User-Symbiotic Speech Enhancement for Hearing Aids

Hoang, Poul

DOI (link to publication from Publisher):
10.54337/aau527273555

Publication date:
2022

USER-SYMBIOTIC SPEECH ENHANCEMENT FOR HEARING AIDS

BY
POUL HOANG

DISSERTATION SUBMITTED 2022

AALBORG UNIVERSITY
DENMARK

User-Symbiotic Speech Enhancement for Hearing Aids

PhD Thesis

Poul Hoang

2022

Dissertation submitted: November 30, 2022

PhD supervisors: Prof. Dr. Zheng-Hua Tan
Aalborg University, Aalborg, Denmark

Prof. Dr. Jesper Jensen
Aalborg University, Aalborg, Denmark
Demant A/S, Smørum, Denmark

Assistant PhD supervisor: Dr. Jan Mark de Haan
Demant A/S, Smørum, Denmark

PhD committee: Professor Søren Bech (chairman)
Aalborg University, Denmark

Professor Sharon Gannot
Bar-Ilan University, Israel

Professor Dr. Reinhold Häb-Umbach
Paderborn University, Germany

PhD Series: Technical Faculty of IT and Design, Aalborg University

Department: Department of Electronic Systems

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-786-4

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Stibo Complete, 2022

About the Author

Poul Hoang

Poul Hoang received his Bachelor's degree in electrical engineering and his Master's degree in signal processing and Computing from Aalborg University, Aalborg, Denmark in 2016 and 2018, respectively. He has since 2018 been working towards his Ph.D. degree at Oticon A/S, Smørum, Denmark in collaboration with Aalborg University, Aalborg, Denmark. Since September 2021, Poul Hoang is employed at Oticon A/S as a DSP specialist working on beamforming and noise reduction for hearing aids. He has a strong research interest in statistical signal processing for audio applications and speech enhancement.

This page intentionally left blank.

Abstract

Speech communication can be challenging for people with a hearing loss, especially in noisy environments that resemble the well-known cocktail-party situation. A hearing loss often leads to a reduced ability to understand speech and therefore requires a greater listening effort for the hearing impaired listener to understand speech, particularly in noisy environments. Hence, it can be highly stressful and exhausting for the hearing impaired listener to participate in noisy events where one is expected to follow conversations.

Hearing aids are electronic devices that help reduce the impact of a hearing loss. Hearing aids cannot fully compensate a hearing loss, but in many cases, they can provide support by improving the audibility and speech intelligibility in noisy environments. A modern hearing aid is typically equipped with at least two microphones to pick-up sound. Based on spatial-temporal filtering, i.e. beamforming and temporal filtering, a hearing aid can suppress the noise as a function of its location and its spectral content. However, in order to achieve optimal and robust noise reduction performance, the parameters of the spatial-temporal filter, e.g. noise field statistics must be provided. Since, noise is often highly non-stationary and might involve competing speakers, the noise statistics need to be estimated online as they are constantly changing.

In this thesis, we propose that better noise reduction can be achieved by taking advantage of behavioral patterns between the hearing aid user and a target speaker who are engaged in a conversation. For example, we demonstrate that these behavioral patterns can provide crucial information about where the desired talker is located. These cues could for example be conversational turn-taking behavior between the user and the target talker. Specifically, we demonstrate how to 1) utilize a prior distribution on the target direction to design robust beamformers, 2) utilize the user's own voice to identify and suppress competing speakers, and 3) exploit an expected conversational turn-taking behavior between the user and target speaker to identify the target speaker in situations with many competing speakers.

This page intentionally left blank.

Resumé

Talekommunikation kan være udfordrende for mennesker med høretab, specielt i støjfyldte omgivelser der minder om den velkendte cocktail-party situation. Et høretab kan ofte føre til en reduceret evne i at forstå tale, og derfor kræves der ofte en større lytteindsats fra den hørehæmmede i at forstå tale i støjfyldte omgivelser. Dermed kan det være yderst stressende og udmattende for en hørehæmmet at deltage i støjfyldte begivenheder, hvor det er forventet at følge og forstå samtaler.

Høreapparater er elektroniske enheder, der kan hjælpe med at reducere effekten af et høretab. Høreapparater kan ikke kurere et høretab, men i mange tilfælde kan de hjælpe med at forbedre hørbarhed og taleforståelighed i støjfyldte miljøer. Et moderne høreapparat er typisk udstyret med to mikrofoner til at opfange lyd. Baseret på rumlig-tidslig filtrering, dvs. beamforming og tidsfiltrering, kan et høreapparat dæmpe baggrundsstøjen som funktion af støjens lokation og dets spektrale indhold. For at opnå optimal og robust støjreduktionsperformance, skal parametrene af det rumlige-tidslige filter vides. Støj er dog ofte yderst ustationært og kan involvere uønskede samtidige talere. Derfor skal parametrene såsom støjstatistikker, estimeres online, da støjen kan ændrer sig hurtigt.

I denne afhandling undersøger vi om bedre støjreduktion kan opnås ved at tage fordel af adfærdsmønstre mellem en høreapparatsbruger og en ønsket taler, når de er engageret i en samtale sammen. For eksempel, forstiller vi at disse adfærdsmønstre kan give afgørende information om, hvor den ønskede taler befinder sig. Disse adfærdsmønstre kunne for eksempel komme fra brugerens øje-retning eller turtagningsmønstre i samtaler. Helt konkret, undersøger vi hvordan man kan 1) udnytte en prior sandsynlighedsfordeling af den ønsket tale retning til at designe robuste beamformers, 2) udnytte brugerens egen-stemme til at identificere og dæmpe uønskede samtidige talere og 3) udnytte et forventet turtagningsmønster mellem en bruger og en ønsket taler til at identificere hvem den ønskede taler er i situationer med mange uønskede samtidige talere.

This page intentionally left blank.

Contents

About the Author iii

Abstract v

Resumé vii

List of Publications xiii

Preface xv

Acknowledgment xvii

I Introduction 1

Introduction 3

1 Introduction to Hearing Loss 4
1.1 The Hearing Sense 4
1.2 Consequences of a Hearing Loss 5
1.3 Turn-taking during conversations 7
2 Hearing Assistive Devices 7
2.1 Overview of Signal Processing in Hearing Aids 9
3 Multichannel Noise Reduction 10
3.1 Overview of Multichannel Noise Reduction 11
3.2 Signal Model of Speech in Noise 12
3.3 Beamforming Methods for Hearing Aids 13
3.4 Post-filtering Techniques 17
4 Parameter Estimation 18
4.1 Noise Estimation and Voice Activity Detection 19
4.2 RTF vector estimation 21
5 Scientific Contributions 22
5.1 Research Challenges 22
5.2 Research topics 23
5.3 Contributions 23
6 Future Research and Direction 27
6.1 Beyond microphones 27

6.2 Behavior-controlled noise reduction 27
6.3 Determining multiple talkers-of-interest exploiting turn-taking 28
6.4 Exploring the speech enhancement paradigm proposed in paper E 28
References 29

II Papers 35

A Robust Bayesian and Maximum a Posteriori Beamforming for Hearing Assistive Devices 37
1 Introduction 39
2 Signal Model and Assumptions 40
3 Wideband Bayesian and MAP Beamforming 41
3.1 Wideband Bayesian beamformer 41
3.2 Posterior probability of the target direction 42
3.3 MAP beamformer 43
4 Experimental Setup 43
5 Results and discussion 46
5.1 Experiment 1: Varying P_t for fixed θ_t 46
5.2 Experiment 2: Varying θ_t for fixed P_t 47
6 Conclusion 47
References 48

B Maximum Likelihood Estimation of the Interference-Plus-Noise Cross Power Spectral Density Matrix for Own Voice Retrieval 51
1 Introduction 53
2 Signal Model 54
3 Maximum Likelihood Interference-Plus-Noise PSD Estimation 55
3.1 Concentrated log-likelihood 56
3.2 ML estimate of $\lambda_c(k,n)$, $\lambda_e(k,n)$, and θ_c 57
4 Performance Evaluation 58
4.1 Simulation setup 58
4.2 Experiment 1: Performance when varying INR 60
4.3 Experiment 2: Performance when varying SNR 60
5 Conclusion 61
References 61

C Joint Maximum Likelihood Estimation of Power Spectral Densities and Relative Acoustic Transfer Functions for Acoustic Beamforming 63
1 Introduction 65

2 Signal Model 66
3 Joint Maximum Likelihood Estimation 67
3.1 Whitening Transform and Eigenvalue Decomposition . 67
3.2 Maximum Likelihood Estimation of the PSDs and RATFs 68
4 Voice Activity Detection Based on Concentrated Log-Likelihood ratios 70
5 Performance Evaluation 71
5.1 Beamforming Methods 72
5.2 Experimental Setup 73
5.3 Simulation Results 73
6 Conclusion 75
References 75

D Multichannel Speech Enhancement with Own Voice-Based Interfering Speech Suppression for Hearing Assistive Devices 79
1 Introduction 81
2 Multi-Microphone Signal Model 85
2.1 Multichannel Wiener filter beamforming 88
2.2 Target and interference identification 89
3 Maximum Likelihood Estimation 90
3.1 Joint ML using grid search 91
3.2 ML in the blocked domain 93
3.3 Unconstrained Joint ML 96
3.4 Robust wideband estimation of the RATF vector 98
4 Performance Evaluation of Proposed Beamforming Systems . . 99
4.1 Acoustic impulse response and sound databases 99
4.2 Simulation of acoustic scenes 100
4.3 Baseline noise reduction systems 102
4.4 Simulation results for canteen and car noise 103
4.5 Simulation results with AIR mismatch and reverberation 106
5 Conclusion 110
A Speech Enhancement Performance Evaluation of the Maximum Likelihood Estimators 111
References 112

E The Minimum Overlap-Gap Algorithm for Speech Enhancement 117
1 Introduction 119
2 Speech interaction in conversations 123
3 The Minimum Overlap-Gap Algorithm 124
3.1 Minimum probability of speech overlap and gap 125
3.2 Bayesian MOG for probability-based speaker ranking . 127
4 Parameter Estimation from Conversational Speech Database . . 131
4.1 Setup and speech database 131

4.2 Parameter Estimation for the Beta-Binomial Distribution 131
4.3 Theoretical Performance of the MOG algorithm 133
5 Evaluation in Speech Enhancement Applications 137
5.1 Speech enhancement system using speaker ranking . . . 137
5.2 Implementation of the MOG and BMOG algorithms . . 138
5.3 State-of-the-art methods for speaker ranking 139
5.4 Application 1: Wireless hearing aid network 143
5.5 Application 2: Beamforming System in Hearing Aids . . 144
6 Conclusion 154
A Appendix 154
A.1 Proof of minimizing speech overlap and gap, and maximizing mutual exclusion 154
A.2 Proof of minimizing speech overlap and gap, and maximizing mean-square-error 155
A.3 Expected misclassification rate for MOG 157
References 159

List of Publications

The main body of the thesis consists of the following publications and are included in Part II:

[A] **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," 2019 *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2019, pp. 1-5, doi: 10.1109/GlobalSIP45357.2019.8969234

[B] **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," ICASSP 2020 - *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6939-6943, doi: 10.1109/ICASSP40776.2020.9053988

[C] **P. Hoang**, Z. -H. Tan, J. M. de Haan and J. Jensen, "Joint Maximum Likelihood Estimation of Power Spectral Densities and Relative Acoustic Transfer Functions for Acoustic Beamforming," ICASSP 2021 - *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6119-6123, doi: 10.1109/ICASSP39728.2021.9414252.

[D] **P. Hoang**, Z.-H. Tan, J. M. de Haan, and J. Jensen, "Multichannel Speech Enhancement with Own Voice-Based Interfering Speech Suppression for Hearing Assistive Devices," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 30, pp. 706-720, 2022.

[E] **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. 14698-14716, 2022.

Co-authors statements for each paper are available to the assessment committee. A summary of each paper is given in Sec. 5.3 in the thesis introduction Part I. Additionally, three patents have been granted in relation to the Ph.D. project:

[F] **P. Hoang**, J. M. de Haan, and J. Jensen, "Hearing device comprising a noise reduction system", pub. no.: US 2021/076124 A1, pub. date: 11. Marts 2021.

[G] **P. Hoang**, J. M. de Haan, M. S. Pedersen, and J. Jensen, "A hearing aid system comprising a database of acoustic transfer functions", pub. no.: US 2022/0174428 A1, pub. date: 2. June, 2022.

[H] **P. Hoang**, J. M. de Haan, M. S. Pedersen, and J. Jensen, "A hearing aid determining talkers of interest", pub. no.: US 2022/0295191 A1, pub. date: 15. September, 2022.

And the following European patent application has been filed:

[I] **P. Hoang**, S. B. B. Thomsen, and J. Jensen, "A hearing aid system comprising a sound source localization estimator", European Patent Appl: 22176296.6, June 2022.

The patents are not included in the thesis.

Preface

The Ph.D. project "User Symbiotic Speech Enhancement for Hearing Aids" was a collaboration between Aalborg University, Aalborg, Denmark and Oticon A/S, Smørum, Denmark. The results as presented in this thesis were partly funded by the Innovation Fund Denmark under the grant application 8053-00011A. The project was carried out between August 2018 to August 2022 and the thesis has been submitted to the Technical Doctoral School of IT and Design at Aalborg University, Aalborg, Denmark as part of the requirements for the Ph.D. degree. The thesis is divided into two parts where part I serves as an extended introduction to the papers, presented in part II. Part II is the main body of the thesis and consists of peer-reviewed and published papers throughout the Ph.D. study.

This page intentionally left blank.

Acknowledgment

First, I would like to thank and express my sincere gratitude for the guidance and support from my supervisors Prof. Jesper Jensen, Dr. Jan Mark de Haan, and Prof. Zheng-Hua Tan. Without their insights, advice, and expertise, the research presented in this thesis would not have been possible. I consider them as my academic role-models, and I am exceptionally grateful for being able to work closely with them throughout my Ph.D. project. Moreover, I would like to thank my colleagues at Oticon especially Armand Myrda, Daniel Michelsanti, Erik Lund, Gary Jones, Jan Mark de Haan, Lena Nilsson, Lucca Julie Nellemann, Michael Gätke, Michael Syskind Pedersen, Mohammad El-Sayed, Robert Rehr, and Vasudha Sathyapriyan for their daily source of inspiration, encouragement, coffee breaks, and support while I was writing this thesis. Thanks to Aalborg University and the professors, researchers, and fellow Ph.D. students at the section of AI and Sound (AIS) for the enlightening discussions we have had. Thanks to the people at Eriksholm Research Centre especially Martha Shiell, and Sergi Rotger Griful, and my former supervisor Thomas Lunner, as they have provided me with valuable feedback, ideas, and insights from their own field of expertise. Special thanks my friends for their unending source of enjoyment and excitement, in particular Rasmus Gundorff Sæderup and Duy Khanh Nguyen. Last but not least, a very special thanks to my family for their everlasting support and comfort.

This page intentionally left blank.

Part I

Introduction

This page intentionally left blank.

Introduction

Our hearing is without doubt one of our most important senses and is crucial for speech communication which a majority of humans use for expressing thoughts and emotions. The auditory system is constantly stimulated by the sound and speech that surrounds us every day. Exposure to speech is particularly important during infancy and early childhood, where the auditory system will mature to perceive speech, which will make it easier to learn spoken language [1–3]. The inability to understand spoken language can impact later education, employment, and social relations for children. Likewise, hearing impairments that arise later in life that reduce speech perception can negatively impact social life. Several types of hearing impairments can fortunately be treated to some degree with hearing assistive devices such as hearing aids. Although, hearing aids cannot restore a normal hearing, they can increase audibility, improve speech understanding, and reduce the listening effort in noisy environments [1, 4]. To do so, many modern hearing aids have access to at least two microphones and a sound processor to deploy advanced signal processing algorithms to reduce the background noise, perform hearing loss compensation, and reduce acoustic feedback [4, 5].

In Part I of this thesis, we will introduce background material which are essential for this thesis, regarding auditory perception and sound processing used in hearing aids. Each of these elements encompasses a large range of aspects, which we cannot cover in this Introduction. However, they will serve as a motivation for the research presented in the papers in Part II of this thesis. In Sec. 1, we introduce the basics of the human hearing sense, the consequences of hearing impairment, and hearing aids. In Secs. 2, 3 and 4, we introduce the multichannel noise reduction algorithms and their parameter estimation commonly used in hearing aids, respectively. Finally in Sec. 5, we motivate and summarize the scientific contributions presented in Part II of this thesis.

1 Introduction to Hearing Loss

A hearing impairment can occur at any time in life, but becomes increasingly common with age [5, 6]. Early symptoms related to hearing impairment are usually mild, e.g., sounds becoming more faint. Furthermore, a hearing impairment usually does not appear over night but may start mildly and then gradually worsen over time, which makes it more difficult to notice. Although, a hearing impairment may not appear life-threatening, hearing impairments are known to reduce quality-of-life [7–9] and potentially cause health issues, e.g., depression and dementia [10, 11]. Many people with mild hearing loss may still be able to participate in normal conversations in quiet and less noisy environments without any help from hearing aids. Since a hearing loss rarely appears over night, a person with a hearing loss may have adapted to the hearing loss. For example, during face-to-face communication, a person with a hearing loss may have learned to efficiently use lip-reading to understand speech better [1]. However, there exist situations which can become overwhelmingly difficult for the person with a hearing loss to participate and follow conversations in. These environments typically resemble the well-known cocktail-party situation, where the speech signal of the talker-of-interest is contaminated by noise and undesired competing speakers, which are known to severely degrade speech intelligibility, especially for persons with a hearing loss [12, 13]. In particular, for a person with hearing loss, such situations may be stressful, unpleasant, and mentally exhausting. Over time, the person with hearing loss may choose to completely avoid these environments, potentially leading to social isolation [1].

1.1 The Hearing Sense

A hearing loss can be caused by many different problems in the auditory system. This section serves as a brief introduction to the human auditory system and types of hearing losses related to it. From a simplified point-of-view, the ear is the organ responsible for converting the acoustic vibrations in the air to electrical pulses sent to the brain in order to perceive these vibrations as being sound [1]. From a very simplified perspective, the ear is comparable to a "biological microphone", whereas the brain is the "processor" responsible for making sense of electrical signals sent from the ear organ.

The peripheral part of the ear organ consists of the outer ear, middle ear, and inner ear. The visible part of the ear is the outer ear which consists of the pinna and the ear canal [1, 14]. The shape of the pinna differs from person to person and helps us localize sound sources [1, 14]. The ear canal is essentially a tube with a length of approximately 2.5 cm and a varying diameter in the order of 0.7 cm and has a resonance frequency at around 3 kHz which happens to be in the spectral region of importance for speech perception [1]. After the sound wave has traveled through the ear canal, it reaches the ear drum which translates the acoustic signal into mechanical movements, i.e., vibrations. The vibrations are carried through the middle ear which amplifies the vibrations before reaching the inner ear. The middle ear consists of the ossicles namely the hammer, anvil, and stirrup [1, 13]. The stirrup is attached to the oval window, a membrane which serves as the entrance to the inner ear. The cochlea is in the inner ear and can be divided into three ducts or "chambers" filled with fluid namely the vestibular ramp, cochlea duct, and the tympanic ramp [1, 13]. The vibrations of the oval window lead to pressure differences between the tympanic ramp and the cochlea duct which are separated by the basilar membrane. The pressure differences create fluid waves which cause the basilar membrane to vibrate. The inner hair cells of the inner ear are located in the organ of Corti which is connected to the basilar membrane. Vibrations in the basilar membrane cause the hair cells to deflect and release neurotransmitters which in turn induce electrical pulses on the auditory nerve. The electrical pulses are transmitted via the auditory nerve through various relay stations to the brain [1, 13].

1.2 Consequences of a Hearing Loss

People with a hearing impairment might not seek medical treatment before they experience poor audibility and speech understanding. The person with a hearing loss may only notice that sounds and speech are becoming more faint, and it becomes difficult to participate in conversations in noisy environments. A hearing loss can be due to one or multiple problems in the ear organ, where some are reversible, while others can cause a permanent hearing loss. A temporary and reversible hearing loss could for example be caused by an infection in the middle-ear and is actually relatively common amongst children [5]. These types of hearing losses are generally less severe and often result in a mild hearing loss. A permanent hearing loss can, for example, be caused by long exposure to noise, age-related hearing loss, and health conditions such as genetic disorders, Meniere's disease, head traumas, and some viral infections [1, 5]. These types of hearing losses cannot be restored to normal hearing and are caused by a permanent damage of parts of the ear organ. For example, long exposure to noise can damage the outer and/or inner hair cells in the cochlea, which results in reduced audibility and sensitivity to sounds.

A hearing loss can be described in overall terms such as *mild*, *moderate*, *severe*, and *profound* hearing loss [15]. A *mild* hearing loss is defined as an overall loss of 10 to 35 dB in hearing sensitivity, a moderate hearing loss is in the range 35 to 70 dB, a severe hearing loss is between 70 dB to 90 dB and a profound hearing loss is greater than 90-100 dB [1]. The sensitivity determines the ability to perceive just-noticeable sound pressure levels. It is closely related to the hearing threshold which indicates the lowest sound level where sound is detectable. A loss in sensitivity is typically measured by a pure-tone audiogram, i.e., the sound pressure level of pure tones that is necessary for a subject to notice their presence [1]. However, a hearing loss is usually more complicated than reduced sensitivity and also involves e.g. reduced frequency resolution, reduced dynamic range, etc. For a more complete overview of different complications related to hearing losses see [13]. Frequency resolution refers to the ability of the auditory system to distinguish between frequency components in sounds. As the hearing loss worsen, the critical bandwidth of the auditory system [16] widens, which creates larger leakage of energy between the auditory frequency bands and hence can cause larger spread of masking [13]. In other words, sound energy which is present at one frequency reduces the sensitivity to sounds of other frequencies. The dynamic range refers to the perceivable range in loudness. It is defined as the range in sound level between the hearing threshold (i.e. the threshold in-quiet) and the uncomfortable loudness level [13]. A reduced dynamic range causes an abnormal growth of perceived loudness for increasing sound pressure level. A reduced temporal resolution can have a negative impact on speech perception as speech is formed by vowels and consonants in quick successions, hence making it difficult to distinguish them [13]. Finally, binaural localization refers to the ability to localize sounds with both ears. A reduced ability to localize sounds can make it more difficult to follow conversations in complex environments with many talkers. In these situations, lip-reading can be helpful to understand speech, but this requires the ability to localize the talker and turn the head towards the talker.

In many situations, a conversation occurs in a noisy environment. Perhaps the most difficult situations to follow conversations in, involve the presence of competing speakers such as in so-called cocktail-party problems [12]. Noise and competing speech is known to severely degrade the intelligibility of a desired speech signal and it usually requires a significantly higher listening effort to understand speech which may cause fatigue [17–19]. Since people with a hearing loss already have a reduced ability to understand speech, cocktail-party environments can be highly unpleasant and uncomfortable [1, 12, 17]. One cause for reduced speech intelligibility in noisy environments is due to energetic masking of the desired speech signal by the noise [12, 20, 21] . Noise refers to any unwanted sounds that have a disturbing effect on the desired speech and includes competing speakers, i.e. unwanted speakers. Due to the redundancy of speech signals, most people are able to deduce what is being said if a few phonemes are being masked by noise, but it usually comes at the cost of higher listening effort and cognitive load [17]. It is therefore not difficult to imagine that a hearing loss can make it extremely difficult to make sense of speech in noise. Another type of masking is informational masking, which occurs when a noise source disturbs the attention of the listener [12].

This can for example be a competing speaker who is talking at the same time as the target speaker and causes a large listening effort. Hearing loss already causes poor audibility and difficulties in simple environments where people with normal hearing find it easy understand speech. The overall goal of hearing aids is therefore to restore as much as possible, the lost audibility of the person with a hearing loss, but also suppress the noise and competing speakers in order to decrease the listening effort in noisy environments [4].

1.3 Turn-taking during conversations

In many situations, people with normal hearing communicate seamlessly through speech even in noisy environments. However, a hearing loss can make speech understanding challenging in environments that otherwise would be simple for normal hearing. Speech communication is typically face-to-face and can involve two or more conversational partners. There are several behavioral patterns and cues that are established when people engage in conversations [22]. These behavioral patterns and cues can potentially hint, where the auditory attention of the conversational partner is focused. One of the most apparent behavioral patterns that exists across language and culture, is the presence of turn-taking [23, 24]. Turn-taking is a social behavioral mechanism where conversational partners coordinate taking turns in talking and listening. It is generally considered socially adept to follow the mechanism of turn-taking in conversations. Breaking the turn-taking mechanism by talking, while the communication partner is talking, can be considered rude. On the other hand, being silent when the conversational partner expects you to talk can be considered socially awkward. The overall behavioral pattern in turn-taking is that conversational partners should avoid long speech overlap and speech gaps as much as possible [24]. Another behavioral pattern that can indicate the attention of the conversational partner is the eye-gaze. For example, it is common that the eyes of a person are gazing towards the conversational partner. Especially in noisy environments, lip-reading requires that the eyes are gazing towards the face of the conversational partner [22].

2 Hearing Assistive Devices

The goal of hearing aids is not necessary to completely restore the normal hearing sense, but to reduce the negative consequences of hearing impairment. Hearing aids come in various styles and are chosen based on the user's needs. For example, some hearing aid styles can be placed in the ear canal making them barely visible, while others are placed behind the ear which allows the hearing aid to be larger and, hence, potentially offer improved hearing loss compensation and noise reduction, e.g. via higher

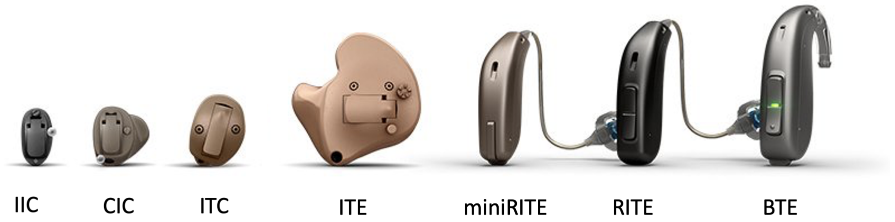


Fig. 1: Different style of hearing aids from Oticon: Behind-the-ear (BTE), receiver-in-the-ear (RITE), mini-receiver-in-the-ear (miniRITE), in-the-ear (ITE), in-the-canal (ITC), completely-in-canal (CIC), invisible-in-canal (IIC).

amplification and more microphones with greater spatial separation. Hearing aid styles that are placed in the ear include *in-the-ear* (ITE), *in-the-canal* (ITC), and *completely-in-canal* (CIC). ITE styles are placed at the entrance of the ear canal whereas ITC and CIC styles are placed in the ear canal with CIC placed deepest in the ear canal. Hearing aids that are placed behind the ear include *behind-the-ear* (BTE) and *receiver-in-the-ear* (RITE). The BTE and RITE hearing aids have their shells, including microphones and electronics placed behind the pinna. However, they have different loudspeaker placements. For BTE hearing aids, the loudspeaker is placed on the hearing aid device behind the pinna and the sound from the loudspeaker is directed through a plastic tube into the ear canal. RITE hearing aids are a hybrid between a BTE and ITC hearing aids. The microphones and electronics are placed behind the pinna, but the loudspeaker is placed in the ear canal. The RITE hearing aid is arguably one of the most popular hearing aids. They usually come with two microphones per device allowing more advanced sound processing compared to one microphone ITE, ITC, and CIC hearing aids.

To help restore audibility and improve speech understanding, modern hearing aids use advanced signal processing algorithms to compensate for the hearing loss by improving audibility and suppress background noise which can improve speech intelligibility [4]. A summary of the basic elements of a hearing aid is shown in Fig. 2. A typical modern hearing aid has one or multiple microphones. In order to process the microphone signals with a digital sound processor, the electrical signals need to be sampled and converted from continuous-time into discrete-time using a analog-to-digital converter (ADC). The discrete-time signal is fed into a sound processor, which typically consists of one or more digital processors which process the discrete-time signal. After processing, the discrete-time signal is converted back to continuous time-domain with a digital-to-analog converter (DAC), then played back through the loudspeaker.

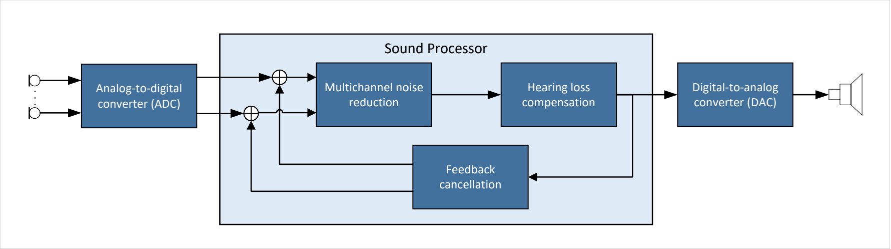

Sound Processor
Analog-to-digital converter (ADC)
Multichannel noise reduction
Hearing loss compensation
Digital-to-analog converter (DAC)
Feedback cancellation

Fig. 2: Simplified overview of the sound processing system in hearing aids. A typical modern hearing aid consists of an ADC which samples the signal to discrete-time which is fed into the sound processor for sound processing. The output of the digital sound processor is converted back into an analog signal which is played back through a loudspeaker. The sound processing system of hearing aids typically consists of multichannel noise reduction, hearing loss compensation, and feedback cancellation.

2.1 Overview of Signal Processing in Hearing Aids

The sound processing system in hearing aids is responsible for manipulating the sound such that audibility, speech intelligibility, and quality are improved. The sound processing algorithms implemented in hearing aids, cf. Fig. 2, can be divided into three categories namely 1) noise suppression, 2) hearing loss compensation, and 3) feedback cancellation [4, 5, 25]. A brief summary of the function of each block is explained in the following.

Multichannel Noise Reduction

Noise reduction is particularly crucial for hearing aid users in noisy environments where speech understanding is poor. The goal of multichannel (or multi-microphone) noise reduction is therefore to enhance the desired speech signal, while attenuating the background noise in order to improve speech intelligibility and quality. This is typically achieved by processing the signal to increase the signal-to-noise ratio (SNR) using spatial and spectro-temporal filters. More details on these types of filters are given in Sec. 3.

Hearing Loss Compensation

After noise reduction, hearing loss compensation algorithms are applied to restore the audibility at the target signal, cf. Fig. 2,. Hearing loss compensation includes amplification, compression and limitation of the sound level [4, 5, 25]. Since the hearing loss is highly individual, the pure-tone audiogram of a hearing aid user may serve as input information for the hearing loss compensation algorithms [5]. The hearing loss compensation algorithm restores lost hearing sensitivity by amplifying sounds and ensuring that the

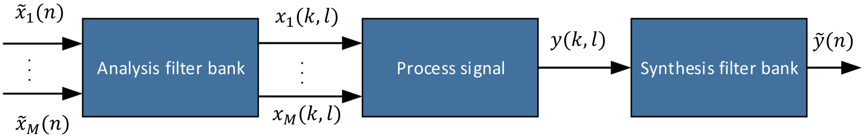


Fig. 3: A typical multichannel noise reduction system, consists of an analysis filter bank that transforms the microphone signals into the time-frequency domain for processing. The synthesis filter bank transforms the processed signal back to the discrete-time domain.

processed sound remains within the auditory range of the user, using a compressor and limiter [4, 5, 25].

