish actors, specifically, two male and two female actors. An example of the visual stimuli presented for the participant during the experiments can be seen in Fig. 7.1. In the videos, a dialogue is taking place between the two actors. The dialog stems from a conversation about a picture of a landscape. The pictures differ from the two actors, and their task is to locate the differences in the two pictures. As seen from Fig. 7.1, the actors are wearing hands-free microphones placed close to their mouths to record the clean speech signals from the two talkers in synchronization with the visual stimuli presented on the TV screen. The acoustic stimuli is sampled at a rate of 48 kHz while the Tobii Pro 2 eye-tracker has a sample rate of approximately 50 Hz. The dataset contains audio-visual stimuli for all single trials, i.e., for all 26 audio-visual scenes, and for each scene, the dataset contains associated eye-gaze data for each of the 24 test participant. Hence, we have access to 24 eye-gaze recordings for each audio-visual scene, leaving us with a total of $26 \cdot 24 = 624$ distinct eye-gaze recordings. However, a preliminary examination of the acoustic stimuli, resulted in a number of the audio-visual scenes as well as test participants had to be disqualified. In the following section, we present our thoughts behind the selection of the trials

and test participants to be used.

7.1.1 Data Cleaning of the Audio-Visual Stimuli

In this section, we comment on the disqualification of a number of audio-visual scenes as well as test participants. Firstly, the eye-gaze data from two test participants has been discarded due to malfunctioning of the test equipment, leaving 22 participants for our study. Secondly, of the 26 audio-visual trials presented in the dataset, we deemed only 10 usable for the purpose of this thesis. The disqualification of the remaining 16 trial was due to multiple factors. First of all, a large amount of noise was present in the clean speech tracks, hence, drastically decreasing the SNR and making the task of using an effective VAD challenging. Secondly, when any of the two talkers was active, i.e., in single-talk situations, the active talkers voice leaked into the second talkers microphone, meaning that both talkers would be present in each of the clean target sound tracks. Hence, a lot of cross-talk was introduced in these audio tracks.

The audio-visual data available for this thesis consists only of two-talker scenarios, however, the data originally contain a third talker performing a monologue in the trials. As a last source of error, in all of the disqualified audio-visual trials a lot of cross-talk originating from the third talker was present.

Note that any balancing of the data may have become invalidated by the exclusion of trials and test participants. We do, however, deem that this is acceptable in order to get rid of the aforementioned defects in the excluded trials and test participants eye-gaze measurements, as such defects could potenitally harm the performance of our beamforming systems.

7.2 Simulating HA Microphone signals

In order to simulate the acoustic stimuli that were presented to the users via their HAs during the experiments, the general framework for simulating acoustic scenes described in Section 5.1.1, will be used. However, to be able to use this framework in the situation where real-world data is considered, some modifications to the framework used in the feasibility study needs to be made. Specifically, in this study, we simulate from a signal model on the form

$$\boldsymbol{x}(n) = (s_1 * \boldsymbol{a}(\bullet, \theta_{s_1}))(n) + (s_2 * \boldsymbol{a}(\bullet, \theta_{s_2}))(n) + \boldsymbol{v}(n), \tag{7.1}$$

where $s_1(n)$ and $s_2(n)$ are the target signals measured at the source locations θ_{s_1} and θ_{s_2}, respectively, the vectors $\boldsymbol{a}(n, \theta_{s_1})$ and $\boldsymbol{a}(n, \theta_{s_2})$contain the AIRs from the target speakers to each of the M microphones on the HA, and $\boldsymbol{v}(n)$ is an additive noise term. To simulate from the signal model in (7.1), we use the audio tracks from the dataset provided by ERH as the clean target speech signals $s_1(n)$ and $s_2(n)$, while we us the AIRs from a database provided by Oticon used to simulate the wave propagation from a sound source to the microphones on the HAs. As described in Section 7.1,

the two loudspeakers playing speech were placed at approximately 14.6° and −12°, respectively, hence, given the angular resolution of our RTF database, we set $\theta_{s_1} = 15°$ and $\theta_{s_2} = -15°$. As the clean speech signals from the audio-visual dataset are sampled at 48 kHz, while the AIRs provided by Oticon are sampled at 44.1 kHz, the clean speech signals are downsampled to 44.1 kHz. This downsampling is done in accordance to the description in Section 5.1.1. After the noisy microphone signals are generated, we downsample the signals to 16 kHz such that sampling frequency of the microphone signals coincides with the sampling frequency at which we process our beamforming systems.

In the feasibility study in Chapter 5, the noise field was created approximately isotropic by placing talkers in 48 evenly spaced points on a circle with the HA user in the center. Here, however, as we aim to reproduce the test setup from the ERH experiment as closely as possible, since changes in the setup may have affected the eye gaze, we let noise only impinge from angles in the range $[-45°, 45°]$, i.e., from 13 directions given the resolution of our RTF database. Note that in this way, our simulated noise field does still not perfectly match the acoustic stimuli from the experiments, however, based on the available RTF database, we deem that this is our best option.

In the generation of each scene, 13 babble noise tracks are chosen at random to be simulated from each of the 13 loudspeakers represented in Fig. 7.3. These babble noise tracks are chosen from a set of pre-computed tracks, each consisting of talk from eight TIMIT audio tracks. Note that in the ERH experiment, the same noise realization was used for all noise sources. Though, we aim to simulate acoustic stimuli as close to those in the ERH experiment, we choose to use different noise realizations for each noise source. This is due to the fact that, if the beamforming system was to be applied in a real world setting, noise impinging from different directions would certainly vary. Also, from a mathematical standpoint, letting the same noise realization impinge from all noise sources, may cause the the noise CPSD to not have full rank, causing it to not be invertible, hence making the beamformer weights unobtainable.

Another place where we choose to simulate slight different compare to the experiment conducted at ERH is that we let speech and noise be co-located, whereas in the ERH experiment speech and noise had dedicated speakers.

Finally, regarding the SNR under which the experiments at ERH was conducted. In the experiment conducted at Eriksholm, the SNR was set to 0 dB. As mentioned earlier, we aim to comply as much as possible with their setup, and hence likewise simulate our acoustic scenes at 0 dB. When simulating the acoustic scenes, the input SNR is controlled in accordance to the description in Section 5.1.1. Do note that when we compute the gain to obtain a SNR of 0 dB, we combine $s_1(n)$ and $s_2(n)$ to a single signal before the average power is computed.

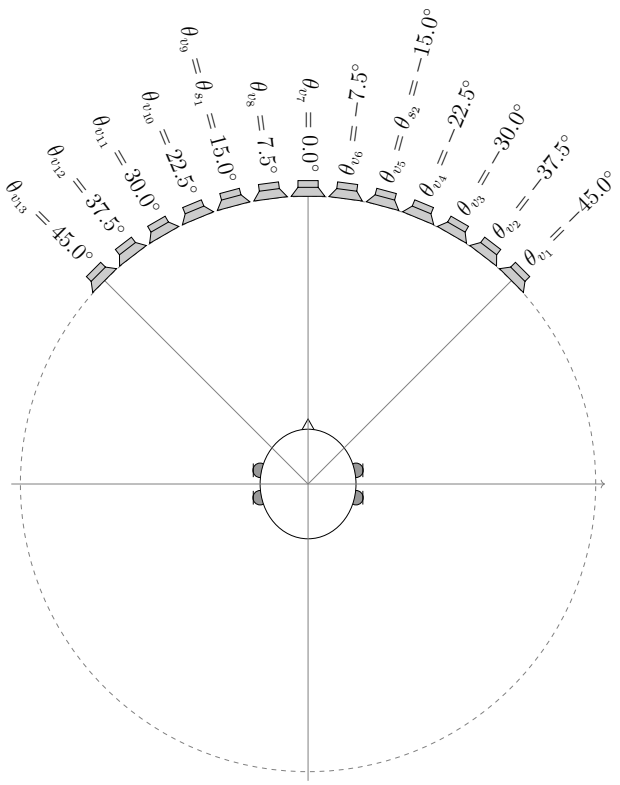


Figure 7.3: Illustration of how the acoustic stimuli presented at the ERH experiments are synthetic simulated in this thesis. The gray loudspeakers indicated the discrete sound source directions.

7.3 Eye-Gaze Data Prepocessing

In order to be able to use the eye-gaze data provided by ERH, a few key aspects must be considered. First of all, the audio-visual material is ensured to be properly time-aligned, and the eye-gaze data have been parsed by ERH such that is correctly aligned with the audio-visual material and is precise up to 20 ms. However, the audio signals, which are measured by microphones, and the eye-gaze measurements, which are recorded by cameras from an eye-tracker, are sampled at different rates. Hence, synchronization of the data is still needed for us to be able to integrated the signals from the two types of sensors to be used in our beamforming systems. Secondly, the eye-gaze data contains missing data points, and lastly, during the first 4 seconds in each trial, the test participants were instructed to look at a fixation cross in the center of the screen, i.e., 0° from the test participants point-of-view, which appeared on the screen preceding the presentation of the audio-visual stimuli in each trial. In general, the participants eye-gaze were steady in the first 4 seconds where the fixations cross was displayed, however, not necessarily at 0°, meaning that an offset of the fixation is observed. As a consequence, this introduces an offset in the eye-gaze measurement angles. In the following, we describe how we choose to preprocess the eye-gaze data,

and in doing so, obtain eye-gaze data which can be applied for the specific purpose of this thesis.

For each of the 10 trials and 22 test participants, the following processing of the eye-gaze measurements from the dataset is applied. The eye-gaze measurements are provided as (x, y, z) coordinates in a three dimensional grid. As described in Chapter 6, for simplicity, we have designed our proposed beamforming systems to take into account the HA user's eye-gaze by letting $\phi(l)$ representing the user's azimuth eye-gaze angle at time l. Therefore, we transform the $3D$-coordinates into azimuth angles in degrees.

After having observed the eye-gaze measurements in degrees with associated timestamps, we seek to correct the offset in measurement angles. In Fig. 7.4, an example of an offset in the eye-gaze measurements is depicted. Worth noting in the figure is the spikes, which occur very suddenly. These may either be caused by sudden eye movements or may as well be measurement errors.

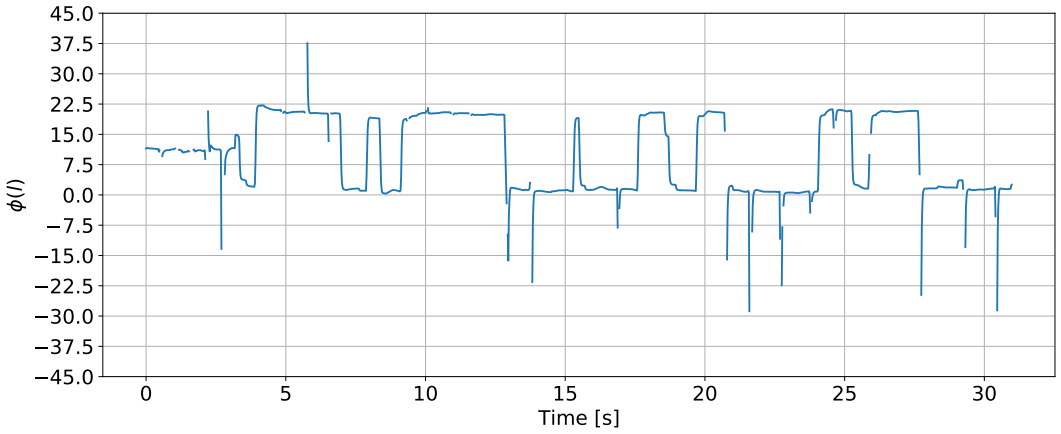


Figure 7.4: Example of eye-gaze measurements as a function of time.

From Fig. 7.4, it is seen that for the first few seconds, the data curve is located around $\approx 12^\circ$ and not at 0°. We correct the offset by finding the mean over the measurements taken during the fixation cross, omitting any missing data points, and subtract this mean from all measurements.

Next, we deal with the problem of missing data. We do this in a forward step and a backward step. First we propagate forwards, in regards to time, in the eye-gaze data and setting any NaN values, i.e., missing data points, equal to the previous non-NaN value. The forward step does, however, not compensate for missing data in the beginning of the data, i.e., in the case where the first data point containing a NaN value. Therefore, beginning from the end of the eye-gaze data, we propagate backwards in regards to time, and set any NaN values equal to the next non-NaN value. This procedure results in no missing data points.

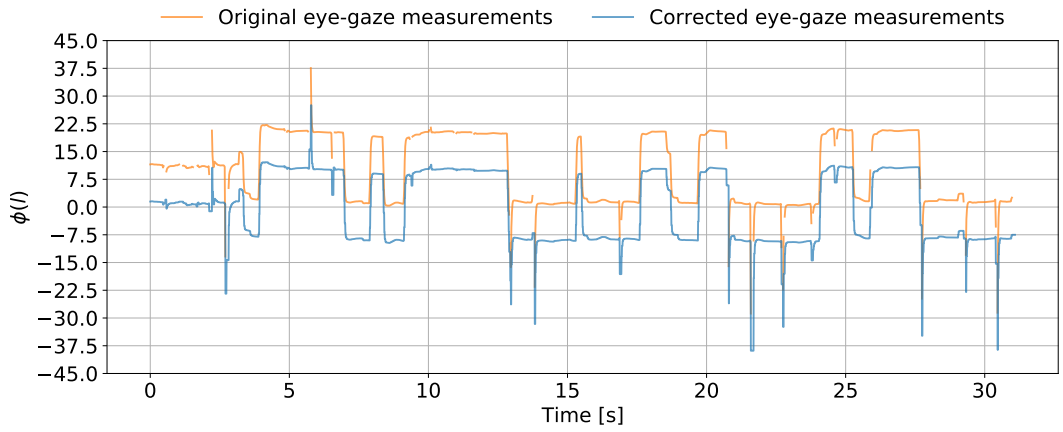
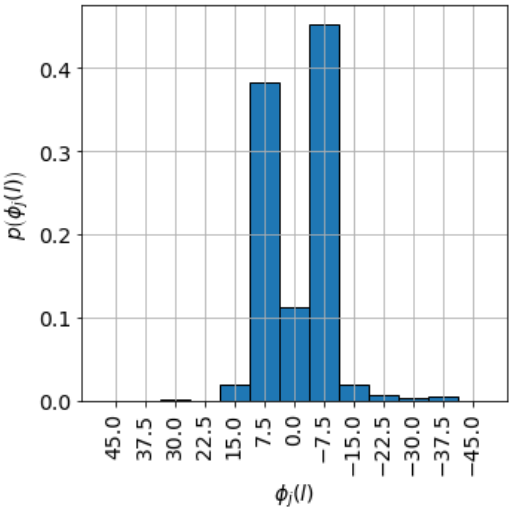


Figure 7.5: Example of eye-gaze measurements with correction as a function of time. The blue signal is the corrected eye-gaze signal while the orange signal is the raw eye-gaze signal.

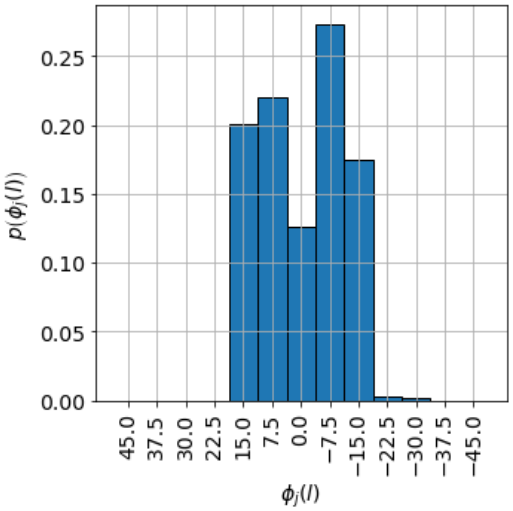
Finally, we deal with the resampling the eye-gaze data. As our beamforming algorithms use a time-frequency representation of the microphone signals, the audio frame rate is determined by the window length and the hop-size chosen for the STFT. As mentioned in Section 7.2, the audio signals from the dataset are downsampled from $f_{s,\text{audio}} = 48$ kHz to $f_{s,\text{audio}} = 16$ kHz, meaning that our temporal resolution of the STFT coefficients corresponds to 125 time frames per second. In order to obtain a samplerate for the eye-gaze measurements which coincide with the rest of our simulation framework, we therefore resample the eye-tracking frames to match this temporal dimension of the STFT of the microphone signals. We do this in the following manner. From the first timestamp, we create new timestamps spaced 8 ms apart, corresponding to a frame rate of 125 Hz. We can now interpolate values at the new timestamps based on the eye-gaze measurements at the original timestamps. We interpolate with a zero-order hold, meaning that the interpolated value at a timestamp is set to the most resent value of the original data. The result of the processing steps are illustrated in Fig. 7.5, where the blue signal is the corrected eye-gaze signal, while the orange signal is the raw eye-gaze signal.

Via the preprocessing of the eye-gaze data described in this section, we have obtained eye-gaze data for 10 scenes and 22 test participants, which has been corrected for offset based on the data during the fixation cross, corrected for missing data, and resampled to a samplerate which coincide with the rest of our simulation framework.