Feedback Cancellation

Hearing aid microphones are typically placed relatively close to the loudspeaker - when combined with a potentially large signal amplification in the hearing aid, an unstable acoustic feedback loop may be created. This unstable system creates a clearly audible "howling" sound, which may completely mask the target signal - it may even be audible to people standing next to the hearing aid user. The problem with acoustic feedback becomes increasingly severe, if the hearing aid user needs large amounts of sound amplification. Furthermore, popular types of ear domes, i.e. the silicon tip attached at the end of the hearing aid, such as "open ear domes" are more prone to cause acoustic feedback as sound can leak more easily to the microphones [4]. Since acoustic feedback reduces the intelligibility of speech and significantly lowers the sound quality, feedback cancellation algorithms are introduced, which aim to reduce the effect of acoustic feedback by designing adaptive filters that minimize the feedback [4].

3 Multichannel Noise Reduction

In this thesis, we primarily focus on the multichannel noise reduction system in hearing aids. In particular, the papers in Part II cover topics within the area of multichannel noise reduction. Therefore, this section serves an an introduction to the multichannel noise reduction as implemented in hearing aids. The signal processing chain typically starts with an analysis filter bank and ends with a synthesis filter bank as depicted in Fig. 3 [4, 5]. The analysis filter bank is used to transform the time-domain microphone signals into the time-frequency domain for sound processing. The synthesis filter bank is used to transform the processed time-frequency domain signal back to the discrete-time domain. Often, the analysis filter bank is based on the short-time Fourier transform (STFT) and the synthesis filter bank is based on the inverse short-time Fourier transform (iSTFT) [4, 5]. Let $\tilde{x}_m(n) \in \mathbb{R}, m \in \{1, \ldots, M\}$ denote the noisy time-domain signal picked up by the m'th microphone where n is the discrete-time index. The time-frequency domain representation of the noisy speech signal, i.e. the output of the analysis filter bank, is given by

$$x_m(k,l) = \sum_{n=0}^{N_{win}-1} \psi_a(n)\tilde{x}_m(n+lD)e^{-2\pi jk\frac{n}{N_{win}}}, \quad x_m(k,l) \in \mathbb{C}, \tag{1}$$

where, $j = \sqrt{-1}$, k is the frequency bin index, l is the frame index, N_{win} is the window length, $\psi_a(n)$ is the analysis window function, and D is the window overlap size. For synthesis, the iSTFT is given as

$$\tilde{y}_m(n+lD) = \frac{1}{N_{win}} \sum_{k=0}^{N_{win}-1} \psi_s(k)y_m(k,l)e^{2\pi jk\frac{n}{N_{win}}}, \tag{2}$$

where $y_m(k,l) \in \mathbb{C}$ is the processed signal in the time-frequency domain, $\psi_s(n)$ is the synthesis window function, and $\tilde{y}_m(n+lD) \in \mathbb{R}$ is the processed time-domain signal. In this case, the square-root Hann window function is often used as analysis and synthesis windows with a window overlap of 50% in order to reduce aliasing artifacts and to assure perfect reconstruction of the input signal in the case that the "process signal" stage does not alter the signal [26].

3.1 Overview of Multichannel Noise Reduction

A typical multi-channel noise reduction system consists of a beamformer and a post filter [27]. The beamformer is able to attenuate sounds as a function of direction (i.e., it is a spatial filter). Beamformers are very effective at suppressing undesired sound sources located at positions in the environment different from the position of the desired talker [28]. Beamformers are typically implemented as a linear combination of noisy microphone signals in the time-frequency (TF) domain:

$$z(k,l) = \sum_{m=1}^{M} w_m^*(k,l)x_m(k,l), \tag{3}$$

where $*$ denotes the complex conjugate of a complex variable, and $w_m(k,l) \in \mathbb{C}$ denotes the beamforming weight for the m'th microphone. For brevity, vector notation is often used such that $\boldsymbol{x}(k,l) = [x_1(k,l), ..., x_M(k,l)]^T \in \mathbb{C}^M$, $\boldsymbol{w}(k,l) = [w_1(k,l), ..., w_M(k,l)]^T \in \mathbb{C}^M$, and $\boldsymbol{z}(k,l) = [z_1(k,l), ..., z_M(k,l)]^T \in \mathbb{C}^M$. In vectorized form, the output of the beamformer can therefore be written as

$$z(k,l) = \boldsymbol{w}^H(k,l)\boldsymbol{x}(k,l), \tag{4}$$

where H denotes the Hermitian transpose. Well-designed beamformers are able to improve speech intelligibility and quality in noisy environments, due to their ability to attenuate noise components that are present in time-frequency tiles where target speech is present [4].

A single-channel spectro-temporal filter, i.e., a post-filter, is then applied to remove any remaining residual noise. The post-filter is typically a scalar gain $g(k,l) \in \mathbb{R}$ that is applied at the output of the beamformer for each frequency bin according to:

$$y(k,l) = g(k,l)z(k,l). \tag{5}$$

Post-filters are particularly known to improve speech quality and can decrease the perceived loudness of background noise improve and listening effort by attenuating noise in noise dominant time-frequency tiles [4, 5].

3.2 Signal Model of Speech in Noise

The most well-known beamformers are formulated based on a signal model of the acoustic environment. In this section, we introduce the multichannel signal model that is commonly used to model a single target speech signal in noise. For each time-frequency tile, we assume that there is a maximum of one target speech signal present at a time. In the presence of speech in a time-frequency tile, the noisy speech signal received at the microphones is modeled using two signal components, namely a desired speech signal and a noise component [27]. The noise can essentially be any undesired sound signal such as reverberation, restaurant background noise, and competing speakers, etc. As before, let $\tilde{x}_m(n)$ be the received signal at the m'th microphone, and let $\tilde{s}'(n)$ be the clean target signal as measured at the location of the sound source. Let $\tilde{a}_m(n)$ be the acoustic impulse response from the target sound source location to the m'th microphone, and let $\tilde{v}_m(n)$ be the additive noise at the m'th microphone. The signal model may then be expressed as

$$\tilde{x}_m(n) = \tilde{s}'(n) * \tilde{a}_m(n) + \tilde{v}_m(n). \tag{6}$$

The STFT is applied to transform the signal to the time-frequency domain such that:

$$x_m(k,l) = s'(k,l)a_m(k,l) + v_m(k,l) \tag{7}$$

where the equivalence between (6) and (7) holds strictly, under the assumption that the STFT window length is longer than the length of the impulse response $\tilde{a}_m(n)$ in (7) [29]. For relatively short window lengths, e.g., 256 samples at a sample rate of 16 kHz, this narrow-band approximation may not hold true in realistic environments where the reverberation time is likely to be longer than the window length. However, this issue can be overcome by redefining the signal model using relative transfer functions (RTFs) [30]. It is common to define the signal-of-interest as the desired speech signal at a pre-selected reference microphone in contrast to the desired speech signal as received at the location of the sound source. Under this redefinition, we define $s(k,l) = s'(k,l)a_1(k,l)$ to be the target speech signal, where, without loss of generality, microphone '1' is selected as the reference microphone. The difference between retrieving $s(k,l)$ and $s'(k,l)$ is that head and torso effects, which are important for signal quality and naturalness, are captured in $s(k,l)$ but not in $s'(k,l)$. The signal model is now

$$x_m(k,l) = s(k,l)\frac{a_m(k,l)}{a_1(k,l)} + v_m(k,l), \quad m = 1, ..., M, \tag{8}$$

where $d_m(k,l) \triangleq \frac{a_m(k,l)}{a_1(k,l)}$ is often referred to as the RTF from the reference microphone to the m'th microphone [27, 30]. Let $\boldsymbol{d}(k,l) = [1, d_2(k,l)..., d_M(k,l)]^T$ be the RTF vector, then the signal model becomes

$$\boldsymbol{x}(k,l) = s(k,l)\boldsymbol{d}(k,l) + \boldsymbol{v}(k,l), \tag{9}$$

where $d(k,l)$ is considered deterministic, while $s(k,l)$, $\boldsymbol{v}(k,l)$, and, hence, $\boldsymbol{x}(k,l)$ are considered random vectors. Linear MMSE estimators of $s(k,l)$ are typically functions of the first- and second-order statistics of the random terms in (8). The first order statistics of $\boldsymbol{x}(k,l)$, $s(k,l)$, and $\boldsymbol{v}(k,l)$ are typically assumed to be zero-mean [31]. The second order statistics are the power spectral density (PSD) of $s(k,l)$, and the (spatial) cross power spectral density (CPSD) matrix of $\boldsymbol{x}(k,l)$ and $\boldsymbol{v}(k,l)$ [27]. The CPSD matrix of $\boldsymbol{x}(k,l)$ is defined as $\boldsymbol{C}_x(k,l) \triangleq \mathbb{E}\left[\boldsymbol{x}(k,l)\boldsymbol{x}^H(k,l)\right]$ which is referred to as the noisy speech CPSD matrix. Under the assumption that the target signal and the noise are uncorrelated, the noisy speech CPSD matrix is given as

$$\boldsymbol{C}_x(k,l) = \boldsymbol{C}_s(k,l) + \boldsymbol{C}_v(k,l). \tag{10}$$

Using the assumption that $\boldsymbol{d}(k,l)$ is deterministic, the target CPSD matrix $\boldsymbol{C}_s(k,l)$ can be expressed as $\boldsymbol{C}_s(k,l) = \lambda_s(k,l)\boldsymbol{d}(k,l)\boldsymbol{d}^H(k,l)$. Furthermore, the noise CPSD matrix $\boldsymbol{C}_v(k,l)$ is decomposed (without loss of generality) into $\boldsymbol{C}_v(k,l) = \lambda_v(k,l)\boldsymbol{\Gamma}_v(k,l)$, where $\lambda_v(k,l) = \mathbb{E}\left[|v_1(k,l)|^2\right]$ is the noise PSD at the reference microphone index and $\boldsymbol{\Gamma}_v(k,l) \triangleq \boldsymbol{C}_v(k,l)/\lambda_v(k,l)$ is the normalized noise CPSD matrix, where the first diagonal element of $\boldsymbol{\Gamma}_v(k,l)$ is equal to '1' [32]. The noisy speech CPSD matrix then becomes

$$\boldsymbol{C}_x(k,l) = \lambda_s(k,l)\boldsymbol{d}(k,l)\boldsymbol{d}^H(k,l) + \lambda_v(k,l)\boldsymbol{\Gamma}_v(k,l). \tag{11}$$

3.3 Beamforming Methods for Hearing Aids

This section will cover beamforming methods that are commonly used in hearing aids. These beamformers are typically low in complexity and perform robustly in most acoustic environments. The beamformers that are covered in this section are the delay-and-sum beamformer (DSB) [28], minimum power distortion-less response (MPDR) beamformer [27, 33], minimum variance distortion-less response (MVDR) beamformer [34], and the multichannel Wiener filter (MWF) [28]. The beamformer coefficients of these beamformers are found through solving a constrained optimization problem of the general form:

$$\begin{aligned} \hat{\boldsymbol{w}}(k,l) &= \arg\min_{\boldsymbol{w}(k,l)} f(\boldsymbol{w}(k,l)) \\ &\text{subject to } g(\boldsymbol{w}(k,l)) = 0, \end{aligned} \tag{12}$$

where $f(\boldsymbol{w}(k,l))$ is a cost function reflecting noise reduction performance, and $g(\boldsymbol{w}(k,l))$ is a constraint function. An overview of cost functions, constraint functions, and expressions for the corresponding solutions to (12) are given in Table 1. The DSB, MPDR, and MVDR beamformers have a distortion-less constraint, i.e., $\boldsymbol{w}^H(k,l)\boldsymbol{d}^H(k,l) - 1 = 0$, which ensures that the beamformer 1) does not alter the magnitude and phase of the target speech signal when it originates from a pre-selected direction, and 2) the solution to (12) is not the trivial solution $\boldsymbol{w}(k,l) = \mathbf{0}$.

	$f(\boldsymbol{w})$	$g(\boldsymbol{w})$	$\hat{\boldsymbol{w}}$
DSB	$\lvert\lvert\boldsymbol{w}\rvert\rvert_2^2$	$\boldsymbol{w}^H\boldsymbol{d} - 1$	$\frac{\boldsymbol{d}}{\lvert\lvert\boldsymbol{d}\rvert\rvert_2^2}$
MPDR	$\boldsymbol{w}^H\mathbf{C}_x\boldsymbol{w}$	$\boldsymbol{w}^H\boldsymbol{d} - 1$	$\frac{\mathbf{C}_x^{-1}\boldsymbol{d}}{\boldsymbol{d}^H\mathbf{C}_x^{-1}\boldsymbol{d}}$
MVDR	$\boldsymbol{w}^H\mathbf{C}_v\boldsymbol{w}$	$\boldsymbol{w}^H\boldsymbol{d} - 1$	$\frac{\mathbf{C}_v^{-1}\boldsymbol{d}}{\boldsymbol{d}^H\mathbf{C}_v^{-1}\boldsymbol{d}}$
MWF	$\mathbb{E}\left[\lvert s - \boldsymbol{w}^H\boldsymbol{x}\rvert^2\right]$	0	$\mathbf{C}_x^{-1}\,\mathbf{c}_{xs}$

Table 1: List of beamformers with cost functions $f(\boldsymbol{w})$, constraint functions $g(\boldsymbol{w})$, and their solutions $\hat{\boldsymbol{w}}$. The frequency and frame indices (k,l) have been omitted for brevity.

3.3.1 Delay-and-sum beamformer

The DSB only requires the RTF vector, and can be shown to achieve optimal noise reduction performance, in the special case, when the noise is spatially white, i.e., the noise CPSD matrix is a scalar times an identity matrix [35, 36]. The DSB does not adapt to the noise field of the acoustic environment, hence it is mostly well-suited for attenuating microphone self-noise which may be considered uncorrelated across microphones, and spatially weakly correlated noise components from the environment [36].

3.3.2 Minimum power distortion-less response beamformer

The MPDR beamformer seeks to minimize the noisy speech signal power, while ensuring that signals arriving from the location encoded in $\boldsymbol{d}(k,l)$ re-

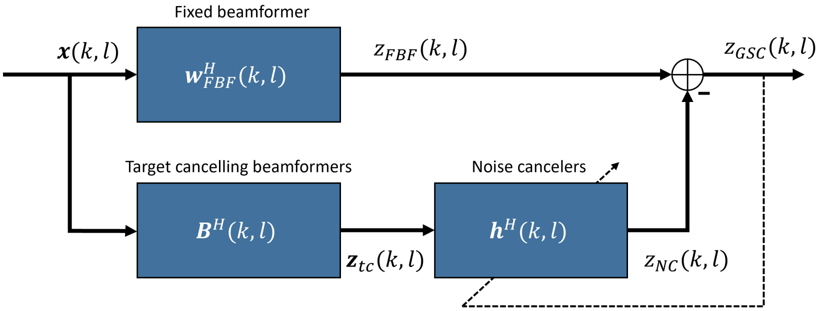

Fixed beamformer
$\boldsymbol{x}(k,l)$
$\boldsymbol{w}_{FBF}^{H}(k,l)$
$z_{FBF}(k,l)$
$z_{GSC}(k,l)$
−
Target cancelling beamformers
$\boldsymbol{B}^{H}(k,l)$
$\boldsymbol{z}_{tc}(k,l)$
Noise cancelers
$\boldsymbol{h}^{H}(k,l)$
$z_{NC}(k,l)$

Fig. 4: The generalized sidelobe canceller (GSC) structure, which consists of a fixed beamformer, target cancelling beamformer, and adaptive noise cancellers. The noise cancellers can optionally be made as adaptive filters.

main distortion-less [36]. Compared to the DSB, the MPDR beamformer is more effective at suppressing spatially correlated noise such as interfering point sources or diffuse noise fields [36]. However, mismatches in the RTF vector $\boldsymbol{d}(k,l)$, can lead to severe distortion and attenuation of the target speech signal for the MPDR beamformer [34].

3.3.3 Minimum variance distortion-less response beamformer

The sensitivity to RTF vector mismatches in the MPDR can be reduced by the closely related MVDR beamformer [34]. The MVDR beamformer is similar to the MPDR beamformer, but requires the noise CPSD matrix, $\boldsymbol{C}_v(k,l)$, to be known. It can easily be shown that the MVDR and MPDR beamformers return the identical solution of $\hat{\boldsymbol{w}}(k,l)$, when the target speech and noise are assumed uncorrelated, the true RTF vector is known in advance, and true CPSD matrices of $\boldsymbol{C}_x(k,l)$ and $\boldsymbol{C}_v(k,l)$ are known. However, these assumptions are generally not satisfied in practice, and the MVDR beamformer typically reveals more robust performance than the MPDR beamformer [34], e.g. less target speech distortion. This is especially apparent in situations where there is a mismatch between the assumed RTF vector and the actual RTF vector which may lead to suppression of the target speech signal.

3.3.4 Multichannel Wiener filter

The MWF beamformer is the multichannel extension to the well-known single-channel Wiener filter (discussed i Sec. 3.4). The MWF is a linear minimum mean-square-error (LMMSE) estimator of the target speech DFT coefficient $s(k,l)$ [28]. In contrast to the other beamformers, the MWF does not involve a distortion-less constraint implicitly. The MWF beamformer can be expressed as the the matrix-vector product between the noisy speech CPSD matrix and the cross CPSD vector between the noisy and target speech signal as shown in Table 1. However, it can be shown that the MWF beamformer can be decomposed into an MVDR beamformer followed by a Wiener post-filter[1] under the standard assumption that the noisy signal consists of exactly one target signal in noise that is uncorrelated with the target [28]. Specifically, the decomposed MWF beamformer is given as

$$\boldsymbol{w}_{\text{MWF}}(k,l) = \boldsymbol{w}_{\text{MVDR}}(k,l) \cdot g_{\text{WF}}(k,l), \tag{13}$$

where $g_{\text{WF}}(k,l)$ is the Wiener gain.

3.3.5 Generalized sidelobe canceller structure

The MPDR and MVDR (and MWF) beamformers can be decomposed into a set of orthogonal beamformers and implemented using the generalized sidelobe canceller (GSC) structure which is shown in Fig. 4 [37–39]. The GSC structure consists of an upper branch and a lower branch. The upper branch in Fig. 4 projects the noisy speech signal onto the target subspace using a fixed beamformer $\boldsymbol{w}_{\text{FBF}}(k,l)$. The fixed non-adaptive beamformer ensures the distortion-less constraint on the target speech signal. The lower branch projects the noisy speech signal onto the noise subspace using a so-called blocking matrix $\boldsymbol{B}(k,l)$ [27] where each column vector of the blocking matrix can be considered as target canceling beamformers. Target cancelling beamformers in $\boldsymbol{B}(k,l)$ are followed by the noise cancelers $\boldsymbol{h}(k,l)$, which seek to minimize the residual noise in the fixed beamformer branch. The GSC filter weights can then be expressed as [27, 40]:

$$\boldsymbol{w}_{\text{GSC}}(k,l) = \boldsymbol{w}_{\text{FBF}}(k,l) - \boldsymbol{B}(k,l)\boldsymbol{h}(k,l), \tag{14}$$

where $\boldsymbol{B}(k,l) \in \mathbb{C}^{M \times M-1}$ is the blocking matrix where column vectors are linear-independent a target canceling beamformer. The output of the GSC structure is given as

$$\begin{aligned} z_{\text{GSC}}(k,l) &= \boldsymbol{w}_{\text{FBF}}^{H}(k,l)\boldsymbol{x}(k,l) - \boldsymbol{h}^{H}(k,l)\boldsymbol{B}^{H}(k,l)\boldsymbol{x}(k,l) \\ &= z_{\text{FBF}}(k,l) - \boldsymbol{h}^{H}(k,l)\boldsymbol{z}_{\text{tc}}(k,l) \end{aligned} \tag{15}$$

where $z_{\text{FBF}}(k,l)$ is the output of the fixed beamformer and $\boldsymbol{z}_{\text{tc}}(k,l) \in \mathbb{C}^{M-1}$ is the output of the $M-1$ target canceling beamformers. The output of the target canceling beamformers ideally only contain signal components related to the noise $\boldsymbol{v}(k,l)$ but are unrelated to the target signal. Hence, the noise canceller are designed to minimize the power of the residual noise at the output of the fixed beamformer [27, 40]. This minimization can be expressed

[1] A Wiener post-filter refers to a single-channel Wiener filter being applied at the output of the MVDR beamformer.

in terms of minimizing the MSE between the output of the fixed beamformer and the output of the noise canceller i.e. [37]

$$\arg\min_{\boldsymbol{h}(k,l)} \mathbb{E}\left[|\boldsymbol{w}_{\mathrm{FBF}}^{H}(k,l)\boldsymbol{x}(k,l) - \boldsymbol{h}^{H}(k,l)\boldsymbol{B}^{H}(k,l)\boldsymbol{x}(k,l)|^{2}\right]. \tag{16}$$

The closed-form solution to (16) is given by

$$\boldsymbol{h}_{\mathrm{GSC},1}(k,l) = \left(\boldsymbol{B}^{H}(k,l)\boldsymbol{C}_{x}(k,l)\boldsymbol{B}(k,l)\right)^{-1}\boldsymbol{B}^{H}(k,l)\boldsymbol{C}_{x}(k,l)\boldsymbol{w}_{\mathrm{FBF}}(k,l). \tag{17}$$

It can be shown that $\boldsymbol{h}_{\mathrm{GSC},1}(k,l)$ is identical to the MPDR beamformer when setting $\boldsymbol{w}_{\mathrm{FBF}}$ equal to the DSB and by inserting (17) into (14) [40]. To obtain the MVDR beamformer, $\boldsymbol{C}_{x}(k,l)$ is substituted with $\boldsymbol{C}_{v}(k,l)$ in (17) i.e.

$$\boldsymbol{h}_{\mathrm{GSC},2}(k,l) = \left(\boldsymbol{B}^{H}(k,l)\boldsymbol{C}_{v}(k,l)\boldsymbol{B}(k,l)\right)^{-1}\boldsymbol{B}^{H}(k,l)\boldsymbol{C}_{v}(k,l)\boldsymbol{w}_{\mathrm{DSB}}(k,l). \tag{18}$$

The GSC structure has desirable properties related to computational complexity and adaptive constrained optimization [27]. For example, MVDR beamformers should be adaptive as noise statistics can change over time in realistic environments e.g. due to changing acoustic environments and head movements of the user.

The GSC structure is more convenient to implement as an adaptive system, as it only requires minimization of an unconstrained optimization problem, whereas the original problem posed by the MVDR beamformer has a linear constraint [27, 41].

3.4 Post-filtering Techniques

Post-filter techniques include methods ranging from heuristically motivated techniques such as spectral subtraction [42–46] to statistical-model based methods, e.g., relying on the minimum mean-square error [31, 45, 47–52]. The spectral-subtraction method has existed for several decades and is amongst the earliest single-channel speech enhancement techniques [50]. It can be shown that the spectral-subtraction method is a maximum likelihood estimator of the clean speech PSD under complex Gaussian assumptions of the target and noise DFT coefficients [31, 48]. It is based on estimating the clean speech magnitude spectrum by subtracting the noise spectrum from the noisy speech spectrum. For each time-frequency tile, an estimate of the noise spectrum is obtained during speech absence and subsequently used to estimate the clean speech spectrum during speech presence. The estimated clean speech spectrum is then used to compute the post-filter gains, which are applied by point-wise multiplication with the noisy STFT coefficients. Although, the spectral subtraction method is straightforward to implement and computationally cheap, it has a tendency to create undesired sound artifacts such as musical noise [31].

The LMMSE estimator of the clean speech DFT i.e. the Wiener filter (or Wiener gain in the frequency domain), is often seen used in the context of post-filtering. Let $g_{\text{wf}}(k,l)$ be the Wiener gain, then the optimization problem of the LMMSE estimator of the clean speech DFT is formulated as [50]

$$g_{\text{WF}}(k,l) = \arg\min_{g(k,l)} \mathbb{E}\left[|s(k,l) - g(k,l)z(k,l)|^2\right], \tag{19}$$

where $z(k,l) = s(k,l) + v_{\text{res}}(k,l)$ is the output of the MVDR beamformer and $v_{\text{res}}(k,l)$ is the residual noise after applying the MVDR beamformer to the noisy microphone signal. The solution to (19) is given by:

$$g_{\text{WF}}(k,l) = \frac{\mathbb{E}[s^*(k,l)z(k,l)]}{\mathbb{E}[|z(k,l)|^2]}. \tag{20}$$

It is often assumed that the clean speech signal and residual noise are uncorrelated, which results in the more common form of the Wiener filter given as

$$g_{\text{WF}}(k,l) = \frac{\mathbb{E}[|s(k,l)|^2]}{\mathbb{E}[|z(k,l)|^2]} = \frac{\lambda_s(k,l)}{\lambda_s(k,l) + \lambda_{v_{\text{res}}}(k,l)}, \tag{21}$$

where $\lambda_{v_{\text{res}}}(k,l)$ is the PSD of the residual noise after the beamformer. The Wiener filter may also be expressed as a function of signal-to-noise ratio (SNR), i.e.,

$$g_{\text{WF}}(k,l) = \frac{\text{SNR}(k,l)}{1 + \text{SNR}(k,l)}, \tag{22}$$

where $\text{SNR}(k,l) \triangleq \frac{\lambda_s(k,l)}{\lambda_{v_{\text{res}}}(k,l)}$ is the SNR of the output signal of the beamformer.

4 Parameter Estimation

In practice, the parameters of the beamformer and the post-filter are unknown and must therefore be estimated from the noisy microphone signals. Parameters that need to be estimated include the noisy speech CPSD matrix $\boldsymbol{C}_x(k,l)$, the noise CPSD matrix $\boldsymbol{C}_v(k,l)$, and target RTF vector $\boldsymbol{d}(k,l)$. Parameters such as the target PSD $\lambda_s(k,l)$ and residual noise PSD $\lambda_{v_{\text{res}}}(k,l)$ can be estimated through $\boldsymbol{C}_x(k,l)$, $\boldsymbol{C}_v(k,l)$, and $\boldsymbol{d}(k,l)$. In many practical noisy environments, these parameters must be estimated online. In particular, the noisy and noise CPSD matrix are likely time-varying. In many situations, $\boldsymbol{d}(k,l)$ is also a time-varying parameters. e.g., because the position of the target speaker changes over time, because the target position changes or due to head movements of the hearing aid user.

4.1 Noise Estimation and Voice Activity Detection

Noise in realistic environments is often non-stationary, i.e., the noise statistics change over time. Hence, it is important that the noise reduction system is able to track these changes to ensure good performance. In many cases, estimation of the noise statistics is done using a voice activity detector (VAD). The task of the VAD is to detect time-frequency tiles where target speech is present (or absent). The VAD is particularly crucial for estimation of the noise statistics, since the VAD can inform the noise reduction system to update the noise statistics during speech absence. A first order recursive update is often used for this purpose, i.e.,

$$\boldsymbol{C}_v(k,l) = \begin{cases} \alpha \boldsymbol{C}_v(k,l-1) + (1-\alpha)\boldsymbol{x}(k,l)\boldsymbol{x}^H(k,l) & \text{if speech absent,} \\ \boldsymbol{C}_v(k,l-1) \quad \text{otherwise,} \end{cases} \tag{23}$$

If the VAD decision is binary, the conventional approach is to estimate the noise statistics during speech absence and reuse the most recent noise estimate during speech presence [27].

However, the decision of speech presence or absence involves uncertainty, and hence noise statistics may be updated using a speech presence probability (SPP) estimate instead [27, 39, 53–55]. The SPP VAD has a similar function as the binary VAD, but returns a probability value between 0 and 1 that describes the a posteriori probability that speech is present in a particular time frequency tile given the noisy observation. The SPP is commonly used to control the adaptation factor α, used in the recursive estimation of the noise statistics of (23), such that when the SPP is close to zero the noise reduction algorithm rapidly updates the noise statistics, whereas a SPP close to one means that the noise statistics should remain constant [39, 53–58]. Specifically, in this case the smoothing factor α is in (23) controlled by the SPP estimator. Estimation of the SPP is typically formulated as a binary hypothesis test, where H_0 is the hypothesis that speech is absent, and H_1 is the hypothesis that speech is present [46]. We estimate the probability that hypothesis H_1 is true given observations $\boldsymbol{x}(k,l)$, i.e., $P(H_1|\boldsymbol{x}(k,l))$. To obtain $P(H_1|\boldsymbol{x}(k,l))$, we use Bayes theorem such that

$$P(H_1|\boldsymbol{x}(k,l)) = \frac{1}{1+\Lambda(k,l)}, \tag{24}$$

where $\Lambda(k,l)$ is the generalized likelihood ratio given as [48]

$$\Lambda(k,l) = \frac{P(H_0)f(\boldsymbol{x}(k,l)|H_0)}{P(H_1)f(\boldsymbol{x}(k,l)|H_1)}. \tag{25}$$

In (25) $P(H_0)$ and $P(H_1)$ are prior probabilities of speech being absent and present in a particular TF tile, respectively, and $f(\boldsymbol{x}(k,l)|H_0)$ and $f(\boldsymbol{x}(k,l)|H_1)$

are likelihood functions under H_0 and H_1. The SPP based estimator of the noise CPSD matrix is given as [57]

$$\boldsymbol{C}_v(k,l) = \hat{\alpha}(k,l)\boldsymbol{C}_v(k,l-1) + (1-\hat{\alpha}(k,l))\boldsymbol{x}(k,l)\boldsymbol{x}^H(k,l), \tag{26}$$

where $\hat{\alpha}(k,l)$ is a smoothing factor obtained through the estimated SPP. For example, $\hat{\alpha}(k,l)$ may be estimated as [57]

$$\hat{\alpha}(k,l) = \beta + (1-\beta)P(H_1|\boldsymbol{x}(k,l)). \tag{27}$$

where $0 < \beta < 1$ is a constant and set close to 1 if the noise is a slowly time-varying process [53, 57]. The concept of using SPPs to update the noise statistics originates from the single-channel speech processing world, and is often derived under complex Gaussian signal assumptions [53]. It is however, possible to generalize the concept of SPP to the multichannel case [56, 57].

However, some SPP methods may depend on a noise tracking algorithm in the initialization to obtain an initial estimate of the noise statistics [57]. One common approach to track the noise is using minimum statistics methods [59]. These methods do not rely on a VAD to track the background noise but instead use the assumption that the noise can be estimated by tracking the minimum level of the noisy speech [59]. This approach works well in many situations where the background noise is slowly time-varying, e.g., car noise. However, the approach can suffer from underestimation of the noise level leading to less noise reduction and potentially also more musical noise unless a bias compensation algorithm is used [59].

The binary VAD, SPP VAD, and noise tracking algorithm discussed above are, however, not able to update the noise statistics when speech is surely present. Until now, the methods we have outlined are based on an estimate of the noise statistics when speech is absent, i.e., (26), and use the estimated noise statistics for noise reduction when speech becomes present. However, methods exist that are able to update the noise statistics even during speech presence. Specifically, these methods often rely on an assumption about the structure of the noise CPSD matrix or on the target RTF vector assumed being known [60–65]. For example, one may assume that the noise CPSD matrix has the form as presented in (11) according to

$$\boldsymbol{C}_v(k,l) = \lambda_v(k,l)\boldsymbol{\Gamma}_v(k,l_0), \tag{28}$$

where l_0 denotes the time index of the most recent noise-only time-frequency tile. The model in (28) implies that the spatial distribution of the noise, at a time instant l_0 prior to speech activity remains constant during speech activity, but the noise PSD $\lambda_v(k,l)$ is time-varying [60]. If the target RTF vector is known, a blocking matrix can formed and used to cancel the target speech and estimate the noise PSD during speech presence. These approaches do obviously not influence the performance of the MVDR beamformer as the structure of the noise CPSD matrix captured by $\mathbf{\Gamma}_v(k, l_0)$ remains unchanged and the MVDR beamformer is immune to scalings of $\mathbf{\Gamma}_v(k, l_0)$ [61]. However, since the noise PSD $\lambda_v(k, l)$ is estimated during speech presence, this can lead to improved noise reduction performance in the post-filter, particularly when the noise is highly non-stationary [60, 61].

4.2 RTF vector estimation

In addition to the noise CPSD matrix $\boldsymbol{C}_v(k, l)$, the RTF vector $\boldsymbol{d}(k, l)$ is another parameter that generally needs to be estimated. It is often assumed for hearing aid applications that the target speaker is located in front of the hearing aid user [66], and hence the RTF vector associated with frontal direction (i.e. 0°) of the user is often used. However, this assumption can potentially lead to severe performance degradation and speech distortion, for example if the target is located to the sides or the rear of the hearing aid user. In worst cases, the beamformer will suppress the target. Furthermore, a RTF vector mismatch can have severe consequences for methods that rely on the target RTF vector for noise PSD estimators. Therefore, RTF vector estimation is crucial to ensure both good beamforming and post-filter performance.

RTF vector estimation can either be done directly or through direction-of-arrival (DOA) estimation. Direct RTF estimators are usually based on computing the target RTF vector from observations of the noisy speech signal and estimates of the noise CPSD matrix. Amongst the common methods are the covariance subtraction method [67], the eigenvalue decomposition method [67], and the covariance whitening method [68]. All three methods typically require an estimate of the noise CPSD matrix in order to estimate the target RTF vector. The other main class of approaches is to use a DOA estimation algorithm. RTF estimation based on DOA estimation, i.e., first estimates the angle to the target speakers with respect to the front direction, and then maps the estimated location to an RTF vector. For example, if the target speaker is localized at 180°, i.e. directly towards the rear, the RTF vector associated with the rear direction is used. For simple array geometries and under free-field assumptions it is possible, to derive a mathematical expression of the mapping from a DOA to an RTF vector. However for hearing aid applications,where the microphones are placed on a human head, an accurate representation of the mathematical function is rarely available for hearing aid applications, hence the mapping from a DOA to a RTF vector is performed using a pre-defined RTF vector dictionary such as in [69].

5 Scientific Contributions

In this section, we outline some of the research challenges that are present in the area of multichannel noise reduction for hearing aid systems. The research challenges are used as motivation for the research topics, which will be the theme of the remaining part of this thesis. We then outline the scientific contributions that are presented in part II in the thesis.

5.1 Research Challenges

Although significant advances have been made in the area of speech enhancement and noise reduction, people with a hearing loss still find it difficult to communicate in noisy environments with hearing aids [70]. Some of the most challenging acoustic environments which a multichannel noise reduction system can face are environments with competing speakers, i.e. the so-called cocktail party problem [12]. There are several aspects that make environments with competing speaker particularly difficult. First, spectral features of the target speech can often not be used to distinguish the target sound from the competing speakers since they tend to be similar. Secondly, speech is highly non-stationary, meaning that "noise statistics" representing a competing speaker would require fast adaptation of the noise statistics, that may render many of the conventional VAD and SPP estimators less effective. This is because the statistics of a competing speaker can change rapidly such that the estimated statistics quickly become outdated. This, in turn, can lead to over- and underestimation of the noise statistics when target speech is present which results in under- and over-suppression, respectively of the noise.

Similarly, estimation of the RTF vector becomes significantly more difficult in environments with competing speakers. Essentially, RTF vector estimation faces the same issues as noise estimation - it is difficult to discriminate between the target RTF vectors and competing speaker RTF vectors. One approach to improve target RTF estimation is to provide prior information about the target RTF vector [71]. For example, one may assume that the target speaker is in a known position (e.g. frontal) or a known spatial region (e.g. the frontal half-plane). In practice, the difficulty in distinguishing competing speakers from the target speaker may cause existing noise reduction systems to treat competing speakers as if they were desired sound sources, meaning that they would not be suppressed to avoid the risk of accidentally suppressing a desired speech signal. Therefore, multichannel noise reduction in current hearing aids is still mostly efficient in handling simple acoustic environments where the noise field and noise statistics change at a slower rate than speech signals do.

5.2 Research topics

In this thesis, we envision noise reduction systems for hearing aid systems which try to retrieve information from the hearing aid user in order to work better. For example, the noise CPSD matrix and RTF vector estimation could potentially be improved if the user provided information about, for example, the target source location or information about the background noise. An *active involvement* of the user is usually less desirable, e.g., if the user needs to adjust the noise reduction settings of the hearing aid through a smartphone. Such approaches may be simple from an algorithm point of view, but the hearing aid user has to spend time and energy, and might find it inconvenient to make such manual adjustments. Instead, we hypothesize that cues about the target can be extracted from the user's natural behavior during conversations. These include additional modalities besides sound that may indicate the position of the target speaker relative to the user head-direction and can provide crucial information to the noise reduction system about the location of the target in a multi-talker situation [22]. Furthermore, the users' own voice may provide information about the absence or presence of target speech by exploiting the so-called turn-taking mechanism which takes place in most natural conversations. Specifically, the target is likely to be absent when the own voice is present and vice versa. Overall, the research presented in this thesis revolves around the general topic of noise reduction systems for hearing aids which exploit side information which may be assumed available a priori or which may be gathered from the hearing aid user. Specifically, we focus on noise reduction systems, which:

1. Use prior information of the target location to improve noise reduction performance. The proposed algorithms are general in the sense that they do not rely on any assumptions about the origin of the this prior knowledge. This could, e.g., be achieved from explicit manual user feedback or by additional sensors, e.g., accelerometers and cameras.

2. Exploit the user's own voice activity to improve noise estimation and competing speaker identification.

5.3 Contributions

The main body and research contributions follows in part II of this thesis. Part II consists of a collection of five accepted peer-reviewed papers. Paper [A-C] are conference papers whereas paper [D-E] are journal papers. The papers that form part II of this thesis are as follows:

[A] **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," 2019 *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*

[B] **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," ICASSP 2020 - *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*

[C] **P. Hoang**, Z. -H. Tan, J. M. de Haan and J. Jensen, "Joint Maximum Likelihood Estimation of Power Spectral Densities and Relative Acoustic Transfer Functions for Acoustic Beamforming," ICASSP 2021 - *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*

[D] **P. Hoang**, Z.-H. Tan, J. M. de Haan, and J. Jensen, “Multichannel Speech Enhancement with Own Voice-Based Interfering Speech Suppression for Hearing Assistive Devices,” IEEE/ACM Trans. Audio Speech Lang. Process. 2022.

[E] **P. Hoang**, Z.-H. Tan, J. M. de Haan, and J. Jensen, “The Minimum Overlap-Gap Algorithm for Speech Enhancement,” IEEE Access, 2022.

A more in detail summary of the papers are given in the following part of this section.

[A] Robust Bayesian and Maximum a Posteriori Beamforming for Hearing Assistive Devices

In paper A, we propose a Bayesian and maximum a posterior (MAP) beamformer to improve the robustness against situations where the target DOA is unknown but a prior distribution of the target DOA is given, e.g., from other external sensors such as cameras and accelerometers. We use the prior distribution in a Bayesian framework, where we assume that the noisy signal is distributed according a multivariate complex circular symmetric Gaussian distribution. The proposed Bayesian and MAP beamformers are compared against other multichannel noise reduction methods used for hearing aids in terms of segmental SNR and extended short-time objective intelligibility (ESTOI).

[B] Maximum Likelihood Estimation of the Interference-Plus-Noise Cross Power Spectral Density Matrix for Own Voice Retrieval

In paper B, we propose a maximum likelihood estimator of the interference-plus-noise CPSD matrix to improve own voice pick-up for telephony application in headsets and hearing aids. The situation, we consider consists of an own voice source, an interferer (e.g. a competing speaker), and background noise, where the statistics of the competing speaker are unknown. We model the noise CPSD matrix to include an interferer such that the interference-plus-noise CPSD matrix is given as

$$\boldsymbol{C}_{qv}(k,l) = \lambda_q(k,l)\boldsymbol{d}_q(k,l)\boldsymbol{d}_q^H(k,l) + \boldsymbol{C}_v(k,l). \tag{29}$$

We assume that the RTF vector of the own voice source is known and time-invariant, and an oracle own voice VAD is given. These assumptions can be motivated by the fact that 1) the location of the microphone array relative to the own voice source location remains approximately fixed in a hearing aid or headset application, and 2) the distance between the own voice source and the microphone array can be considered closer than the between the own voice source and an interferer. The proposed method is able to update both the interferer statistics and noise statistics while, own voice is present in contrast to conventional noise estimation methods. This is highly important since interferer signal can be highly non-stationary, particularly if the interferer is a competing speaker. Our simulation results also indicate the proposed estimation of the interference-plus-noise CPSD matrix performs either on par or significantly better than conventional methods, especially in those situations where the interferer is highly non-stationary and dominant.

[C] Joint Maximum Likelihood Estimation of Power Spectral Densities and Relative Acoustic Transfer Functions for Acoustic Beamforming

In paper C, we derive the joint maximum likelihood estimators of the target and noise PSDs, and the target RTF vector, given an estimate of the normalized CPSD matrix. Among the findings, we show that the well-known covariance whitening method is in fact a maximum likelihood estimator of the target RTF vector. We also show that the target and noise PSDs can be found from the eigenvalues of the whitened noisy speech CPSD matrix. We use the jointly estimated parameters to implement MWF beamformers and implement a SPP based on the estimated parameters. We evaluate the proposed multichannel noise reduction system in simulations against similar methods where the noise reduction algorithm is particularly good at handling situations where a pre-defined dictionary of target RTF vectors are not given in advance.

[D] Multichannel Speech Enhancement with Own Voice-Based Interfering Speech Suppression for Hearing Assistive Devices

Paper D can be seen as an extension of paper B, i.e. using the own voice activity to provide cues on what is desired and undesired. In this situation, we consider an own voice source, a target speaker, a competing speaker, and noise. Only the own voice RTF vector and the normalized CPSD matrix of the background noise is known. However, the RTF vectors of the target speaker and the competing speaker need to be estimated. The location of the target speaker and competing speaker can, in principle, be completely arbitrarily positioned. Hence, this situation will cause a conventional noise reduction system to perform poorly since there is typically no way to decide which of them is target. In this paper, we propose a noise reduction system that exploits the own voice activity to discriminate between the target speaker and the competing speaker. We assume that the user and the target speaker are engaged in a conversation, and hence take turns. Specifically, during own voice presence we assume that the target speech is absent. Hence, we estimate the RTF vector and PSD of the competing speaker. When the user's own voice is absent, we assume that the target speech may become present. To estimate the competing speaker and noise statistics, we propose maximum likelihood estimators of these. We compare the proposed methods with conventional beamforming systems in simulation, where the results show that the proposed methods perform significantly better than the conventional beamforming system in situations with a dominant competing speaker.

[E] The Minimum Overlap-Gap Algorithm for Speech Enhancement

In paper E, we address the problem of enhancing a target speaker in an environment of multiple competing speakers. To this day, there exists a limited number of methods that are able to enhance the talker-of-interest when multiple speakers are present in the noisy speech. In particular, exists deep learning methods for separating the speakers into candidate speakers, but there is usually no way of telling which one of the speakers is the target. In fact, one could argue that without additional prior knowledge, it is impossible to decide which of the speakers is the target speaker. We propose, a speech enhancement paradigm which addresses this problem and is able to effectively solve this problem. The speech enhancement paradigm consists of three elements which are speech separation, speaker ranking, and enhancement. The speech separation system separates the mixture of multiple talkers from the noisy speech signal into candidate speakers, where each candidate speaker represents a talker present in the environment. The speaker ranking is performed by the proposed algorithm which ranks the candidate talkers according to how likely they are of being the talker of interest. Finally, the enhancement is a linear combination of the candidate talkers whose coefficients are determined by the ranking algorithm. In this paper, we also propose methods for ranking the candidate talker based on the turn-taking behavior between a user and a conversational partner, i.e., the target speaker. In contrast to most existing schemes, the proposed speaker ranking algorithms only relies on microphone signals, which is highly desirable since current hearing aids do not have access to more sophisticated sensors, e.g. EEG and cameras (see Sec. 5.2). We demonstrate the proposed speech enhancement paradigm and the speaker ranking algorithm in hearing aid applications in simulation experiments. We find that they are highly effective and provide a significant improvement over conventional noise reduction systems in multi-talker environments.

6 Future Research and Direction

The research presented in this thesis has covered the potential use of the user's own voice, conversational turn-taking, and prior knowledge about the target source to improve noise estimation and target RTF vector estimation. However, there are several aspect that we did not cover in this thesis that could be of potential interest for future research.

6.1 Beyond microphones

In this thesis, the proposed methods for estimating noise and target source, were primarily based on the noisy speech signals picked up by microphones. However, in principle, other sensor besides microphones can be used to improve estimation of noise and target statistics. Since, hearing aids are in constant development, future hearing aids might have access to more sophisticated sensors such as in-ear electrodes and cameras [72–75]. Accessibility to other sensor signals will allow sensor fusion and can potentially improve noise reduction performance considerably. For example, access to the user's eye-gaze from cameras faced towards the user's eyes [76] can be used as a cue to detect where the target speaker is located in very noisy environments.

6.2 Behavior-controlled noise reduction

Obviously, the use of the user's own voice to discriminate between the target speaker and the competing speakers only works if the user is actively participating in a conversation with the target speaker. However, there are many situations where the user most of the time takes the role of a listener. For example, during conversations which involve multiple people, the hearing aid user might not actively participate in the conversation. Also, some users may simply be less talkative, and tend to engage in conversations less often. Therefore, other behavioral characteristics of the user could be involved in order to obtain a more robust discrimination between the target speaker and competing speakers. Examples include the use of eye-gaze behavior [72] and features in EEG signals obtained from electrodes, e.g., placed at/in the ear of the user [72, 73].

One may also analyze the behavior of the candidate target speakers and develop a discriminator function to distinguish between the target speaker and competing speakers. For example, one could envision the use of cameras, mounted on the hearing aid user to form an audio-visual enhancement system [74, 77], which could analyze which person(s) are gazing and talking to the user.

6.3 Determining multiple talkers-of-interest exploiting turn-taking

In paper D and E, we assume that there is only one target speaker. However, in many situations a group of people engaged in the same conversation can all be considered as potential targets. Neither paper D and E can handle situations with more than a single target efficiently and therefore it would be of interest to generalize the methodology to situations with multiple targets. For example, one may examine the turn-taking behavior in conversations between not only two but multiple talkers.

6.4 Exploring the speech enhancement paradigm proposed in paper E

The proposed speech enhancement paradigm in paper E consists of the three processing stages: speech separation, speaker ranking, and speech enhancement. It can be referred to as the separation, ranking, and enhancement (SRE) paradigm. The SRE paradigm can efficiently solve the multi-talker situations with multiple competing speakers and a single target speaker, which is often considered as an otherwise extremely difficult situation to solve. However, we believe that the proposed instance of the paradigm might only be the first of several to come. For example, we believe that performance can be improved beyond what is reported in paper E by the use of more advanced deep learning algorithms. Specifically, complex conversational patterns between the user and the talkers of interest can hypothetically be learned by a deep learning algorithm to make better predictions of who the conversational partner is.

References

[1] C. Elberling, K. Worsoe, B. Diemer, and Oticon Foundation, *Fading sounds: about hearing and hearing aids*. Herlev, Denmark: Bording A/S, 2006, oCLC: 180767009.

[2] J. R. Hurford, "The evolution of the critical period for language acquisition," *Cognition*, vol. 40, no. 3, pp. 159–201, Sep. 1991.

[3] A. R. Luria and F. I. Yudovich, *Speech and the development of mental processes in the child: an experimental investigation*, ser. Penguin papers in education. Harmondsworth: Penguin, 1971.

[4] H. Dillon, *Hearing aids*, 2nd ed. Sydney: Boomerang Press [u.a.], 2012, oCLC: 803977183.

[5] G. R. Popelka, B. C. J. Moore, R. R. Fay, and A. N. Popper, *Hearing Aids*. Cham: Springer International Publishing : Imprint: Springer, 2016, oCLC: 967694134.

[6] Q. Huang and J. Tang, "Age-related hearing loss or presbycusis," *European Archives of Oto-Rhino-Laryngology*, vol. 267, no. 8, pp. 1179–1191, Aug. 2010.

[7] C. D. Mulrow, C. Aguilar, J. E. Endicott, R. Velez, M. R. Tuley, W. S. Charlip, and J. A. Hill, "Association Between Hearing Impairment and the Quality of Life of Elderly Individuals," *Journal of the American Geriatrics Society*, vol. 38, no. 1, pp. 45–50, Jan. 1990.

[8] C. Carabellese, I. Appollonio, R. Rozzini, A. Bianchetti, G. B. Frisoni, L. Frattola, and M. Trabucchi, "Sensory Impairment and Quality of Life in a Community Elderly Population," *Journal of the American Geriatrics Society*, vol. 41, no. 4, pp. 401–407, Apr. 1993.

[9] J. E. Crews and V. A. Campbell, "Vision Impairment and Hearing Loss Among Community-Dwelling Older Americans: Implications for Health and Functioning," *American Journal of Public Health*, vol. 94, no. 5, pp. 823–829, May 2004.

[10] R. F. Uhlmann, "Relationship of Hearing Impairment to Dementia and Cognitive Dysfunction in Older Adults," *JAMA: The Journal of the American Medical Association*, vol. 261, no. 13, p. 1916, Apr. 1989.

[11] C.-M. Li, X. Zhang, H. J. Hoffman, M. F. Cotch, C. L. Themann, and M. R. Wilson, "Hearing Impairment Associated With Depression in US Adults, National Health and Nutrition Examination Survey 2005-2010," *JAMA Otolaryngology–Head & Neck Surgery*, vol. 140, no. 4, p. 293, Apr. 2014.

[12] A. Bronkhorst, "The cocktail-party problem revisited: early processing and selection of multi-talker speech." *Atten Percept Psychophys*, vol. 77 (5), pp. 1465–1487, 2015.

[13] C. J. Plack, *The sense of hearing*, third edition ed. London ; New York: Routledge, Taylor & Francis Group, 2018.

[14] J. O. Pickles, *An introduction to the physiology of hearing*, 4th ed. London: Emerald, 2012.

[15] W. H. Organization. (2021) Deafness and hearing loss. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[16] H. Fletcher, *Speech and Hearing in Communication*, ser. Bell Telephone Laboratories series. Van Nostrand, 1953. [Online]. Available: https://books.google.dk/books?id=f4xZAAAAMAAJ

[17] K. L. Payton, R. M. Uchanski, and L. D. Braida, "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," *The Journal of the Acoustical Society of America*, vol. 95, no. 3, pp. 1581–1592, Mar. 1994.

[18] P. A. Gosselin and J.-P. Gagné, "Older adults expend more listening effort than young adults recognizing audiovisual speech in noise," *International Journal of Audiology*, vol. 50, no. 11, pp. 786–792, Nov. 2011.

[19] C. B. Hicks and A. M. Tharpe, "Listening Effort and Fatigue in School-Age Children With and Without Hearing Loss," *Journal of Speech, Language, and Hearing Research*, vol. 45, no. 3, pp. 573–584, Jun. 2002.

[20] M. Cooke, M. L. Garcia Lecumberri, and J. Barker, "The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception," *The Journal of the Acoustical Society of America*, vol. 123, no. 1, pp. 414–427, Jan. 2008.

[21] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Release from informational masking by time reversal of native and non-native interfering speech," *The Journal of the Acoustical Society of America*, vol. 118, no. 3, pp. 1274–1277, Sep. 2005.

[22] L. V. Hadley, W. O. Brimijoin, and W. M. Whitmer, "Speech, movement, and gaze behaviours during dyadic conversation in noise," *Scientific Reports*, vol. 9, no. 1, p. 10451, Dec. 2019.

[23] H. Sacks, E. A. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation," *Language*, vol. 50, no. 4, p. 696, Dec. 1974.

[24] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, Oct. 2010.

[25] A. Schaub, *Digital hearing aids*. New York: Thieme, 2008, oCLC: 244290628.

[26] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer handbook of speech processing*. Berlin ; London: Springer, 2008, oCLC: ocn190966783.

[27] 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*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[28] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, 2001, oCLC: 990622835.

[29] M. Farmani, "Informed Sound Source Localization for Hearing Aid Applications," *PhD Series*, vol. Technical Faculty of IT and Design, p. Aalborg University, 2017, medium: PDF Publisher: Aalborg University Press.

[30] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[31] 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*, ser. Synthesis lectures on speech and audio processing. Williston, VT: Morgan & Claypool, 2013, no. 11, oCLC: 935190351.

[32] A. Kuklasinski, S. Doclo, T. Gerkmann, S. Holdt Jensen, and J. Jensen, "Multi-channel PSD estimators for speech dereverberation - A theoretical and experimental comparison," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 91–95.

[33] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[34] 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, Oct. 1987.

[35] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, Jul. 1996.

[36] H. L. Van Trees, *Optimum array processing*, ser. Detection, estimation, and modulation theory / Harry L. Van Trees. New York: Wiley, 2002, no. 4, oCLC: 248233642.

[37] K. Buckley and L. Griffiths, "An adaptive generalized sidelobe canceller with derivative constraints," *IEEE Transactions on antennas and propagation*, vol. 34, no. 3, pp. 311–319, 1986.

[38] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Theoretical noise reduction limits of the generalized sidelobe canceller (gsc) for speech enhancement," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, vol. 5. IEEE, 1999, pp. 2965–2968.

[39] I. Cohen, "Analysis of two-channel generalized sidelobe canceller (gsc) with post-filtering," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 684–699, 2003.

[40] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[41] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*, 2018, oCLC: 1059285919.

[42] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[43] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *IEEE International Conference on Acoustics Speech and Signal Processing*. Orlando, FL, USA: IEEE, May 2002, pp. IV–4164–IV–4164.

[44] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in *ICASSP ’79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. Washington, DC, USA: Institute of Electrical and Electronics Engineers, 1979, pp. 208–211.

[45] J. Lim and A. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[46] R. McAulay and M. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[47] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[48] ——, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[49] R. C. Hendriks, R. Heusdens, and J. Jensen, “MMSE based noise PSD tracking with low complexity,” in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. Dallas, TX, USA: IEEE, 2010, pp. 4266–4269.

[50] P. C. Loizou, *Speech enhancement: theory and practice*. Boca Raton, Fla.: CRC Press, 2013, oCLC: 958799095.

[51] R. C. Hendriks and R. Heusdens, “On linear versus non-linear magnitude-DFT estimators and the influence of super-Gaussian speech priors,” in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. Dallas, TX, USA: IEEE, 2010, pp. 4750–4753. [Online]. Available: http://ieeexplore.ieee.org/document/5495172/

[52] R. Martin, “Speech enhancement based on minimum mean-square error estimation and supergaussian priors,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[53] I. Cohen and B. Berdugo, “Noise estimation by minima controlled recursive averaging for robust speech enhancement,” *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[54] I. Cohen, “Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[55] ——, “Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator,” *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113–116, Apr. 2002.

[56] M. Souden, Jingdong Chen, J. Benesty, and S. Affes, “Gaussian Model-Based Multichannel Speech Presence Probability,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1072–1077, Jul. 2010.

[57] M. Souden, J. Chen, J. Benesty, and S. Affes, “An Integrated Solution for Online Multichannel Noise Tracking and Reduction,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2159–2169, Sep. 2011.

[58] M. Souden, M. Delcroix, K. Kinoshita, T. Yoshioka, and T. Nakatani, "Noise Power Spectral Density Tracking: A Maximum Likelihood Perspective," *IEEE Signal Processing Letters*, vol. 19, no. 8, pp. 495–498, Aug. 2012.

[59] 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, Jul. 2001.

[60] J. Jensen and U. Kjems, "Maximum Likelihood Based Noise Covariance Matrix Estimation For Multi-Microphone Speech Enhancement," *EUSIPCO*, Aug. 2012.

[61] J. Jensen and M. S. Pedersen, "Analysis of beamformer directed single-channel noise reduction system for hearing aid applications," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 5728–5732.

[62] S. Braun, A. Kuklasinski, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and Comparison of Late Reverberation Power Spectral Density Estimators," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1056–1071, Jun. 2018.

[63] Hao 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, Apr. 1995.

[64] A. Kuklasinski, 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, Sep. 2016.

[65] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using dft domain subspace decompositions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 541–553, 2008.

[66] A. Kuklasinski, "Multichannel Wiener Filter for Speech Dereverberation in Hearing Aids -Sensitivity to DoA Errors," in *AES Convention*, 2016.

[67] M. Taseska and E. A. P. Habets, "Relative transfer function estimation exploiting instantaneous signals and the signal subspace," in *2015 23rd European Signal Processing Conference (EUSIPCO)*. Nice: IEEE, Aug. 2015, pp. 404–408.

[68] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 544–548.

[69] 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, Mar. 2018.

[70] S. Kochkin, "MarkeTrak V: "Why my hearing aids are in the drawer"," *The Hearing Journal*, vol. 53, no. 2, p. 34, Feb. 2000.

[71] 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, Feb. 2000.

[72] 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, p. 233121651881438, Jan. 2018.

[73] A. Favre-Felix, 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)*. Seogwipo: IEEE, Jul. 2017, pp. 4086–4089.

[74] 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," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2021.

[75] D. Michelsanti, Z.-H. Tan, S. Sigurdsson, and J. Jensen, "Deep-learning-based audio-visual speech enhancement in presence of Lombard effect," *Speech Communication*, vol. 115, pp. 38–50, Dec. 2019.

[76] S. Thomsen, "Eye-Gaze Steered Beamforming for Hearing Aids," Master's thesis, Aalborg University, Mar. 2022.

[77] Aalborg University and D. Michelsanti, "Audio-Visual Speech Enhancement Based on Deep Learning," Ph.d, Aalborg University, 2021.

Part II

Papers

This page intentionally left blank.

Paper A

Robust Bayesian and Maximum a Posteriori Beamforming for Hearing Assistive Devices

Poul Hoang, Zheng-Hua Tan, Jan Mark de Haan, Thomas Lunner, and Jesper Jensen

The paper has been published in
IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 1-5, 2019.

Abstract

Multi-microphone speech enhancement systems often apply beamforming to enhance one or multiple desired signals in a noisy environment. Common for many beamforming methods, is that they require the direction-of-arrival (DOA) of the target sound source to be known in order to achieve optimal noise reduction performance. To improve robustness against DOA uncertainty, we propose maximum a posteriori (MAP) and Bayesian beamformers that are able to take advantage of prior information on the target direction. We compare the proposed MAP and Bayesian beamformers to state-of-the-art beamforming methods for noise reduction in hearing assistive devices. We evaluate the proposed beamformers in isotropic babble noise in terms of segmental SNR (SSNR) and extended short-time objective intelligibility (ESTOI). Results show that the proposed methods outperform current state-of-the-art beamformers used for noise reduction in hearing aids in most scenarios.

1 Introduction

Noise reduction systems in e.g. hearing assistive devices and hands-free communication devices use adaptive beamforming [1–3] to enhance one or multiple desired target signals from a noisy environment. Noise reduction performance of beamformers typically used in hearing aid (HA) systems, such as the minimum variance distortionless response (MVDR) beamformer and the multichannel Wiener filter (MWF), depends greatly on the robustness of the direction-of-arrival (DOA) estimation since DOA mismatches can potentially degrade the speech intelligibility and sound quality [4]. Robust DOA estimation has therefore historically been a significant and important research topic for multichannel speech enhancement systems [3] and the research area remains highly active, see e.g. [5, 6] for recent contributions.

In [6] a stochastic maximum likelihood DOA estimator is proposed, which is used in the context of MVDR beamforming (we will refer to this beamformer as the ML beamformer). The ML beamformer relies on a novel ML estimator of the noise cross-power spectral density (CPSD) matrix that assumes that the underlying structure of the inter-microphone noise CPSD matrix in speech absence remains fixed during speech presence [7–9] but can be updated by a time-varying scalar multiplier. The method in [6], however, do not incorporate potential any prior information available on the target DOA.

Existing beamforming methods for HAs often make the rather strict assumption that the target is located in a fixed (known) position, e.g. frontal to the user [4, 7, 10]. Alternatively, they make no prior assumptions about the target position whatsoever [6]. In practice, some potentially vague prior information about the target location may be available. For example, for HAs it may be known a priori that the target is located in the frontal plane with respect to the user. Bayesian beamforming [11–13], however, offers an elegant framework for incorporating such prior information to form optimal beamformers under DOA uncertainty,

In this paper, we propose a maximum a posteriori (MAP) and a Bayesian beamformer based on the noise CPSD matrix model presented in [14] and derive the likelihood function of the noisy observations. In contrast to [6], we utilize prior information of the target DOA and derive the a posteriori distribution, which is used for the proposed MAP and Bayesian beamformers. We moreover demonstrate the advantage of incorporating prior information and compare the proposed beamformers to competing methods.

2 Signal Model and Assumptions

For a microphone array with M microphones, where each microphone picks up the sound from the noisy acoustic environment, the noisy observation $x_m(t)$, $t \in \mathbb{N}_0$, at the m'th microphone can be modeled as

$$x_m(t) = s(t) * d_m(t, \theta_s) + v_m(t), \quad m = 1, ..., M, \tag{A.1}$$

where we have assumed a single target, $*$ is the convolution operator, $d_m(t, \theta_s)$ denotes the relative impulse response from the target to the microphone with respect to a pre-selected reference microphone, $s(t)$ is the target signal at the reference microphone and impinges on the array from direction $\theta_s \in]-180^\circ, 180^\circ]$, and the noise, $v_m(t)$, is a sum of all undesired signals e.g. competing speakers, reverberation, and microphone self-noise. The short-time Fourier transform (STFT) is used to transform $x_m(t)$ into the time-frequency (TF) domain. Let k and n be the frequency bin index and the frame index respectively. Then the TF domain representation of the noisy signal, $x_m(k,n) \in \mathbb{C}$, is

$$x_m(k, n) = s(k, n) d_m(k, n, \theta_s) + v_m(k, n). \tag{A.2}$$

Collect the noisy observations, $x_m(k, n)$, in a vector $\mathbf{x}(k, n) = [x_1(k, n), ..., x_M(k, n)]^T$ and define the cross power spectral density (CPSD) matrix of the noisy signal as $\mathbf{C}_x(k, n) = \mathbb{E}\{\mathbf{x}(k, n)\mathbf{x}^H(k, n)\}$, where $\mathbb{E}\{\cdot\}$ is the expectation operator. Under the standard assumption that $s(k, n)$ is uncorrelated with the noise, $\mathbf{C}_x(k, n)$ is given as

$$\mathbf{C}_x(k, n) = \lambda_s(k, n) \mathbf{d}(k, n, \theta_s) \mathbf{d}^H(k, n, \theta_s) + \mathbf{C}_v(k, n), \tag{A.3}$$

where $\mathbf{C}_v(k, n) = \mathbb{E}\{\mathbf{v}(k, n)\mathbf{v}^H(k, n)\}$ is the noise CPSD matrix with $\mathbf{v}(k, n) = [v_1(k, n), ..., v_M(k, n)]^T$, $\lambda_s(k, n)$ is the target power spectral density (PSD) at the reference microphone, and $\mathbf{d}(k, n, \theta_s) = [d_1(k, n, \theta_s), ..., d_M(k, n, \theta_s)]^T$ is the relative transfer function (RTF) vector. In many acoustic scenarios, such as car cabins or cocktail-party scenarios, the noise CPSD matrix model presented in [7, 14] can be used to estimate the noise CPSD matrix in speech presence. The model assumes that the underlying structure of the noise CPSD matrix observed during speech absence, $\mathbf{C}_v(k, n_0)$, remains identical during speech presence. In other words, the signal model in (A.3) may be written as

$$\mathbf{C}_x(k,n) = \lambda_s(k,n)\mathbf{d}(k,n,\theta_s)\mathbf{d}^H(k,n,\theta_s) + \lambda_v(k,n)\mathbf{C}_v(k,n_0), \tag{A.4}$$

where $\lambda_v(k, n)$ is a scalar and $\mathbf{C}_v(k, n_0)$ is the noise CPSD matrix from the most recent speech absent frame n_0.