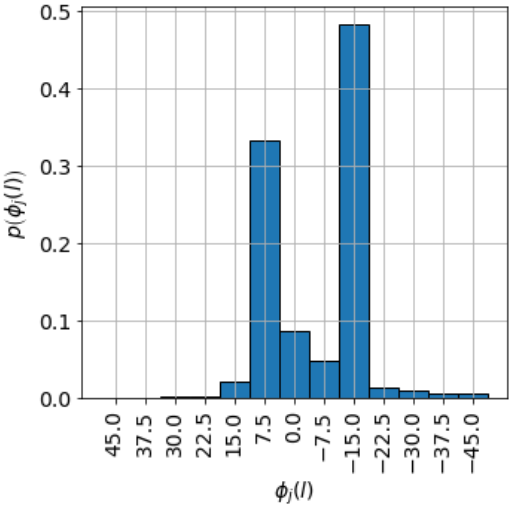
As we in this thesis are interested in the distribution of the eye-gaze measurements, we in Fig. 7.6 depict examples of histograms of eye-gaze measurements taken over whole trials. Specifically, in order to give insight into the variance of the eye-gaze measurements over trials and test participants, we illustrate histograms of two test participants' eye-gaze measurements for two trials. From Fig. 7.6, we clearly see two



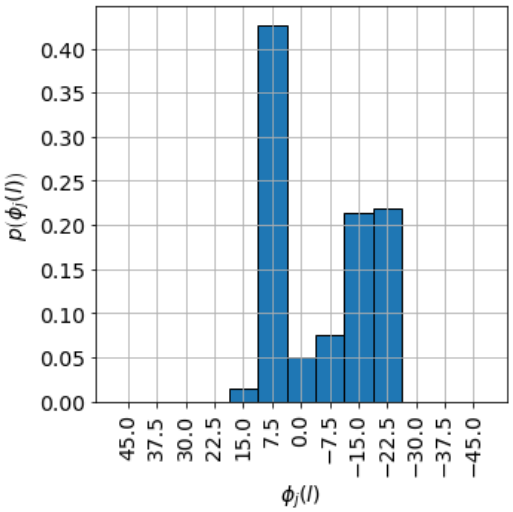
(a) Histogram density estimate of $p(\phi_j(l))$ for example test participant 1, example trial 1.



(b) Histogram density estimate of $p(\phi_j(l))$ for example test participant 1, example trial 2.



(c) Histogram density estimate of $p(\phi_j(l))$ for example test participant 2, example trial 1.



(d) Histogram density estimate of $p(\phi_j(l))$ for example test participant 2, example trial 2.

Figure 7.6: Histogram density estimate of $p(\phi_j(l))$ for e of two test participants' eye-gaze measurements for two trials.

modes in the histograms, corresponding to the two targets. Interesting is that we see quite different histograms both across test participants but also across trials.

In the following section, we will describe how we use the preprocessed eye-gaze signals to construct the pre-computed look-up table of conditional probabilities used in the one of our proposed eye-gaze based Bayesian beamforming methods.

7.4 Computed Look-Up Table for Proposed Joint Audio-Gaze Bayesian Beamformer

As described in Section 6.3.1, in this thesis, we propose a joint audio-gaze Bayesian beamforming method in which the HA user's eye-gaze is incorporated by means of a pre-computed look-up table with conditional PMFs $p(\theta_i|\phi_j(l))$, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. We have, however, not discussed how to actual compute this look-up table using the real-world eye-gaze data. Having presented the eye-gaze data, we are able to compute the look-up table which is used to implement the proposed joint audio-gaze Bayesian beamformer. As mentioned, the conditional PMF $p(\theta_i|\phi_j(l))$ describes the probability that the target signal arrives from direction θ_i given that the HA user's eye-gaze angle is $\phi_j(l)$. This conditional PMFs can be obtained from the other conditional PMF $p(\phi_j(l)|\theta_i)$, for $i = 1 \ldots, I$ and $j = 1, \ldots, J$. In contrast to $p(\theta_i|\phi_j(l))$, $p(\phi_j(l)|\theta_i)$ describes the probability that the HA user's eye-gaze is towards $\phi_j(l)$ given that the target signal arrives from direction θ_i. In order to obtain the value of $p(\phi_j(l)|\theta_i)$, we may count all fraction of situations where the HA user's eye-gaze is towards $\phi_j(l)$ when the target signal arrives from direction θ_i. Hence, to be able to compute $p(\phi_j(l)|\theta_i)$, we are in the need of an ideal VAD to classify a given time frame l as being speech-absent, single-talk, or double-talk in order to determine a unique θ_i. We begin this section by a description of the VAD that we have implemented in order to make this classification.

In the proceeding, we consider time frames of speech absence, single-talk and double-talk and hence introduce the following notation. Let $s_p^l(n)$ for $p = 1, 2$ and $n = 0, \ldots, N-1$, denote the l'th segmented time domain signal obtained from the windowing process of the STFT. We then introduce the following notation for the l'th time frame:

- $\mathcal{S}_0$: Neither $s_1^l(n)$ nor $s_2^l(n)$ contains speech.
- $\mathcal{S}_1$: $s_1^l(n)$ contains speech but $s_2^l(n)$ does not (single-talk for).
- $\mathcal{S}_2$: $s_1^l(n)$ does not contain speech but $s_2^l(n)$ does.
- $\mathcal{S}_3$: Both $s_1^l(n)$ and $s_2^l(n)$ contains speech.

As these states are identified by the VAD, we refer to them as VAD states. The VAD used to identify these states are described in the following.

7.4.1 Voice Activity Detection

This section serves the purpose of introducing a frame based VAD which is used to detect speech-presence in a two-talker scenario. The VAD assumes access to the clean time-domain target signals $s_1(n)$ and $s_2(n)$, in order to classify each time frame as belonging to one of the four states $\mathcal{S}_0$, $\mathcal{S}_1$, $\mathcal{S}_2$, or $\mathcal{S}_3$. Important to note is that the beamformers are applied for each time-frequency tile individually. However, we choose to assign one state to entire time frames, i.e., across frequency. Alternatively, the VAD could have classified each time-frequency tile individually. This approach could potentially have lessened the amount of double-talk detected, as two talkers would most likely not occupy the exact same frequency bins simultaneously. Though this could prove beneficial, the process of threshold calibration in the VAD would become more challenging, hence we choose a frame based VAD.

The proceeding outline of the VAD describes application for a single trial, i.e., a single acoustic scene. The VAD operates on the time domain signals $s_1(n)$ and $s_2(n)$, however, in frames corresponding to the frames of the spectro-temporal signals $\tilde{s}_1(k,l)$ and $\tilde{s}_2(k,l)$. Therefore, we apply the same windowing process as in the STFT, which is described in the following.

For $s_p(n), p = 1, 2$ the following procedure is performed. First the clean signal is zero-padded in the front by zeros corresponding to one second, as to make the length of the clean speech signal match the length of the noise signal. Afterwards, the signal is zero-padded at the front and end by half the window length, i.e., $N/2 = 128$, which in terms of the STFT, centers the first window segment on the first data point. Thereafter, a number of zeros is appended to the end of the signal in order to make the signal fit precisely into an integer value of window segments. With an overlap of 50%, the number of necessary zeros can be computed as $N/2 - (N_s \bmod (N/2))$, where N_s is the number of samples in $s_p(n)$, and mod is the modulo operator. Next, $s_p(n)$ is segmented with the square-root Hann window sequence $w(n)$ and the hop size $D = N/2 = 128$, i.e., we let $s_p^l(n) = s_p(n + lD)w(n)$, for $n = 0, \ldots, N-1$ denote the l'th segment of $s_p(n)$. Then, the energy of the l'th frame of $s_p(n)$, which we denote $E_{s_p}(l)$, is computed as

$$E_{s_p}(l) = 10 \log_{10} \left(\sum_{n=0}^{N-1} \left(s_p^l(n) \right)^2 + \epsilon \right) \quad \text{dB}, \quad \forall l,$$

where ϵ is a small number to avoid taking the logarithm of zero. Let $E_{\max,p}$ denote the largest $E_{s_p}(l)$ across all window segments, we then define the interim binary VAD outputs as

$$\alpha_p(l) = \begin{cases} 1, & \text{if } E_{\max,p} - E_{s_p}(l) < 42 \ \text{ dB} \\ 0, & \text{otherwise} \end{cases}, \quad \forall l,$$

where the threshold of 42 dB has been chosen based on visual inspection of example eye-gaze signals.

Based on $\alpha_p(l)$, for $p = 1, 2$, the VAD state for the l'th time frame is determined as

$$\begin{aligned}
\mathcal{S}_0 &: \text{ if } \alpha_1(l) = 0 \text{ and } \alpha_2(l) = 0 \quad \text{(speech absence)} \\
\mathcal{S}_1 &: \text{ if } \alpha_1(l) = 1 \text{ and } \alpha_2(l) = 0 \quad \text{(single-talk for } s_1^l(n)\text{)} \\
\mathcal{S}_2 &: \text{ if } \alpha_1(l) = 0 \text{ and } \alpha_2(l) = 1 \quad \text{(single-talk for } s_2^l(n)\text{)} \\
\mathcal{S}_3 &: \text{ if } \alpha_1(l) = 1 \text{ and } \alpha_2(l) = 1 \quad \text{(double-talk).}
\end{aligned}$$

The computation of the VAD is summarized in Algorithm 6 as pseudo-code.

Algorithm 6 Voice Activity Detector

Input:

$s_1(n)$, $s_2(n)$, for $n = 0, \ldots, N_s - 1$: Clean target speech signals in the time domain.

Output:

$\mathcal{S}_i, i \in \{0, 1, 2, 3\}$: VAD state for each time frame l.

1: **for** $p = 1, 2$ **do**
2: Zero-pad $s_p(n)$ to comply with STFT.
3: Segment $s_p(n)$ into overlapping segments as $s_p^l(n) = s_p(n + lD)w(n)$.
4: **for all** l **do**
5: Compute frame energy as $E_{s_p}(l) = 10 \log_{10} \left(\sum_{n=0}^{N-1} \left(s_p^l(n) \right)^2 + \epsilon \right)$.
6: **end for**
7: Compute maximum frame energy: $E_{\max,p} = \max_l (E_{s_p}(l))$.
8: **for all** l **do**
9: Compute interim binary VAD as

$$\alpha_p(l) = \begin{cases} 1, & \text{if } E_{\max,p} - E_{s_p}(l) < 42, \\ 0, & \text{otherwise.} \end{cases}$$

10: **end for**
11: **end for**
12: **for all** l **do**
13: Determine VAD state:

$$\begin{aligned}
\mathcal{S}_0 &: \text{ if } \alpha_1(l) = 0 \text{ and } \alpha_2(l) = 0 \\
\mathcal{S}_1 &: \text{ if } \alpha_1(l) = 1 \text{ and } \alpha_2(l) = 0 \\
\mathcal{S}_2 &: \text{ if } \alpha_1(l) = 0 \text{ and } \alpha_2(l) = 1 \\
\mathcal{S}_3 &: \text{ if } \alpha_1(l) = 1 \text{ and } \alpha_2(l) = 1.
\end{aligned}$$

14: **end for**

Having introduced to the VAD, we finally move on to the actual computation

of the look-up table with conditional PMFs used in the proposed joint audio-gaze beamforming system.

7.4.2 Computation of Conditional PMF

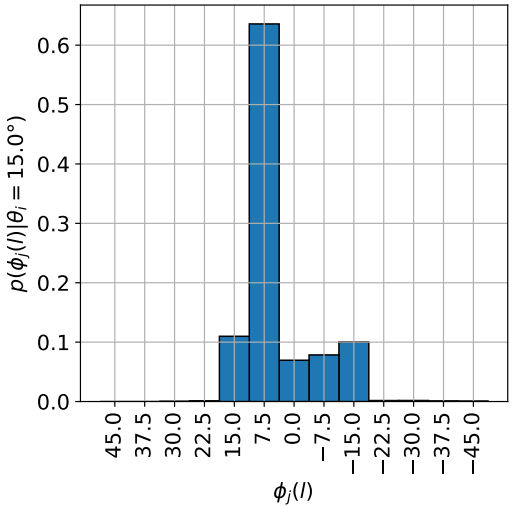
To construct the conditional PMFs $p(\phi_j(l)|\theta_i)$, ideally, we would estimate the distribution of $\phi_j(l)$ given θ_i for all i. This would be done by collecting all available eye-gaze angles $\phi_j(l)$ for each target DOA θ_i, and use these measurements for the density estimation. However, as apparent from Section 7.2, in the experiment conducted at ERH, the participants were only placed in two positions, namely at $\theta_i = \pm 15^\circ$, hence based on the data, we are only able to fill in the two rows of Table 6.2 corresponding to $\theta_i = \pm 15^\circ$. In a case where the available data have had target talkers located in each of the I directions, we would be able to construct $p(\phi_j(l)|\theta_i)$ for each of the target DOAs θ_i, for $i = 1, \ldots, I$, based on real-world eye-gaze data. However, since we are not able to estimate $p(\phi_j(l)|\theta_i)$ for $\theta_i \in \Theta \backslash \{-15^\circ, 15^\circ\}$ directly from the available data, we have to decide upon how to obtain the remaining rows of the PMF matrix in Table 6.2. Out of necessity, we have decided to synthetically fill the remaining rows of $p(\phi_j(l)|\theta_i)$ based on the actual observed densities $p(\phi_j(l)|\theta_i)$ for $\theta_i = \pm 15^\circ$. In the proceeding, we first describe how $p(\phi_j(l)|\theta_i = 15^\circ)$ and $p(\phi_j(l)|\theta_i = -15^\circ)$ are computed directly from the eye-gaze data. Secondly, we describe how we have decided to compute $p(\phi_j(l)|\theta_i = 0^\circ)$, and following this, we describe the computation of $p(\phi_j(l)|\theta_i = 7.5^\circ)$ and $p(\phi_j(l)|\theta_i = -7.5^\circ)$, respectively. Afterwards, we describe how $p(\phi_j(l)|\theta_i)$ for $\theta_i \in \{\pm 22.5^\circ, \pm 30^\circ, \pm 37.5^\circ, \pm 45^\circ\}$ is computed, and finally, we discuss how to form $p(\phi_j(l)|\theta_i)$ when the target arrives from outside the DOA range $\theta_i = \pm 45^\circ$.

As described in Section 7.1.1, the data consists of 10 trial each with 22 associated eye-gaze measurements. To determine how to best utilize this data in the look-up table, we have experimented with three different methods. Firstly, eye-gaze data from all scenes and all test participants have been used to construct one "global" look-up table. Secondly, we have created a look-up table for each scene in which data from all test participants was used, with the intent of having a tailored look-up table for each scene. Lastly, we have tested with a participant-specific look-up table so that each participant has their own tailored look-up table. Based on experiments where each of the three different look-up tables have been used to implement the proposed Bayesian beamformer, the highest performance scores were obtained when using the participant-specific look-up tables. Therefore, we choose to include this look-up table in the beamformer evaluation in Chapter 8.

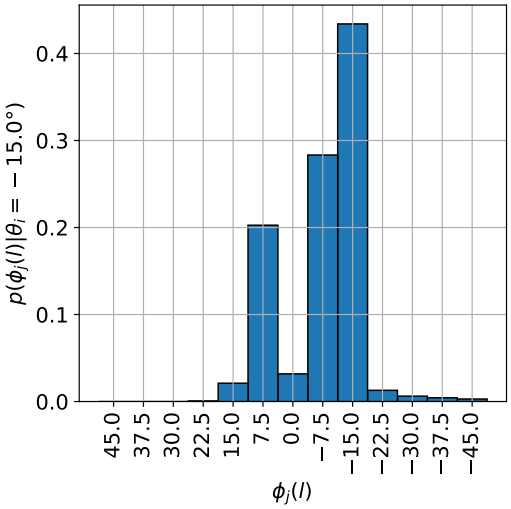
Computation of $p(\phi_j(l)|\theta_i)$ for $\theta_i = \pm 15^\circ$

To find the density $p(\phi_j|\theta_i = 15^\circ)$, we gather the eye-gaze measurements from all 10 trials with VAD state $\mathcal{S}_1$ according to Algorithm 6. We repeat this procedure for the VAD state $\mathcal{S}_2$, which gives two sets of data from which we can estimate the

probabilities $p(\phi_j|\theta_i = 15°)$ and $p(\phi_j|\theta_i = -15°)$. Note that eye-gaze measurements at time frames with VAD state $\mathcal{S}_3$, i.e., double-talk, is excluded from both groupings of data. As described in Section 6.3, we choose to estimate the densities via the histogram approach described in Section 6.2.2. Examples of the estimated densities $p(\phi_j|\theta_i = 15°)$ and $p(\phi_j|\theta_i = -15°)$ can be seen in Fig. 7.7a and Fig. 7.7b, respectively. From the figures it is seen that the histogram of $p(\phi_j|\theta_i = 15°)$ has a clear spike in 7.5° while $p(\phi_j|\theta_i = 15°)$ has a spike in $-15°$ but does, however, also contain noticeably mass around $-15°$. As mentioned, there is a mismatch between the visual position of the target talkers on the TV screen and the acoustic positions, i.e., the positions of the loudspeakers playing the sound signal from the target talkers. According to the simulated target DOA, in Fig. 7.7a, we would expect a spike in 15°, hence, the deviation from this spike may be a consequence of this mismatch between the acoustic and visual stimuli.



(a) Histogram density estimates of $p(\phi_j(l)|\theta = 15°)$ for a specific test participant.



(b) Histogram density estimates of $p(\phi_j(l)|\theta = -15°)$ for a specific test participant.