The performance of a beamformer depends on the estimated DOA, as DOA mismatches can degrade the performance significantly [4]. In most acoustic scenarios, the target DOA θ_s is not known exactly. Instead, it may be considered as a random variable with probability distribution $P(\theta)$. For example, in a cocktail-party-like environment, the target might be more likely to arrive from the front rather than from the rear of the HA user. In the proposed method we assume that $P(\theta)$ is a known probability mass function (PMF) and that $\theta_s \sim P(\theta)$ in order to evaluate the advantage of integrating prior target location knowledge into the beamformer.

3 Wideband Bayesian and MAP Beamforming

A beamformer is a linear combination of the noisy signal $\mathbf{x}(k, n)$ with the beamformer weights $\mathbf{w}(k, n) \in \mathbb{C}^M$ such that the output of the beamformer is

$$\hat{s}(k, n) = \mathbf{w}^H(k, n)\mathbf{x}(k, n). \tag{A.5}$$

Typically, the beamformer weights are derived through an optimization problem, e.g. by minimizing the noise output power with a distortionless constraint (MVDR beamformer) or minimizing the mean-square-error (MSE) between the target signal and the estimated one (MWF). While these beamformers assume perfect knowledge of the target DOA, a *Bayesian* beamformer refers to an optimal beamformer that minimizes the MSE between the target signal $s(k, n)$ and the estimated target signal $\hat{s}(k, n)$ under DOA uncertainty [11].

3.1 Wideband Bayesian beamformer

Let $\mathbf{X}(k, n) = [\mathbf{x}(k, n - N + 1), ..., \mathbf{x}(k, n)] \in \mathbb{C}^{M \times N}$ and $\underline{\mathbf{X}}(n) = [\mathbf{X}(1, n), ..., \mathbf{X}(K, n)] \in \mathbb{C}^{M \times NK}$. To derive the Bayesian beamformer, consider the minimum mean-square-error (MMSE) estimator of the target signal at a particular TF tile. The MMSE estimator is the conditional expectation of the target signal $\hat{s}(k, n) = \mathbb{E}\{s(k, n)|\underline{\mathbf{X}}(n)\}$ [11] i.e.

$$\hat{s}(k,n) = \int_{\mathbb{C}} s(k,n) f\left(s(k,n)|\underline{\mathbf{X}}(n)\right) \, \mathrm{d}s(k,n), \tag{A.6}$$

where $f\left(s(k,n)|\underline{\mathbf{X}}(n)\right)$ is the posterior of the $s(k,n)$, given the noisy observations $\underline{\mathbf{X}}(n)$. As θ_s is modeled as a random variable with PMF $P(\theta)$, the law of total probability is used to further expand $f\left(s(k,n)|\underline{\mathbf{X}}(n)\right)$:

$$\hat{s}(k,n) = \int_{\mathbb{C}} s(k,n) \sum_{i=1}^{I} P(\theta_i|\underline{\mathbf{X}}) f\left(s(k,n)|\underline{\mathbf{X}}(n), \theta_i\right) \, \mathrm{d}s(k,n),$$

where $P(\theta_i|\underline{\mathbf{X}}(n))$ is the a posteriori probability of the target direction and $\theta_i \in \Theta$, Θ is a discrete candidate set of directions from which the target can arrive, and I is the cardinality of Θ. Assuming further that $s(k,n)$ is only dependent on the noisy observation at frequency bin k [15], it follows that

$$\hat{s}(k,n) = \sum_{i=1}^{I} P(\theta_i|\underline{\mathbf{X}}(n)) \mathbb{E}\{s(k,n)|\mathbf{X}(k,n), \theta_i\},$$

where $\mathbb{E}\{s(k,n)|\mathbf{X}(k,n),\theta_i\} = \int_{\mathbb{C}} s(k,n) f(s(k,n)|\mathbf{X}(k,n),\theta_i) \mathrm{d}s(k,n)$. Assuming that $s(k,n)$ has a uniform prior [16], it follows that the Bayesian MMSE estimator of $s(k,n)$ is a linear combination of MVDR beamformers:

$$\mathbf{w}_{\mathrm{B}}(k,n) = \sum_{i=1}^{I} P(\theta_i|\underline{\mathbf{X}}(n)) \mathbf{w}_{\mathrm{MVDR}}(k,n,\theta_i), \tag{A.7}$$

where

$$\mathbf{w}_{\mathrm{MVDR}}(k,n,\theta_i) = \frac{\mathbf{C}_v^{-1}(k,n_0)\mathbf{d}(k,n,\theta_i)}{\mathbf{d}^H(k,n,\theta_i)\mathbf{C}_v^{-1}(k,n_0)\mathbf{d}(k,n,\theta_i)}. \tag{A.8}$$

3.2 Posterior probability of the target direction

In order to derive an expression for the posterior probability $P(\theta_i|\underline{\mathbf{X}}(n))$, some key assumptions will be made. First, it is assumed that the noisy observations $\mathbf{x}(k,n)$ are temporally uncorrelated e.g. [6, 11] and distributed according to a circular symmetric complex Gaussian distribution such that $\mathbf{x}(k,n) \sim \mathcal{N}_C\left(\mathbf{0}, \mathbf{C}_x(k,n,\theta_s)\right)$ [11]. Next, $\mathbf{x}(k,n)$ is assumed uncorrelated across frequency i.e. $\mathbb{E}\{\mathbf{x}(k,n)\mathbf{x}^H(j,n)\} = \mathbf{0}$ for $j \neq k$, and finally, it is assumed that $\mathbf{C}_x(k,n,\theta_s)$ may be modeled according to (A.4). With the first assumption, the likelihood function of the noisy observations $\mathbf{X}(k,n)$ is

$$f(\mathbf{X}(k,n)|\theta_i) = \frac{\exp\left(-N\mathrm{tr}\left(\hat{\mathbf{C}}_x(k,n)\mathbf{C}_x^{-1}(k,n,\theta_i)\right)\right)}{\pi^{MN}|\mathbf{C}_x(k,n,\theta_i)|^N},$$

where $\hat{\mathbf{C}}_x(k,n) = \frac{1}{N}\mathbf{X}(k,n)\mathbf{X}(k,n)^H$ is the sample estimate of the noisy CPSD matrix and $\mathrm{tr}(\cdot)$ is the trace operator. We substitute $\mathbf{C}_x(k,n,\theta_i)$ with the expression found in (A.4), and observe that $\lambda_s(k,n)$ and $\lambda_v(k,n)$ are unknown parameters. Closed-form expressions for ML estimators of $\lambda_s(k,n)$ and $\lambda_v(k,n)$ are derived in [8, 14]. Inserting these ML estimators $\hat{\lambda}_s(k,n,\theta_i)$ and $\hat{\lambda}_v(k,n,\theta_i)$ into the likelihood function, we obtain the concentrated likelihood function $\bar{f}(\mathbf{X}(k,n)|\theta_i) = f(\mathbf{X}(k,n)|\theta_i, \hat{\lambda}_s(k,n,\theta_i), \hat{\lambda}_v(k,n,\theta_i))$ [14]

$$\begin{aligned}\bar{f}(\mathbf{X}(k,n)|\theta_i) = \pi^{-MN}|\hat{\lambda}_s(k,n,\theta_i)\mathbf{d}(k,n,\theta_i)\mathbf{d}^H(k,n,\theta_i)\\ + \hat{\lambda}_v(k,n,\theta_i)\mathbf{C}_v(k,n_0)|^{-N}\exp\left(-MN\right),\end{aligned}$$

where $|\cdot|$ denotes the matrix determinant. We may derive a joint concentrated likelihood function across frequency bins as

$$\bar{f}(\underline{\mathbf{X}}(n)|\theta_i) = \prod_{k=1}^{K} \bar{f}(\mathbf{X}(k,n)|\theta_i). \tag{A.9}$$

We arrive at the posterior probability using Bayes theorem:

$$P(\theta_i|\underline{\mathbf{X}}(n)) = c \cdot P(\theta_i)\bar{f}(\underline{\mathbf{X}}(n)|\theta_i), \tag{A.10}$$

where $c = \sum_{i=1}^{I} P(\theta_i)\bar{f}(\underline{\mathbf{X}}(n)|\theta_i)$ is a normalization constant and $P(\theta_i)$ is the target DOA prior. Inserting (A.10) into (A.7) leads to the proposed wideband Bayesian beamformer.

3.3 MAP beamformer

The MAP beamformer is an MVDR beamformer steered towards the MAP estimate of the DOA. The DOA is first estimated by

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta_i} P(\theta_i)\bar{f}(\underline{\mathbf{X}}(n)|\theta_i), \tag{A.11}$$

and afterwards used in (A.8) to form an MVDR beamformer.

In order to compute an upper bound performance, an oracle beamformer was implemented. The oracle beamformer is an MVDR beamformer, which uses the true DOA, i.e. $\mathbf{w}_{\text{MVDR}}(k,n,\theta_s)$ cf. (A.8).

4 Experimental Setup

Simulation experiments are conducted where a target speaker is placed in an approximately isotropic noise field, where the sound is picked up by microphones placed on behind-the-ear (BTE) HAs worn by a HA user. We evaluate the performance with two different microphone arrays; one where a BTE HA is placed behind the left ear, and another with two BTE HAs with one placed

	P_t	1.00				0.75				0.50				0.35			
	SNR (dB)	-12	-6	0	6	-12	-6	0	6	-12	-6	0	6	-12	-6	0	6
SSNR	Unprocessed	-19.83	-14.05	-8.21	-2.16	-19.60	-13.66	-7.81	-1.38	-19.50	-13.61	-7.43	-1.41	-19.48	-12.95	-7.68	-1.28
	Fixed-MVDR	**-12.87**	-7.96	-3.94	-0.89	**-12.85**	-8.12	-4.35	-0.98	**-12.91**	-8.33	-4.65	-1.72	**-12.97**	-7.93	-4.84	-2.29
	ML-MVDR	-14.09	-8.12	-1.65	4.70	-13.97	-8.09	-1.59	4.93	-13.94	-7.12	-1.77	4.55	-13.84	-7.81	-1.67	4.39
	MAP-MVDR	-13.28	-7.23	-1.14	4.81	-13.91	-8.03	-1.56	4.94	-13.92	-7.11	-1.76	4.55	-13.84	-7.81	-1.67	4.39
	Bayesian	-12.91	**-6.99**	**-1.05**	**4.83**	-13.61	**-7.82**	**-1.46**	**4.97**	-13.63	**-7.88**	**-1.67**	**4.58**	-13.55	**-7.61**	**-1.57**	**4.41**
	Oracle	-12.50	-6.53	-0.85	4.90	-12.93	-6.91	-0.92	5.07	-12.98	-6.11	-1.24	4.70	-13.08	-7.11	-1.04	4.53
ESTOI	Unprocessed	0.07	0.17	0.33	0.55	0.07	0.18	0.35	0.57	0.08	0.18	0.36	0.57	0.08	0.20	0.35	0.57
	Fixed-MVDR	**0.13**	0.28	0.47	0.67	**0.11**	0.23	0.43	0.60	**0.09**	0.20	0.34	0.53	0.08	0.18	0.33	0.48
	ML-MVDR	0.08	0.25	0.52	0.77	0.08	0.25	**0.54**	**0.77**	**0.09**	**0.25**	**0.53**	**0.77**	**0.09**	0.25	**0.53**	**0.76**
	MAP-MVDR	**0.13**	**0.33**	**0.57**	**0.78**	0.08	**0.26**	**0.54**	**0.77**	**0.09**	**0.25**	**0.53**	**0.77**	**0.09**	0.25	**0.53**	**0.76**
	Bayesian	**0.13**	**0.33**	**0.57**	**0.78**	0.08	**0.26**	**0.54**	**0.77**	**0.09**	**0.25**	**0.53**	**0.77**	**0.09**	**0.26**	**0.53**	**0.76**
	Oracle	0.21	0.39	0.60	0.79	0.20	0.38	0.61	0.78	0.20	0.38	0.59	0.79	0.20	0.38	0.60	0.78

Table A.1: SSNR and ESTOI scores for different beamformers when varying P_t for $\theta_t = 60°$.

	θ_t	90°				60°				30°				0°			
	SNR (dB)	-12	-6	0	6	-12	-6	0	6	-12	-6	0	6	-12	-6	0	6
SSNR	Unprocessed	-19.76	-13.50	-7.59	-1.76	-19.83	-14.05	-8.21	-2.16	-20.34	-14.30	-8.13	-2.17	-19.87	-14.03	-8.11	-1.98
	Fixed-MVDR	**-13.06**	-7.88	-4.12	-1.22	**-12.87**	-7.96	-3.94	-0.89	-13.10	-7.78	-2.93	0.73	**-12.32**	**-6.46**	**-0.54**	**5.54**
	ML-MVDR	-14.23	-8.21	-2.07	4.16	-14.09	-8.12	-1.65	4.70	-14.49	-8.05	-1.03	5.43	-14.09	-8.13	-1.30	5.38
	MAP-MVDR	-14.16	-7.89	-1.83	4.20	-13.28	-7.23	-1.14	4.81	-12.96	-6.87	-0.49	5.48	**-12.32**	**-6.46**	**-0.54**	**5.54**
	Bayesian	-13.78	**-7.66**	**-1.75**	**4.23**	-12.91	**-6.99**	**-1.05**	**4.83**	**-12.68**	**-6.71**	**-0.44**	**5.49**	**-12.32**	**-6.46**	**-0.54**	**5.54**
	Oracle	-13.48	-7.45	-1.52	4.33	-12.50	-6.53	-0.85	4.90	-12.42	-6.44	-0.31	5.63	-12.32	-6.46	-0.54	5.54
ESTOI	Unprocessed	0.07	0.19	0.36	0.57	0.07	0.17	0.33	0.55	0.06	0.17	0.34	0.55	0.07	0.17	0.35	0.57
	Fixed-MVDR	0.11	0.25	0.42	0.63	**0.13**	0.28	0.47	0.67	**0.15**	0.31	0.52	0.71	**0.19**	**0.37**	**0.58**	**0.77**
	ML-MVDR	0.07	0.25	0.51	0.76	0.08	0.25	0.52	0.77	0.08	0.26	0.54	0.78	0.07	0.23	0.52	0.76
	MAP-MVDR	**0.12**	**0.31**	**0.55**	**0.78**	**0.13**	**0.33**	**0.57**	**0.78**	**0.15**	**0.35**	**0.59**	**0.79**	**0.19**	**0.37**	**0.58**	**0.77**
	Bayesian	**0.12**	**0.31**	**0.55**	**0.78**	**0.13**	**0.33**	**0.57**	**0.78**	**0.15**	**0.35**	**0.59**	**0.79**	**0.19**	**0.37**	**0.58**	**0.77**
	Oracle	0.19	0.38	0.59	0.78	0.21	0.39	0.60	0.79	0.20	0.39	0.60	0.79	0.19	0.37	0.58	0.77

Table A.2: SSNR and ESTOI scores for different beamformers when varying θ_t for $P_t = 1$.

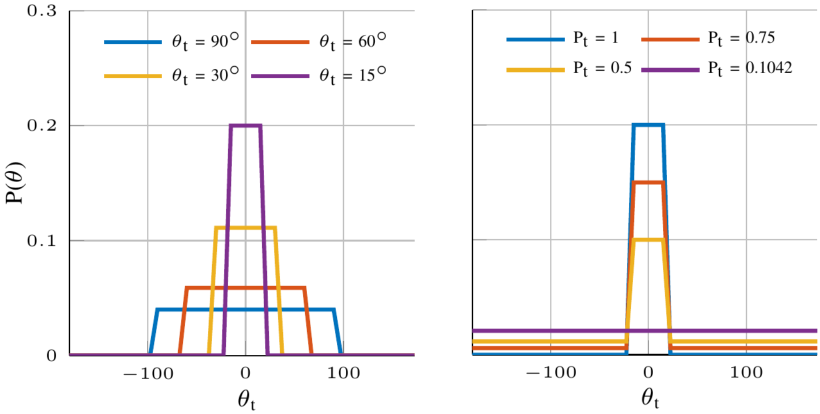


Fig. A.1: Examples of the a priori distributions, $P(\theta)$, for different P_t and θ_t. On the left figure $P_t = 1$ while varying θ_t, and on the right figure $\theta_t = 15^\circ$ while varying P_t.

on each ear of the user. Each BTE HA has two microphones placed at a distance of 1.3 cm, thus the first array has two microphones while the second array has four. We assume perfect wireless signal exchange between the HAs if two are used. Head-related transfer functions are obtained from measurements with HAs placed on a real head and used to derive the RTF vectors with the front microphone of the left BTE HA used as reference microphone. The RTF vectors are available with a resolution of 7.5° around a circle in the horizontal plane.

The target signal is speech from a range of different male and female speakers from the TIMIT corpus [17]. The noise type is babble and speech-shaped noise (SSN). We construct approximately isotropic babble or SSN noise fields by letting different speech signals or SSN sequences impinge on the array from all $\frac{360^\circ}{7.5^\circ}$=48 directions. The babble noise is constructed using speech, while the SSN is created by filtering Gaussian white noise through a spectral shaping filter.

The duration of an acoustic scene is 5 seconds and for each scene realization, new speech signals are randomly chosen as the target and the babble noise, while new signal realizations are drawn for SSN. For acoustic scene realizations, the target direction is drawn from a discrete circular distribution such that $\theta_s \sim P(\theta)$ and is kept fixed during the acoustic scene. We assume that the circular distribution, with an angular resolution of 7.5°, is available as a prior distribution for the Bayesian and MAP beamformers. We choose to model the circular distribution as a piecewise linear function such that the target probability, P_t, of arriving from an interval is $P_t = P(-\theta_t \leq \theta \leq \theta_t)$ as shown in Fig A.1.

To transform the noisy signal into the TF domain we use the STFT with a frame length of 256 samples, sampling frequency of 16 kHz, and a square-root Hanning window with 50% overlap for analysis and synthesis. The first second of each acoustic scene has no speech activity in order to initialize the noise CPSD matrices which are updated by recursive averaging with a time constant of 50 ms.

5 Results and discussion

To quantify the performance of the beamformers, we use the segmental SNR (SSNR) [18] and extended short-time objective intelligibility (ESTOI) [19] for evaluating the noise reduction and speech intelligibility performance respectively. The scores are averaged over 200 realizations for each type of acoustic scene. The Bayesian and MAP beamformers are compared to the ML beamformer [6] and a fixed MVDR beamformer which is often used in practical HA systems [20–22]. In our experiments, the fixed MVDR beamformer is steered towards 0° i.e. the front of the hearing aid user. Finally, to estimate an upper performance bound, we also compare the performance to an oracle beamformer which for each realization is steered towards the true target DOA.

In the following, due to space limitations, we show results using four microphones and acoustic scenes with babble noise. The simulation experiments with SSN and the two microphone array, lead to essentially identical conclusions, although for two microphones, the width of the beam is wider, so that the performance difference between the methods is less significant.

5.1 Experiment 1: Varying P_t for fixed θ_t

In this experiment we vary P_t, while keeping θ_t fixed, to examine how changing the certainty of the target arriving from $[-\theta_t, \theta_t]$ affects the beamforming performance. We expect that as P_t approaches 1, the advantage of incorporating the prior distribution in the beamformer, as done for the MAP and Bayesian beamformers, becomes more significant. We test for $P_t = (1, 0.75, 0.5, 0.35)$ of the target arriving in the interval $[-60^\circ, 60^\circ]$. With $P_t = 1$, the target certainly arrives from $[-\theta_t, \theta_t]$, while for $P_t = 0.35$ the distribution becomes uniform i.e. the target is equally likely to arrive from any direction. To examine the robustness against different SNR levels, the beamformers are evaluated for SNRs in $[-12 \text{ dB}, 6 \text{ dB}]$.

The SSNR and ESTOI scores for the first experiment are reported in Table A.1. As P_t approaches 1, we observe that the improvement of the Bayesian and MAP beamformers over the ML beamformer increases. For $P_t = 0.35$, the MAP and ML beamformers give identical scores, since the prior distribution is uniform and the MAP beamformer reduces to the ML beamformer. At high SNRs, e.g. 6 dB, the Bayesian, MAP and ML beamformers perform approximately identically in terms of SSNR, but with the Bayesian beamformer performing close to the oracle beamformer. At low SNRs, e.g. -12 dB, the fixed MVDR beamformer performs best in terms of SSNR. The reason for this is that DOA estimation at such low SNR is very challenging, and that the fixed MVDR beamformer steers a beam towards a fixed DOA of $\theta = 0°$, which is centered in the actual DOA range $[-\theta_t, \theta_t]$ used in the simulation. However, at higher SNRs, the MAP and Bayesian beamformers always return a higher SSNR than the fixed MVDR beamformer.

In terms of ESTOI, similar patterns as for the SSNR are observed. The MAP and Bayesian beamformers perform best when P_t is close to 1 and approach the oracle beamformer performance at high SNRs. Compared to the ML beamformer, the greatest improvement is when the SNR is between -6 dB and 0 dB. For $P_t = 0.35$, the Bayesian, MAP, and ML beamformers perform approximately identically.

5.2 Experiment 2: Varying θ_t for fixed P_t

In the second experiment, P_t is kept fixed at $P_t = 1$ for $\theta_t = (90°, 60°, 30°, 0°)$ to examine the effect of changing the width of the prior distribution. The results are shown in Table A.2. For $\theta_t = 0°$, the target always arrives from the front, so the fixed MVDR, MAP, and Bayesian beamformers give identical SSNR and ESTOI scores. As θ_t approaches 0°, the SSNR and ESTOI improvement over the ML beamformer increase and the performance of the MAP and Bayesian beamformers approaches the performance of the oracle beamformer. Hence, as expected, when the interval θ_t of the prior distribution decreases, the advantage of integrating a prior distribution for the MAP and Bayesian beamformers becomes more significant.

6 Conclusion

We proposed a beamformer that integrates a priori information about the potential DOA of the target sound source and utilizes a novel noise CPSD model to derive the a posteriori distribution for MAP and Bayesian beamformers. We compare the proposed beamformers to a state-of-the-art ML beamformer and fixed-MVDR beamformer commonly used for hearing assistive devices. We demonstrate in simulation experiments the advantage of having a realistic a priori distribution of the target direction available for the proposed beamformers. The beamformers were tested on simulated, but realistic acoustic scenes with binaural hearing aids. In terms of SSNR and ESTOI scores, results indicated that the proposed beamformers outperform the ML beamformer in both low and high SNRs. At high SNRs, they approach the performance of an oracle beamformer that has access to the true target DOA.

References

[1] H. L. Van Trees, *Optimum array processing*, ser. Detection, estimation, and modulation theory / Harry L. Van Trees. New York: Wiley, 2002, no. 4.

[2] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*, ser. Springer topics in signal processing. Berlin: Springer, 2008, no. 1.

[3] 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*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[4] A. Kuklasinski and J. Jensen, "Multichannel Wiener Filters in Binaural and Bilateral Hearing Aids — Speech Intelligibility Improvement and Robustness to DoA Errors," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, pp. 8–16, Feb. 2017. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=18539

[5] S. Braun, W. Zhou, and E. A. P. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA: IEEE, Oct. 2015, pp. 1–5.

[6] 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, Mar. 2018. [Online]. Available: http://ieeexplore.ieee.org/document/8186189/

[7] J. Jensen and M. S. Pedersen, "Analysis of beamformer directed single-channel noise reduction system for hearing aid applications," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 5728–5732.

[8] A. Kuklasinski, S. Doclo, T. Gerkmann, S. Holdt Jensen, and J. Jensen, "Multi-channel PSD estimators for speech dereverberation - A theoretical and experimental comparison," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 91–95. [Online]. Available: http://ieeexplore.ieee.org/document/7177938/

[9] J. Jensen and A. Kuklasiński, "Multi-microphone method for estimation of target and noise spectral variances for speech degraded by reverberation and optionally additive noise," Aug. 1 2017, uS Patent 9,723,422.

[10] 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*, ser. Synthesis lectures on speech and audio processing. Williston, VT: Morgan & Claypool, 2013, no. 11.

[11] 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, Feb. 2000. [Online]. Available: http://ieeexplore.ieee.org/document/823966/

[12] 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, Nov. 2006. [Online]. Available: http://ieeexplore.ieee.org/document/1710387/

[13] S. Chakrabarty and E. A. P. 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, no. 1, pp. 145–160, Jan. 2018.

[14] Hao 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, Apr. 1995. [Online]. Available: http://ieeexplore.ieee.org/document/376846/

[15] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984. [Online]. Available: http://ieeexplore.ieee.org/document/1164453/

[16] L. Svensson and M. Lundberg, "On posterior distributions for signals in Gaussian noise with unknown covariance matrix," *IEEE Transactions on Signal Processing*, vol. 53, no. 9, pp. 3554–3571, Sep. 2005. [Online]. Available: http://ieeexplore.ieee.org/document/1495890/

[17] "TIMIT: acoustic-phonetic continuous speech corpus." Philadelphia, Pa., 1993.

[18] J. H. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Fifth international conference on spoken language processing*, 1998.

[19] 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, Nov. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7539284/

[20] A. Schaub, *Digital hearing aids*. New York: Thieme, 2008.

[21] J. M. Kates, *Digital hearing aids*, ser. Audiology. San Diego, Calif.: Plural Publ, 2008.

[22] H. Dillon, *Hearing aids*, 2nd ed. Sydney: Boomerang Press [u.a.], 2012.

This page intentionally left blank.

Paper B

Maximum Likelihood Estimation of the Interference-Plus-Noise Cross Power Spectral Density Matrix for Own Voice Retrieval

Poul Hoang, Zheng-Hua Tan, Jan Mark de Haan, Thomas Lunner, and Jesper Jensen

The paper has been published in
2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6939-6943, 2020.

Abstract

In headset and hearing aid applications, it is of interest to retrieve the user's own voice in a noisy environment, e.g. for telephony applications. To do so, the cross-power spectral density (CPSD) of the interference-plus-noise is required. In this paper, a novel maximum likelihood (ML) estimator of the interference-plus-noise CPSD matrix is proposed. The proposed method is able to estimate the interference-plus-noise CPSD matrix, even during signal regions with own voice activity. The method uses a novel procedure for estimating the interference-plus-noise CPSD matrix by first estimating the interference PSD and afterwards the noise PSD in a maximum likelihood sense. Simulation experiments, where the proposed method is compared to other noise CPSD matrix estimators, show that it performs on par or better than competing methods, particularly, in situation where the interference-to-noise ratio is large.

1 Introduction

Speech enhancement and noise reduction algorithms are often needed in real-world audio applications, where noise from the acoustic environment masks a desired speech signal. Examples include hands-free wireless communication devices, e.g. headsets, automatic speech recognition systems, and hearing aids (HA) [1]. In particular for applications such as headset communication devices, where the user's voice is picked up by the headset microphones and transmitted to a far-end conversational partner, noise can greatly reduce sound quality and speech intelligibility making conversations more difficult.

Noise reduction algorithms in multi-microphone devices are often implemented as spatio-temporal filters [1–3]. To find the optimal filter coefficients, it is usually necessary to know the characteristics of the acoustic environment, e.g. in the form of spatial, spectral and/or temporal noise and target statistics. Typically, these statistics are unknown and must be estimated online from the noisy speech signal.

Statistics that are often necessary for multi-microphone noise reduction algorithms include the cross-power spectral densities (CPSDs) of the noise. To estimate these statistics, a variety of estimators have been proposed, e.g. [4–9]; see also [3]. In [4, 7] a maximum likelihood (ML) estimator of the noise CPSD matrix is proposed using the assumption, that the noise CPSD matrix remains identical up to a scalar multiplier during speech presence. This estimator performs well, when the structure of the noise CPSD matrix does not change significantly over signal regions with speech activity. However, in many realistic acoustic scenes, the underlying structure of the noise CPSD matrix cannot be assumed fixed, e.g., when a prominent speech-like interferer is present in the acoustic scene. In this case, many existing noise reduction systems fail at efficiently suppressing the interferer, as it is harder to determine if the target or the interferer is the desired speech.

In this paper we propose a novel maximum likelihood estimator of the noise CPSD matrix. The method avoids the assumption of a fixed noise PSD matrix structure made in [4, 7]. It also overcomes the problem of distinguishing the target from competing speakers. Additionally, in cases when the noise CPSD matrix structure is actually time-invariant, the proposed method performs as well as existing methods.

2 Signal Model

We consider an acoustic scene consisting of a user equipped with hearing aids or a headset with access to $M > 2$ microphones. The microphones pick up the user's own voice in the presence of noise and interfering speakers, and the noisy signal is sampled into a discrete-time sequence $x_m(t) \in \mathbb{R}$ for all $m = 1, ..., M$ microphones. Fig. B.1 shows an example situation, where the user's *own voice* signal is denoted $s_o'(t)$, and the ambient noise $v_e'(t)$. In this example, a single point-like interferer is present as a competing speaker $v_c'(t)$ arriving from an angle θ_c. Let $x_m(t)$ denote the noisy signal picked up at microphone m, and let $x_m(k, n)$ denote its short-time Fourier Transform (STFT), where k, n denote a frequency bin and a frame index, respectively. We vectorize $x_m(t)$ such that $\mathbf{x}(k,n) = [x_1(k,n),..., x_M(k,n)]^T$ and

$$\mathbf{x}(k,n) = s_o(k,n)\mathbf{d}_o(k,n) + v_c(k,n)\mathbf{d}_c(k,n,\theta_c) + \mathbf{v}_e(k,n), \tag{B.1}$$

where $\mathbf{d}_o(k, n)$ denotes the relative transfer function (RTF) [10] from the users' mouth to the HA microphones, with respect to a pre-selected reference microphone. The vector $\mathbf{d}_c(k, n, \theta_c)$ is the RTF from an interferer arriving from an unknown direction $\theta_c \in \Theta$, where we assume that $\Theta = \theta_1, ...\theta_I$ to the HA microphones, and where we have used the notation $\theta_c \triangleq \theta_c(k, n)$ for brevity. In other words, the proposed signal model in (B.1) is more general than examplified in Fig. 1 in that it allows for different interferer angles in different time-frequency units. Furthermore, $s_o(k, n)$ and $v_c(k, n)$ denote the own-voice signal and the interferer signal at a pre-selected reference microphone, respectively. Finally, $\mathbf{v}_c(k, n)$ is the ambient noise signal at the microphones. We assume that the own voice RTF vector is time-invariant i.e. $\mathbf{d}_o(k) \triangleq \mathbf{d}_o(k, n)$. This assumption is reasonable, because the microphones of the headset or hearing aid tend to be located in a fixed position with respect to the mouth of the user. Moreover, we make the standard assumption that $s_o(k, n)$, $v_c(k, n)$, and $\mathbf{v}_\mathrm{e}(k, n)$ are mutually uncorrelated random processes

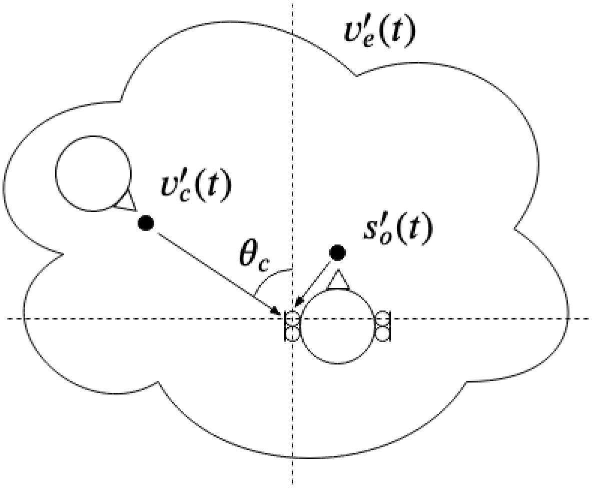


Fig. B.1: An example acoustic scene where $s'_o(t)$ is the own voice. The signal $v'_c(t)$ is the speech signal of a competing speaker arriving from direction θ_c and $v'_e(t)$ is the noise.

meaning that the noisy CPSD matrix, $\mathbf{C}_x(k,n) = \mathbb{E}\{\mathbf{x}(k,n)\mathbf{x}^H(k,n)\}$, is

$$\begin{aligned}\mathbf{C}_x(k,n) &= \lambda_s(k,n)\mathbf{d}_o(k)\mathbf{d}_o^H(k) \\ &+ \underbrace{\lambda_c(k,n)\mathbf{d}_c(k,\theta_c)\mathbf{d}_c^H(k,\theta_c) + \lambda_\mathrm{e}(k,n)\mathbf{\Gamma}_\mathrm{e}(k,n)}_{\mathbf{C}_v(n,k)},\end{aligned} \tag{B.2}$$

where $\lambda_s(k,n)$, $\lambda_c(k,n)$, and $\lambda_e(k,n)$ are power spectral densities (PSDs) of the own-voice, interference, and noise, respectively. The matrix $\mathbf{\Gamma}_\mathrm{e}(k,n)$ is the normalized noise CPSD matrix with a value of 1 at the diagonal element corresponding to the reference microphone index. We assume that $\mathbf{\Gamma}_\mathrm{e}(k,n)$ is a known matrix and may be chosen according to prior knowledge or estimated in noise only regions. The own voice RTF vector $\mathbf{d}_o(k)$ is assumed known as it can be measured before deployment. The parameters that remain to be estimated are $\lambda_c(k,n)$, $\lambda_e(k,n)$, and θ_c. We derive ML estimators of these parameters in the following section.

3 Maximum Likelihood Interference-Plus-Noise PSD Estimation

To estimate the interference and noise PSDs (i.e. $\lambda_c(k,n)$ and $\lambda_e(k,n)$) and the interference direction θ_c, we first apply an own voice canceling beamformer to obtain an interference-plus-noise-only signal. The own voice canceling beamformer is implemented using an own voice blocking matrix $\mathbf{B}_o(k) \in \mathbb{C}^{M\times M-1}$.

To obtain the own voice blocking matrix, we first find the orthogonal projection matrix of $\mathbf{d}_o(k)$ and then select the first $M-1$ column vectors of the projection matrix [7]. Let $\mathbf{I}_{M\times M}$ be an $M \times M$ identity matrix and let $\mathbf{I}_{M\times M-1}$ denote the first $M-1$ column vectors of $\mathbf{I}_{M\times M}$. The own voice blocking matrix is then given as

$$\mathbf{B}_o(k) = \left(\mathbf{I}_{M\times M} - \frac{\mathbf{d}_o(k)\mathbf{d}_o^H(k)}{\mathbf{d}_o^H(k)\mathbf{d}_o(k)}\right)\mathbf{I}_{M\times M-1}. \tag{B.3}$$

The own voice blocked signal, $\mathbf{z}(k,n)$, can be expressed as

$$\begin{aligned}\mathbf{z}(k,n) &= \mathbf{B}_o^H(k)\mathbf{x}(k,n)\\ &= v_c(k,n)\underbrace{\mathbf{B}_o^H(k)\mathbf{d}_c(k,\theta_c)}_{\tilde{\mathbf{d}}_c(k,\theta_c)} + \underbrace{\mathbf{B}_o^H(k)\mathbf{v}_\mathrm{e}(k,n)}_{\tilde{\mathbf{v}}_\mathrm{e}(k,n)},\end{aligned} \tag{B.4}$$

and the own voice blocked CPSD matrix is

$$\begin{aligned}\mathbf{C}_z(k,n) &= \mathbb{E}\{\mathbf{z}(k,n)\mathbf{z}^H(k,n)\}\\ &= \lambda_c(k,n)\tilde{\mathbf{d}}_c(k,\theta_c)\tilde{\mathbf{d}}_c^H(k,\theta_c) + \lambda_\mathrm{e}(k,n)\tilde{\mathbf{\Gamma}}_\mathrm{e}(k,n).\end{aligned}$$

3.1 Concentrated log-likelihood

In order to derive ML-estimates of $\lambda_c(k,n)$, $\lambda_\mathrm{e}(k,n)$, and θ_c, we follow an approach similar to the one presented in [11, 12]. We assume that the own-voice, interference, and noise are independent and identically distributed across a short time duration of N frames, and are distributed according to a circular symmetric complex Gaussian distribution. Hence, the blocked own voice-plus-interference signal is also circular symmetric complex Gaussian distributed, i.e. $\mathbf{z}(k,n) \sim \mathcal{N}_\mathbb{C}\left(\mathbf{0}, \mathbf{C}_z(k,n)\right)$. Let $\mathbf{Z}(k,n) = [\mathbf{z}(k,n-N+1),...,\mathbf{z}(k,n)] \in \mathbb{C}^{M-1\times N}$ denote N observations of $\mathbf{z}(k,n)$, then the likelihood of $\mathbf{Z}(k,n)$ is

$$\begin{aligned}&f(\mathbf{Z}(k,n)|\theta_i, \lambda_c(k,n), \lambda_\mathrm{e}(k,n)) =\\ &\qquad\frac{\exp\left(-N\mathrm{tr}\left(\hat{\mathbf{C}}_z(k,n)\mathbf{C}_z^{-1}(k,n,\theta_i)\right)\right)}{\pi^{MN}|\mathbf{C}_z(k,n,\theta_i)|^N},\end{aligned} \tag{B.5}$$

where $\hat{\mathbf{C}}_z(k,n) = \frac{1}{N}\mathbf{Z}(k,n)\mathbf{Z}^H(k,n)$ is the sample estimate of $\mathbf{C}_z(k,n)$ and $\mathrm{tr}(\cdot)$ is the trace operator. For a given $\tilde{\mathbf{d}}_c(k,\theta_i)$, the ML-estimates of $\lambda_c(k,n)$ and $\lambda_\mathrm{e}(k,n)$ were derived in [7, 13]. Specifically, the ML-estimator of $\lambda_\mathrm{e}(k,n)$ given $\tilde{\mathbf{d}}_c(k,\theta_i)$ is

$$\begin{aligned}&\hat{\lambda}_\mathrm{e}(k,n,\theta_i) = \frac{1}{M-2}\times\\ &\mathrm{tr}\left(\hat{\mathbf{C}}_q(k,n,\theta_i)\left(\tilde{\mathbf{B}}^H(\theta_i)\tilde{\mathbf{\Gamma}}_\mathrm{e}(k,n)\tilde{\mathbf{B}}(\theta_i)\right)^{-1}\right),\end{aligned} \tag{B.6}$$

where $\hat{\mathbf{C}}_q(k,n) = \frac{1}{N}\tilde{\mathbf{B}}^H(\theta_i)\mathbf{Z}(k,n)\mathbf{Z}^H(k,n)\tilde{\mathbf{B}}(\theta_i)$ is the sample CPSD matrix of the own voice-plus-interference blocked signal and $\tilde{\mathbf{B}}(\theta_i) \in \mathbb{C}^{M-1\times M-2}$ is the own-voice-plus-interference blocking matrix, i.e.

$$\tilde{\mathbf{B}}(\theta_i) = \left(\mathbf{I}_{M\text{–}1\times M\text{–}1} - \frac{\tilde{\mathbf{d}}_c(k,\theta_i)\tilde{\mathbf{d}}_c^H(k,\theta_i)}{\tilde{\mathbf{d}}_c^H(k,\theta_i)\tilde{\mathbf{d}}_c(k,\theta_i)}\right)\mathbf{I}_{M\text{–}1\times M\text{–}2}.$$

Furthermore, for a given $\tilde{\mathbf{d}}_c(k,\theta_i)$, the ML estimator of the interference PSD is given by [7, 13]

$$\begin{aligned}&\hat{\lambda}_{\mathrm{c}}(k,n,\theta_i) =\\ &\tilde{\mathbf{w}}^H(k,n,\theta_i)\big(\hat{\mathbf{C}}_z(k,n) - \hat{\lambda}_{\mathrm{e}}(k,n,\theta_i)\tilde{\boldsymbol{\Gamma}}_{\mathrm{e}}(k,n)\big)\tilde{\mathbf{w}}(k,n,\theta_i),\end{aligned}$$

where $\tilde{\mathbf{w}}(\theta_i)$ is the MVDR beamformer constructed from the blocked own voice CPSD matrix as follows

$$\tilde{\mathbf{w}}(k,n,\theta_i) = \frac{\tilde{\boldsymbol{\Gamma}}_{\mathrm{e}}^{-1}(k,n)\tilde{\mathbf{d}}_c(k,\theta_i)}{\tilde{\mathbf{d}}_c^H(k,\theta_i)\tilde{\boldsymbol{\Gamma}}_{\mathrm{e}}^{-1}(k,n)\tilde{\mathbf{d}}_c(k,\theta_i)}. \tag{B.7}$$

Inserting the ML estimates $\hat{\lambda}_{\mathrm{e}}(k,n,\theta_i)$ and $\hat{\lambda}_{\mathrm{c}}(k,n,\theta_i)$ into the likelihood function in (D.15), we obtain the concentrated likelihood function $\bar{f}(\mathbf{Z}(k,n)|\theta_i,\hat{\lambda}_{\mathrm{c}}(k,n,\theta_i),\hat{\lambda}_{\mathrm{e}}(k,n,\theta_i))$ which we denote as $\bar{f}(\mathbf{Z}(k,n)|\theta_i)$. It can be shown that the concentrated log-likelihood function is proportional to [13]

$$\begin{aligned}&\ln \bar{f}(\mathbf{Z}(k,n)|\theta_i) \propto\\ &-\ln|\hat{\lambda}_{\mathrm{c}}(k,n,\theta_i)\tilde{\mathbf{d}}_c(k,\theta_i)\tilde{\mathbf{d}}_c(k,\theta_i)^H + \hat{\lambda}_{\mathrm{e}}(k,n,\theta_i)\tilde{\boldsymbol{\Gamma}}_{\mathrm{e}}(k,n)|.\end{aligned}$$

3.2 ML estimate of $\lambda_c(k,n)$, $\lambda_e(k,n)$, and θ_c

We find ML estimates of $\lambda_c(k,n)$, $\lambda_e(k,n)$, and θ_c by evaluation the concentrated likelihood for a candidate set of vectors $\tilde{\mathbf{d}}_c(k,\theta_i)$, $\theta_i \in \Theta$, and choosing the vector and subsequently the estimates of $\lambda_c(k,n)$ and $\lambda_e(k,n)$ that maximize the concentrated likelihood. While the proposed framework is general and allows potentially different interferers in each time-frequency bin, we assume here that only one single interferer is present in the acoustic environment and that the noisy observations across frequency bins are uncorrelated. Then the wideband concentrated log-likelihood function is [12]

$$\ln \bar{f}(\mathbf{Z}(1,n),...,\mathbf{Z}(K,n)|\theta_i) = \sum_{k=1}^{K} \ln \bar{f}(\mathbf{Z}(k,n)|\theta_i), \tag{B.8}$$

where K is the total number of frequency bins of the one-sided spectrum. To obtain the ML estimate of the interference direction $\hat{\theta}_c$, we maximize the wideband concentrated log-likelihood function with respect to θ_i. As θ_i belongs to a discrete set of directions, the ML estimate of θ_c is obtained through an exhaustive search over θ_i. Finally, to obtain an estimate of the interference-plus-noise CPSD matrix we insert the ML estimates $\hat{\lambda}_c(k,n,\theta_i)$, $\hat{\lambda}_{\mathrm{e}}(k,n,\theta_i))$, and $\hat{\theta}_c$ into the interference-plus-noise CPSD model, i.e.

$$\begin{aligned}\hat{\mathbf{C}}_v(k,n) = \hat{\lambda}_c(k,n,\hat{\theta}_c)\mathbf{d}_c(k,\hat{\theta}_c)\mathbf{d}_c^H(k,\hat{\theta}_c) \\ + \hat{\lambda}_e(k,n,\hat{\theta}_c)\boldsymbol{\Gamma}_{\mathrm{e}}(k,n).\end{aligned} \tag{B.9}$$

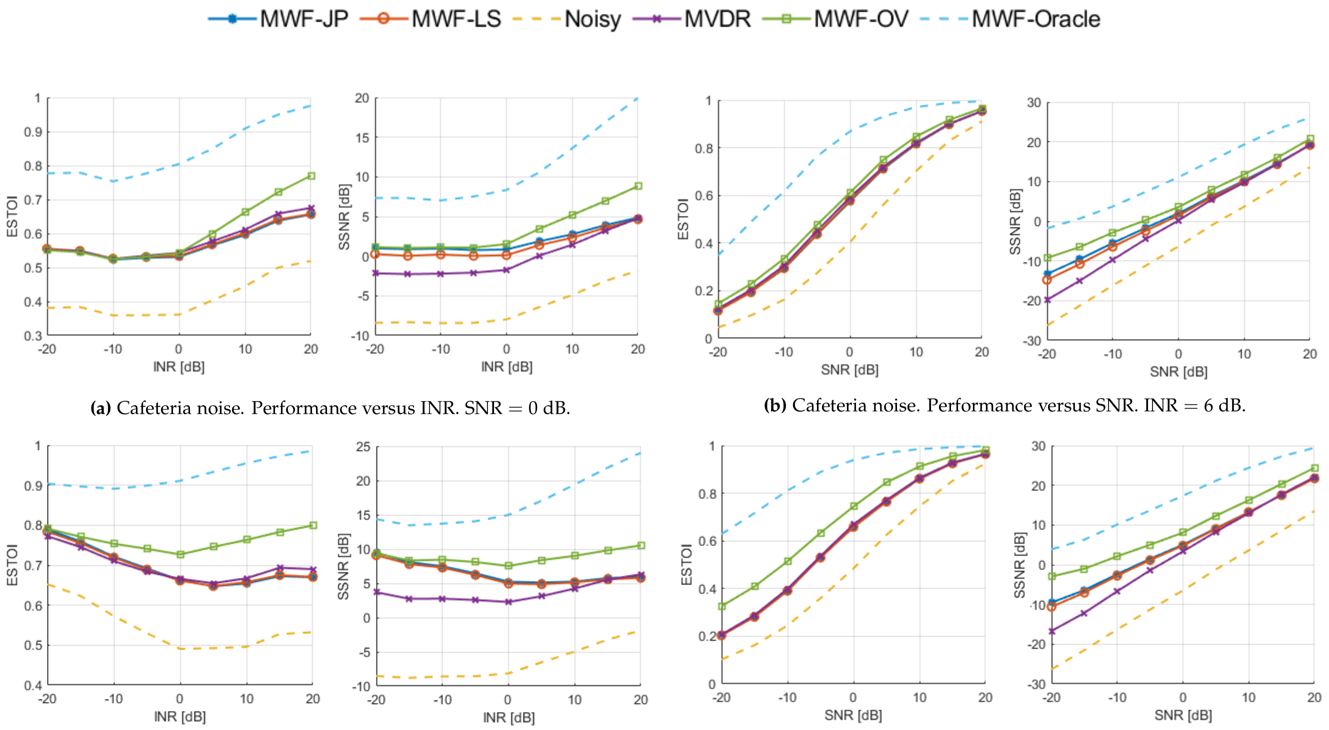
4 Performance Evaluation

To analyse the performance of the proposed ML noise CPSD matrix estimator, we conduct simulation experiments for own voice retrieval in hearing aid (HA) applications. The proposed estimator is used in combination with a multichannel Wiener filter (MWF) beamformer [14, sec. 4.3] to retrieve the own voice. The performance is reported in terms of extended short-time objective intelligibility (ESTOI) [15] and segmental SNR (SSNR) [16] to estimate speech intelligibility and noise reduction performance, respectively. The proposed beamformer (MWF-OV) is compared to three other beamformers, 1) a standard MVDR beamformer, 2) an MWF beamformer, MWF-JP, where the noise CPSD matrix $\mathbf{C}_v(k,n)$ is estimated using [7], and 3) an MWF beamformer, MWF-LS, where $\mathbf{C}_v(k,n)$ is estimated using [17, 18]. The performance scores for the unprocessed noisy observations and an oracle MWF beamformer (MWF-Oracle), where the true signal statistics are known, are included to obtain an lower and upper bound of performance, respectively.

4.1 Simulation setup

Two behind-the-ear (BTE) HA, each with two microphones separated with a distance of 1.3 cm, are mounted behind the left and right ear of a real human, and acoustic transfer functions are measured from the loudspeaker to the microphones where the loudspeaker was placed at a distance of 1.9 meters and angle of θ_c to derive the RTF vectors $\mathbf{d}_o(k)$ and candidate vectors $\mathbf{d}_c(k,\theta_c)$ for $\theta_c = -172.5°, ..., 180°$. The direction of the competing speaker is randomized according to a uniform circular distribution for each acoustic scene.

We simulate an acoustic situation, where a HA user's own voice is picked up by the HA microphones in the presence of an interferer in the form of a competing speaker and ambient noise, c.f. Fig B.1. The clean own voice target signal $s_o(k,n)$ and interference signal $v_c(k,n)$ are speech signals obtained from the TIMIT database [19]. The ambient noise types are car and cafeteria



(a) Cafeteria noise. Performance versus INR. SNR = 0 dB.

(b) Cafeteria noise. Performance versus SNR. INR = 6 dB.

(c) Car noise. Performance versus INR. SNR = 0 dB.

(d) Car noise. Performance versus SNR. INR = 6 dB.

Fig. B.2: Performance scores when varying either INR or SNR for different noise fields and noise types.

noise. These natural noise sources were measured with a spherical microphone and transformed in the simulation to reproduce a realistic noise field at the HA microphones. The speech signals and ambient noise segments are randomly selected from the database for each new realization of an acoustic scene.

A number of 150 different acoustic scenes are created to obtain averaged performance scores. The duration of each acoustic scene is 5 seconds and the own voice is active during the whole simulation. The sampling frequency is 16 kHz and for the STFT, a the square-root Hanning window is used for analysis and synthesis with a window size of 256 samples and an overlap of 50%. To follow an approach that can be used in practice, we make the rough assumption that the noise field is approximately isotropic. Hence we initialize the normalized noise CPSD matrix, $\mathbf{\Gamma}_\mathrm{e}(k,n)$, to be

$$\mathbf{\Gamma}_\mathrm{e}(k,n) = \frac{1}{|\Theta|} \sum_{\theta_c \in \Theta} \mathbf{d}_c(k,\theta_c)\mathbf{d}_c^H(k,\theta_c), \tag{B.10}$$

with $\Theta = \{-172.5°, -165°, ..., 180°\}$ and $|\Theta| = 48$. Note that the natural noise fields used in the simulation do not strictly obey this noise model.

4.2 Experiment 1: Performance when varying INR

In the first experiment, the signal-to-noise ratio (SNR), which is defined as the ratio between the own-voice power and the interference-plus-noise power, is fixed at 0 dB. The interference-to-noise ratio (INR), defined as the ratio between the interference power and the ambient noise power, is varied between -20 dB to 20 dB in intervals of 5 dB. The purpose of the experiment is to examine how the power ratio between the competing speaker and noise influences the performance. The results are shown in Fig. B.2a and Fig. B.2c for car noise and cafeteria noise, respectively. We observe that the proposed method outperforms the competing methods significantly, when the INR is large and for both noise types. Clearly, the proposed explicit modeling of point-like interferers is highly advantageous in this situation. At low INR levels, the performance of the proposed method converges to MWF-JP.

4.3 Experiment 2: Performance when varying SNR

For the second experiment, the INR is fixed at 6 dB while varying the SNR from -20 dB to 20 dB in intervals of 5 dB. In this experiment we examine how the performance is influenced by varying the SNR. The results are shown in Fig. B.2b and Fig. B.2d for car noise and cafeteria noise, respectively. With the proposed method, we observe an improvement in SSNR of 1.6 dB and 3.1 dB at input SNR $= 0$ dB for cafeteria and car noise, respectively, and an improvement of 4.0 dB and 6.6 dB at input SNR $= -20$ dB for cafeteria and car noise. The better performance in car noise may be due to the fact that the underlying N-frame stationarity assumption from section 3.1 is better valid. Furthermore, the performance improvement increases at lower input SNR.

5 Conclusion

Own-voice retrieval in a noisy environment is of interest in headset and hearing aid applications. The task of own-voice retrieval may be solved using multi-channel speech enhancement algorithms. Such algorithms rely on knowledge of the inter-microphone interference-plus-noise CPSD matrix. In this paper, we propose a maximum likelihood estimator of the interference-plus-noise CPSD matrix during own voice signal regions. We conduct simulation experiments for own voice retrieval in a hearing aid application. The results indicate that the proposed method performs on par or better than competing methods for estimating the interference-plus-noise CPSD matrix. The results indicate that the improvement of the proposed method increases as a function of the interference-to-noise ratio and for lower signal-to-noise ratios.

References

[1] 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*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[2] H. L. Van Trees, *Optimum array processing*, ser. Detection, estimation, and modulation theory / Harry L. Van Trees. New York: Wiley, 2002, no. 4.

[3] M. Parchami, W.-P. Zhu, B. Champagne, and E. Plourde, “Recent Developments in Speech Enhancement in the Short-Time Fourier Transform Domain,” *IEEE Circuits and Systems Magazine*, vol. 16, no. 3, pp. 45–77, 2016.

[4] 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)*, Aug 2012, pp. 295–299.

[5] Yujie Gu and A. Leshem, “Robust Adaptive Beamforming Based on Interference Covariance Matrix Reconstruction and Steering Vector Estimation,” *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3881–3885, Jul. 2012.

[6] R. C. Hendriks and T. Gerkmann, “Noise Correlation Matrix Estimation for Multi-Microphone Speech Enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 223–233, Jan. 2012.

[7] J. Jensen and M. S. Pedersen, “Analysis of beamformer directed single-channel noise reduction system for hearing aid applications,” in *2015 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 5728–5732.

[8] M. Souden, J. Chen, J. Benesty, and S. Affes, "An Integrated Solution for Online Multichannel Noise Tracking and Reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2159–2169, Sep. 2011.

[9] M. Taseska and E. A. P. Habets, "Non-Stationary Noise PSD Matrix Estimation for Multichannel Blind Speech Extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2017.

[10] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001. [Online]. Available: http://ieeexplore.ieee.org/document/934132/

[11] 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, Mar. 2018.

[12] P. Hoang, Z.-H. Tan, J. 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) (GlobalSIP 2019)*, Ottawa, Canada, Nov. 2019.

[13] Hao 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, Apr. 1995.

[14] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, 2001.

[15] 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, Nov. 2016.

[16] J. H. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Fifth international conference on spoken language processing*, 1998.

[17] S. Braun and E. P. Habets, "A multichannel diffuse power estimator for dereverberation in the presence of multiple sources," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–14, 2015.

[18] S. Braun, A. Kuklasinski, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and Comparison of Late Reverberation Power Spectral Density Estimators," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1056–1071, Jun. 2018.

[19] "TIMIT: acoustic-phonetic continuous speech corpus." Philadelphia, Pa., 1993.

Paper C

Joint Maximum Likelihood Estimation of Power Spectral Densities and Relative Acoustic Transfer Functions for Acoustic Beamforming

Poul Hoang, Zheng-Hua Tan, Jan Mark de Haan, and Jesper Jensen

The paper has been published in
2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6119-6123, 2021.

Abstract

Acoustic beamforming is crucial for many applications where extraction of a target signal from a noisy environment is required. In order to implement practical beamformers, e.g. the multichannel Wiener filter (MWF), estimation of the target and noise power spectral densities (PSDs), and the relative acoustic transfer functions (RATFs) is essential. Several methods, e.g. the so-called covariance whitening (CW) approach, have been proposed for estimating these parameters. However, it seems largely unknown that the CW approach in fact leads to maximum likelihood (ML) estimates of the RATFs. We use historical results to derive joint ML estimates (MLEs) of the RATFs and PSDs in the context of acoustic beamforming. In addition, based on the MLEs, we propose a basic VAD framework using concentrated likelihood ratios. We use the joint MLEs of the PSDs, RATFs, and the proposed VAD to implement beamformers in a hearing aid application, and compare its performance to competing methods. Simulation results show that the proposed scheme can outperform competing methods, in particular in realistic situations where highly accurate prior RATF knowledge is not available or at higher signal-to-noise ratios.

1 Introduction

Speech is often contaminated by background noise which can make it difficult for humans or human-to-machine interaction systems to extract verbal information when the environment becomes increasingly noisy. Acoustic beamforming is often used due to its ability to effectively suppress background noise [1, 2]. Commonly used beamformers include the minimum variance distortionless response (MVDR) beamformer and multichannel Wiener filter (MWF) beamformer [2]. Often, the implementation of these beamformers requires the knowledge of the relative acoustic transfer functions (RATFs) from the target to the microphones [3, 4] and information about the target and noise statistics, e.g. power spectral densities (PSDs) [4–6].

Estimation of the RATFs and PSDs is often treated as separate problems, e.g. [6–8]. As a result, there exists a great variety of contributions on PSD estimation for known RATFs [5–7, 9], or oppositely, methods that estimate RATFs for known PSDs [8, 10, 11]. Also, methods exist for jointly estimating RATF and PSDs, e.g. [12, 13], where RATFs and PSDs are estimated in ML sense using a dictionary of candidate RATFs. Other methods are based on speech presence probabilities (SPP) [14–16] to update target and noise statistics. Yet, these methods do not explicitly and jointly estimate the PSDs and RATFs.

In [17–19], Anderson et. al. used results from [20] to derive ML estimates of parameters in signal models that resemble those used in the multi-microphone audio processing community, e.g. [4, 6, 12, 13, 21]. Based on these historical results, it may be argued that recently proposed methods, e.g. the covariance whitening (CW) method for RATF estimation [8, 10, 21, 22] is in fact ML optimal. Moreover, applying the results from [17–19] in a beamforming context, enables derivation of a joint ML estimator (MLE) of the PSDs and RATFs. The fact, which seems largely unknown in the audio processing community, that this joint estimator is ML optimal allows its use in statistical procedures, e.g. likelihood-ratios for voice activity detection.

In this paper, we first derive the joint MLEs of the PSDs and RATFs based on [17–19]. Secondly, we propose a basic VAD framework based on likelihood ratios from the joint MLEs. The joint MLEs and proposed VAD are compared in simulation experiments to competing methods. We see that in realistic situations, where limited a priori information is available on the RATFs or at higher SNRs, the proposed method can outperform the competing methods.

2 Signal Model

The noisy acoustic signal is picked up by M microphones with arbitrary array geometry. The signal is transformed into the time-frequency (TF) domain using the short-time Fourier transform (STFT) with frequency index k and frame index l. The signal model of the noisy signal, $\boldsymbol{x}(k,l) \in \mathbb{C}^M$, may be expressed as [4, 13, 15]

$$\boldsymbol{x}(k,l) = s(k,l)\boldsymbol{d}(k,l) + \boldsymbol{v}(k,l), \tag{C.1}$$

where $s(k,l)$ is the target speech at a pre-selected reference microphone, $\boldsymbol{d}(k,l)$ is the RATF vector with a value of 1 at the reference microphone index [3, 4], and $\boldsymbol{v}(k,l)$ is the noise vector. We assume that $s(k,l)$ and $\boldsymbol{v}(k,l)$ are uncorrelated such that the noisy cross-power spectral density (CPSD) matrix, $\mathbf{C}_x(k,l) \triangleq \mathbb{E}\{\boldsymbol{x}(k,l)\boldsymbol{x}^H(k,l)\}$, is given as

$$\mathbf{C}_x(k,l) = \phi_s(k,l)\boldsymbol{d}(k,l)\boldsymbol{d}^H(k,l) + \mathbf{C}_v(k,l), \tag{C.2}$$

where H is the Hermitian transpose, $\phi_s(k,l)$ is the target PSD at the reference microphone and $\mathbf{C}_v(k,l)$ is the noise CPSD matrix. We assume that the noise CPSD matrix has the form $\mathbf{C}_v(k,l) = \phi_v(k,l)\mathbf{\Gamma}_v(k,l_0)$ during speech presence [7, 12], where l_0 is the most recent speech absent frame index and $\mathbf{\Gamma}_v(k,l_0)$ is the noise CPSD matrix updated during speech absence and normalized with respect to the reference microphone. We model $\boldsymbol{x}(k,l)$ as a circularly symmetric complex Gaussian distributed random vector [23]. Given N observations of $\boldsymbol{x}(k,l)$, we define the sample noisy CPSD matrix for each TF tile as $\mathbf{R}(k,l) = \frac{1}{N}\boldsymbol{X}(k,l)\boldsymbol{X}^H(k,l)$ where $\boldsymbol{X}(k,l) = [\boldsymbol{x}(k,l-N+1), ..., \boldsymbol{x}(k,l)]$ is a data matrix of N noisy observations. The likelihood, parameterized by ϕ_s,

ϕ_v, and $\boldsymbol{d}$, is given as

$$f(\mathbf{X};\phi_s,\phi_v,\boldsymbol{d}) = \frac{\exp\left(-N\mathrm{tr}\left(\mathbf{C}_x^{-1}(\phi_s,\phi_v,\boldsymbol{d})\mathbf{R}\right)\right)}{\pi^{MN}|\mathbf{C}_x(\phi_s,\phi_v,\boldsymbol{d})|^N}, \tag{C.3}$$

where $\mathrm{tr}(\cdot)$ and $|\cdot|$ denote the trace and determinant operators, and the frequency and frame indices are omitted for brevity.