Computation of $p(\phi_j(l)|\theta_i)$ for $\theta_i = 0°$

To synthetically create $p(\phi_j(l)|\theta_i = 0°)$, we use $p(\phi_j(l)|\theta_i)$ for $\theta_i = \pm 15°$. From visual inspection of $p(\phi_j|\theta_i = 15°)$ and $p(\phi_j|\theta_i = -15°)$ in Fig. 7.7a and Fig. 7.7b, respectively, though difficult to see due to the limited resolution, it might seem plausible that the densities takes on the shape of bell curves, though slightly skewed with longer tails towards to center. Since $p(\phi_j|\theta_i = 15°)$ and $p(\phi_j|\theta_i = -15°)$ are skewed in opposite direction, it seems fair to assume that $p(\phi_j|\theta_i = 0°)$ will be symmetric around 0°, and given the shape of $p(\phi_j|\theta_i = 15°)$ and $p(\phi_j|\theta_i = -15°)$, we choose to model $p(\phi_j|\theta_i = 0°)$ as a Gaussian distribution. To find the standard deviation of this Gaussian distribution, we find the sample standard deviation of the two sets of data for $\theta_i = 15°$ and $\theta_i = -15°$, respectively, and take the average. Next, we sample a

zero-mean Gaussian distribution with the aforementioned standard deviation in the J discrete angles. We assume that the eye-gaze cannot assume values beyond 45° in either direction, therefore, we set the probability outside this range to zero. At last we normalize $p(\phi_j|\theta_i = 0°)$ so that the entries sum to one. Do note that after the truncation, the distribution is not Gaussian anymore, however, we simulated from the Gaussian distribution due to the general shape, hence the exact distribution is not important. An example of the estimated density $p(\phi_j|\theta_i = 0°)$ is illustrated in Fig. 7.8.

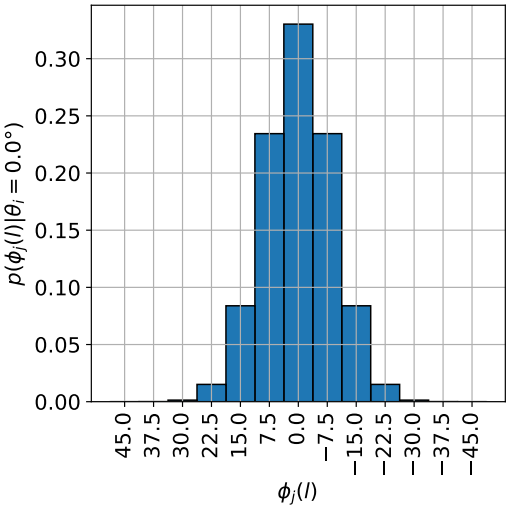
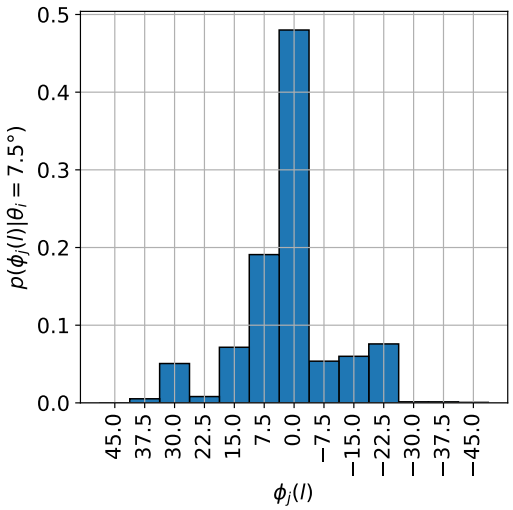


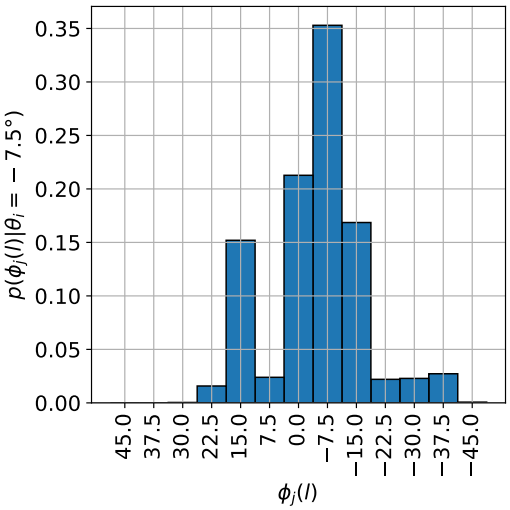
Figure 7.8: Histogram density estimates of $p(\phi_j(l)|\theta = 0°)$ for a specific test participant.

Computation of $p(\phi_j(l)|\theta_i)$ for $\theta_i = \pm 7.5°$

In the construction of $p(\phi_j(l)|\theta_i)$ for $\theta_i = \pm 7.5°$, we wish to retain the information observed in $p(\phi_j(l)|\theta_i)$ for $\theta_i = \pm 15°$, yet reflect the shift in angles. We choose to do this in the following manner. The density $p(\phi_j|\theta_i = 7.5°)$ is constructed as a linear combination of $p(\phi_j|\theta_i = 15°)$ and $p(\phi_j|\theta_i = -15°)$. To do so, we first rotate $p(\phi_j|\theta_i = 15°)$ by $-7.5°$ and weight the density by $\frac{3}{4}$. Next, we rotate $p(\phi_j|\theta_i = -15°)$ by 22.5° and weight the density by $\frac{1}{4}$. The weighting is chosen due to the distance between the angles. These two rotated and weighted densities are then added together. Afterwards, the density is normalized so the entries sum to one, which gives the density $p(\phi_j|\theta_i = 7.5°)$. Likewise, $p(\phi_j|\theta_i = -7.5°)$ is constructed in the same manner where $p(\phi_j|\theta_i = 15°)$ is rotated $-22.5°$ and weighted by $\frac{1}{4}$ while $p(\phi_j|\theta_i = -15°)$ is rotated 7.5° and weighted by $\frac{3}{4}$. Examples of the densities $p(\phi_j|\theta_i = 7.5°)$ and $p(\phi_j|\theta_i = -7.5°)$ are depicted in Fig. 7.9a and Fig. 7.9b, respectively. Form the figures, it is seen that the histograms span a more wide range of directions, which was to be expected due the they way they are constructed.



(a) Histogram density estimates of $p(\phi_j(l)|\theta = 7.5°)$ for a specific test participant.



(b) Histogram density estimates of $p(\phi_j(l)|\theta = -7.5°)$ for a specific test participant.

Computation of $p(\phi_j(l)|\theta_i)$ for $\theta_i \in \{\pm 22.5°, \pm 30°, \pm 37.5°, \pm 45°\}$

Due to restricted possible rotation of the human eye, one could easily imagine that the skewness of the distribution of the look direction would increase as the target talker departed further from the frontal direction. However, since it is difficult to model the behaviour of the user's eye-gaze in these situations such that this model would reflect the eye-gaze behaviour properly, we choose to naively estimate the density $p(\phi_j|\theta_i = 22.5°)$ as $p(\phi_j|\theta_i = 15°)$ rotated by 7.5°, likewise we estimate $p(\phi_j|\theta_i = 30°)$ by rotating $p(\phi_j|\theta_i = 7.5°)$ by 15°, etc. The distributions for the target directions $\theta_i = -22.5°, -30°, -37.5°, -45°$ are constructed by negative rotation of $p(\phi_j|\theta_i = -15°)$. It is important emphasize that we do not finds this approach to be ideal, however, necessary due to the limitations regarding the data as well as due to time constraints. Note that we have chosen to use this approach to fill conditional PMFs only up to ±45°. This is due to the fact that, in order to be able to compare our proposed systems with each other and with competing methods, it is desirable to have a fixed DOA range. From all eye-gaze measurements, we found that 99.99% had a magnitude below 48.75°, which is the outer edge of the bin associated with 45°.

Computation of $p(\phi_j(l)|\theta_i)$ for $|\theta_i| > 45°$

Lastly, we consider the computation of $p(\phi_j(l)|\theta_i)$ when the target arrives from outside ±45°. In the case where the target is located at a direction exceeding ±45°, it does not make much sense to try to locate mass around the target position. Therefore, we choose to form a PMF with mass inside ±45°. Next, comes the question of which shape this PMF should take. As our eye-gaze data provides no such information, we choose to let $p(\phi_j(l)|\theta_i)$ have equal mass in the directions $-45°, -37.5°, \ldots, 45°$ and

zero probability elsewhere.

Now, we have obtained the conditional PMFs $p(\phi_j(l)|\theta_i)$, for $I = 1, \ldots, I$ and $j = 1, \ldots, J$, and hence, we can finally determine the values of the pre-computed look-up table $p(\theta_i|\phi_j(l))$, for $I = 1, \ldots, I$ and $j = 1, \ldots, J$.

Computation of $p(\theta_i|\phi_j(l))$

In order to obtain the values of $p(\theta_i|\phi_j(l))$ in (6.25) of the look-up table, we are in need of the prior probability $p(\theta_i)$, for $i = 1, \ldots, I$. We choose to construct this prior probability as having equal mass in the angles $\theta_i \in \{-45°, \ldots, 45°\}$ and zero probability elsewhere. Finally, the normalization constant $p(\phi_j(l))$ for $j = 1, \ldots, J$, can be computed from the joint PMF $p(\phi_j|\theta_i)p(\theta_i)$ as $p(\phi_j) = \sum_{i=1}^{I} p(\phi_j|\theta_i)p(\theta_i)$.

In Fig. 7.10, an example of a column of the look-up table $p(\theta_i|\phi_j(l))$ is depicted. Specifically, the plot shows the estimated density $p(\theta_i|\phi_j(l) = 15°)$. From the figure,

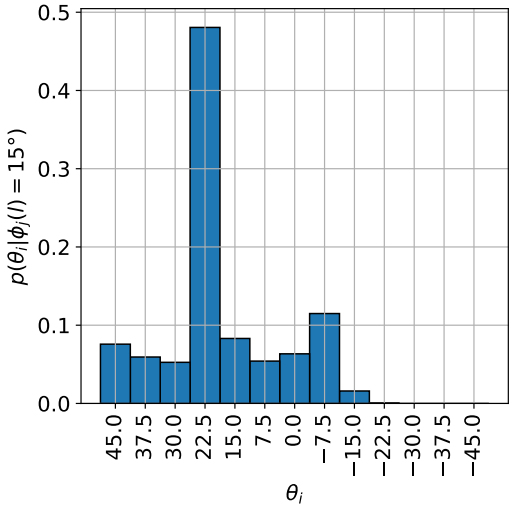


Figure 7.10: Example of conditional probability from the look-up table $p(\theta_i|\phi_j(l))$, specifically, $p(\theta_i|\phi_j(l) = 15°)$.

we unexpectedly see a large spike in $\theta_i = 22.5°$. However, since $p(\phi_j(l)|\theta_i = 15°)$ has a large spike in $\phi_j(l) = 7.5°$, the joint PMF $p(\phi_j(l)|\theta_i = 22.5°)$ will have the same spike in $\phi_j(l) = 15°$, due to the synthetic rotation. This indicates that, using $p(\phi_j(l)|\theta_i)$ for $\theta_i = \pm 15°$ to synthetically fill $p(\phi_j(l)|\theta_i)$, is far from ideal, which in fact may influence the performance of the proposed beamforming system which rely on this pre-computed look-up table to enhance the target speech signal. However, we still decide to use this method, as our options have been limited due to the data available.

Based on the audio-visual simulation study presented in this chapter, in the next chapter, we will evaluate the performance of our proposed beamforming methods.

8. Performance Evaluation of Proposed Audio-Gaze Beamforming Methods

The main purpose of the simulation experiments conducted in this chapter is to examine if an eye-gaze based approach is able to outperform audio-only ones in realistic numerical simulations. The chapter is organized by first summarizing the implemented beamforming systems that will be evaluated and compared in this chapter through numerical experiments. Following this, we present the experiential setup for the performance evaluation, and finally, the results from the simulation experiments are presented. In Chapter 9, we will discuss the results found in this chapter.

8.1 Beamformer Evaluation

In this section, various beamforming systems are compared in order to access the potential of using eye-gaze information in beamforming systems, in addition to acoustic information, to enhance a target sound signal. We begin this section by summarizing the implemented beamforming systems.

- **MVDR-Ideal** To get an indication of an upper bound performance on the noise reduction an oracle MVDR beamformer, denoted MVDR-Ideal, is implemented according to Section 5.1.2. MVDR-Ideal has access to the true target DOA and optimal noise statistics, however, due to the fact that the acoustic scenes comprise two-talker scenarios, some choices have to me made in relation to determining the true target DOA θ_s for each time l frame in

$$\hat{\boldsymbol{w}}_{\text{MVDR}}(k,l,\theta_s) = \frac{\hat{\boldsymbol{C}}_{\boldsymbol{v}}^{-1}(k,l)\boldsymbol{d}(k,l,\theta_s)}{\boldsymbol{d}^H(k,l,\theta_s)\hat{\boldsymbol{C}}_{\boldsymbol{v}}^{-1}(k,l)\boldsymbol{d}(k,l,\theta_s)}. \tag{8.1}$$

  The MVDR-Ideal assumes access to the ideal VAD, described in Section 7.4.1, to classify time frames as belonging to one of the four VAD states $\mathcal{S}_0$: speech absence, $\mathcal{S}_1$: single-talk for $s_1^l(n)$, $\mathcal{S}_2$: single-talk for $s_2^l(n)$, and $\mathcal{S}_3$: double-talk.

The RTF vector $\boldsymbol{d}(k,l,\theta_s)$ used to implement the ideal MVDR beamformer in (8.1) is determined by the VAD states as follows:

$$\begin{aligned}
\mathcal{S}_0: &\quad \boldsymbol{d}(k,l,\theta_s), \quad \theta_s = 0^\circ \\
\mathcal{S}_1: &\quad \boldsymbol{d}(k,l,\theta_s), \quad \theta_s = 15^\circ \\
\mathcal{S}_2: &\quad \boldsymbol{d}(k,l,\theta_s), \quad \theta_s = -15^\circ \\
\mathcal{S}_3: &\quad \boldsymbol{w}_{\text{MVDR}}(k,l) = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^T \in \mathbb{C}^M.
\end{aligned}$$

As seen from the four conditions, if the VAD classifies a time frame as being speech-absent, the beamformer is steered towards the frontal direction. This choice is solely based on the fact that speech-absent frames are not included in the computation of the performance scores, and hence, the choice of RTF vector in the VAD state $\mathcal{S}_0$ should not affect the results, yet, some choice have to be made. Moreover, if the VAD classifies double-talk, it is seen that a 1-microphone beamformer $\boldsymbol{w}_{\text{MVDR}} = [1, 0, \dots, 0]^T \in \mathbb{C}^M$ is applied. The motivation behind using a 1-microphone beamformer in time frames with double-talk is due to the fact that the MVDR beamformer allows for only one beam, due to its distortionless constraint in a signle target direction. Having to choose which target to point the beamformer at would be a challenging task and would also suppress target talk, and hence, invalidating the idea of an ideal beamformer. It should be notet that, in free-field, a 1-microphone beamformer will let audio pass undistorted from all directions, however, since the HAD is placed on a human head, the propagation of the audio will still cause distortion due to the head-shadow effect. Specifically, sound sources originating from the right-hand side of the HA user will be attenuated when reaching the reference microphone of the left HAD, whereas sound sources originating from the left-hand side are not affected by the head-shadow effect, and hence, not attenuated due to the user's head.

- **Bayesian-Audio-Gaze:** The proposed beamforming system Bayesian-Audio-Gaze (Section 6.2) uses the Bayesian beamformer weights

$$\hat{\boldsymbol{w}}_{\text{B}}(k,l) = \sum_{i=1}^{I} p(\theta_i|\boldsymbol{X}(k,l), \phi_j)\hat{\boldsymbol{w}}_{\text{MVDR}}(k,l),$$

in which the posterior probability $p(\theta_i|\boldsymbol{X}(k,l), \phi_j(l))$ uses the audio-based likelihood function $f(\boldsymbol{X}(k,l)|\theta_i)$ in (6.19) as well as the conditional PMF $p(\theta_i|\phi_j(l))$ from the pre-computed loop-up table $p(\theta_i|\phi_j(l))$. The implementaiton of this beamforming system is summarizedd in Algorithm 5, while the offline computation of the look-up table $p(\theta_i|\phi_j(l))$ is described in Section 7.4.

- **Bayesian-Gaze-Prior:** The proposed beamforming system Bayesian-Gaze-

Prior uses the Bayesian beamformer weights

$$\hat{\boldsymbol{w}}_{\mathrm{B}}(k,l) = \sum_{i=1}^{I} p(\theta_i|\boldsymbol{X}(k,l))\hat{\boldsymbol{w}}_{\mathrm{MVDR}}(k,l),$$

in which the posterior probability $p(\theta_i|\boldsymbol{X}(k,l))$ uses the audio-based likelihood function $f(\boldsymbol{X}(k,l)|\theta_i)$ in (6.19) as well as the prior probability distribution $p(\theta_i)$ obtained as a histogram over the T most resent time frames of eye-gaze measurements. A value of $T = 250$ time frames, i.e., using the previous 2 seconds of eye-gaze measurements, have been chosen through an iterative test procedure with trial and error. The implementation of the beamforming system is summarized in Algorithm 4.

- **Bayesian-Uniform-Prior:** In order to examine the performance of our two proposed eye-gaze based beamforming methods (Byesian-Gaze-Prior and Bayesian-Audio-Gaze), we will, as mentioned, compare them to an audio-only baseline Bayesian beamforming system. The beamforming system Bayesian-Uniform-Prior is used to emulate such a baseline system that only relies on the noisy microphone signals, i.e., only acoustic information. The purpose of comparing with this baseline system is to see what the benefit of incorporating the user's eye-gaze in a Bayesian beamforming strategy is. This baseline beamforming system uses the Bayesian beamformer weights

$$\hat{\boldsymbol{w}}_{\mathrm{B}}(k,l) = \sum_{i=1}^{I} p(\theta_i|\boldsymbol{X}(k,l))\hat{\boldsymbol{w}}_{\mathrm{MVDR}}(k,l),$$

in which the posterior probability $p(\theta_i|\boldsymbol{X}(k,l))$ uses the audio-based likelihood function $f(\boldsymbol{X}(k,l)|\theta_i)$ in (6.19) and a prior probability $p(\theta_i)$ with mass equally spread in the DOA range $\{-45°, 45°\}$ and zero-probability elsewhere. We choose this specific DOA range, as to compare the methods properly, the range for the target DOA parameter θ_i must be the same for all the Bayesian beamforming methods. Since the two proposed Bayesian beamformers are design in such a way that there is a underlying assumption that the target signal arrives from one out of the 13 possible target directions in the range $\{-45°, \ldots, 45°\}$, this assumption is also incorporated in the baseline audio-only Bayesian beamformer.

- **MVDR-ML:** The beamforming system MVDR-ML is used to emulate an MVDR beamformer steered using a microphone-only DOA estimator. For this method, the target RTF vector $\boldsymbol{d}(k,l,\theta_s)$ is estimated using the directional-based maximum likelihood DOA estimation method in Algorithm 1, and is used to implement the MVDR beamformer. I.e., for this method $\hat{\boldsymbol{d}}(k,l,\hat{\theta}_s)$ is the target RTF vector associated with the dictionary maximum likelihood DOA estimate of θ_s, i.e., $\hat{\theta}_s = \hat{\theta}_{s,\mathrm{ML}}$. This audio-only beamforming system is included in the simulation experiments, as to relate to the feasibility test carried out in Chapter 5.

- **MVDR-Fixed:** The beamforming system MVDR-Fixed is used to emulate an MVDR beamformer that assumes a frontal target. For this method, $\boldsymbol{d}(k, l, \theta_s)$ is the RTF vector associated with the frontal direction, i.e., $\theta_s = 0^\circ$ and is used to implement the MVDR beamformer. This audio-only beamforming system is included in the simulation experiments, as to again relate to the feasibility study carried out in Chapter 5.

8.1.1 Experimental Setup

The performance of the beamforming systems is evaluated on noisy microphone signals which are simulated in accordance to the description in Section 7.2. As described in Section 7.1.1, the audio-visual data set provides 10 trials from which we simulate acoustic scenes. For each of these trials, eye-gaze measurements from the 22 participants are used.

From the results of the feasibility test carried out in Section 5.3, it was concluded that using $M = 2$ microphones, the performance gain of using an oracle eye-gaze steered MVDR beamformer is marginal when compared to a frontal MVDR beamformer. We also established that by solely relying on the upper bound performance results obtained from the oracle study, we cannot confirm whether using $M = 4$ or $M = 6$ would possibly yield a gain in performance using eye-gaze steered beamforming in future HAs. In the simulation experiments carried out in this chapter, we therefore likewise explore the influence of the number of microphones used in a HA system on the beamformer performance. Specifically, we again consider the three microphone array configurations; a $M = 2$ microphone monaural configuration, a $M = 4$ microphone binaural configuration, and a $M = 6$ microphone binaural configuration. For each microphone array configuration, ESTOI and segSNR scores are obtain for all pairs of trials and test participants. This means that for the proposed beamforming systems, which utilize eye-gaze information, 220 ESTOI and segSNR scores are obtained for each microphone array configuration, whereas 10 ESTOI and segSNR scores are obtained for the audio-only beamforming systems. For the proposed eye-gaze based beamforming systems, the performance scores are averaged over participants as to obtain 10 ESTOI and segSNR scores, respectively. Subsequently, the performance scores for all beamforming methods are averaged over trials, yielding a single ESTOI and segSNR score for each beamforming system for all the considered microphone array configurations.

The settings for analysis and synthesis, for the implementation of the MVDR beamformers, and for the maximum likelihood estimation of the RTF vectors are all specified in accordance with the descriptions in Section 5.1.2.

8.2 Simulation Results

In this section the experimental results will be displayed. These results will seek to shed light on the performance of the proposed eye-gaze based beamforming methods and to which extend the incorporating of information provided by the HA user's eye-gaze in a beamforming method is beneficial when compared to audio-only beamforming systems. This will be done by using the proposed beamformers as well as the audio-only beamformers to solve the same task of retrieving a target signal contaminated with noise. As mentioned, the performance of the beamformers will be reported in terms of ESTOI and segSNR scores. For reference, performance scores are computed and included for the unprocessed noisy microphone signal at the reference microphone, denoted "Noisy", to act as a lower bound on the noise reduction. The ESTOI and segSNR performance measures are described in Section 5.2.

In Fig. 8.1, average ESTOI scores as well the improvement in ESTOI scores in relation to the unprocessed noisy microphone signal are presented for the various beamforming systems for $M = 2$, $M = 4$, and $M = 6$ microphones. Likewise, Fig. 8.2 illustrates the average segSNR scores and the improvement in segSNR scores in relation the unprocessed noisy microphone signal.

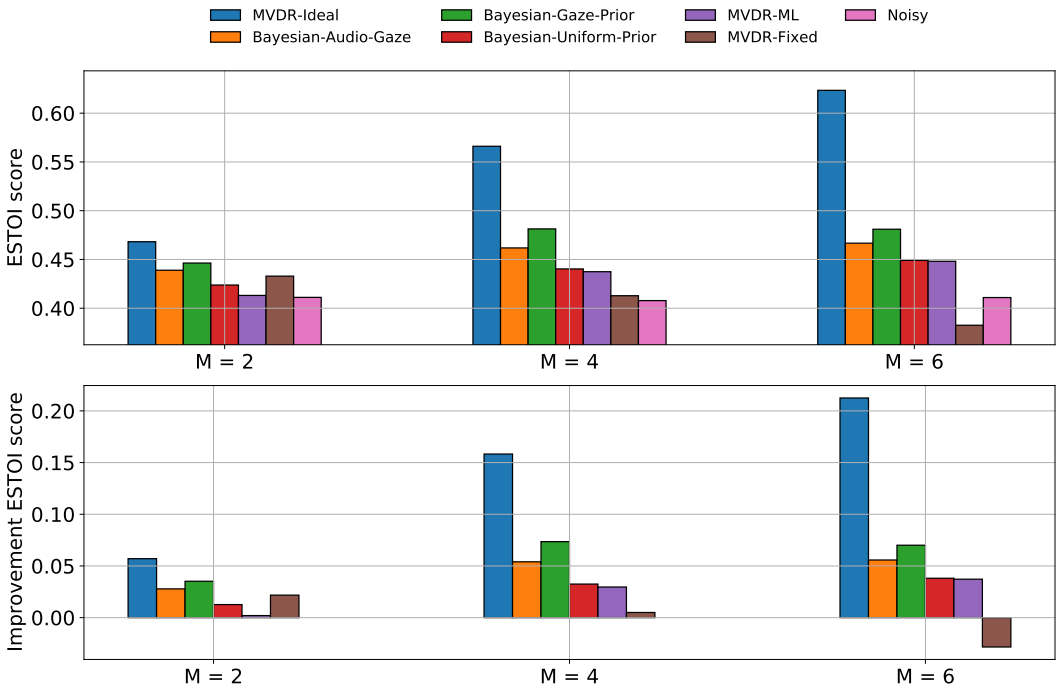


Figure 8.1: Average ESTOI scores and average improvement in ESTOI scores depicted for $M = 2$, $M = 4$ and $M = 6$ microphones.

The specific values plotted in Fig. 8.1 and Fig. 8.2 are displayed in Table 8.1.

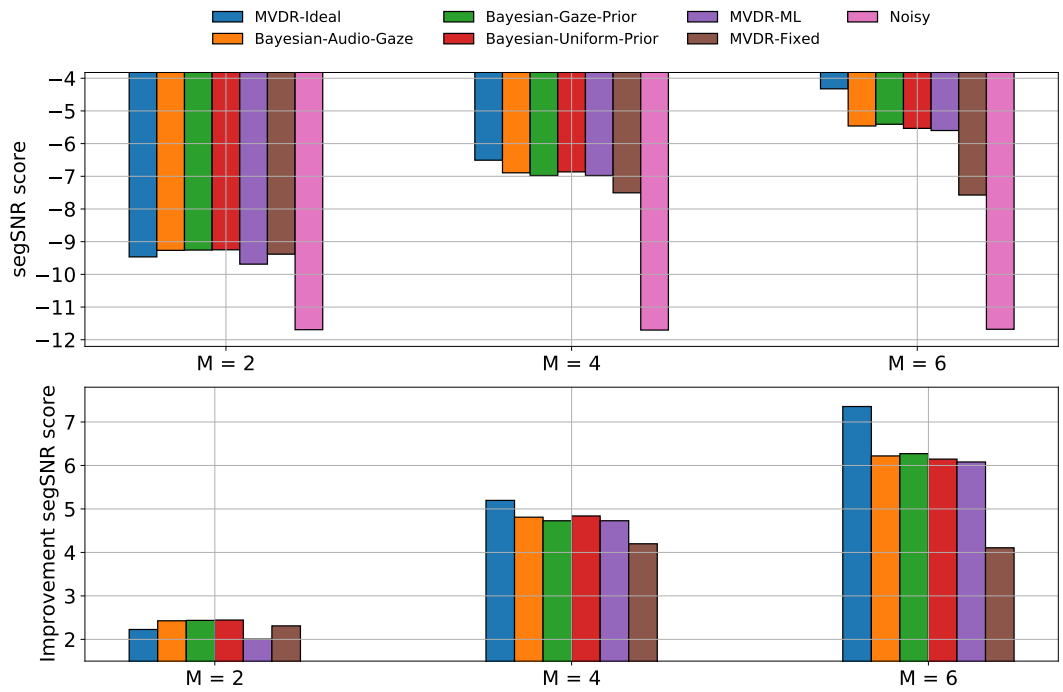


Figure 8.2: Average segSNR scores and average improvement in segSNR scores depicted for $M = 2$, $M = 4$ and $M = 6$ microphones.

Furthermore, to directly compare the performance of the two proposed eye-gaze based beamforming systems, Bayesian-Gaze-Prior and Bayesian-Audio-Gaze, to each other, as well as to compare their performance against their audio-only counterpart, Bayesian-Uniform-Prior, the difference in performance scores are depicted in Table 8.2.

From Figs. 8.1 and 8.2, it is seen that the ESTOI and segSNR scores obtained for the MVDR-Ideal improve as the number of microphone increases. This is expected as for each time-frequency tile, the MVDR-Ideal is steered towards the true target direction, identified by the VAD, and since the beam of the MVDR beamformer becomes slimmer as M increases, the MVDR-Ideal may suppress more noise while still maintaining the distortionless constraint in the target direction. For the frontal MVDR beamformer, MVDR-Fixed, the opposite behaviour is observed for ESTOI and partly for segSNR, in that we see a decrease in performance as the number of microphones increases. Again, this is in line with our expectations. As we saw in Chapter 5, when $M = 2$, the frontal MVDR beamformer performed very close to the upper performance bound of the MVDR beamformer when the target arrived from directions in the DOA range $\theta_s \in \{-45°, \ldots 45°\}$, whereas their performance differences became more significant as the number of microphones increased. Hence, when the target signal arrives from $\pm 15°$, as is the case for the simulations in Figs. 8.1 and 8.2, it

	ESTOI			segSNR		
#Microhones	$M = 2$	$M = 4$	$M = 6$	$M = 2$	$M = 4$	$M = 6$
MVDR-Ideal	0.468	0.566	0.623	-9.466	-6.508	-4.324
Bayesian-Audio-Gaze	0.439	0.462	0.467	-9.265	-6.894	-5.461
Bayesian-Gaze-Prior	**0.446**	**0.481**	**0.481**	-9.256	-6.976	**-5.408**
Bayesian-Uniform-Prior	0.424	0.44	0.449	**-9.248**	**-6.865**	-5.534
MVDR-ML	0.413	0.437	0.448	-9.688	-6.975	-5.6
MVDR-Fixed	0.433	0.413	0.383	-9.383	-7.505	-7.573
Noisy	0.411	0.408	0.411	-11.693	-11.704	-11.68

Table 8.1: Average performance scores for the beamforming systems presented in Section 8.1, as well as for the unprocessed noisy microphone signal. The performance scores are depicted for three microphone array configurations using $M = 2$, $M = 4$ and $M = 6$ microphones.

	ESTOI		
#Microhones	$M = 2$	$M = 4$	$M = 6$
Bayesian-Audio-Gaze - Bayesian-Gaze-Prior	-0.007	-0.019	-0.014
Bayesian-Audio-Gaze - Bayesian-Uniform-Prior	0.015	0.022	0.018
Bayesian-Gaze-Prior - Bayesian-Uniform-Prior	0.022	0.041	0.032
	segSNR		
#Microhones	$M = 2$	$M = 4$	$M = 6$
Bayesian-Audio-Gaze - Bayesian-Gaze-Prior	-0.009	0.082	-0.053
Bayesian-Audio-Gaze - Bayesian-Uniform-Prior	-0.017	-0.017	0.73
Bayesian-Gaze-Prior - Bayesian-Uniform-Prior	-0.008	-0.111	0.126

Table 8.2: Difference in performance scores for the beamforming systems Bayesian-Audio-Gaze, Bayesian-Gaze-Prior and Bayesian-Uniform-Prior.

is expected that the performance difference between MVDR-Fixed, which assumes a frontal target direction, and MVDR-Ideal is smaller when $M = 2$ microphones is considered, than is the case with $M > 2$.

An important observation is that for $M = 2$ microphones we see that the MVDR-Ideal is outperformed in terms of segSNR by every beamforming system except MVDR-ML. Similar unreliable behavior was also observed in the simulations in Section 5.3.2, and even though the exact reason remains unknown, a possible explanation was given in Section 5.3.2. Furthermore, another possibility is that the MVDR-Ideal is steered using the VAD described in Algorithm 6. When the VAD state is $\mathcal{S}_3$, i.e., in time frames with double-talk, the MVDR-Ideal uses a 1-microphone beamformer, hence, not performing any noise reduction. This may also contribute to the bizarre segSNR observations.

For the beamforming systems, Bayesian-Audio-Gaze, Bayesian-Gaze-Prior, Bayesian-Uniform-Prior, and MVDR-ML, it is seen that the ESTOI scores increases from $M = 2$ to $M = 4$ microphones, however, the increase in performance scores is much less significant than is the case for the MVDR-Ideal. Comparing the results for $M = 4$

and $M = 6$ microphones, almost no increase in ESTOI scores is observed for Bayesian-Audio-Gaze, Bayesian-Gaze-Prior, Bayesian-Uniform-Prior, and MVDR-ML. Since we see that the performance of the MVDR-Ideal does not stagnate as M increases, we know that this gain is obtainable. The reason why Bayesian-Audio-Gaze, Bayesian-Gaze-Prior, Bayesian-Uniform-Prior, and MVDR-ML does not increase from $M = 4$ to $M = 6$ microphones remains unknown, but a likely explanation may be that the requirements of accurate beamformer parameter estimation increases as the number of microphones increases. In other words, the beam becomes slimmer and hence inaccurate estimates are punished more heavily. Since, the MVDR-Ideal has access to the true target DOA and optimal noise statistics, this would explain why the performance of this beamforming system does not stagnate like other beamforming systems.

What is remarkably from from Fig. 8.1 is that the ESTOI scores for the MVDR-ideal is significant lower than what was observed from the results obtained in the feasibility test in Chapter 5. This observation is likely explained by that fact for the acoustic scenes simulated in the current chapter, noise is only impinging on the microphone array from the frontal quarter-plane, whereas we considered isotropic noise fields in the feasibility test. As mentioned in Chapter 7, the choice of only simulating noise from the frontal quarter-plane stems from the choice of resemble the acoustic stimuli that the HA users in the ERH was presented, as closely as possible. Had noise also impinged on the microphone array from the rear, this noise may have been almost suppressed by using an ideal MVDR beamformer, and hence, potentially having increased the performance scores across all the beamforming methods. Recall that the Bayesian beamformers are implemented as linear combinations of MVDR beamformers pointing in different directions. However, this hypothesis has not been verified, as our focus in this thesis primarily is to investigating the influence of incorporating eye-gaze into beamformers, hence we are mostly interested in the performance gain between eye-gaze based and audio-based beamforming systems.

When comparing the performance of the two proposed Bayesian beamforming methods in Table 8.2 in terms of ESTOI, an interesting observation is that the Bayesian-Gaze-Prior is superior to the Bayesian-Audio-Gaze for all of the considered number of microphones. We do however, not see the same consistent pattern for segSNR.