3 Joint Maximum Likelihood Estimation

In this section, we derive the joint MLEs of ϕ_s, ϕ_v, and $\boldsymbol{d}$ in a beamforming context. The derivation is based on the results from [17–19].

3.1 Whitening Transform and Eigenvalue Decomposition

The MLEs of ϕ_s, ϕ_v, and $\boldsymbol{d}$, rely on a whitening transform of $\boldsymbol{x}(k,l)$. First, consider the inverse matrix square-root of $\mathbf{\Gamma}_v$

$$\mathbf{\Gamma}_v^{-\frac{1}{2}} = \mathbf{U}_v\mathbf{D}_v^{-\frac{1}{2}}\mathbf{U}_v^H \text{ and } \mathbf{\Gamma}_v^{\frac{1}{2}} = \mathbf{U}_v\mathbf{D}_v^{\frac{1}{2}}\mathbf{U}_v^H, \tag{C.4}$$

where $\mathbf{\Gamma}_v^{\frac{1}{2}}$ and $\mathbf{\Gamma}_v^{-\frac{1}{2}}$ are Hermitian matrices, $\mathbf{U}_v$ is a unitary matrix with eigenvectors of $\mathbf{\Gamma}_v$ as columns, and $\mathbf{D}_v$ is a diagonal matrix with real-valued and non-negative eigenvalues of $\mathbf{\Gamma}_v$. We define $\mathbf{D}_v^{\frac{1}{2}}$ as the element-wise non-negative square-root of $\mathbf{D}_v$ and $\mathbf{D}_v^{-\frac{1}{2}}$ as the inverse of $\mathbf{D}_v^{\frac{1}{2}}$. Let $\tilde{\mathbf{C}}_x$ and $\tilde{\mathbf{R}}$ be whitened CPSD matrices defined as

$$\tilde{\mathbf{C}}_x = \mathbf{\Gamma}_v^{-\frac{1}{2}}\mathbf{C}_x\mathbf{\Gamma}_v^{-\frac{H}{2}} \text{ and } \tilde{\mathbf{R}} = \mathbf{\Gamma}_v^{-\frac{1}{2}}\mathbf{R}\mathbf{\Gamma}_v^{-\frac{H}{2}}, \tag{C.5}$$

$$\mathbf{C}_x = \mathbf{\Gamma}_v^{\frac{1}{2}}\tilde{\mathbf{C}}_x\mathbf{\Gamma}_v^{\frac{H}{2}} \text{ and } \mathbf{R} = \mathbf{\Gamma}_v^{\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{\Gamma}_v^{\frac{H}{2}}. \tag{C.6}$$

In the whitened domain, it follows that

$$\tilde{\mathbf{C}}_x = \phi_s\tilde{\boldsymbol{d}}\tilde{\boldsymbol{d}}^H + \phi_v\mathbf{I}, \tag{C.7}$$

where $\mathbf{I}$ is the identity matrix and $\tilde{\boldsymbol{d}} \triangleq \mathbf{\Gamma}_v^{-\frac{1}{2}}\boldsymbol{d}$. The eigenvalue decomposition (EVD) of $\tilde{\mathbf{C}}_x$ is

$$\tilde{\mathbf{C}}_x = \tilde{\mathbf{U}}_x\tilde{\mathbf{D}}_x\tilde{\mathbf{U}}_x^H = \tilde{\mathbf{U}}_s\tilde{\mathbf{D}}_s\tilde{\mathbf{U}}_s^H + \tilde{\mathbf{U}}_v\tilde{\mathbf{D}}_v\tilde{\mathbf{U}}_v^H, \tag{C.8}$$

where $\tilde{\mathbf{U}}_s\tilde{\mathbf{D}}_s\tilde{\mathbf{U}}_s^H = \phi_s\tilde{\boldsymbol{d}}\tilde{\boldsymbol{d}}^H$. Note that since $\phi_s\tilde{\boldsymbol{d}}\tilde{\boldsymbol{d}}^H$ is a positive semi-definite rank-1 matrix, $\phi_s\tilde{\boldsymbol{d}}\tilde{\boldsymbol{d}}^H$ has one positive eigenvalue, $\tilde{\lambda}_s$, such that

$\tilde{\mathbf{D}}_s = \text{diag}(\tilde{\lambda}_s, 0, .., 0)$. Let $\tilde{\boldsymbol{u}}_{s,1}$ be the unit-norm eigenvector corresponding to the largest eigenvalue $\tilde{\lambda}_s$. Then

$$\tilde{\lambda}_s \tilde{\boldsymbol{u}}_{s,1} \tilde{\boldsymbol{u}}_{s,1}^H = \phi_s \tilde{\boldsymbol{d}} \tilde{\boldsymbol{d}}^H, \tag{C.9}$$

and because $||\tilde{\boldsymbol{u}}_{s,1}|| = 1$, it follows that $\tilde{\lambda}_s = \phi_s ||\tilde{\boldsymbol{d}}||^2$. As the whitened noise CPSD matrix is $\phi_v \mathbf{I}$, see (C.7), its eigenvalues are all ϕ_v, i.e., $\tilde{\mathbf{D}}_v = \phi_v \mathbf{I}$ and $\tilde{\mathbf{U}}_v$ can be chosen as any unitary matrix. Hence, we may express

$$\tilde{\mathbf{U}}_x \tilde{\mathbf{D}}_x \tilde{\mathbf{U}}_x^H = \tilde{\mathbf{U}}_s \left(\tilde{\mathbf{D}}_s + \tilde{\mathbf{D}}_v \right) \tilde{\mathbf{U}}_s^H. \tag{C.10}$$

Let the EVD of $\tilde{\mathbf{R}}$ be $\tilde{\mathbf{R}} = \tilde{\mathbf{U}}_R \tilde{\mathbf{D}}_R \tilde{\mathbf{U}}_R^H$ and the unitary matrix

$$\mathbf{Q} \triangleq \tilde{\mathbf{U}}_x^H \tilde{\mathbf{U}}_R. \tag{C.11}$$

We now substitute (C.6) and (C.11) into (C.3) to find

$$f(\boldsymbol{X}; \phi_s, \phi_v, \boldsymbol{d}) = \frac{\exp\left(-N \text{tr}\left(\tilde{\mathbf{C}}_x^{-1} \tilde{\mathbf{R}}\right)\right)}{\pi^{MN} |\mathbf{\Gamma}_v^{\frac{1}{2}} \tilde{\mathbf{C}}_x \mathbf{\Gamma}_v^{\frac{H}{2}}|^N} \tag{C.12}$$

$$f(\boldsymbol{X}; \phi_s, \phi_v, \boldsymbol{d}) = \frac{\exp\left(-N \text{tr}\left(\tilde{\mathbf{D}}_x^{-1} \mathbf{Q} \tilde{\mathbf{D}}_R \mathbf{Q}^H\right)\right)}{\pi^{MN} |\mathbf{\Gamma}_v^{\frac{1}{2}} \tilde{\mathbf{U}}_x \tilde{\mathbf{D}}_x \tilde{\mathbf{U}}_x^H \mathbf{\Gamma}_v^{\frac{H}{2}}|^N},$$

where $\tilde{\mathbf{D}}_x \triangleq \tilde{\mathbf{D}}_x(\phi_s, \phi_v)$ is a function of ϕ_s and ϕ_v, and $\mathbf{Q} \triangleq \mathbf{Q}(\boldsymbol{d})$ is a function of $\boldsymbol{d}$ only.

3.2 Maximum Likelihood Estimation of the PSDs and RATFs

Let $\mathcal{L}(\phi_s, \phi_v, \boldsymbol{d}) \triangleq \log f(\boldsymbol{X}; \phi_s, \phi_v, \boldsymbol{d})$ denote the log-likelihood. To derive the joint MLEs of ϕ_s, ϕ_v, and $\boldsymbol{d}$, we maximize $\mathcal{L}$ with respect to ϕ_s, ϕ_v, and $\boldsymbol{d}$

$$\hat{\phi}_s, \hat{\phi}_v, \hat{\boldsymbol{d}} = \arg \max_{\phi_s, \phi_v, \boldsymbol{d}} \mathcal{L}(\phi_s, \phi_v, \boldsymbol{d}). \tag{C.13}$$

Using (C.12), it can be shown that

$$\mathcal{L}(\phi_s, \phi_v, \boldsymbol{d}) \propto -\log |\tilde{\mathbf{D}}_x| - \text{tr}\left(\tilde{\mathbf{D}}_x^{-1} \mathbf{Q} \tilde{\mathbf{D}}_R \mathbf{Q}^H\right). \tag{C.14}$$

To find the MLE of $\boldsymbol{d}$, we maximize $\mathcal{L}(\phi_s, \phi_v, \boldsymbol{d})$ with respect to $\mathbf{Q}(\boldsymbol{d})$. Since $\mathbf{Q}$ is unitary by definition, cf. (C.11), the maximization must be performed over the set $\mathcal{Z}$ of unitary matrices, i.e.,

$$\hat{\mathbf{Q}} = \arg \max_{\mathbf{Q} \in \mathcal{Z}} -\text{tr}\left(\tilde{\mathbf{D}}_x^{-1} \mathbf{Q} \tilde{\mathbf{D}}_R \mathbf{Q}^H\right). \tag{C.15}$$

It has been found in [17–19], that the MLE of $\mathbf{Q}$ is $\hat{\mathbf{Q}} = \mathbf{I}$. This implies that $\tilde{\mathbf{U}}_x^H \tilde{\mathbf{U}}_R = \mathbf{I}$, such that the MLE of $\tilde{\mathbf{U}}_x$ is $\hat{\tilde{\mathbf{U}}}_x = \tilde{\mathbf{U}}_R$ and hence the MLE of $\tilde{\boldsymbol{u}}_{s,1}$ is $\hat{\tilde{\boldsymbol{u}}}_{s,1} = \tilde{\boldsymbol{u}}_{R,1}$, where $\tilde{\boldsymbol{u}}_{R,1}$ is the eigenvector to the corresponding largest eigenvalue of $\tilde{\mathbf{R}}$. By dewhitening $\hat{\tilde{\boldsymbol{u}}}_{s,1}$, the MLE of $\boldsymbol{d}$, i.e. $\hat{\boldsymbol{d}}$, can therefore be found as

$$\hat{\boldsymbol{d}} = \frac{\mathbf{\Gamma}_v^{\frac{1}{2}} \tilde{\boldsymbol{u}}_{R,1}}{\boldsymbol{e}^H \mathbf{\Gamma}_v^{\frac{1}{2}} \tilde{\boldsymbol{u}}_{R,1}}, \tag{C.16}$$

where $\boldsymbol{e}$ is a unit vector with a value of 1 at the reference microphone index. Thus, the CW method [8] is also the MLE of $\boldsymbol{d}$. Note that the MLE of $\hat{\boldsymbol{d}}$ in the whitened domain is $\hat{\tilde{\boldsymbol{d}}} = \tilde{\boldsymbol{u}}_{R,1} \cdot (\boldsymbol{e}^H \mathbf{\Gamma}_v^{\frac{1}{2}} \tilde{\boldsymbol{u}}_{R,1})^{-1}$. Concentrating the likelihood in (C.14) by inserting $\hat{\mathbf{Q}} = \mathbf{I}$ gives

$$\begin{aligned} \mathcal{L}(\phi_s, \phi_v | \hat{\boldsymbol{d}}) &\propto -\log |\tilde{\mathbf{D}}_x| - \operatorname{tr}\left(\tilde{\mathbf{D}}_x^{-1} \tilde{\mathbf{D}}_R\right) \\ &= -\sum_{m=1}^{M} \log \tilde{\lambda}_{x,m} - \sum_{m=1}^{M} \frac{\tilde{\lambda}_{R,m}}{\tilde{\lambda}_{x,m}}, \end{aligned} \tag{C.17}$$

where $\tilde{\lambda}_{R,1} \geq \tilde{\lambda}_{R,2} \geq, ..., \geq \tilde{\lambda}_{R,M}$ are the diagonal elements of $\tilde{\mathbf{D}}_R$, i.e., the eigenvalues of $\tilde{\mathbf{R}}$. From (C.10) we conclude that $\tilde{\lambda}_{x,1} = \tilde{\lambda}_s + \phi_v$ and $\tilde{\lambda}_{x,m} = \phi_v$ for $m = 2, .., M$. Using $\tilde{\lambda}_s = \phi_s ||\tilde{\boldsymbol{d}}||^2$, and expanding $\tilde{\lambda}_{x,m}$ in (C.17) yield

$$\begin{aligned} \mathcal{L}(\phi_s, \phi_v | \hat{\boldsymbol{d}}) \propto &-\log\left(\phi_s ||\hat{\tilde{\boldsymbol{d}}}||^2 + \phi_v\right) - (M-1)\log \phi_v \\ &- \phi_v^{-1} \sum_{m=2}^{M} \tilde{\lambda}_{R,m} - \frac{\tilde{\lambda}_{R,1}}{\phi_s ||\hat{\tilde{\boldsymbol{d}}}||^2 + \phi_v}. \end{aligned} \tag{C.18}$$

Taking the partial derivative with respect to ϕ_s gives

$$\frac{\partial \mathcal{L}(\phi_s, \phi_v | \hat{\boldsymbol{d}})}{\partial \phi_s} = -\frac{1}{\phi_s ||\hat{\tilde{\boldsymbol{d}}}||^2 + \phi_v} + \frac{\tilde{\lambda}_{R,1}}{(\phi_s ||\hat{\tilde{\boldsymbol{d}}}||^2 + \phi_v)^2}, \tag{C.19}$$

and setting equal to zero, the MLE of ϕ_s is found as

$$\hat{\phi}_s = (\tilde{\lambda}_{R,1} - \phi_v) ||\hat{\tilde{\boldsymbol{d}}}||^{-2}, \tag{C.20}$$

where $||\hat{\tilde{\boldsymbol{d}}}||^{-2} = (\boldsymbol{e}^H \mathbf{\Gamma}_v^{\frac{1}{2}} \tilde{\boldsymbol{u}}_{R,1})^2$. Inserting (C.20) into (C.18) and taking the partial derivative with respect to ϕ_v yields

$$\frac{\partial \mathcal{L}(\phi_v | \hat{\phi}_s, \hat{\boldsymbol{d}})}{\partial \phi_v} = -(M-1)\phi_v^{-1} + \phi_v^{-2} \sum_{m=2}^{M} \tilde{\lambda}_{R,m}, \tag{C.21}$$

and solving for ϕ_v, the MLE of ϕ_v is

$$\hat{\phi}_v = \frac{1}{M-1} \sum_{m=2}^{M} \tilde{\lambda}_{R,m}. \tag{C.22}$$

Hence, MLEs of $\boldsymbol{d}$, ϕ_s, and ϕ_v are given by (C.16), (C.20), and (C.22).

4 Voice Activity Detection Based on Concentrated Log-Likelihood ratios

In noise reduction systems, voice activity detectors (VADs) are, e.g. used to detect noise dominant TF tiles to update the noise structure matrix, $\mathbf{\Gamma}_v$. Using the theoretical results from Sec. 3, we derive a simple log-likelihood ratio (LLR) based VAD. First, we define two hypotheses regarding voice activity in a given TF-tile where

$$\begin{aligned} &H_0 : \text{Speech absence} \\ &H_1 : \text{Speech presence.} \end{aligned} \tag{C.23}$$

Given the noisy observations $\boldsymbol{X}$, a speech presence probability (SPP) $P(H_1|\boldsymbol{X})$, can be derived and used to form a binary VAD, i.e.,

$$\text{VAD} = \begin{cases} 0, & \text{if } P(H_1|\boldsymbol{X}) < \gamma \\ 1, & \text{otherwise,} \end{cases} \tag{C.24}$$

where γ is a pre-selected threshold. Let $P(H_0)$ denote the prior probability for speech absence, and let $\beta \triangleq \frac{P(H_0)}{1-P(H_0)}$. Then, it can be shown that the LLR is

$$\text{LLR} = \log f(\boldsymbol{X}|H_0) - \log f(\boldsymbol{X}|H_1) + \log \beta. \tag{C.25}$$

Using Bayes theorem, $P(H_1|\boldsymbol{X})$ may be expressed as

$$P(H_1|\boldsymbol{X}) = \frac{1}{1+\exp(\text{LLR})}. \tag{C.26}$$

Under speech absence (H_0), the noisy CPSD matrix is $\mathbf{C}_x = \phi_v \mathbf{\Gamma}_v$. It can be shown that the likelihood under H_0 is

$$\begin{aligned} f(\boldsymbol{X}|H_0;\phi_v) &= \frac{\exp\left(-N\phi_v^{-1}\text{tr}\left(\mathbf{R}\mathbf{\Gamma}_v^{-1}\right)\right)}{\pi^{MN}|\phi_v\mathbf{\Gamma}_v|^N} \\ &\propto \phi_v^{-M}\exp\left(-\phi_v^{-1}\text{tr}\left(\tilde{\mathbf{R}}\right)\right), \end{aligned} \tag{C.27}$$

where ϕ_v^{-M} comes from the scalar product property of the determinant. $\text{tr}(\tilde{\mathbf{R}}) = \sum_m \tilde{\lambda}_{R,m}$ is the sum of eigenvalues of $\tilde{\mathbf{R}}$. Taking the partial derivative of the log-likelihood of (C.27) with respect to ϕ_v and solving for ϕ_v, the MLE of ϕ_v under H_0 is

$$\hat{\phi}_{v|H_0} = \frac{1}{M}\sum_{m=1}^{M}\tilde{\lambda}_{R,m}. \tag{C.28}$$

Inserting (C.28) into (C.27), the concentrated log-likelihood under H_0 is

$$\begin{aligned} &\log f(\boldsymbol{X}, \hat{\phi}_{v|H_0}|H_0) = \\ &- MN\log\pi - MN\log\hat{\phi}_{v|H_0} - N\log|\boldsymbol{\Gamma}_v| - MN. \end{aligned} \tag{C.29}$$

The concentrated log-likelihood under H_1 is found by inserting (C.16), (C.20), and (C.22) into the log-likelihood under speech presence (D.15)

$$\begin{aligned} &\log f(\boldsymbol{X}, \hat{\boldsymbol{d}}, \hat{\phi}_{s|H_1}, \hat{\phi}_{v|H_1}|H_1) = \\ &\quad - MN\log\pi - N\log|\boldsymbol{\Gamma}_v| - N\log|\hat{\tilde{\mathbf{D}}}_{x|H_1}| - MN. \end{aligned} \tag{C.30}$$

where $\hat{\tilde{\mathbf{D}}}_{x|H_1}$ is a diagonal matrix whose elements are eigenvalues of the concentrated noisy CPSD matrix $\hat{\tilde{\mathbf{C}}}_{x|H_1}$ under H_1. Inserting (C.29) and (C.30) into the log-likelihood ratio in (C.25) yields

$$\mathrm{LLR} = N\log|\hat{\tilde{\mathbf{D}}}_{x|H_1}| - MN\log\hat{\phi}_{v|H_0} + \log\beta. \tag{C.31}$$

The determinant of $\hat{\tilde{\mathbf{D}}}_{x|H_1}$ is the product of its eigenvalues, i.e.,

$$\begin{aligned} \log|\hat{\tilde{\mathbf{D}}}_{x|H_1}| &= \sum_{m=1}^{M} \log\hat{\tilde{\lambda}}_{x,m|H_1} \\ &= \log(\hat{\phi}_{s|H_1}||\hat{\tilde{\boldsymbol{d}}}||^2 + \hat{\phi}_{v|H_1}) + (M-1)\log\hat{\phi}_{v|H_1} \\ &= \log\tilde{\lambda}_{R,1} + (M-1)\log\hat{\phi}_{v|H_1}. \end{aligned} \tag{C.32}$$

Inserting (C.32) into (C.31), the concentrated LLR is

$$\mathrm{LLR} = N\left[\log\tilde{\lambda}_{R,1} + (M-1)\log\hat{\phi}_{v|H_1} - M\log\hat{\phi}_{v|H_0}\right] + \log\beta, \tag{C.33}$$

where $\hat{\phi}_{v|H_1}$ and $\hat{\phi}_{v|H_0}$ are given by (C.22) and (C.28), respectively. We see that the LLR and consequently the VAD in (C.24) can be expressed as a function of eigenvalues, $\tilde{\lambda}_{R,m}$, of $\tilde{\mathbf{R}}$ only.

5 Performance Evaluation

In order to demonstrate the efficiency of the joint RATF and PSD estimators as well as the proposed VAD, we use them in a simple beamforming system for hearing aids (HAs) and compare it to competing methods that solve similar problems.

5.1 Beamforming Methods

5.1.1 Proposed joint maximum likelihood method

The MLEs of the PSDs, and RATF vector, i.e. (C.16), (C.20), and (C.22), are used to implement an MWF beamformer given as [2]

$$w_{\text{MWF}} = \frac{\mathbf{\Gamma}_v^{-1}\boldsymbol{d}}{\boldsymbol{d}^H\mathbf{\Gamma}_v^{-1}\boldsymbol{d}} \cdot \frac{\phi_s}{\phi_s + \phi_v\left(\boldsymbol{d}^H\mathbf{\Gamma}_v^{-1}\boldsymbol{d}\right)^{-1}}. \tag{C.34}$$

The proposed VAD using (C.24), (C.26), and (C.33), is used to detect speech dominated TF tiles, where the MWF beamformer is applied. In noise dominant TF tiles, the output of the beamformer is the unprocessed signal at the reference microphone attenuated by 20 dB. We refer to the proposed beamforming scheme as J-ML.

5.1.2 Dictionary-based maximum likelihood method

The dictionary-based maximum likelihood (D-ML) method proposed in [12, 24] is similar to J-ML as it uses MLEs of the target and noise PSDs and the RATF vector. However, the MLE of the RATF vector is based on maximizing the likelihood function with an exhaustive search of a predefined dictionary of RATF vectors. Hence, the MLEs of the target and noise PSDs are therefore not identical to those derived in Sec. 3. The D-ML can be advantageous as it considers only physically plausible RATF vectors, but may suffer performance degradation because it selects among a finite and potenially mismatched set of RATF vectors. Hence, we have included two different implementations of the D-ML to show the effect of the accuracy of RATF dictionaries. In the first implementation, an accurate dictionary of RATF vectors is available (referred to as D-ML) where the RATF vectors in the dictionary are measured on the actual HA users. In practice, however, it is rare that such person dependent RATFs would be available. Instead, a RATF dictionary measured on an average head would be used. In the second implementation, we therefore simulate this more realistic non-person-dependent setup referred to as D-ML-NP.

The D-ML methods are used in combination with an MWF beamformer and a VAD obtained from speech presence probabilities estimated from the likelihood ratios derived in [12]. As for J-ML, the MWF beamformer is applied in speech dominant TF tiles while a fixed attenuation of 20 dB is applied in noise dominant TF-tiles.

5.1.3 Oracle multichannel Wiener filter

For completeness, we also evaluate the performance of an oracle MWF where the true PSDs and RATF are known.

5.2 Experimental Setup

We conducted two experiments with different microphone arrays. A monaural HA placed on the left ear of the user was used with $M = 2$ microphones. The microphone distance is approximately 1.3 cm. In the second experiment, we used $M = 4$ microphones and simulated a binaural beamformer. We assume instantaneous and error-free signal exchange between the HAs.

Acoustic transfer functions (ATFs) from loudspeakers to each microphone were available to simulate acoustic scenes [25]. The loudspeakers were placed at a distance of approximately 1.9 m to each microphone and ATFs were measured at an angular resolution of 7.5° on the horizontal axis. The ATFs are available for the D-ML methods in the form of RATFs at an angular resolution of 30°.

An acoustic scene was constructed to consist of one target source in background noise. The target is speech obtained from the TIMIT database [26], and the location of the target was randomized uniformly over a circle with a resolution of 7.5 degrees. The noise was recorded in a busy canteen with a spherical microphone. The noise recordings were used to reproduce the identical noise field from the canteen at the HA microphones in the simulation. Additionally, the noise recordings were used to initialize $\mathbf{\Gamma}_v$ before the simulation. Experiments with other noise types, e.g. noise measured in a bar, were also conducted and showed essentially similar results and were therefore not included in the paper.

Each performance score in our experiments is an average over 20 different realizations of acoustic scenes. The duration of an acoustic scene is 5 seconds and a different target speech source, target position, and noise source realization were used for each scene.

The sampling frequency is 16 kHz and a square-root Hann window with a length of 256 samples was used for analysis and synthesis in the STFT with an overlap of 128 samples. The number of observations used to form the likelihood function is $N = 10$ and the prior probabilities used to form the likelihood ratios for J-ML and D-ML were adjusted to make both methods perform as good as possible.

5.3 Simulation Results

The beamforming performance is reported in Fig. C.1 as extended short-time objective intelligibility (ESTOI) [27] and perceptual evaluation of speech

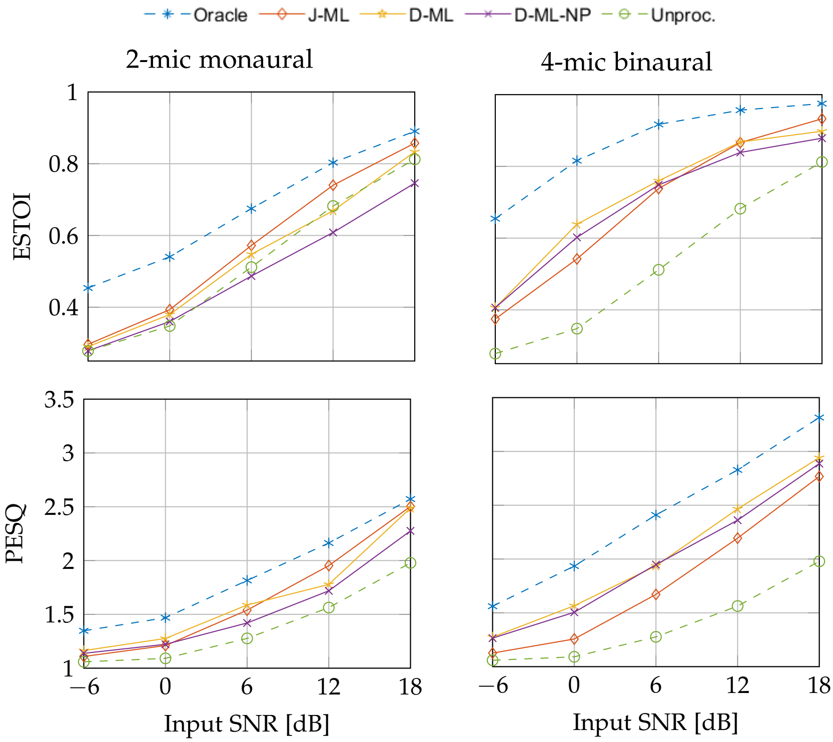


Fig. C.1: Beamforming performance in acoustic scenes with a single-target source and canteen noise.

quality (PESQ) [28] scores.

For the $M = 2$ experiment, the proposed method, J-ML, performs well at higher SNRs, while performing on par with the D-ML methods at lower SNRs. In particular, comparing J-ML to the more realistic D-ML-NP method, reveals the disadvantage of relying on a dictionary of RATF vectors as the SNR increases.

For the $M = 4$ experiment, results suggest better performance with J-ML at higher SNRs, whereas D-ML and D-ML-NP tend to work better at low SNRs. The results indicate that the prior knowledge of the RATF vectors provided by the dictionary, improves robustness at low SNRs. Conversely, at higher SNRs, we see that the dictionary of the D-MLs becomes a restriction, because the RATF vector must be selected as an element of the RATF dictionary, which generally does not reflect the actual position of the target source.

6 Conclusion

In this paper, we derived joint maximum likelihood estimators (MLEs) of the target and noise power spectral densities and relative acoustic transfer functions (RATF). Elements of the derived joint MLEs are based on historical results that appear largely unknown to the audio processing community. Furthermore, we proposed a basic voice activity detector (VAD) based on speech presence probabilities estimated by concentrating likelihood ratios with the joint MLEs.

We apply the joint MLEs and VAD in a multichannel Wiener filter beamformer and compare it to state-of-the-art beamformers whose parameters were estimated using RATF dictionaries. Results indicate that the proposed method can outperform the dictionary-based methods in situations with high SNRs and realistic situations where mismatches are present in the RATF dictionary.

References

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*, ser. Springer topics in signal processing. Berlin: Springer, 2008.

[2] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, 2001.

[3] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[4] S. Gannot, E. Vincent, S. M.-Golan, and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE TASLP*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[5] A. Kuklasinski, S. Doclo, T. Gerkmann, S. Holdt Jensen, and J. Jensen, "Multichannel PSD estimators for speech dereverberation - A theoretical and experimental comparison," in *2015 IEEE ICASSP*. IEEE, Apr. 2015, pp. 91–95.

[6] S. Braun, A. Kuklasinski, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and Comparison of Late Reverberation Power Spectral Density Estimators," *IEEE TASLP*, vol. 26, no. 6, pp. 1056–1071, Jun. 2018.

[7] J. Jensen and M. S. Pedersen, "Analysis of beamformer directed single-channel noise reduction system for hearing aid applications," in *2015 IEEE ICASSP*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 5728–5732.

[8] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *2015 IEEE ICASSP*. IEEE, Apr. 2015.

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

[10] S. Markovich, S. Gannot, and I. Cohen, "Multichannel Eigenspace Beamforming in a Reverberant Noisy Environment With Multiple Interfering Speech Signals," *IEEE TASLP*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[11] M. Taseska and E. A. P. Habets, "Joint maximum likelihood estimation of power spectral densities and relative acoustic transfer functions for acoustic beamforming," in *2015 23rd EUSIPCO*. IEEE, Aug. 2015.

[12] M. Zohourian, G. Enzner, and R. Martin, "Binaural Speaker Localization Integrated Into an Adaptive Beamformer for Hearing Aids," *IEEE TASLP*, vol. 26, no. 3, pp. 515–528, Mar. 2018.

[13] 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 GlobalSIP*. Ottawa, ON, Canada: IEEE, Nov. 2019, pp. 1–5.

[14] M. Souden, Jingdong Chen, J. Benesty, and S. Affes, "Gaussian Model-Based Multichannel Speech Presence Probability," *IEEE TASLP*, vol. 18, no. 5, pp. 1072–1077, Jul. 2010.

[15] M. Souden, J. Chen, J. Benesty, and S. Affes, "An Integrated Solution for Online Multichannel Noise Tracking and Reduction," *IEEE TASLP*, vol. 19, no. 7, pp. 2159–2169, Sep. 2011.

[16] M. Taseska and E. A. P. Habets, "Non-Stationary Noise PSD Matrix Estimation for Multichannel Blind Speech Extraction," *IEEE TASLP*, 2017.

[17] B. M. Anderson, T. W. Anderson, and I. Olkin, "Maximum likelihood estimators and likelihood ratio criteria in multivariate components of variance," *The Annals of Statistics*, pp. 405–417, 1986.

[18] T. W. Anderson, *An introduction to multivariate statistical analysis*, ser. Wiley series in probability and mathematical statistics. New York, NY: Wiley, 1958, oCLC: 180226.

[19] T. Anderson and I. Olkin, "Maximum-likelihood estimation of the parameters of a multivariate normal distribution," *Linear Algebra and its Applications*, vol. 70, pp. 147–171, Oct. 1985.