Comparing the MVDR-Ideal and Noisy for $M = 6$ microphones, we obtain an improvement of 0.212 in terms of ESTOI, while we for segSNR see an improvement of 7.356. Furthermore, it is seen that the segSNR scores obtained for the proposed Bayesian-Audio-Gaze and Bayesian-Gaze-Prior are closer to performance of MVDR-ideal than to Noisy, as the difference between MVDR-Ideal and Bayesian-Gaze-Prior is 1.084 while it is 6.272 between Bayesian-Gaze-Prior and Noisy, and the difference between MVDR-Ideal and Bayesian-Audio-Gaze is 1.137 while it is 6.219 between Bayesian-Audio-Gaze and Noisy. Comparing Bayesian-Gaze-Prior to its audio-only counterpart Bayesian-Uniform-Prior, we see a performance improvement regardless

of M, in fact the greatest improvement in terms of ESTOI is 0.041, which is obtained for $M = 4$ microphones. Furthermore, from Table 8.2, we observe that, in terms of ESTOI, both the proposed eye-gaze based beamforming systems outperform their audio-only based counterpart Bayesian-Uniform-Prior for all M, specifically, across M, we see a general improvement of 0.032 for Bayesian-Gaze-Prior over Bayesian-Uniform-Prior and an improvement of 0.018 for Bayesian-Audio-Gaze over Bayesian-Uniform-Prior. Furthermore, from Table 8.1, the same observations are made when comparing the eye-gaze based systems to the audio-only based system MVDR-ML, in fact, there we see greater improvements. This is in line with our expectations.

Lastly, we compare the audio-only based beamforming systems Bayesian-Uniform-Prior and MVDR-ML. For all values of M, we observe that the former yields ever so slightly better ESTOI and segSNR results. This may be explained by the fact that Bayesian-Uniform-Prior employs a linear combination of MVDR beamformers, whereas MVDR-ML simply points a beam in one direction. Furthermore, Bayesian-Uniform-Prior is restricted the DOAs $-45°, -37.5°, \ldots, 45°$, while MVDR-ML is allowed to search DOAs on the entire circle. However, due to the fact that all sound arrives from the frontal quarter half-plane in our simulated acoustic scenes, the fact that MVDR-ML search in the entire DOA range may not influence its performance as much as if sound was arriving from all directions. Hence, it may be reasonable to suppose that if the sound was not restricted to arrive from the front, but from all around, the performance difference between the Bayesian beamformers and MVDR-ML would be much more significant, as the DOA estimation conducted by the dictionary-based maximum likelihood DOA estimator in MVDR-ML would struggle to estimate the target direction.

In this chapter, we have presented results obtained by using the proposed beamformers as well as the audio-only beamformers to solve the same task of retrieving a target signal in realistic synthetic acoustic scenes which are simulated using the acoustic stimuli from the audio-visual data set from ERH. These results will lay the foundation for the discussion of the proposed beamforming systems in the proceeding chapter.

9. Discussion

In this chapter, we will discuss the results found in Chapter 8. In that regard, we will discuss how our specific choices of using the HA user's eye-gaze in combination with the microphone signals to compute the posterior DOA probabilities, may impact the results and thereby access the performance of the proposed audio-gaze beamforming methods in relation to speech enhancement. Furthermore, we will discuss possible sources of error as well as limitations and issues which may be connected to the proposed methods.

9.1 Beamformer Performance

The results presented in Chapter 8 indicate the potential for using the HA user's eye-gaze in combination with the HA microphone signals to enhance the target speech signal by the use of a Bayesian beamforming system. The acoustic data used for simulation and validation of the proposed eye-gaze steered beamforming methods were realistic synthetic noisy microphone signals generated by the constructed framework for simulating acoustic scenes. The generated noisy microphone signals were simulated to resemble the acoustic stimuli that the participants from the ERH experiment were presented. Therefore, as clean target speech signals, we used the audio tracks from the audio-visual dataset provided by ERH. The AIRs used to simulate the wave propagation from the sound sources to the microphones on the HAs were obtained from the database provided by Oticon, and the babble noise used to create the noisy microphone signals were synthetically generated using multiple speech signals from the TIMIT corpus.

The results obtained in Chapter 8, suggested that the inclusion of eye-gaze information improved performance, as in terms of ESTOI an improvements were seen for both proposed beamforming systems for $M = 2$, $M = 4$ and $M = 6$ microphones, when compared to their audio-only counterpart Bayesian-Uniform-Prior. In fact, for $M = 4$ microphones, we obtained a performance improvement for the Bayesian-Gaze-Prior 0.041 in terms of ESTOI. For segSNR, we did, however, not see that same results as the differences in scores between Bayesian-Audio-Gaze, Bayesian-Gaze-Prior

and Bayesian-Uniform-Prior were marginal. These results suggests that in terms of estimated speech intelligibility, we are indeed able to improve performance of beamforming systems by incorporating the HA users eye-gaze information, whereas, in terms of estimated speech quality, no definite conclusion can be made as to whether incorporation of eye-gaze is beneficial.

In regards to small differences in performance scores, it is worth noting what we observe that the ESTOI and segSNR scores for the noisy microphone signals are similar for $M = 2$, $M = 4$ and $M = 6$ microphones, however, with minor variations of at most 0.003 for ESTOI and 0.024 dB for SegSNR. As no processing is applied to the noisy microphone signal, the microphone array configuration should not affect performance scores. Note that the same trials are used to construct the scenes for different microphone configurations, however, most certainly with different noise realizations, hence small fluctuations for performance scores are expected. Therefore, we consider small differences to be inconsiderable, as the difference in performance scores across M might in fact be due to variantion in the noise.

In Chapter 5, we found that when the target arrived from directions in the DOA range $\theta_s \in \{-45°, \ldots, 45°\}$ there was almost no performance difference between MVDR-Fixed and MVDR-Ideal when $M = 2$ microphones was considered. Likewise, in Section 8.2, we examined the difference in performance between MVDR-Fixed and the proposed eye-gaze based beamforming systems for $M = 2$ microphones. Again we found only small differences in performance scores, which suggests that when using $M = 2$ microphones, there is no remarkably gain in incorporating eye-gaze compared to simply using a frontal steered MVDR beamformer. On the other hand, for $M = 4$ and $M = 6$ microphones, we observe a gain in performance of using eye-gaze based beamforming system compared to MVDR-Fixed. This indicates that there is a potential of using eye-gaze steered beamforming in future HAs.

Comparing the two proposed beamforming methods Bayesian-Gaze-Prior and Bayesian-Audio-Gaze, which utilize the HA user's eye-gaze information in different manners, we found that, in terms of ESTOI, the beamforming system Bayesian-Gaze-Prior is superior to Bayesian-Audio-Gaze for all values of M with an average ESTOI improvement across M of 0.013. The reason why Bayesian-Gaze-Prior performs better than the Bayesian-Audio-Prior may be explained by the fact that the Bayesian-Gaze-Prior adapts as new measurements are observed, while the Bayesian-Audio-Gaze chooses a predefined probability distribution based on the most resent eye-gaze measurement which might not be as good a fit to the live data as the one constructed by Bayesian-Gaze-Prior. Furthermore, since the pre-computed conditional probability look-up table $p(\theta_i|\phi_j)$, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, is constructed by only using data for the situations where $\theta_i = \pm 15°$, even minor inaccuracies and outliers may introduce undesirable artefacts in the entire look-up table, possibly affecting the overall performance of the Bayesian-Audio-Gaze systems. From the results obtained in Chapter 8, this indicates that in terms of estimated speech intelligibility, the beamforming system

Bayesian-Gaze-Prior succeeds over Bayesian-Audio-Gaze.

In regards to the difference between Bayesian-Audio-Gaze and Bayesian-Gaze-Prior, we made some choices during the development of these systems. In the following, we will address these differences and how different choices might have affected the performance of the eye-gaze based beamforming systems.

Bayesian-Gaze-Prior incorporates eye-gaze information by estimating the prior probability $p(\theta_i)$ as a histogram over the most recent T eye-gaze measurements. Based on initial testing, we settle on a value of $T = 250$ corresponding to 2 seconds of previous eye-gaze data, as this provided the best results in terms of beamforming performance scores. Had a smaller value of T been chosen, the beamforming system might have become more adaptive, and had we chosen T larger, the beamforming system might have become less susceptible to e.g., measurement errors. Had the value of the parameter T been optimized further, we might had seen better performance for Bayesian-Gaze-Prior.

For the beamforming system Bayesian-Audio-Gaze, the instantaneous eye-gaze measurement $\phi_j(l)$ is mapped to a predefined probability density from the look-up table created in advance of application of the beamforming system. Based on preliminary experiments, we have chosen to construct a look-up tables specifically tailored to each test participants. However, in the experiment at ERH, at which the eye-gaze measurements were recorded, a mismatch between the actor's position on the TV screen and the loudspeaker position from which the acoustic stimuli was played, was observed. Furthermore, as the position of the actors was not fixated during the trials, the actors might have sat or moved differently during different trials. Since, the participant-specific look-up tables are obtained over all scenes, these potentially varying mismatches will be captured in the look-up tables. If we were to have created look-up tables for each combination of test participant and scene, we might have seen better performance for Bayesian-Audio-Gaze. However, this has not been tested due to the fact that when histograms are used for density estimation, the amount of data on which the histograms are computed over is of crucial importance as to obtain a reliable estimate. As the amount of data available to us is fairly limited, creating look-up tables for each individual combinations of trials and participants, might not be feasible from a mathematical point-of-view.

Another important aspect of the Bayesian-Audio-Gaze is the assumption of the noisy microphone signals $\tilde{\boldsymbol{x}}(k,l)$ being conditional independent of the HA user's eye-gaze angle $\phi_j(l)$, given that the target DOA θ_i is known. Though we argued that the assumption seems fair, this might after all be to simple a model, which may indicate the need for a more complex model linking the eye-gaze measurements and the acoustic signal.

In general, we did not see as big an improvement for the eye-gaze based beamforming systems compared to the competing audio-only systems, that we might have hoped, which may very well be explained by multiple factors. In a noisy environment where

the SNR is fairly high, audio-only beamforming systems generally perform well, as the target DOA can be estimated from the acoustic signals alone. However, as the SNR decreases the task of using on acoustic signal for reliable DOA estimation becomes challenging. Therefore, at lower SNRs, we would expected our proposed eye-gaze beamforming systems to perform much better compared to audio-only beamforming systems in such scenario. However, the eye-gaze measurements were recorded at a SNR of 0 dB, and as such, our acoustic scenes were simulated at a SNR of 0 dB. This is due to the fact that we find it important to mimic the setup at which the eye-gaze measurements were recorded, as the eyes might behave differently under different conditions. Furthermore, audio-only beamforming systems performs well in the case when only a single talker is active at a time. In contrast, the eye-gaze is often steered towards the target of interest, hence eye-gaze based beamforming system, should be able to decipher which direction to look at even when competing speakers are present.

9.2 Lack of Realism

In terms of ESTOI, we do see a benefit of incorporating eye-gaze into beamforming systems. However, it is important to emphasize that the performance presented in these simulations are conditioned on our specific setup. Due to the natural biological behavior of the eye, we would likely see different behavior in varying setups. For example, it is easy to imagine that the rate at which people speak as well as how articulate they are will subject the eye-gaze of the HA user. Furthermore, in our simulations, the test participants were seated 1.8 m away from the TV, which in relation to a realistic setup might be considered a long distance. In addition, in the RTF database which we have used throughout this thesis, the HAs which recorded the incoming sounds, was not placed 1.8 m from the speakers. This means that the RTFs in the database may not accurately describe the wave propagation from the target talker to the HA user.

As presented in Section 8.2, noise was simulated as only impinging from the frontal quarter-plane, which does not resemble to noise in a real situation. In addition, we argue that the recorded eye-gaze measurements does not truly reflect the noise field in which is was recorded. This is due to the fact that babble noise is simulated from each direction, yet there are no visual indication of this for the HA user. If eye-gaze were to be recorded in a real situation with background talkers, one could easily imagine that the HA user, from time to time, would gaze at these talkers, hence, changing the behavior of the eye-gaze. With these considerations in mind, we argue that our simulations may deviate from a completely realistic setup, and that the results are dependent on the specific setup. However, under these conditions, we observe a performance improvement of using eye-gaze based beamforming system, which suggests potentials for such methods, and that further effort should be put into this research area.

10. Conclusion

We have in this thesis investigated the possibility of using the HA user's eye-gaze in combination with the HA microphone signals to enhance the target speech signal by the means of a HA beamforming algorithm that exploit this combined information. To this end, the goal of this thesis was to seek an answer to the following main question:

How can information provided by the HA user's eye-gaze and by the HA microphone signals be combined and used to construct a beamformer for HA applications, and can such a beamformer potentially outperform current audio-only beamforming methods in terms of predicted speech intelligibility and predicted speech quality in noisy acoustic scenes?

In order to answer this question, we proposed two beamforming systems for hearing aids which incorporates the user's eye-gaze, namely a Bayesian beamformer with the posterior probabilities estimated based on a prior probability derived from the user's eye-gaze direction (Bayesian-Gaze-Prior), and a Bayesian beamformer with the posterior DOA probabilities jointly estimated from the HA microphone signals and the HA user's eye-gaze signal (Bayesian-Audio-Gaze). To allow the incorporating of the HA user's eye-gaze through a Bayesian approach, we considered the the target DOA as a discrete random variable with a prior distribution that reflects its uncertainty, relied on a statistical model linking the eye-gaze information and the acoustic information to the random variable describing the target direction/DOA. Specifically, inspired by [24] and [23], the idea of the proposed methods was to estimate a probability distribution of the target talker location, i.e. for each possible target direction assign a probability, and then use this estimated probability distribution for a Bayesian beamformer.

We compared the performance of the proposed Bayesian beamforming methods with state-of-the-art audio-only methods used to solve the same task of enhancing a target speech signal. Specifically, the proposed methods were compared to an MVDR beamformer steered using an audio-only maximum likelihood DOA estimator, a fixed beamformer that is always steered towards the front of the user, and an audio-only Bayesian beamformer. From the results presented in Chapter 8, it was seen that,

in general, there is an improvement in the performance of the proposed Bayesian-Gaze-Prior and Bayesian-Audio-Gaze for the simulated acoustic condition in terms of ESTOI compared to the audio-only counterpart Bayesian-Uniform-Prior. Specifically, for $M = 4$ microphones, we see an ESTOI score improvement of 0.41 when comparing Bayesian-Gaze-Prior and Bayesian-Uniform-Prior. Based on the results, it was discussed in Chapter 9 that there is room for improvement in the performance of the proposed eye-gaze based beamformers in terms of both ESTOI and segSNR for $M = 4$ and $M = 6$ microphones, as there is a gab between the upper bound performance obtained by the MVDR-Ideal and the proposed Bayesian-Gaze-Prior and Bayesian-Audio-Gaze.

Furthermore, it was seen that, in general, the proposed Bayesian-Gaze-Prior was superior to the proposed Bayesian-Audio-Gaze in terms of ESTOI and segSNR, however, with a maximal ESTOI score improvement of 0.19. Following this, it was discussed that this presumably may be due to the way we have chosen to incorporate the eye-gaze information in our proposed beamforimng systems.

From results obtained from the feasibility test carried out in Chapter 5, we found that when the target arrived from directions in the DOA range $\theta_s \in \{-45°, \ldots, 45°\}$ there was almost no performance difference between the frontal steered MVDR beamformer and the upper bound performance of the eye-gaze steered MVDR beamformer (MVDR-Eye-Gaze) when $M = 2$ microphones was considered. Likewise, in Section 8.2, we found only small differences in performance scores between our proposed methods and MVDR-Fixed. In Chapter 9, it was discussed that these observations suggest that, when using $M = 2$ microphones, there is no remarkably gain in incorporating eye-gaze compared to simply using a frontal steered MVDR beamformer under the acoustic conditions considered in this thesis. On the other hand, for $M = 4$ and $M = 6$ microphones, we observed a gain in performance of using eye-gaze based beamforming system compared to the frontal steered MVDR beamformer MVDR-Fixed. This is an important result, as many current HAs are only equipped with $M = 2$ microphones.

It can be concluded that by combining information provided by the HA user's eye-gaze and the HA microphone signals in a Bayesian framework, by estimating a probability distribution of the target speaker location, a beamformer can be constructed which is able to outperform audio-only beamforming methods, at least in terms of predicted speech intelligibility in noisy acoustic scenes.

10.1 Further Work

For all the simulations carried out in Chapter 8, the input SNR was fixed to 0 dB to simulate a noisy acoustic scene that resemble the acoustic stimuli the test participants at the ERH experiments experienced. However, at low SNRs, eyes tend to be at the target more often than at high SNRs [60]. Hence, it seem reasonable to as-

sume that, at low SNRs, eye-gaze will have greater significance, while at high SNRs, eye-gaze will have less significance. In further work, it would therefore be interesting to examine the proposed eye-gaze based beamformers under different input SNRs. To be able to do so, would require access to eye-gaze signals recorded under different acoustic conditions, i.e., at different input SNRs. However, in lack of such data being available, another option might be to make a synthetic simulation at lower input SNRs. It should however be noted that in this case, the eye-gaze data used in this thesis would not match the synthesized low SNR. However, as there is evidence that at low SNRs, eyes tend to be at target more often than at high SNRs, it would be interesting to examine if an improvement can be obtained by simulating scenes with input SNRs lower than 0 dB. Following this, it could be interesting to examine the effect of making a SNR-dependent look-up table $p(\boldsymbol{d}(k,l,\theta_i),\phi_j(n))$, where we take into account that at low input SNRs, the eye-gaze tends to be at the target speaker more often than at high SNRs.

Besides varying SNRs, we would in further work, find it very interesting to also let other parameters vary in the simulation experiments. First of all, we expect eye-gaze based beamformers to perform particularly well when competing speakers are present, hence the inclusion of competing speakers in the simulated acoustic scenes would be interesting to consider. Furthermore, the positions of the actors presented on the TV screen were fixed throughout the experiments conducted at ERH. Letting the target positions vary, would provide us with deeper insight into a more versatile use of the eye-gaze information in beamforming for HA applications. Moreover, we would find it interesting in further work to have data available in which the HA user interacted with the target talkers, as this is a common event in real world applications. In our simulation experiments, babble noise was simulated as impinging from within the range $\pm 45°$, which does not very well reflect a realistic acoustic environment, as opposed to letting noise impinge from all directions. Varying the aforementioned conditions will certainly affect the behavior of the eye-gaze, and hence, we have not been able to do so in this thesis. However, for further research in the field of eye-gaze based beamforming, we deem that recording a dataset of eye-gaze measurements with the aforementioned varying conditions is attractive.

In regards to the proposed eye-gaze based beamforming systems, in further work, we would find it interesting to extend upon the systems based on the knowledge obtained in this thesis. For the proposed method Bayesian-Gaze-Prior, an estimate of the density $p(\theta_i)$ of the eye-gaze over a fixed period of time is obtained as a histogram. An interesting extension to this method would be to use a VAD in order to only include actual speech in the density estimation, as periods of silence certainly does not contribute any information, but may erroneously have introduced unwanted switching artifact. Furthermore, for the proposed beamforming systems, the target sound source is assumed to arrive from the DOA range $\theta_s \in \{-45°,\ldots,45°\}$. In a realistic acoustic environment it may likely be that the target is able to arrive from any

location. Hence, in a future study, where the noise field is simulated approximately isotropic, it would be interesting to see how our proposed method fare, compared to the audio-only systems. Lastly, our proposed beamforming systems are build on the fact that the eye-gaze measurements are discretized to a particular moment in time. However, as eye-gaze measurements are definitely not discrete, we see potentials in modelling them through a continuous statistical model. Given the circularly nature of eye-gaze data, the model of the data would likewise have to be circular. However, such models might be assigned mass in directions for which the eye cannot possibly gaze. As an alternative to a parametric modelling, the continuous eye-gaze model could be learned by a data-driven method, e.g., a deep neural network, if large amounts of data with synchronized audio-visual stimuli and associated eye-gaze behaviour becomes available for future studies.

Bibliography

[1] C. Elberling and K. Worsøe, *Fading sounds : about hearing and hearing aids*. Oticon Foundation, 2006, ISBN: 87-991301-0-6.

[2] World Health Organization (WHO), *Deafness and hearing loss*, `https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss`, Accessed: 08/05-2022.

[3] G. R. Popelka, B. C. J. Moore, R. R. Fay, and A. N. Popper, *Hearing Aids*, 1st ed. Springer International Publishing, 2016, ISBN: 9783319330341.

[4] Oticon, `https://www.oticon.global/hearing-aid-users`, Accessed: 31/05-2022.

[5] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*. John Wiley & Sons, Ltd, 2010, ch. 9, pp. 269–302, ISBN: 9780470487068. DOI: `https://doi.org/10.1002/9780470487068.ch9`.

[6] P. Hoang, "A hybrid approach for speech enhancement with dnn supported acoustic beamforming," M.S. thesis, Aalborg University, 2018.

[7] S. Chakrabarty and E. Habets, "A bayesian approach to informed spatial filtering with robustness against doa estimation errors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, Jan. 2018. DOI: `10.1109/TASLP.2017.2752364`.

[8] A. H. Andersen, S. Santurette, M. S. Pedersen, E. Alickovic, L. Fiedler, J. Jensen, and T. Behrens, "Creating clarity in noisy environments by using deep learning in hearing aids," in *Seminars in Hearing*, Thieme Medical Publishers, Inc., vol. 42, 2021, pp. 260–281.

[9] S. Chakrabarty and E. A. P. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 136–140.

[10] M. Brandstein and D. Ward, *Microphone Arrays*, 1st ed. Springer, 2001, ISBN: 978-3-662-04619-7.

[11] H. Ye and D. DeGroat, "Maximum likelihood doa estimation and asymptotic cramer-rao bounds for additive unknown colored noise," *IEEE Transactions on Signal Processing*, vol. 43, no. 4, pp. 938–949, 1995. DOI: 10.1109/78.376846.

[12] U. Kjems and J. Jensen, "Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 295–299.

[13] J. Jensen and M. Pedersen, "Analysis of beamformer directed single-channel noise reduction system for hearing aid applications," IEEE Signal Processing Society, 2015, pp. 5728–5732. DOI: 10.1109/ICASSP.2015.7179069.

[14] P. Hoang, Z.-H. Tan, J. M. De Haan, and J. Jensen, "The minimum overlap-gap algorithm for speech enhancement," *IEEE Access*, vol. 10, pp. 14 698–14 716, 2022. DOI: 10.1109/ACCESS.2022.3147514.

[15] J. Hart, D. Onceanu, C. Sohn, D. Wightman, and R. Vertegaal, "The attentive hearing aid: Eye selection of auditory sources for hearing impaired users," in *INTERACT*, 2009.

[16] M. Harrison, "Evaluating the use of steering a hearing aid in a dynamic multi-talker environment using body signals," *University of Glasgow, Eriksholm Research Centre*, 2018.

[17] A. Favre-Félix, C. Graversen, T. Dau, and T. Lunner, "Real-time estimation of eye gaze by in-ear electrodes," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017, pp. 4086–4089. DOI: 10.1109/EMBC.2017.8037754.

[18] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, *An overview of deep-learning-based audio-visual speech enhancement and separation*, 2021. arXiv: 2008.09586 [eess.AS].

[19] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt, "Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, 2001, 301–308, ISBN: 1581133278. DOI: 10.1145/365024.365119.

[20] A. Favre-Félix, C. Graversen, R. K. Hietkamp, T. Dau, and T. Lunner, "Improving speech intelligibility by hearing aid eye-gaze steering: Conditions with head fixated in a multitalker environment," *Trends in Hearing*, vol. 22, 2018. DOI: 10.1177/2331216518814388.

[21] J. Gerald Kidd, C. R. Mason, V. Best, and J. Swaminathan, "Benefits of acoustic beamforming for solving the cocktail party problem," *Trends in Hearing*, vol. 19, 2015, PMID: 26126896. DOI: 10.1177/2331216515593385.

[22] V. Best, E. Roverud, T. Streeter, C. R. Mason, and J. Gerald Kidd, "The benefit of a visually guided beamformer in a dynamic speech task," *Trends in Hearing*, vol. 21, p. 2 331 216 517 722 304, 2017, PMID: 28758567. DOI: 10.1177/2331216517722304. [Online]. Available: https://doi.org/10.1177/2331216517722304.

[23] K. Bell, Y. Ephraim, and H. Van Trees, "A bayesian approach to robust adaptive beamforming," *IEEE Transactions on Signal Processing*, vol. 48, no. 2, pp. 386–398, 2000. DOI: 10.1109/78.823966.

[24] P. Hoang, Z.-H. Tan, J. M. de Haan, T. Lunner, and J. Jensen, "Robust bayesian and maximum a posteriori beamforming for hearing assistive devices," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2019, pp. 1–5. DOI: 10.1109/GlobalSIP45357.2019.8969234.

[25] M. Taboga, *Lectures on Probability Theory and Mathematical Statistics*, Second. CreateSpace Independent Publishing Platform, 2012, ISBN-13: 9781480215238.

[26] A. Kuklasinski, "Multi-channel dereverberation for speech intelligibility improvement in hearing aid applications," Ph.D. dissertation, 2016. DOI: 10.5278/vbn.phd.engsci.00129.

[27] B. Xie, *Head-related transfer function and virtual auditory display*, 2nd ed. J. Ross Publishing, 2013, ISBN: 978-1-60427-070-9.

[28] J. O. Smith, *"Spherical Waves from a Point Source" in Physical Audio Signal Processing*. http://ccrma.stanford.edu/~jos/pasp/, accessed 08/05-2022, online book, 2010 edition.

[29] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.

[30] J. Benesty, M. M. Sondhi, and Y. A. Huang, *Springer Handbook of Speech Processing*, 1st ed. Springer International Publishing, 2016, ISBN: 9783540491255.

[31] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time fourier transform domain," *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, 2007. DOI: 10.1109/LSP.2006.888292.

[32] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001. DOI: 10.1109/89.928915.

[33] A. Moore, J. Haan, M. Pedersen, P. Naylor, M. Brookes, and J. Jensen, "Personalized signal-independent beamforming for binaural hearing aids," *The Journal of the Acoustical Society of America*, vol. 145, pp. 2971–2981, May 2019. DOI: 10.1121/1.5102173.

[34] P. Hoang, Z.-H. Tan, T. Lunner, J. M. de Haan, and J. Jensen, "Maximum likelihood estimation of the interference-plus-noise cross power spectral density matrix for own voice retrieval," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6939–6943. DOI: 10.1109/ICASSP40776.2020.9053988.

[35] "Parameter estimation ii," in *Optimum Array Processing*. John Wiley & Sons, Ltd, 2002, ch. 9, pp. 1139–1317, ISBN: 9780471221104. DOI: https://doi.org/10.1002/0471221104.ch9.

[36] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365 –1376, 1987. DOI: 10.1109/TASSP.1987.1165054.

[37] S. Haykin, *Adaptive Filter Theory*, Fifth. Pearson, 2014, ISBN: 978-0-273-76408-3.

[38] B. A. D. H. Brandwood, "A complex gradient operator and its applica-tion in adaptive array theory," 1983.

[39] M. K. Steven, "Fundamentals of statistical signal processing," *PTR Prentice-Hall, Englewood Cliffs, NJ*, vol. 10, p. 151 045, 1993.

[40] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood psd estimation for speech enhancement in reverberation and noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1599–1612, 2016. DOI: 10.1109/TASLP.2016.2573591.

[41] M. Zohourian, G. Enzner, and R. Martin, "Binaural speaker localization integrated into an adaptive beamformer for hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 515–528, 2018.

[42] P. Hoang, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Multichannel speech enhancement with own voice-based interfering speech suppression for hearing assistive devices," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 706–720, 2022. DOI: 10.1109/TASLP.2022.3145294.

[43] C. Bishop, "Pattern recognition and machine learning," in. Springer-Verlag New York, 2006, p. 227, ISBN: 0-387-31073-8.

[44] P. Hoang, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Multichannel speech enhancement with own voice-based interfering speech suppression for hearing assistive devices," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 706–720, 2022. DOI: 10.1109/TASLP.2022.3145294.

[45] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, *TIMIT acoustic-phonetic continuous speech corpus LDC93S1*, https://catalog.ldc.upenn.edu/LDC93S1, Philadelphia: Linguistic Data Consortium, 1993.

[46] M. Kolbœk, Z. Tan, and J. Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 305–311. DOI: 10.1109/SLT.2016.7846281.

[47] *Scipy.signal.resample_poly*, 2020. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.resample_poly.html.

[48] R. W. S. Alan V. Oppenheim, *Discrete-Time Signal Processing*, 3rd ed. Pearson, 2014, ISBN: 9781292025728.

[49] R. C. Hendriks, T. Gerkmann, and J. Jensen, "Dft-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art," *Synthesis Lectures on Speech and Audio Processing*, vol. 9, no. 1, pp. 1–80, 2013. DOI: 10.2200/S00473ED1V01Y201301SAP011.

[50] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016. DOI: 10.1109/TASLP.2016.2585878.

[51] K. Kondo, "Speech quality," in *Subjective Quality Measurement of Speech: Its Evaluation, Estimation and Applications*. Springer Berlin Heidelberg, 2012, pp. 7–20, ISBN: 978-3-642-27506-7. DOI: 10.1007/978-3-642-27506-7_2.

[52] P. C. Loizou, *Speech Enhancement: Theory and Practice*, Second. CRC Press, 2017, ISBN: 978-1-138-07557-3.

[53] M. Brookes *et al.*, "Voicebox: Speech processing toolbox for matlab," *Software, available [Mar. 2011] from www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html*, vol. 47, p. 45, 1997.

[54] C. Lam and A. Singer, "Bayesian beamforming for doa uncertainty: Theory and implementation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4435–4445, 2006. DOI: 10.1109/TSP.2006.880257.

[55] L. A. Dalton and E. R. Dougherty, "Bayesian minimum mean-square error estimation for classification error—part i: Definition and the bayesian mmse error estimator for discrete classification," *IEEE Transactions on Signal Processing*, vol. 59, no. 1, pp. 115–129, 2011. DOI: 10.1109/TSP.2010.2084572.

[56] M. Vetterli, J. Kovačević, and V. K. Goyal, *Foundations of Signal Processing*, v1.1. Creative Commons, 2014, ISBN: 110703860X.

[57] P. Olofsson and M. Andersson, *Probability, Statistics, and Stochastic Processes*, Second. Wiley, 2012, ISBN: 978-0-470-88974-9.

[58] P. Drábek and J. Milota, *Methods of Nonlinear Analysis - Applications to Differential Equations*. Birkhäuser Verlag AG, 2007, ISBN: 978-3-7643-8146-2.

[59] M. Taseska, "Informed spatial filters for speech enhancement noise and interference reduction, blind source separation, and acoustic source tracking," Ph.D. dissertation, Jan. 2018.

[60] Šabić E, H. D, M. H, M. A, H. MC, and M. JA, "Examining the role of eye movements during conversational listening in noise," 2020, PMID: 32116975. DOI: 10.3389/fpsyg.2020.00200. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7033431/.

Appendix A

In this appendix, we provide mathematical definitions and results which are used throughout the thesis.

A.1 Linear Convolution

In Definition A.1, we define the linear convolution between two sequences, which is used in Chapter 2.

Definition A.1 (Linear Convolution)
Let $p, q \in [1, \infty]$ satisfy $1/p+1/q = 1$, assuming the convention that $1/\infty$ is allowed. The convolution between two sequences $h \in \ell^p(\mathbb{Z})$ and $x \in \ell^q(\mathbb{Z})$ is defined as [56, p. 206]

$$(x * h)(n) = \sum_{m \in \mathbb{Z}} x(m)h(n-m), \quad n \in \mathbb{Z}. \tag{A.1}$$

Note that the dependency on the variable n in (A.1) is used to denote the argument over which we perform the convolution.

Remark A.1 (Convergence of the convolution sum) Recall that a doubly infinite sum, as the the convolution sum in (A.1), is said to converge when it converges absolutely [56, pp. 136-137]. Hence, we know that the linear convolution $(x * h)(n)$, for $n \in \mathbb{Z}$, on the left hand side of (A.1) is well defined when the sum on the right hand side is absolute convergent for every value of n. To ensure that that the convolution is well defined, the condition for p and q to be satisfied in the definition, which follows from Hölder's inequality for sequences [56, p. 139], is included, as it can be shown that the convolution sum in (A.1) is guaranteed to converge absolutely when $h \in \ell^p(\mathbb{Z})$ and $x \in \ell^q(\mathbb{Z})$ for some p and q in $[1, \infty]$ satisfying $1/p + 1/q = 1$, again with the convention that $1/\infty$ is allowed [56, p. 316]. In signal processing, we often employ the case were $p = 1$ and $q = \infty$, due to the fact that restriction of the impulse response h to $\ell^1(\mathbb{Z})$ allows the input x to be any sequence in $\ell^\infty(\mathbb{Z})$ [56, p. 316].

Definition A.1 assumes that the underlying domain is infinite. However, in practice, we observe only a finite portion of an infinite-length sequence, and therefore, we have to consider how to handle results that apply to infinite-length sequences when only a finite amount of data is observed. To this end, we are forced to apply methods that are used to embed the finite-length sequences in the infinite-length sequences. Specifically, in this thesis, we choose to apply zero-padding for that purpose. Specifically, for a sequence of length N for some finite $N \in \mathbb{N}_0$, we set $x(n) = 0$ for all n outside of $\{0, 1, \ldots, N-1\}$, and thereby extent the finite-length sequence to an infinite-length sequence. Throughout this thesis, we do most often consider sequences with support in $\mathbb{N}_0$, and obviously the definition of the convolution applies, sine we through the use of zero-padding consider infinite-length sequences that are zero-valued at negative times. [56, pp. 182-184]

A.2 The Short-Time Fourier Transform

This section provides a definition of the STFT given by [30, p. 230].

Definition A.2 (Short-time Fourier Transform)
Let $x = \{x(n)\}_{n\in\mathbb{N}_0}$ be a discrete time signal. Let furthermore $w = \{w(n)\}_{n=0}^{N-1}$ be a chosen discrete window sequence such that the l'th time frame of x is given by

$$x_l(n) = x(n + lD)w(n), \quad n = 0, \ldots, N-1,$$

where N is the length of the window sequence and D is the hop size. Then the short-time Fourier transform of x_l is defined as

$$\tilde{x}(k, l) = \sum_{n=0}^{N-1} x_l(n) \exp^{\frac{-j2\pi nk}{N}}, \quad k = 0, \ldots, N-1, \tag{A.2}$$

where k is the frequency bin index and j is the imaginary unit. [30, p. 230]

Hence, the STFT can be interpreted as applying a N-point discrete Fourier transform on a sequence of windowed signals of length N.