[20] J. von Neumann, *Some matrix-inequalities and metrization of matrix space*. New York: Tomsk University Review, 1937, vol. 4.

[21] I. Kodrasi and S. Doclo, "Analysis of Eigenvalue Decomposition-Based Late Reverberation Power Spectral Density Estimation," *IEEE TASLP*, vol. 26, no. 6, pp. 1106–1118, Jun. 2018.

[22] J. Zhang, *Energy-aware noise reduction for wireless acoustic sensor networks*, 2020, oCLC: 8457116491.

[23] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[24] Hao 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, Apr. 1995.

[25] A. Moore, H. J. de, M. Pedersen, D. Brookes, P. Naylor, and J. Jensen, “Personalized signal-independent beamforming for binaural hearing aids,” *Journal of the Acoustical Society of America*, vol. 145, pp. 2971–2981, 2019.

[26] “TIMIT: acoustic-phonetic continuous speech corpus.” Philadelphia, Pa., 1993.

[27] 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, Nov. 2016.

[28] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in *2001 IEEE ICASSP.*, vol. 2. IEEE, 2001, pp. 749–752.

This page intentionally left blank.

Paper D

Multichannel Speech Enhancement with Own Voice-Based Interfering Speech Suppression for Hearing Assistive Devices

Poul Hoang, Zheng-Hua Tan, Jan Mark de Haan, and Jesper Jensen

The paper has been published in
IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 706-720, 2022.

The layout has been revised.

Abstract

Enhancement of a desired speech signal in the presence of competing or interfering speech remains an unsolved problem, as it can be hard to determine which of the speech signals is the one of interest. In this paper, we propose a multichannel noise reduction algorithm which uses the presence of the user's own voice signal, e.g. during conversations with the target speaker, as an asset to efficiently identify interfering speech and noise. Specifically, following the typical speech pattern in natural conversations, the presence of an own voice may indicate the absence of the target speech, hence undesired speech and noise can be identified and estimated during own voice presence.

In contrast to conventional noise reduction systems, the proposed noise reduction systems use the user's own voice to identify interfering speech that otherwise could be confused with the target speech. We demonstrate the performance of the proposed noise reduction systems in a comparison against state-of-the-art noise reduction systems in terms of beamforming performance for hearing assistive devices. The results show that the proposed beamforming scheme in particular outperforms state-of-the-art methods in terms of ESTOI and PESQ in situations with a target speaker and a strong interfering speaker.

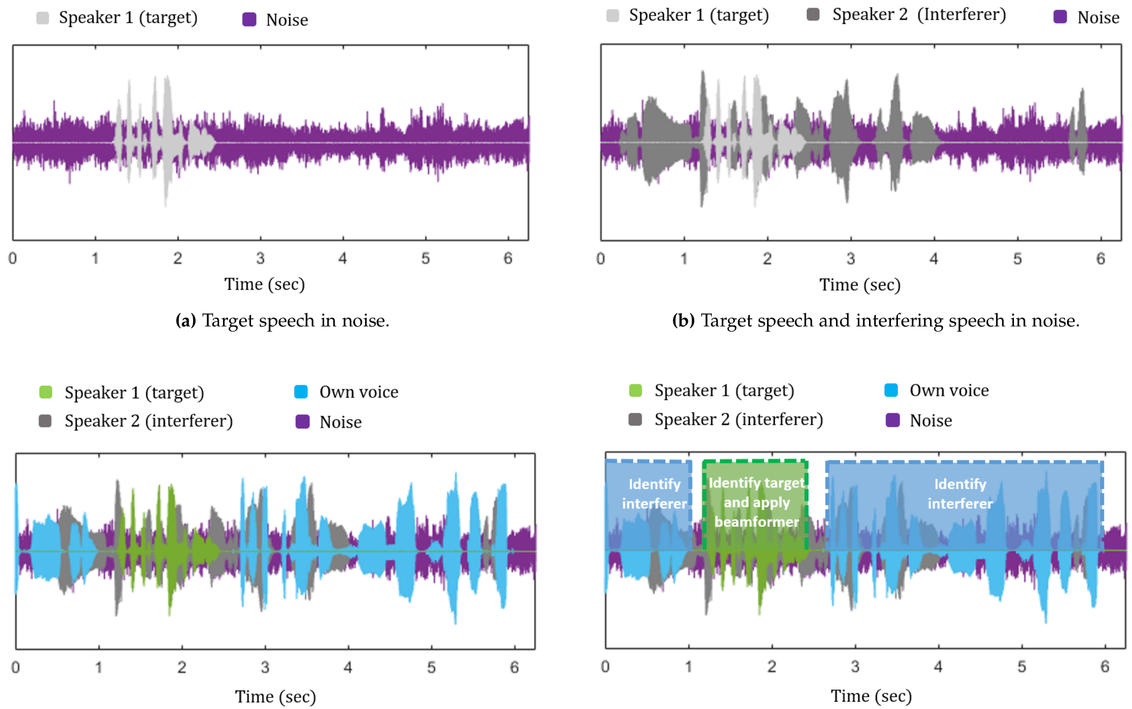
1 Introduction

Spoken language is for most people their primary way of communicating in many social situations. Speech, however, may become challenging to understand, when the acoustic environment becomes increasingly noisy. Especially, when the acoustic environment is contaminated with many competing speakers or interferers, speech intelligibility is often poor.

One of the purposes of hearing assistive devices (HADs), e.g. hearing aids (HAs), is to increase speech intelligibility and quality by reducing the background noise. This is commonly achieved with the use of noise reduction algorithms such as beamformers, when multiple microphones are accessible [1–3]. Examples of well-known beamformers are the minimum-variance distortion-less response (MVDR), the multichannel Wiener filter (MWF) and the linear constrained minimum variance (LCMV) beamformers [2–4]. Implementation of these beamformers is often done in the time-frequency (TF) domain and the parameters required are typically noise statistics, e.g. the noise cross power spectral density (CPSD) matrix [3] and the relative acoustic transfer function (RATF) vector of the target source[4]. These parameters are, however, rarely known in real-world situations and therefore have to be estimated.

One approach to estimate the noise CPSD matrix is to use noise dominant TF tiles to update the noise CPSD matrix, and use the resulting estimate during speech presence, e.g. [5]. Detecting noise dominant TF tiles requires a voice activity detector (VAD) or, more generally, speech presence probabilities (SPPs) estimated from the noisy microphone signals. Multichannel methods for estimating the speech presence probability have been proposed in [6–11]. These methods update the noise CPSD matrix in a soft-decision manner using a multichannel extension of the minima controlled recursive average procedure in [12, 13]. These methods may perform less well if only few noise dominant TF tiles can be identified or if the noise is highly non-stationary during speech dominated TF tiles. To overcome this issue, several methods have been proposed to update the noise statistics using speech dominant TF tiles as well. For example, methods presented in [14–21] are maximum likelihood estimators (MLEs) of the noise CPSD matrix under the assumption that the spatial coherence of the noise field remains fixed during speech presence. As a consequence, these methods may perform less well when the spatial properties of the noise field change during speech presence. An example where this occurs, is when a non-stationary interfering source, e.g. a competing speaker emerges in the noise field. In [22], an MLE of the interference-plus-noise CPSD matrix was proposed to handle situations with strong interfering speech and noise. However, the method requires that the target RATF vector is known in advance.

Accurate target localization and target RATF vector estimation are crucial for beamformers to steer the acoustic beam towards the target speaker [4]. In acoustic scenarios with interfering speakers, target RATF vector estimation can be particularly difficult. The problem of identifying a target speaker amongst a set of interfering speakers and background noise is essentially ill-posed: without any additional information, it is very difficult to single out the target speaker from the set of active speakers. Hence, in order to identify the target speaker, existing methods have applied various prior knowledge. For example, the widely used steered-response methods [3, ch. 8] implicitly rely on the assumption that the target source is closer in distance, and hence more powerful, than other sound sources. These methods identify the target source by directing beamformers to all possible directions, and selecting the beamformer with the highest output power. However, in many practical situations the target speech need not to be loudest, and systems based on this assumption will fail. Other methods rely on prior assumptions of the target location e.g. the methods presented in [23–25]. In HAD applications, the target location is often assumed frontal relative to the user [15, 25, 26]. This assumption is motivated by the observation that for face-to-face conversation, where the HAD-user uses eye-contact and lip-reading, the target source is often located in the frontal half-plane with respect to the user. However, also this assumption is not always valid, e.g., in situations, where the HAD user is unable to look at the target (e.g., when driving a car). Finally, other RATF vector estimation methods, e.g. [10, 27, 28], can perform well in simple



(a) Target speech in noise.

(b) Target speech and interfering speech in noise.

(c) Target speech and own voice in an environment with an interferer and noise.

(d) The proposed systems identify the interferer during own voice presence and identify the target speech during target presence.

Fig. D.1: Fig. D.1a depicts a simple acoustic situation with a target speaker in background noise but without interfering speakers. Fig. D.1b extends the acoustic situation to include an interferer, in addition to the target speech and noise. In Fig. D.1c, an own voice is conversing with the target speech in an acoustic environment with an interferer and noise. Fig. D.1d shows the basic idea of the proposed noise reduction systems. During own voice presence, the interferer is identified and during target presence the target is identified.

situations where the target source is present in background noise, but where no interfering speakers are present, cf. Fig. D.1a[1].

Unfortunately, in more complex acoustic situations, where one or more interfering speakers are simultaneously present, cf. Fig. D.1b, estimation of the noise CPSD matrix and the target RATF vector can be difficult tasks. The presence of interferers can make it difficult to determine the target speaker, particularly when the interferer is voice-like. A voice-like interferer can make voice activity detection difficult as it is hard to distinguish between desired and interfering speech. This can result in interference and noise statistics being captured poorly and degrade the noise reduction performance significantly. Recent proposed methods can potentially help identifying the target speaker by decoding the direction of the user's auditory attention [30] or the user's eye-gaze direction [31, 32] with the use of EEG signals or eye-trackers. However, these methods require the use of additional sensors which may not be available for the speech enhancement system. Other situations that can be particularly difficult for existing noise reduction systems to handle are conversations between the user and a target speaker. The situation is further complicated, if an interfering speaker is present during the conversation between the user and a target speaker, cf. Fig. D.1c. The presence of the own voice signal will leave few instances of noise dominant TF tiles making the SPP-based methods ineffective.

In this paper, we propose a method which solves these problems by using the presence of the user's own voice signal as an asset. Specifically, we use the fact that the presence of own voice signal often indicates the absence of the target signal due to the avoidance of speech overlap between the user and the target speaker[33–35]. Additionally, the absence of own voice may indicate the presence of a target signals.

The proposed method relies on the assumption that any sound source during own voice presence is of no interest to the user, and can hence be regarded as interfering signals. Therefore, statistics related to the interference and noise can be updated during own voice activity as shown in Fig. D.1d. To demonstrate the idea, we consider the situation where only a single interfering speech source may be present. This problem is already very challenging to solve with state-of-the-art methods, as it is difficult to decide which of sound sources is target and which is the interfering speaker. However, the proposed method can in principle be extended to handle multiple interfering speakers such that any speaker during own voice presence is considered undesired. The acoustic situation, we specifically seek to solve in this paper, is the presence of a target speaker, the user's own voice, an interfering speaker, and noise. Such a situation can be regarded as particularly difficult to solve with the current state-of-the-art methods due to the interfering

[1] Speech signals used in the figures are from the speech database in [29]

speaker and the very few instances, where noise dominates the noisy signal. As shown in Fig. D.1d, the proposed systems identify the interferer during own voice, i.e. estimate the interferer RATF vector and use this estimate to support the implementation of a beamforming system during target speech presence. Specifically, the estimated interferer RATF vector from own voice presence is used during own voice absence (presumably target presence) to support the estimation of the interference-plus-noise CPSD matrix and target RATF vector. The estimated interference-plus-noise CPSD matrix and target RATF vector are then used in an MWF beamformer to suppress the interferer and noise.

The paper is structured as follows. In Sec. 2, the signal model of the microphone signals is presented. In Sec. 3, the MLEs of the interference and noise PSDs, and interference and target RATFs, are presented respectively. Sec. 4 presents the simulation setup and evaluates the proposed noise reduction algorithm in simulation experiments. Finally, in Sec. 5, a conclusion of the results is given.

2 Multi-Microphone Signal Model

We consider a HAD with M microphones placed in an arbitrary array geometry. The considered acoustic situation is depicted in Fig. D.2. Each microphone picks up sound from the acoustic environment, and the signals are then sampled into a discrete-time sequence $x_m(n)$ for $m = 1, .., M$. The acoustic scene consists of an own voice signal, $s'_o(n)$, a target signal, $s'_t(n)$, interfering speech signal $s'_q(n)$, and noise denoted as $v(n)$. We assume, for simplicity, the presence of a single interferer per TF tile.

Let $h_{o,m}(n)$, $h_{t,m}(n)$, and $h_{q,m}(n)$ denote the acoustic impulse response (AIR) from the own voice, target, and interferer respectively to the m'th microphone. The signal model of the observed noisy signal is then

$$x_m(n) = \sum_{j \in \{t,o,q\}} s'_j(n) * h_{j,m}(n) + v_m(n), \tag{D.1}$$

where $*$ denotes the linear convolution operator. The proposed noise reduction algorithm is derived and implemented in the TF domain using the short-time Fourier transform (STFT) with window function $\psi(n)$, window size N_{win}, and overlap N_{ov}. The STFT of the noisy signal is [1, 36]

$$x_m(k,l) = \sum_{n=0}^{N_{\text{win}}-1} x_m(n + lN_{\text{ov}})\psi(n)e^{-2\pi ik\frac{n}{N_{\text{win}}}}, \tag{D.2}$$

where $i{=}\sqrt{-1}$, k and l denote the frequency bin and frame index, respectively. We define $\boldsymbol{x}(k,l) = [x_1(k,l), ..., x_M(k,l)]^T$ as an $M \times 1$ complex vector

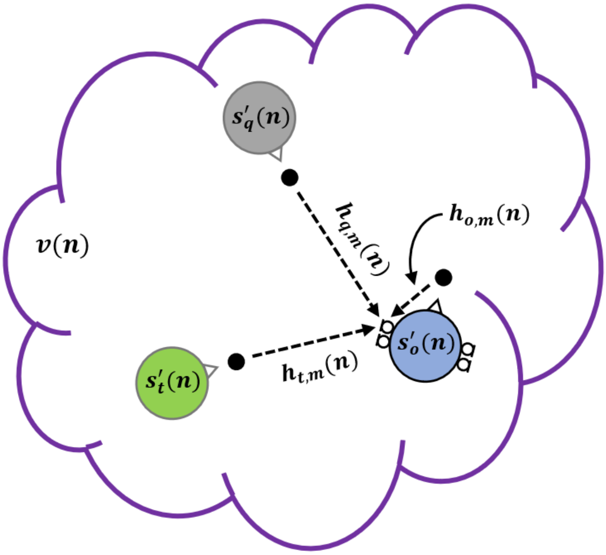


Fig. D.2: Example of an acoustic scene with an own voice $s'_o(n)$, target $s'_t(n)$, interference $s'_q(n)$, and noise $v(n)$ where the microphones are mounted on the user's head. The acoustic impulse response from the j'th source ($j \in \{t, o, q\}$) to the m'th microphone is denoted as $h_{j,m}(n)$.

containing the noisy TF observations for all M microphones. In the TF domain, the signal model becomes

$$\boldsymbol{x}(k,l) = \sum_{j\in\{t,o,q\}} s'_j(k,l)\boldsymbol{h}_j(k,l) + \boldsymbol{v}(k,l), \tag{D.3}$$

where $\boldsymbol{h}_j(k,l)$ and $\boldsymbol{v}(k,l)$ are the stacked acoustic transfer functions (ATFs) and noise, respectively. We assume that the AIRs are shorter than the STFT analysis window $\psi(n)$ [37]. The signals $s'_j(k,l)$ for $j \in \{t, o, q\}$ denote the speech signals of the target, own voice, and interferer at their respective locations. Let m^* denote a pre-selected reference microphone, then we may normalize the ATFs with respect to the reference microphone such that

$$\boldsymbol{x}(k,l) = \sum_{j\in\{t,o,q\}} s_j(k,l)\boldsymbol{d}_j(k,l) + \boldsymbol{v}(k,l), \tag{D.4}$$

where

$$\boldsymbol{d}_j(k,l) = \left[\frac{h_{1,j}(k,l)}{h_{m^*,j}(k,l)}, \ldots, \frac{h_{M,j}(k,l)}{h_{m^*,j}(k,l)}\right]^T, \tag{D.5}$$

is the RATF vector [38] and $s_j(k,l)$ is the j'th signal as captured at the reference microphone.

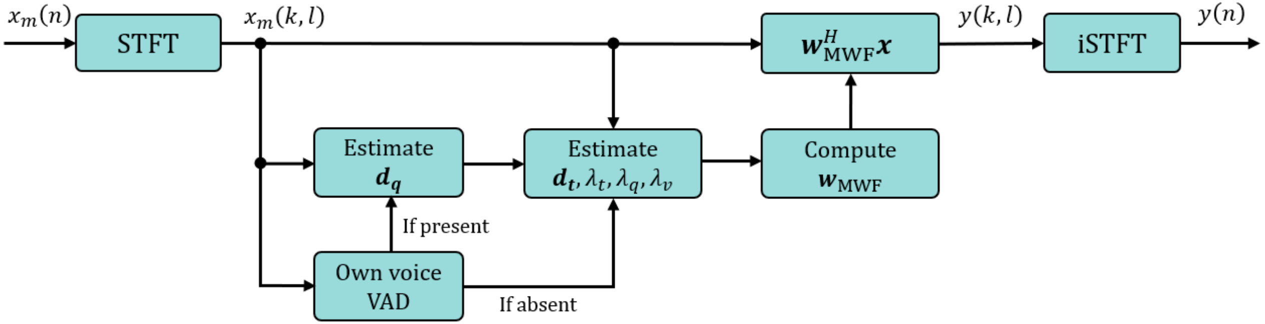

$x_m(n)$
STFT
$x_m(k,l)$
Estimate $\boldsymbol{d_q}$
Own voice VAD
If present
If absent
Estimate $\boldsymbol{d_t}, \lambda_t, \lambda_q, \lambda_v$
Compute $\boldsymbol{w}_{\text{MWF}}$
$\boldsymbol{w}^H_{\text{MWF}}\boldsymbol{x}$
$y(k,l)$
iSTFT
$y(n)$

Fig. D.3: Overview of the proposed noise reduction systems, where the interferer is identified during own voice presence, and the target is identified during own voice absence.

We assume that the presence of the own voice signal and the target signal are mutually exclusive. This assumption is based on the conversational model in [34], where interlocutors in conversations avoid speech overlaps and pauses. This assumption is supported by results found in human experiments in [33, 35, 39]. These results suggest that interlocutors during conversations avoid speech overlap and pauses in noisy environment. Hence, the signal model may be divided to reflect two situations, namely, 1) when own voice is present and target is absent,

$$\boldsymbol{x}(k,l) = s_o(k,l)\boldsymbol{d}_o(k,l) + s_q(k,l)\boldsymbol{d}_q(k,l) + \boldsymbol{v}(k,l), \tag{D.6}$$

and 2) when the target is present, but own voice is absent

$$\boldsymbol{x}(k,l) = s_t(k,l)\boldsymbol{d}_t(k,l) + s_q(k,l)\boldsymbol{d}_q(k,l) + \boldsymbol{v}(k,l). \tag{D.7}$$

In the sequel, we omit the frequency bin and frame index, e.g. $\boldsymbol{x} \triangleq \boldsymbol{x}(k,l)$, for brevity.

2.1 Multichannel Wiener filter beamforming

The task of the beamformer is to retrieve the target speech s_t, while suppressing the interference and noise. The output of a linear beamformer is given by [3]

$$y = \boldsymbol{w}^H \boldsymbol{x}, \tag{D.8}$$

where $\boldsymbol{w}$ is the vector of beamformer weights. The multichannel Wiener filter (MWF) is the linear minimum mean square error (LMMSE) estimator of the target signal with beamformer weights $\boldsymbol{w}_{\text{MWF}}$ which are found by solving the following optimization problem [3]:

$$\boldsymbol{w}_{\text{MWF}} = \arg\min_{\boldsymbol{w}} \mathbb{E}\left[|s_t - \boldsymbol{w}^H\boldsymbol{x}|^2\right], \tag{D.9}$$

where H is the Hermitian transpose. Assuming that s_t, s_q and $\boldsymbol{v}$ are uncorrelated random variables, the MWF can be shown [3] to be dependent on the target RATF $\boldsymbol{d}_t$, the target power spectral density (PSD) $\lambda_t = \mathbb{E}\left[|s_t|^2\right]$, and the interference-plus-noise CPSD matrix, $\mathbf{C}_{qv}$. The interference-plus-noise CPSD matrix is defined to be

$$\begin{aligned} \mathbf{C}_{qv} &= \mathbb{E}\left[(s_q\boldsymbol{d}_q + \boldsymbol{v})(s_q\boldsymbol{d}_q + \boldsymbol{v})^H\right] \\ &= \lambda_q \boldsymbol{d}_q \boldsymbol{d}_q^H + \mathbf{C}_v, \end{aligned} \tag{D.10}$$

where $\lambda_q = \mathbb{E}\left[|s_q|^2\right]$ is the interference PSD and $\mathbf{C}_v = \mathbb{E}\left[\boldsymbol{v}\boldsymbol{v}^H\right]$ is the noise CPSD matrix. Then the MWF beamformer can be expressed as [3]

$$\boldsymbol{w}_{\text{MWF}} = \frac{\mathbf{C}_{qv}^{-1}\boldsymbol{d}_t}{\boldsymbol{d}_t^H\mathbf{C}_{qv}^{-1}\boldsymbol{d}_t} \cdot \frac{\lambda_t}{\lambda_t + (\boldsymbol{d}_t^H\mathbf{C}_{qv}^{-1}\boldsymbol{d}_t)^{-1}}, \tag{D.11}$$

where the first factor is known as the minimum variance distortion-less response (MVDR) beamformer and the second factor is known as the single-channel post Wiener filter. We see that the MWF beamformer in the form of (D.11) requires $\boldsymbol{d}_t$, $\boldsymbol{d}_q$, λ_t, λ_q, and $\mathbf{C}_v$ to be known and in the following, we propose methods to estimate these parameters for each time-frequency tile by exploiting the own voice of the user. For simplification, we assume that the noise, $\boldsymbol{v}$, is a time-varying random process and that its CPSD matrix can be expressed as $\mathbf{C}_v = \lambda_v \boldsymbol{\Gamma}_v$. Here, $\boldsymbol{\Gamma}_v$ is a known noise CPSD matrix which is normalized with respect to the reference microphone and obtained from the most recent noise-only observation [14].

2.2 Target and interference identification

During own voice presence, we estimate $\boldsymbol{d}_q$. Following (D.6), the noisy CPSD matrix during own voice is modeled as

$$\mathbf{C}_x = \lambda_o \boldsymbol{d}_o \boldsymbol{d}_o^H + \lambda_q \boldsymbol{d}_q \boldsymbol{d}_q^H + \lambda_v \boldsymbol{\Gamma}_v, \tag{D.12}$$

and likewise, during own voice absence, we assume that target is present cf. (D.7), such that the noisy CPSD matrix is modeled as

$$\mathbf{C}_x = \lambda_t \boldsymbol{d}_t \boldsymbol{d}_t^H + \lambda_q \boldsymbol{d}_q \boldsymbol{d}_q^H + \lambda_v \boldsymbol{\Gamma}_v. \tag{D.13}$$

In applications such as HADs, the microphone array is commonly mounted in a fixed position on the user's head. Therefore, the acoustic transfer function, $\boldsymbol{d}_o$, from the user's mouth to the microphones can be considered approximately time-invariant. This allows offline estimation of the own voice RATF vector $\boldsymbol{d}_o$, which can be used during online deployment of the noise reduction algorithm. Additionally, the microphones are placed close to the user's mouth hence the own voice signal can be considerably louder than the target and interference speech signals, especially at lower frequencies, when the own voice is active [40, p. 251]. For these reasons, we consider the own voice RATF vector, $\boldsymbol{d}_o$, as known and assume that an own voice activity detector (OVAD) is available. The estimated $\boldsymbol{d}_q$ is then used during own voice absence (but target presence) to estimate the remaining parameters $\boldsymbol{d}_t$, λ_t, λ_q, and λ_v per TF tile and the resulting MWF beamformer can then be applied.

In practice, it may occur that the signal models in (D.12) or (D.13) are violated, for example due to speech overlap and gaps. A worst case example is speech overlap between the user and the target speaker. Such situations can potentially lead to suppression of the target, as the target might be identified as the interfering speaker. One potential solution is to use several seconds of noisy observations during own voice presence. Since speech overlaps between the user and the target are often short and brief (e.g. 250 ms) [33, 34, 41], increasing the number of observations from own voice presence can reduce the likelihood of the target being identified as interference.

Algorithm 1 MWF beamformer with proposed target and interference identification.

```
Input: $\boldsymbol{d}_o$.
if own voice is present then
    Estimate $\boldsymbol{d}_q$.
else if own voice is absent then
    Estimate $\boldsymbol{d}_t$, $\lambda_t$, $\lambda_q$, and $\lambda_v$ given $\boldsymbol{d}_q$.
    Form the interference-plus-noise CPSD matrix
    $\mathbf{C}_{qv} = \lambda_q \boldsymbol{d}_q \boldsymbol{d}_q^H + \lambda_v \boldsymbol{\Gamma}_v$.
    Compute the MWF beamformer weights $\boldsymbol{w}_{\mathrm{MWF}}$ in (D.11).
    Apply the beamformer $y = \boldsymbol{w}_{\mathrm{MWF}}^H \boldsymbol{x}$.
end if
```

Furthermore, the use of own voice to identify an interfering speaker can be generalized to multiple interfering speakers. Specifically, any speakers that are present for minimum duration during own voice presence can be considered undesired. A potential procedure could for example involve a model-order selection algorithm e.g. minimum description length, Akaike or Bayesian information criterion to first determine the number of interfering speakers [42, 43]. This is then followed by an estimation procedure of the RATF vectors for all the interfering speakers and finally estimation of the interference-plus-noise CPSD matrix.

The proposed noise reduction scheme is summarized in Fig. D.3 and as pseudo-code in Algorithm 1.

3 Maximum Likelihood Estimation

In order to implement MWF beamformers for the considered acoustic situation, the parameters $\boldsymbol{d}_q$, $\boldsymbol{d}_t$, λ_t, λ_q, and λ_v must be estimated. In the following, we present several MLE-based schemes for estimation of the parameters of interest.

It is widely known that MLEs of the RATF vectors and PSDs perform well when used in a beamforming context e.g. in [10, 14, 18, 22]. Comparative and theoretical performance of these estimators, e.g. in terms of Cramer-Rao bounds, have been derived and presented in [15, 17, 18, 44]. Let us first note that the signal model in (D.12), where $\boldsymbol{d}_o$ and $\boldsymbol{\Gamma}_v$ are assumed known, and the signal model in (D.13) where $\boldsymbol{d}_q$ and $\boldsymbol{\Gamma}_v$ are assumed known, both can be written in the following general form

$$\mathbf{C} \triangleq \mathbf{C}(\lambda_1, \lambda_2, \phi, \boldsymbol{d}_1) = \lambda_1 \boldsymbol{d}_1 \boldsymbol{d}_1^H + \lambda_2 \boldsymbol{d}_2 \boldsymbol{d}_2^H + \phi \boldsymbol{\Gamma}. \tag{D.14}$$

Hence, finding estimates of the parameters of interest for both (D.12) and (D.13), corresponds to finding MLEs of λ_1, λ_2, ϕ, and $\boldsymbol{d}_1$ in (D.14). In particular, let $\hat{\lambda}_1$, $\hat{\lambda}_2$, $\hat{\phi}$, and $\hat{\boldsymbol{d}}_1$ denote the MLEs of these parameters. Then for (D.12), we estimate the interference RATF vector as $\hat{\boldsymbol{d}}_q = \hat{\boldsymbol{d}}_1$, (the MLEs $\hat{\lambda}_q = \hat{\lambda}_1$, $\hat{\lambda}_o = \hat{\lambda}_2$, and $\hat{\lambda}_v = \hat{\phi}$ are nuisance parameters and, hence, not used in the subsequent steps). Similarly, when own voice is absent, the parameters of (D.13) are given by $\hat{\boldsymbol{d}}_t = \hat{\boldsymbol{d}}_1$, $\hat{\lambda}_t = \hat{\lambda}_1$, $\hat{\lambda}_q = \hat{\lambda}_2$, and $\hat{\lambda}_v = \hat{\phi}$.

We assume that the noisy observations, $\boldsymbol{x}$, are complex Gaussian distributed [10, 16, 18] such that the likelihood for N observations of $\boldsymbol{x}$, $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]$, is

$$f(\boldsymbol{X}; \lambda_1, \lambda_2, \phi, \boldsymbol{d}_1) = \frac{\exp\left(-N\mathrm{tr}(\mathbf{C}^{-1}\mathbf{R})\right)}{\pi^{NM}|\mathbf{C}|^N}, \tag{D.15}$$

where $\mathbf{R} = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^H$, $|\cdot|$ is the determinant operator, and $\mathrm{tr}(\cdot)$ denotes the trace operator. Furthermore, we assume that $\boldsymbol{d}_1$ is an element of a pre-defined dictionary $\mathcal{D} = \{\boldsymbol{d}^{(1)}, \boldsymbol{d}^{(2)}, ..., \boldsymbol{d}^{(N_D)}\}$, where N_D is the dictionary size [10]. The MLEs can be found by solving the optimization problem

$$\underset{\lambda_1, \lambda_2, \phi, \boldsymbol{d}_1 \in \mathcal{D}}{\arg\max} \; \log f(\boldsymbol{X}; \lambda_1, \lambda_2, \phi, \boldsymbol{d}_1). \tag{D.16}$$

Closed-form solutions for this optimization problem seem not to exist [44]. Instead, we use a numerical approach to solve (D.16) in Sec. 3.1 which involves a two-dimensional search. In Secs. 3.2 and 3.3, we adapt MLEs from [22] and [19] to estimate λ_1, λ_2, ϕ, and $\boldsymbol{d}_1$. The estimators in [22] and [19] are not strictly MLEs of the problem posed in (D.16). However, they are computationally much less expensive as they only involve a one-dimensional search and – as we show – to perform essentially on par with the true computational comples MLEs of (D.16) in terms of speech enhancement performance.