Appendix B

B.1 Test of Python Implementation of Dictionary-Based Maximum Likelihood DOA Estimation Method

An example of the comparison is shown in Fig. B.1 where the estimated target direction is plotted as a function of time frame. In this specific example, we have

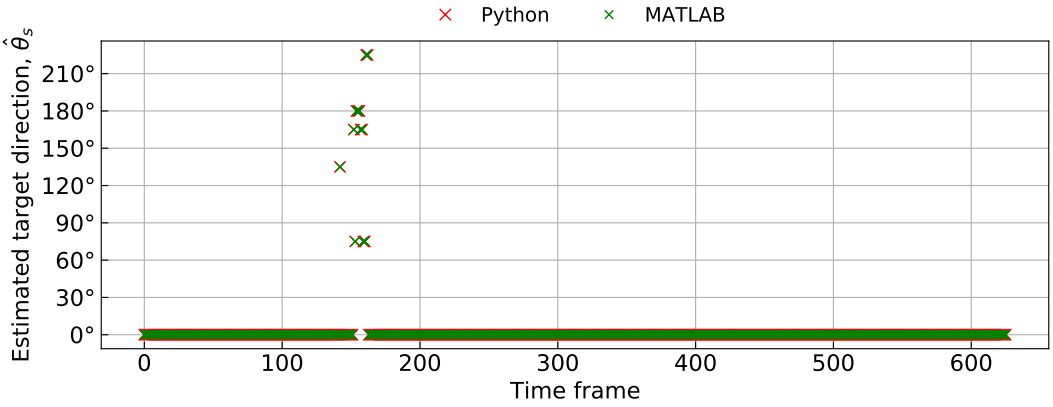


Figure B.1: Comparison of MATLAB function and Python implementation of the dictionary based maximum likelihood DOA estimation of the RTF vectors. Estimated target direction is plotted as a function of time-frames.

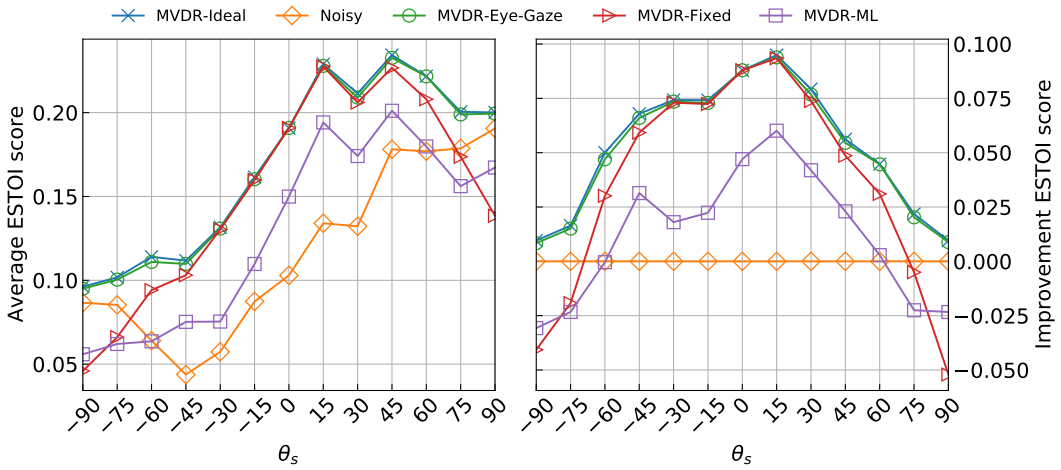
simulated an acoustic scene containing a single target speaker placed in an approximately isotropic SSN noise field with an input SNR of 10 dB. The target direction is $\theta_s = 0^\circ$, i.e, the target speaker is placed in the front of the user. Furthermore, the noisy microphone signals was formed using $M = 4$ microphones in a binaural HA configuration where we have used the front and rear microphones. The initial first second of noise-only samples were used to estimate $\mathbf{\Gamma}_{\boldsymbol{v}}(k, l)$ and a number of $L = 15$ noisy observations were used to perform the maximum likelihood estimation.

From Fig. B.1 it is seen that not all estimated target directions are equal to the true target direction. This may be explained by the fact that directions are estimated even in speech-absent regions. Since no speaker is active in these regions, the most prominent noise source is probably selected. However, the two implementations

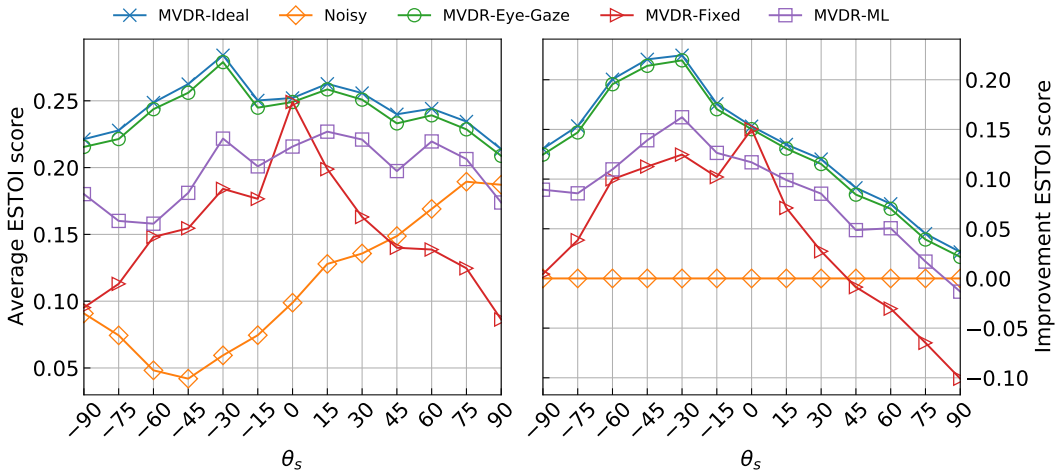
provide identical results.

B.2 Additional Results

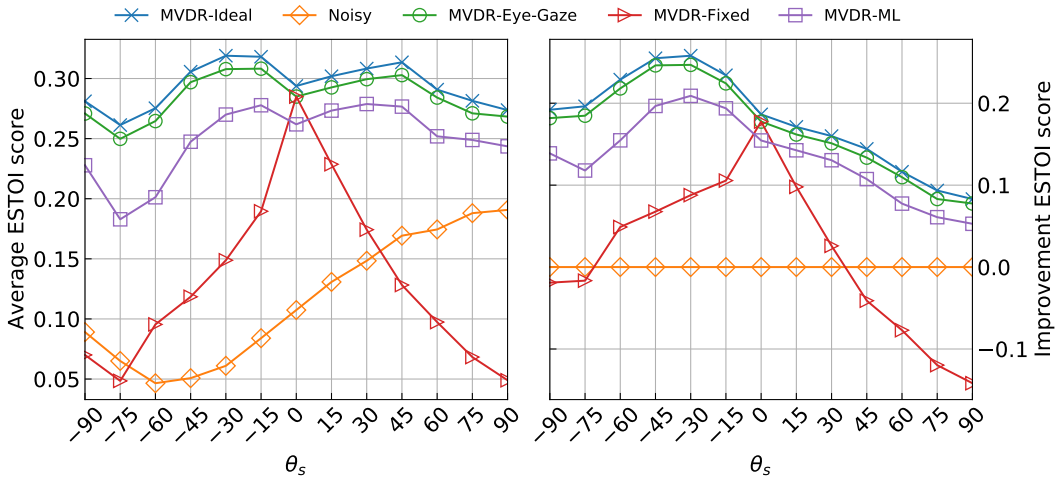
In this appendix, we provide the remaining results from the beamformer performance evaluations in Section 5.3. The results are depicted in Figs. B.2 to B.7.



(a) A 2-microphone monaural configuration, using the front and rear microphone on the left HA.

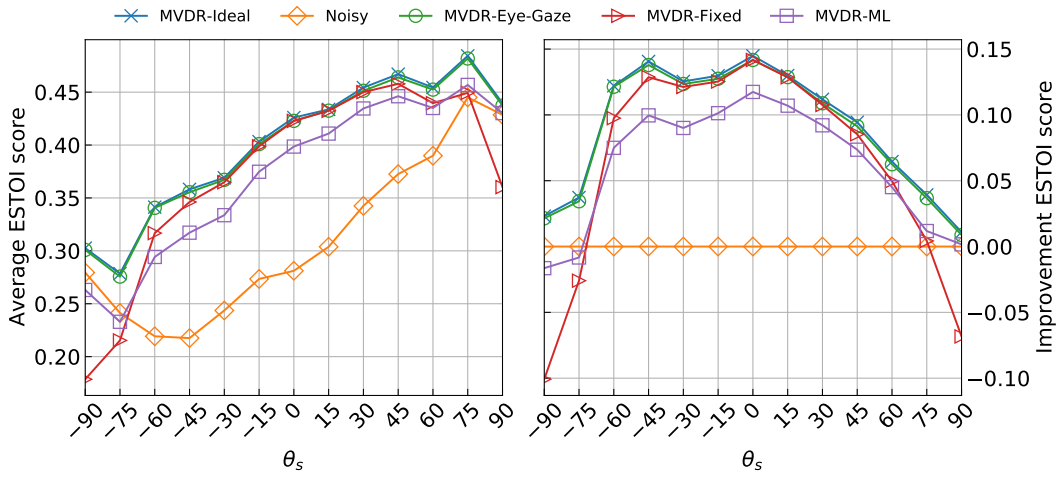


(b) A 4-microphone binaural configuration, using the front and rear microphones on both the left and right HA.

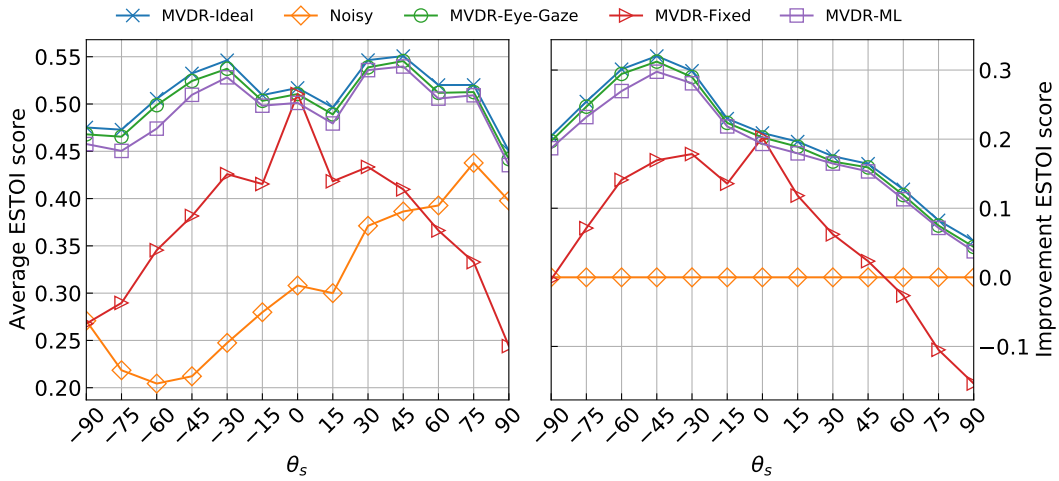


(c) A 6-microphone binaural configuration, using all three microphones on both HAs.

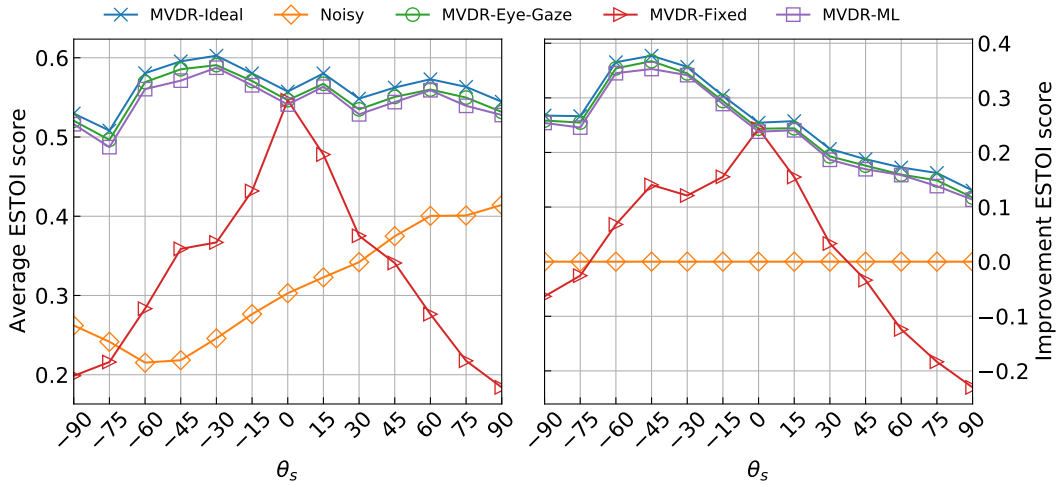
Figure B.2: Average ESTOI scores and average improvement ESTOI scores as a function of target DOA. The noise type is speech shaped noise in an approximately isotropic noise field with an input SNR of −10 dB.



(a) A 2-microphone monaural configuration, using the front and rear microphone on the left HA.

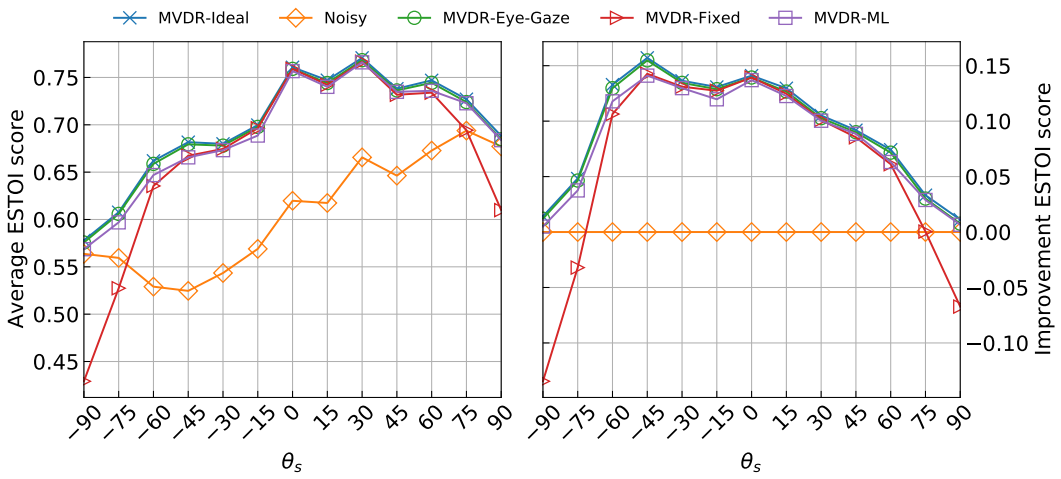


(b) A 4-microphone binaural configuration, using the front and rear microphones on both the left and right HA.

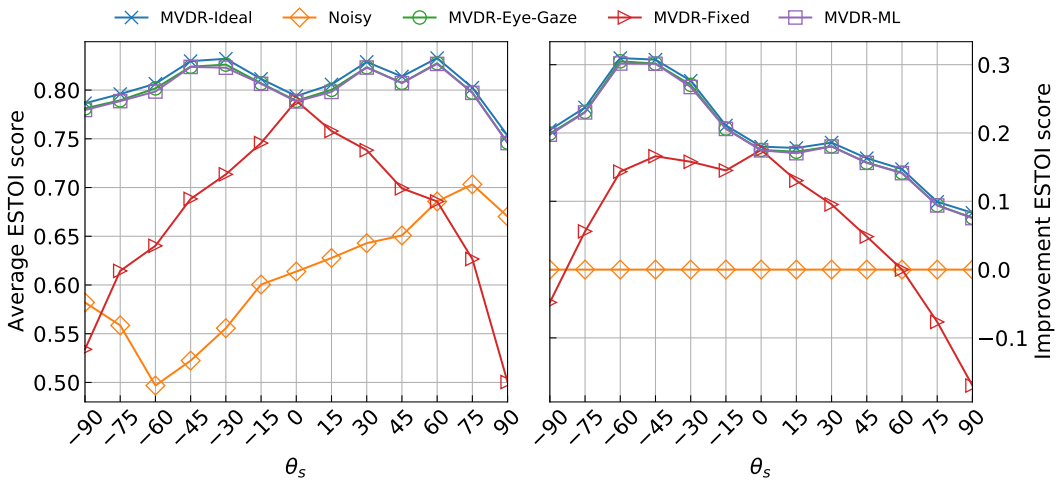


(c) A 6-microphone binaural configuration, using all three microphones on both HAs.

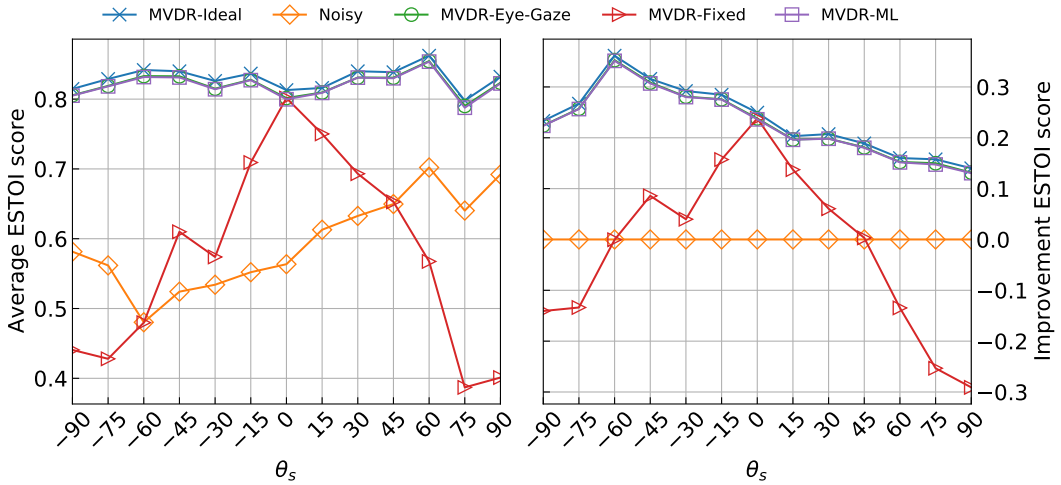
Figure B.3: Average ESTOI scores and average improvement ESTOI scores as a function of target DOA. The noise type is speech shaped noise in an approximately isotropic noise field with an input SNR of 0 dB.



(a) A 2-microphone monaural configuration, using the front and rear microphone on the left HA.

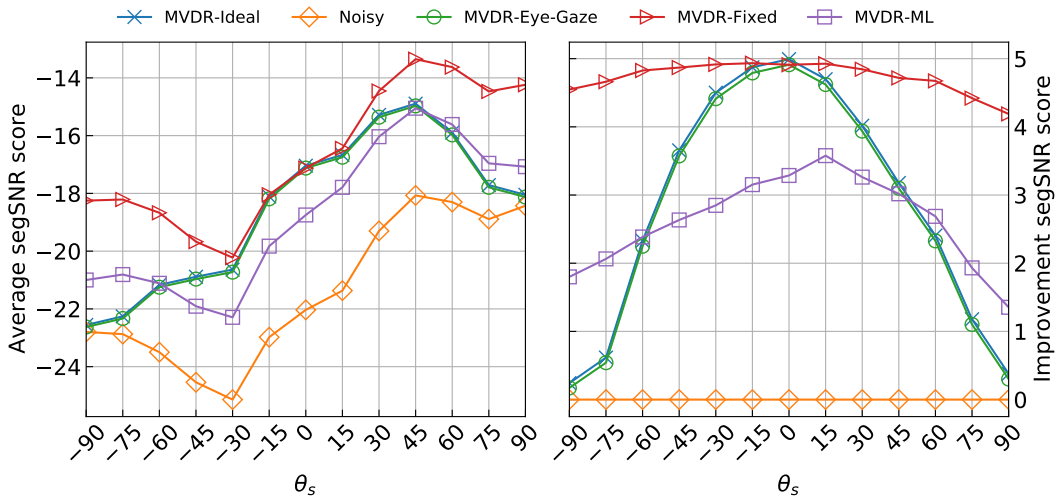


(b) A 4-microphone binaural configuration, using the front and rear microphones on both the left and right HA.

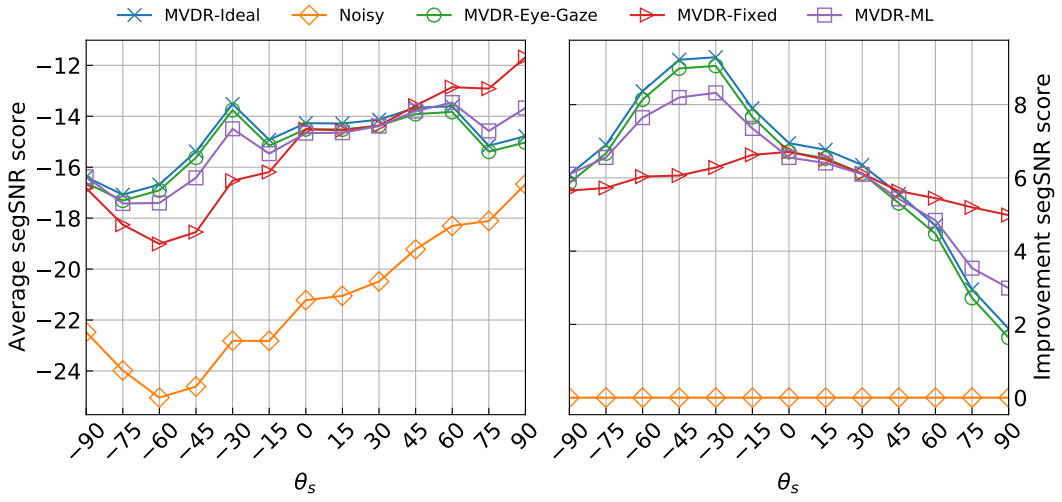


(c) A 6-microphone binaural configuration, using all three microphones on both HAs.

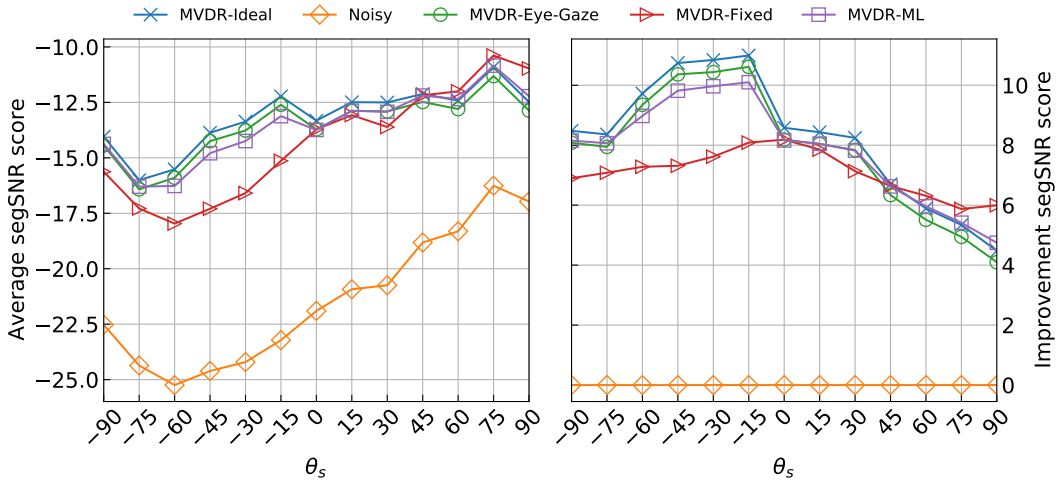
Figure B.4: Average ESTOI scores and average improvement ESTOI scores as a function of target DOA. The noise type is speech shaped noise in an approximately isotropic noise field with an input SNR of 10 dB.



(a) A 2-microphone monaural configuration, using the front and rear microphone on the left HA.

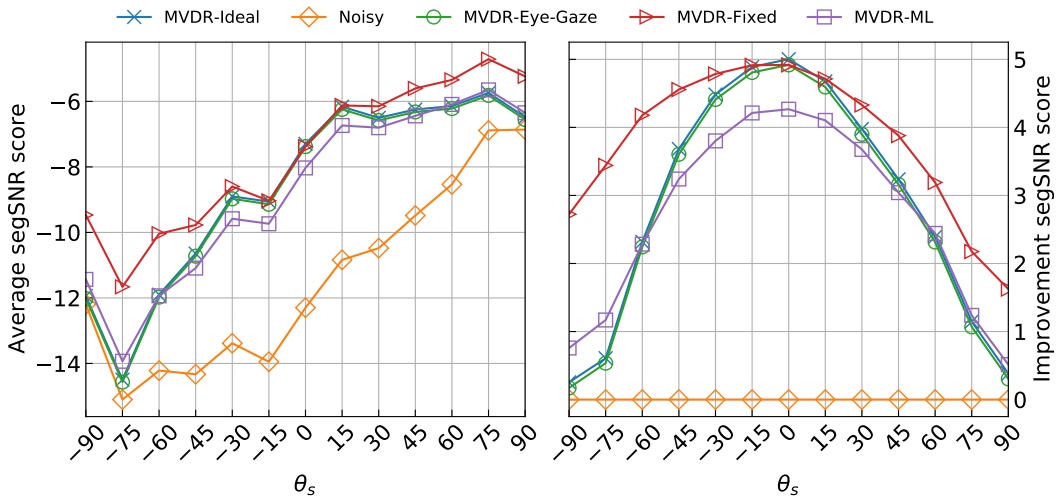


(b) A 4-microphone binaural configuration, using the front and rear microphones on both the left and right HA.

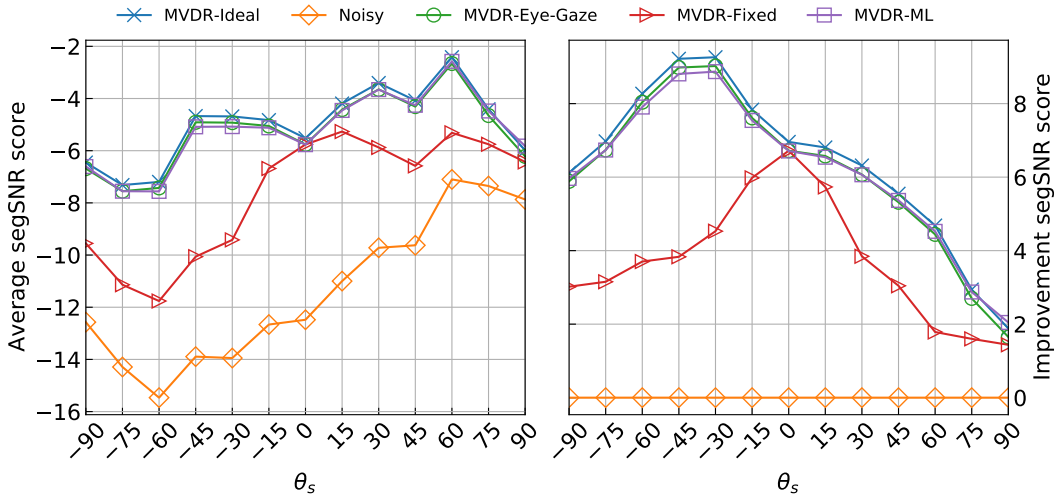


(c) A 6-microphone binaural configuration, using all three microphones on both HAs.

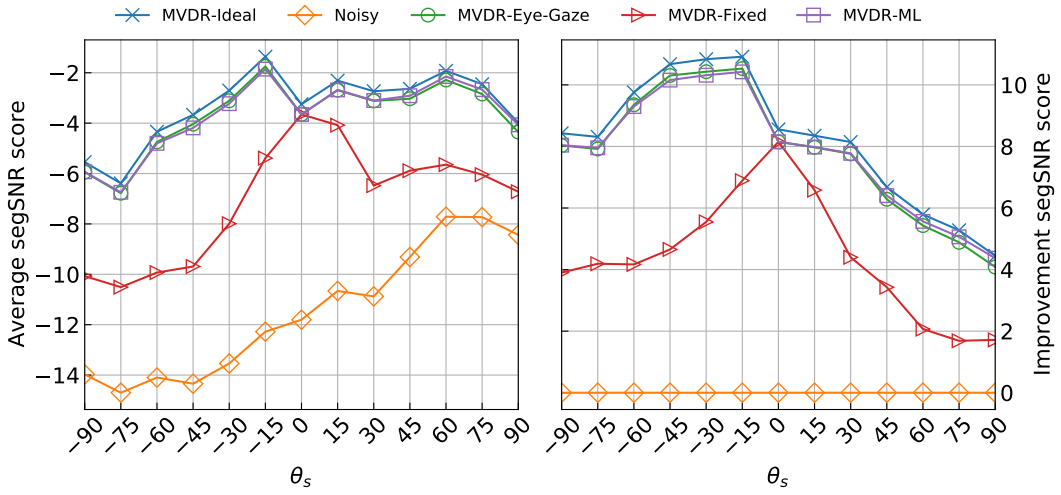
Figure B.5: Average segSNR scores and average improvement segSNR scores as a function of target DOA. The noise type is speech shaped noise in an approximately isotropic noise field with an input SNR of −10 dB.



(a) A 2-microphone monaural configuration, using the front and rear microphone on the left HA.

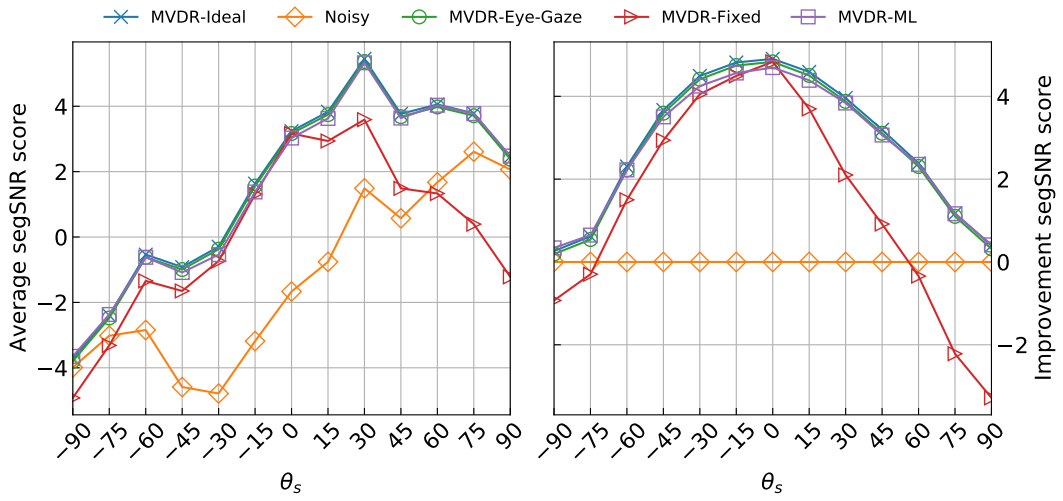


(b) A 4-microphone binaural configuration, using the front and rear microphones on both the left and right HA.

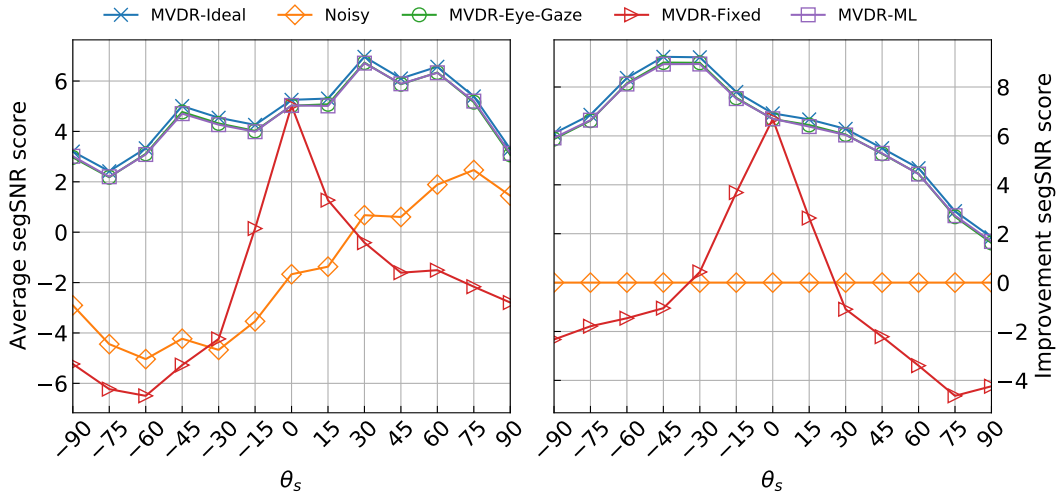


(c) A 6-microphone binaural configuration, using all three microphones on both HAs.

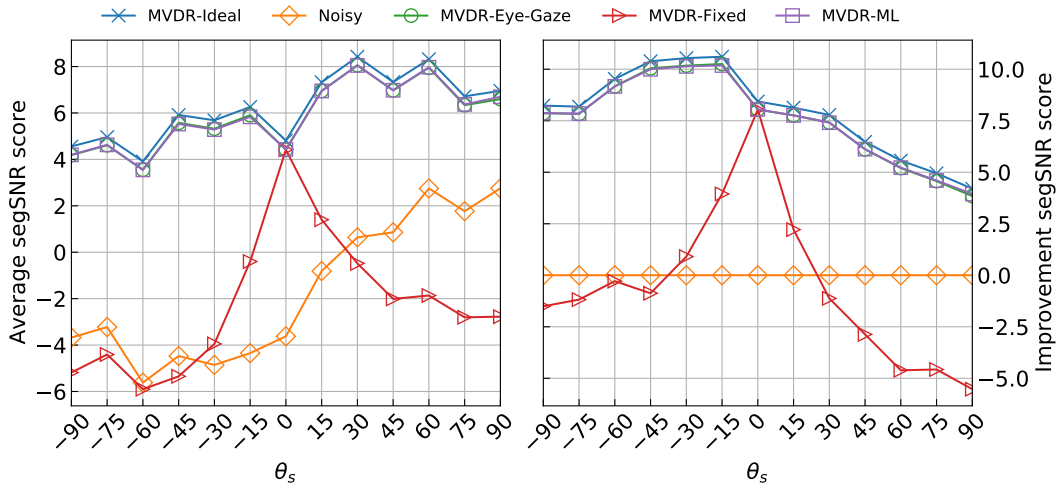
Figure B.6: Average segSNR scores and average improvement segSNR scores as a function of target DOA. The noise type is speech shaped noise in an approximately isotropic noise field with an input SNR of 0 dB.



(a) A 2-microphone monaural configuration, using the front and rear microphone on the left HA.



(b) A 4-microphone binaural configuration, using the front and rear microphones on both the left and right HA.



(c) A 6-microphone binaural configuration, using all three microphones on both HAs.

Figure B.7: Average segSNR scores and average improvement segSNR scores as a function of target DOA. The noise type is speech shaped noise in an approximately isotropic noise field with an input SNR of 10 dB.