3.1 Joint ML using grid search

Let us rewrite equation (D.14) as

$$\begin{aligned}\mathbf{C} &= \lambda_1 \boldsymbol{d}_1 \boldsymbol{d}_1^H + \phi\left(\frac{\lambda_2}{\phi}\boldsymbol{d}_2\boldsymbol{d}_2^H + \boldsymbol{\Gamma}\right), \\ &= \lambda_1 \boldsymbol{d}_1 \boldsymbol{d}_1^H + \phi\boldsymbol{\Phi}(\psi(\phi)),\end{aligned} \tag{D.17}$$

where $\psi(\phi) = \frac{\lambda_2}{\phi}$, and $\boldsymbol{\Phi}\left(\psi\right) \triangleq \psi\boldsymbol{d}_2\boldsymbol{d}_2^H + \boldsymbol{\Gamma}$. For notational convenience, we define $\psi \triangleq \psi(\phi)$. For a given value of ψ, closed-form MLEs of λ_1 and ϕ exist, while conditioned on $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ [15, 19]. Hence, conditioned on $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$, estimating the remaining parameters λ_1, ϕ, and ψ involves a one-dimensional search procedure over ψ or implicitly λ_2. In principle any numerical solver, e.g. a grid-search or gradient ascent method, can be used to solve the optimization problem. However, as a proof of concept, we use a grid-search based solver as the grid is only over ψ and $\boldsymbol{d}_1$. The grid-search procedure can be simplified to be over the dictionary, $\Psi = \{\psi^{(1)}, ..., \psi^{(N_\psi)}\}$, where N_ψ denotes the cardinality of Ψ. Obviously, for a sufficiently fine grid, Ψ, the proposed approach will return estimates arbitrarily close to the true MLE.

The first step in the procedure is to obtain an MLE for ϕ for a particular grid point, $\psi \in \Psi$, while conditioned on $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$. To do so, we define the MVDR beamformer with distortion-less constraint on $\boldsymbol{d}_1$,

$$\boldsymbol{w}(\psi, \boldsymbol{d}_1) = \boldsymbol{\Phi}^{-1}(\psi)\boldsymbol{d}_1 \left(\boldsymbol{d}_1^H \boldsymbol{\Phi}^{-1}(\psi)\boldsymbol{d}_1\right)^{-1}. \tag{D.18}$$

Furthermore, let

$$\mathbf{Q}(\psi, \boldsymbol{d}_1) = \mathbf{I} - \boldsymbol{d}_1 \boldsymbol{w}^H(\psi, \boldsymbol{d}_1), \tag{D.19}$$

where $\mathbf{I}$ is the identity matrix. The MLE of ϕ is then given by [19]

$$\hat{\phi}(\psi, \boldsymbol{d}_1) = \frac{1}{M-1} \operatorname{tr}\left(\mathbf{Q}(\psi, \boldsymbol{d}_1)\mathbf{R}\boldsymbol{\Phi}^{-1}(\psi)\right), \tag{D.20}$$

where $\mathbf{R} = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^H$ is the sample noisy CPSD matrix, and the MLE of λ_1 is [19]

$$\hat{\lambda}_1(\psi, \boldsymbol{d}_1) = \boldsymbol{w}(\psi, \boldsymbol{d}_1)^H \left(\mathbf{R} - \hat{\phi}(\psi, \boldsymbol{d}_1)\boldsymbol{\Phi}(\psi)\right) \boldsymbol{w}(\psi, \boldsymbol{d}_1). \tag{D.21}$$

The MLEs, $\hat{\phi}$ and $\hat{\lambda}_1$, are then used to concentrate the log-likelihood in (D.16), such that the optimization problem is reduced to

$$\hat{\psi}, \hat{\boldsymbol{d}}_1 = \underset{\psi \in \Psi, \boldsymbol{d}_1 \in \mathcal{D}}{\arg\max} \ \log f(\boldsymbol{X}, \hat{\phi}, \hat{\lambda}_1; \psi, \boldsymbol{d}_1). \tag{D.22}$$

Given the MLE $\hat{\psi}$, the MLE of λ_2 can be found as

$$\hat{\lambda}_2(\hat{\psi}, \hat{\boldsymbol{d}}_1) = \hat{\phi}(\hat{\psi}, \hat{\boldsymbol{d}}_1) \cdot \hat{\psi}. \tag{D.23}$$

The whole procedure is summarized in Algorithm 2.

Algorithm 2 Joint ML using grid search

```
Input: $\Psi = \{\psi^{(1)}, ..., \psi^{(N_\psi)}\}$, $\mathcal{D} = \{\boldsymbol{d}_1^{(1)}, ..., \boldsymbol{d}_1^{(N_D)}\}$, $\boldsymbol{d}_2$, $\mathbf{\Gamma}$.
for $i = 1, 2, ..., N_D$ do
    for $j = 1, 2, ..., N_\psi$ do
        Compute $\boldsymbol{w}(\psi^{(j)}, \boldsymbol{d}_1^{(i)})$ using (D.18).
        Compute $\mathbf{Q}(\psi^{(j)}, \boldsymbol{d}_1^{(i)})$ using (D.19).
        Estimate $\hat{\phi}(\psi^{(j)}, \boldsymbol{d}_1^{(i)})$ using (D.20).
        Estimate $\hat{\lambda}_1(\psi^{(j)}, \boldsymbol{d}_1^{(i)})$ using (D.21).
        Evaluate $\log f(\boldsymbol{X}, \hat{\phi}, \hat{\lambda}_1; \psi^{(j)}, \boldsymbol{d}_1^{(i)})$ using (D.22) and (D.15).
        Compute $\hat{\lambda}_2(\psi^{(j)}, \boldsymbol{d}_1^{(i)}) = \psi^{(j)} \cdot \hat{\phi}(\psi^{(j)}, \boldsymbol{d}_1^{(i)})$.
    end for
end for
Find $i^*$ and $j^*$ that maximize $\log f(\boldsymbol{X}, \hat{\phi}, \hat{\lambda}_1; \psi^{(j)}, \boldsymbol{d}_1^{(i)})$.
The joint MLEs are then $\hat{\phi}(\psi^{(j^*)}, \boldsymbol{d}_1^{(i^*)})$, $\hat{\lambda}_1(\psi^{(j^*)}, \boldsymbol{d}_1^{(i^*)})$, $\hat{\lambda}_2(\psi^{(j^*)}, \boldsymbol{d}_1^{(i^*)})$, and
    $\hat{\boldsymbol{d}}_1 := \boldsymbol{d}_1^{(i^*)}$.
```

3.2 ML in the blocked domain

As an alternative to the joint ML method, which requires a two-dimensional dictionary search, we propose in the following a simpler ML estimation procedure in the blocked domain [16, 44]. Specifically, the MLEs are not guaranteed to be ML optimal for the problem posed in (D.16), but have been demonstrated to perform well in terms of beamforming performance in [22]. The ML estimation of the parameters in (D.14), i.e. λ_1, λ_2, ϕ, and $\boldsymbol{d}_1$, in the blocked domain is adapted from [22] and consists of two stages. The first stage is ML estimation of λ_1 and ϕ conditioned on $\boldsymbol{d}_1$ in the blocked domain of $\boldsymbol{d}_2\boldsymbol{d}_2^H$, i.e. the null-space of $\boldsymbol{d}_2\boldsymbol{d}_2^H$. The second stage is ML estimation of λ_2 where the MLEs of λ_1 and ϕ conditioned on $\boldsymbol{d}_1$ are used to concentrate the log-likelihood in (D.16). The rationale behind this ML estimation in the blocked domain, is to simplify the estimation problem by canceling one of the speech components with a blocking matrix $\mathbf{B}$. Specifically, the speech components λ_2 and $\boldsymbol{d}_2$ are eliminated in the first stage by projecting $\boldsymbol{x}$ to the null-space of $\boldsymbol{d}_2\boldsymbol{d}_2^H$. In the second stage, only λ_1, ϕ, and $\boldsymbol{d}_1$ remain and are estimated using the MLEs in [19].

3.2.1 ML estimation of λ_1 and ϕ

To map the noisy observations into the blocked domain, we form a blocking matrix, which cancels the $\lambda_2 \boldsymbol{d}_2 \boldsymbol{d}_2^H$ term from (D.14). The blocking matrix, $\mathbf{B}$, is given as [22]

$$\mathbf{B} = \left(\mathbf{I}_{M\times M} - \frac{\boldsymbol{d}_2 \boldsymbol{d}_2^H}{\boldsymbol{d}_2^H \boldsymbol{d}_2} \right) \mathbf{I}_{M\times M-1}. \tag{D.24}$$

where $\mathbf{I}_{M\times M}$ is an $M \times M$ identity matrix and $\mathbf{I}_{M\times M-1}$ is the first $M-1$ column vectors of $\mathbf{I}_{M\times M}$. Applying the blocking matrix to the input vector $\mathbf{B}^H \boldsymbol{x}$, the CPSD matrix in the blocked domain is

$$\tilde{\mathbf{C}} = \mathbf{B}^H \mathbf{C} \mathbf{B} = \tilde{\lambda}_1 \tilde{\boldsymbol{d}}_1 \tilde{\boldsymbol{d}}_1^H + \tilde{\phi} \tilde{\boldsymbol{\Gamma}}, \tag{D.25}$$

where $\mathbf{C}$ is the CPSD matrix from (D.14), $\tilde{\boldsymbol{\Gamma}} = \mathbf{B}^H \boldsymbol{\Gamma} \mathbf{B}$, and $\tilde{\boldsymbol{d}}_1 = \mathbf{B}^H \boldsymbol{d}_1$. The parameters to estimate in (D.25) are the blocked domain PSDs $\tilde{\lambda}_1$ and $\tilde{\phi}$, and the RATF vector $\tilde{\boldsymbol{d}}_1$. The CPSD matrix in (D.25) has a form that is identical to the CPSD matrix in (D.17). Therefore, estimating $\tilde{\lambda}_1$, $\tilde{\phi}$, and $\tilde{\boldsymbol{d}}_1$ follows a similar procedure as found in Sec. 3.1. In the first stage, the likelihood function in the blocked domain is

$$f(\tilde{\boldsymbol{X}}; \tilde{\lambda}_1, \tilde{\phi} | \tilde{\boldsymbol{d}}_1) = \frac{\exp\left(-N \mathrm{tr}(\tilde{\mathbf{C}}^{-1} \tilde{\mathbf{R}}) \right)}{\pi^{NM} |\tilde{\mathbf{C}}|^N}, \tag{D.26}$$

while conditioned on $\tilde{\boldsymbol{d}}_1$, and $\tilde{\boldsymbol{X}} = \mathbf{B}^H \boldsymbol{X}$ and $\tilde{\mathbf{R}} = \mathbf{B}^H \mathbf{R} \mathbf{B}$. The optimization problem is

$$\underset{\tilde{\lambda}_1, \tilde{\phi}}{\arg\max} \ \log f(\tilde{\boldsymbol{X}}; \tilde{\lambda}_1, \tilde{\phi} | \tilde{\boldsymbol{d}}_1). \tag{D.27}$$

In the following, the MLEs of $\tilde{\lambda}_1$ and $\tilde{\phi}$ are adaptations of the MLEs derived in [19]. The ML estimate of $\tilde{\phi}$ can be shown to be a function of an MVDR beamformer in the blocked domain with a distortion-less constraint on $\tilde{\boldsymbol{d}}_1$ [19] i.e.

$$\tilde{\boldsymbol{w}}_1(\tilde{\boldsymbol{d}}_1) = \tilde{\boldsymbol{\Gamma}}^{-1} \tilde{\boldsymbol{d}}_1 \left(\tilde{\boldsymbol{d}}_1{}^H \tilde{\boldsymbol{\Gamma}}^{-1} \tilde{\boldsymbol{d}}_1 \right)^{-1}, \tag{D.28}$$

and

$$\tilde{\mathbf{Q}}_1(\tilde{\boldsymbol{d}}_1) = \mathbf{I}_{M-1\times M-1} - \tilde{\boldsymbol{d}}_1 \tilde{\boldsymbol{w}}_1^H(\tilde{\boldsymbol{d}}_1), \tag{D.29}$$

where $\mathbf{I}_{M-1\times M-1}$ is an $M-1 \times M-1$ identity matrix. The MLE of ϕ in the blocked domain is [19]

$$\hat{\tilde{\phi}}(\tilde{\boldsymbol{d}}_1) = \frac{1}{M-2} \mathrm{tr}\left(\tilde{\mathbf{Q}}_1(\tilde{\boldsymbol{d}}_1) \tilde{\mathbf{R}} \tilde{\boldsymbol{\Gamma}}^{-1} \right), \tag{D.30}$$

where $\tilde{\mathbf{R}} = \mathbf{B}^H \mathbf{R} \mathbf{B}$ and the MLE of λ_1 is [19, 22]

$$\hat{\tilde{\lambda}}_1(\tilde{\boldsymbol{d}}_1) = \tilde{\boldsymbol{w}}_1^H(\tilde{\boldsymbol{d}}_1) \left(\tilde{\mathbf{R}} - \hat{\tilde{\phi}}(\tilde{\boldsymbol{d}}_1) \tilde{\boldsymbol{\Gamma}} \right) \tilde{\boldsymbol{w}}_1(\tilde{\boldsymbol{d}}_1). \tag{D.31}$$

Algorithm 3 ML in the blocked domain

```
Input: $\mathcal{D} = \{\boldsymbol{d}_1^{(1)}, ..., \boldsymbol{d}_1^{(N_D)}\}$, $\boldsymbol{d}_2$, $\mathbf{\Gamma}$.
Obtain the blocking matrix $\mathbf{B}$ from (D.24)
Compute blocked domain $\mathbf{\Gamma}$ as $\tilde{\mathbf{\Gamma}} = \mathbf{B}^H\mathbf{\Gamma}\mathbf{B}$
for $i = 1, 2, ..., N_D$ do
    $\tilde{\boldsymbol{d}}_1^{(i)} = \mathbf{B}^H \boldsymbol{d}_1^{(i)}$
    Compute $\tilde{w}_1(\tilde{\boldsymbol{d}}_1{}^{(i)})$ in (D.28).
    Compute $\tilde{\mathbf{Q}}_1(\tilde{\boldsymbol{d}}_1{}^{(i)})$ in (D.29).
    Estimate $\hat{\phi}(\tilde{\boldsymbol{d}}_1{}^{(i)})$ using (D.30).
    Estimate $\hat{\lambda}_1(\tilde{\boldsymbol{d}}_1{}^{(i)})$ using (D.31).
    Set $\hat{\mathbf{\Phi}}(\boldsymbol{d}_1^{(i)}) := \hat{\lambda}_1 \boldsymbol{d}_1^{(i)} (\boldsymbol{d}_1^{(i)})^H + \hat{\phi}\mathbf{\Gamma}$
    Compute $w_2(\boldsymbol{d}_1^{(i)})$ in (D.35).
    Compute $\mathbf{Q}_2(\boldsymbol{d}_1^{(i)})$ in (D.36).
    Estimate $\hat{\gamma}(\boldsymbol{d}_1^{(i)})$ using (D.37).
    Estimate $\hat{\lambda}_2(\boldsymbol{d}_1^{(i)})$ using (D.38).
    Evaluate $\log f(\boldsymbol{X}, \hat{\lambda}_1, \hat{\lambda}_2, \hat{\phi}; \boldsymbol{d}_1^{(i)})$ in (D.41) using (D.15)
end for
Find $i^* = \arg\max_i \log f(\boldsymbol{X}, \hat{\lambda}_1, \hat{\lambda}_2, \hat{\phi}; \boldsymbol{d}_1^{(i)})$
The joint MLEs are then $\hat{\phi}(\boldsymbol{d}_1^{(i^*)})$, $\hat{\lambda}_1(\boldsymbol{d}_1^{(i^*)})$, $\hat{\lambda}_2(\boldsymbol{d}_1^{(i^*)})$, $\hat{\boldsymbol{d}}_1 := \boldsymbol{d}_1^{(i^*)}$.
```

3.2.2 ML estimation of λ_2

Given $\hat{\hat{\lambda}}_1(\tilde{d}_1)$ and $\hat{\hat{\phi}}(\tilde{\boldsymbol{d}}_1)$, these may be inserted into the noisy CPSD matrix in (D.14) such that it becomes

$$\begin{aligned} \mathbf{C}(\lambda_2, \boldsymbol{d}_1) &= \lambda_2 \boldsymbol{d}_2 \boldsymbol{d}_2^H + \left(\hat{\hat{\lambda}}_1 \boldsymbol{d}_1 \boldsymbol{d}_1^H + \hat{\hat{\phi}}\mathbf{\Gamma}\right) \\ &= \lambda_2 \boldsymbol{d}_2 \boldsymbol{d}_2^H + \hat{\mathbf{\Phi}}(\boldsymbol{d}_1), \end{aligned} \tag{D.32}$$

where $\hat{\mathbf{\Phi}}(\boldsymbol{d}_1) = \hat{\hat{\lambda}}_1 \boldsymbol{d}_1 \boldsymbol{d}_1^H + \hat{\hat{\phi}}\mathbf{\Gamma}$. For ML estimation of the remaining parameter, λ_2, we introduce the parameter γ such that the noisy CPSD matrix is

$$\mathbf{C}(\lambda_2, \boldsymbol{d}_1, \gamma) = \lambda_2 \boldsymbol{d}_2 \boldsymbol{d}_2^H + \gamma \hat{\mathbf{\Phi}}(\boldsymbol{d}_1), \tag{D.33}$$

which ensures that (D.33) has a form identical to (D.25), and, hence, the MLEs of γ and λ_2 can be found similarly. The optimization problem is

$$\arg\max_{\lambda_2, \gamma} \ \log\ f(\boldsymbol{X}; \lambda_2, \gamma | \hat{\mathbf{\Phi}}(\boldsymbol{d}_1)), \tag{D.34}$$

where the likelihood function is conditioned on $\hat{\mathbf{\Phi}}(\boldsymbol{d}_1)$, and has the form as in (D.15). To estimate λ_2 and γ, first we form the MVDR beamformer with distortion-less constraint on $\boldsymbol{d}_2$

$$\boldsymbol{w}_2(\boldsymbol{d}_1) = \hat{\mathbf{\Phi}}(\boldsymbol{d}_1)^{-1}\boldsymbol{d}_2 \left(\boldsymbol{d}_2{}^H\hat{\mathbf{\Phi}}(\boldsymbol{d}_1)^{-1}\boldsymbol{d}_2\right)^{-1}, \tag{D.35}$$

such that

$$\mathbf{Q}_2(\boldsymbol{d}_1) = \mathbf{I}_{M\times M} - \boldsymbol{d}_2\boldsymbol{w}_2^H(\boldsymbol{d}_1). \tag{D.36}$$

Then the MLE of γ is [19]

$$\hat{\gamma}(\boldsymbol{d}_1) = \frac{1}{M-1}\mathrm{tr}\left(\mathbf{Q}_2(\boldsymbol{d}_1)\mathbf{R}\hat{\mathbf{\Phi}}(\boldsymbol{d}_1)^{-1}\right), \tag{D.37}$$

and the MLE of λ_2 is [19]

$$\hat{\lambda}_2(\boldsymbol{d}_1) = \boldsymbol{w}_2^H(\boldsymbol{d}_1)\left(\mathbf{R} - \hat{\gamma}(\boldsymbol{d}_1)\hat{\mathbf{\Phi}}(\boldsymbol{d}_1)\right)\boldsymbol{w}_2(\boldsymbol{d}_1). \tag{D.38}$$

The introduction of the variable $\hat{\gamma}(\boldsymbol{d}_1)$, means that the MLE of λ_1 and ϕ becomes

$$\hat{\lambda}_1(\boldsymbol{d}_1) = \hat{\hat{\lambda}}_1(\boldsymbol{d}_1)\cdot\hat{\gamma}(\boldsymbol{d}_1), \tag{D.39}$$

and

$$\hat{\phi}(\boldsymbol{d}_1) = \hat{\hat{\phi}}(\boldsymbol{d}_1)\cdot\hat{\gamma}(\boldsymbol{d}_1). \tag{D.40}$$

Finally, the MLE of $\boldsymbol{d}_1$ is found by evaluating the log-likelihood for each $\boldsymbol{d}_1 \in \mathcal{D}$, and choose the one that maximizes the log-likelihood i.e.

$$\hat{\boldsymbol{d}}_1 = \arg\max_{\boldsymbol{d}_1\in\mathcal{D}} \log f(\boldsymbol{X}, \hat{\lambda}_1, \hat{\lambda}_2, \hat{\phi}; \boldsymbol{d}_1). \tag{D.41}$$

The ML procedure in the blocked domain is summarized in Algorithm 3.

3.3 Unconstrained Joint ML

Let $\mathbf{D} \triangleq [\boldsymbol{d}_1\,\boldsymbol{d}_2]$ and $\mathbf{\Lambda}(\lambda_1, \lambda_2) = \mathrm{diag}(\lambda_1, \lambda_2)$. Then the CPSD matrix in (D.14) can be written as

$$\mathbf{C}(\boldsymbol{d}_1, \mathbf{\Lambda}) = \mathbf{D}\mathbf{\Lambda}\mathbf{D}^H + \phi\mathbf{\Gamma}. \tag{D.42}$$

Note that the signal model is only identical to the one in (D.14) if $\mathbf{\Lambda}$ is a diagonal matrix. For known matrices $\mathbf{D}$ and $\mathbf{\Gamma}$, MLEs of $\mathbf{\Lambda}$ and ϕ were derived in [19]. However, the MLEs presented in [19] do not guarantee that the estimate of $\mathbf{\Lambda}$ is a diagonal matrix. The MLEs in [19], therefore, are not necessary maximum likelihood for the problem posed in (D.16). Nevertheless, as demonstrated in the simulation experiment in Appendix A, using the diagonal elements of the MLE of [19] works essentially as good as the joint ML method from Sec. 3.1.

In [19], the MLE for ϕ is derived by first defining the linearly constrained minimum variance (LCMV) beamformer with distortion-less constraints on $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$,

$$\mathbf{W}(\boldsymbol{d}_1) = \boldsymbol{\Gamma}^{-1}\mathbf{D}\left(\mathbf{D}^H\boldsymbol{\Gamma}^{-1}\mathbf{D}\right)^{-1}, \tag{D.43}$$

where $\mathbf{W}(\boldsymbol{d}_1) \in \mathbb{C}^{M\times 2}$ and

$$\mathbf{Q}(\boldsymbol{d}_1) = \mathbf{I} - \mathbf{D}\mathbf{W}^H(\boldsymbol{d}_1). \tag{D.44}$$

Then the MLE of the noise PSD is given as [19]

$$\hat{\phi}(\boldsymbol{d}_1) = \frac{1}{M-2}\mathrm{tr}\left(\mathbf{Q}(\boldsymbol{d}_1)\mathbf{R}\boldsymbol{\Gamma}^{-1}\right), \tag{D.45}$$

and the ML estimate of $\boldsymbol{\Lambda}$ is then [19]

$$\hat{\boldsymbol{\Lambda}}(\boldsymbol{d}_1) = \mathbf{W}^H(\boldsymbol{d}_1)\left(\mathbf{R} - \hat{\phi}(\boldsymbol{d}_1)\boldsymbol{\Gamma}\right)\mathbf{W}(\boldsymbol{d}_1). \tag{D.46}$$

We propose to find, the estimates of λ_1 and λ_2 as the main diagonal of $\hat{\boldsymbol{\Lambda}}(\boldsymbol{d}_1)$, i.e. $\hat{\lambda}_1 = \hat{\Lambda}_{1,1}$ and $\hat{\lambda}_2 = \hat{\Lambda}_{2,2}$. Finally, in order to estimate $\boldsymbol{d}_1$, we concentrate the log-likelihood with the MLEs of $\boldsymbol{\Lambda}(\boldsymbol{d}_1)$ and $\phi(\boldsymbol{d}_1)$, and search over the dictionary $\mathcal{D}$ until the element that returns the highest log-likelihood is found i.e.

$$\hat{\boldsymbol{d}}_1 = \arg\max_{\boldsymbol{d}_1\in\mathcal{D}}\ \log f(\boldsymbol{X}, \hat{\boldsymbol{\Lambda}}, \hat{\phi}; \boldsymbol{d}_1). \tag{D.47}$$

The unconstrained ML procedure is summarized in Algorithm 4. We have compared the three proposed algorithms in terms of speech enhancement performance in Appendix A. Our experiments demonstrate that the proposed algorithms essentially perform on par in terms of ESTOI and PESQ. 3 and Algorithm 4 perform marginally better than Algorithm 2 in terms of PESQ score, however slightly worse in terms of ESTOI score. This is possibly due to a slightly more aggressive noise reduction for Algorithm 3 and Algorithm 4 than Algorithm 2. This leads to higher PESQ scores but at the cost of more speech distortion and lower ESTOI scores. For this reason, we choose to leave Algorithm 2 out of the evaluation in Sec. 4.4, although Algorithm 3 and 4 do not solve the initial problem posed in (D.16). However, since Algorithm 2 requires a two-dimensional search, Algorithm 2 is likely much more computationally complex compared to Algorithm 3 and 4. The

Algorithm 4 Unconstrained joint ML

```
Input: $\mathcal{D} = \{\boldsymbol{d}_1^{(1)}, ..., \boldsymbol{d}_1^{(N_D)}\}$, $\boldsymbol{d}_2$, $\boldsymbol{\Gamma}$.
for $i = 1, 2, ..., N_D$ do
    Define $\mathbf{D}_i \triangleq [\boldsymbol{d}_1^{(i)} \; \boldsymbol{d}_2]$
    Compute $\mathbf{W}(\boldsymbol{d}_1^{(i)})$ in (D.43).
    Compute $\mathbf{Q}(\boldsymbol{d}_1^{(i)})$ in (D.44).
    Estimate $\hat{\phi}(\boldsymbol{d}_1^{(i)})$ using (D.45).
    Estimate $\hat{\boldsymbol{\Lambda}}(\boldsymbol{d}_1^{(i)})$ using (D.46).
    Evaluate $\log f(\boldsymbol{X}, \hat{\boldsymbol{\Lambda}}, \hat{\phi}; \boldsymbol{d}_1^{(i)})$ in (D.47)
end for
Find $i^* = \arg\max_i \; \log f(\boldsymbol{X}, \hat{\boldsymbol{\Lambda}}, \hat{\phi}; \boldsymbol{d}_1^{(i)})$
The joint MLEs are then $\hat{\lambda}_1 = \hat{\Lambda}_{1,1}(\boldsymbol{d}_1^{(i^*)})$, $\hat{\phi}(\boldsymbol{d}_1^{(i^*)})$, $\hat{\lambda}_2 = \hat{\Lambda}_{2,2}(\boldsymbol{d}_1^{(i^*)})$, $\hat{\boldsymbol{d}}_1 := \boldsymbol{d}_1^{(i^*)}$.
```

experiments in Sec. 4.4 furthermore reveal that the real-time factor of Algorithm 4 is 9.46, Algorithm 3 is 4.98, and Algorithm 2 is 153.46. Although, we did not perform code optimization, the real-time factors give an indication of the computation complexity of the proposed methods and favors Algorithm 4 and 3 over Algorithm 2.

3.4 Robust wideband estimation of the RATF vector

The proposed MLEs estimate the interferer and target RATF vectors independently over frequency bins. This approach allow multiple target and interferers in the acoustic scene, as long as a maximum of one target and interferer is present for a given TF tile. However, for acoustic sound sources, it is plausible to assume that the location of the target is identical across frequency. Therefore, in order to improve performance, estimation of the RATF vector can be done jointly over frequency hence made more robust [10, 25]. The joint MLE of $\boldsymbol{d}_1$ is

$$\underset{i \in \{1,2,...,|\mathcal{D}|\}}{\arg\max} \sum_{k=1}^{K} \log f(\boldsymbol{X}(k), \hat{\lambda}_1(k), \hat{\lambda}_2(k), \hat{\phi}(k); \boldsymbol{d}_1^{(i)}(k)), \tag{D.48}$$

where $|\mathcal{D}|$ is the cardinality of $\mathcal{D}$. Hence, the concentrated log-likelihood for a particular dictionary index i is added across frequency, where i corresponds to a location of the sound source.

4 Performance Evaluation of Proposed Beamforming Systems

The proposed MLEs in Sec. 3 are evaluated in terms of beamforming performance when implemented into noise reduction systems. The beamforming performance of the proposed systems is found through simulation experiments where the task is to retrieve a target speech signal contaminated with interfering speech and noise. We compare the proposed methods against state-of-the-art methods which solve similar problems but do not explicitly model the presence of the own voice and interferer. The parameter estimation of the PSDs and RATF vectors used in the proposed noise reduction systems are based on Algorithm 3 and Algorithm 4 in Sec. 3 and used in an MWF beamformer, Algorithm 1, as shown in Fig. D.3. We refer to the noise reduction systems based on Algorithm 3 as ML-BD and Algorithm 4 as UML.

4.1 Acoustic impulse response and sound databases

4.1.1 Acoustic impulse response database

Acoustic impulse functions (AIRs) are used to simulate the sound waves propagating from sound sources to the HA microphones. The AIRs were measured in an acoustic setup consisting of a circular loudspeaker array with a radius of 1.9 meters placed in an acoustically damped room [45]. A human HA user was seated in the center of the array during the measurements wearing two behind-the-ear (BTE) HAs; one placed on each ear. Each HA has a front and rear microphone separated by 1.3 cm. The AIRs mostly depend on the head and torso acoustics while reverberation has been removed by truncating the AIRs.

All M=4 microphones are used in a binaural HA configuration for the simulations. A direct implementation - as used in our simulations - of the MWF beamformers for a binaural HA configuration will result in a "noise collapse" [46]. In other words, all noise sources will sound as if they were originating from the target location. This is obviously important for a binaural HA application. However, several methods have been developed to mask or avoid this unwanted perceptual effect, e.g., [46–48]. Such methods are outside the scope of the present paper.

We assume instantaneous and error-free signal exchange between the left and right HAs. The AIRs were sampled at a horizontal resolution of 7.5 degrees with 0° defined as the frontal direction from the HA user's point of view, and the azimuth is counterclockwise rotating. Hence, the dictionary of AIRs contains AIRs from 48 different directions. The own voice AIRs were measured using a mouth reference microphone placed in front of the HA user's mouth. The HA user was asked to read a text up loud, and the AIRs from the own voice reference point to the HA microphones were measured [45].

In Sec. 4.5, AIR mismatches are simulated by using two different sets of AIR dictionaries measured on two different human heads. One dictionary is used to simulate the acoustic scene, while the other is used as a dictionary in the noise reduction systems.

The RATF dictionary used for the proposed algorithms is obtained by transforming the AIR dictionary using (D.5). The frontal microphone of the left ear HA is used as the reference microphone.

4.1.2 Speech and noise databases

Speech signals used for the own voice, target, and interference, are obtained from the TIMIT database [49]. Speech pauses are removed with an energy-based VAD to minimize the influence of speech gaps in the evaluation. We do not simulate speech gaps caused by conversation pauses. However, the acoustic scene still include situations where neither the own voice nor the target speech are present in a TF tile due to speech being sparse in the TF domain. Hence, there are TF-tiles where own voice or target speech is absent even if they are detected present.

The noise database used in the simulation is recordings of noise found in realistic acoustic environments (e.g. a busy canteen and car cabin). The recordings of the noise are made with a spherical microphone array to accurately capture the noise field as measured at a reference point of the spherical microphone array. The captured noise is then transformed and convolved with the AIRs, such that the resulting noise field at the HA microphones in the simulation is identical to the one measured with the spherical microphone array [50].

4.2 Simulation of acoustic scenes

4.2.1 Target and noise levels

We define the input signal-to-interference-plus-noise ratio (SINR) as the ratio between the average target speech power and the average interference-plus-noise power. The target speech and interference-plus-noise power are computed prior to convolving the signals with the AIRs. The interference-to-noise ratio (INR) is defined similarly as the ratio between the average interfering speech power and the average noise power prior to convolving with the AIRs. The own voice and target speech are set to have equal power prior to convolving with the AIRs.

4.2.2 Target and interferer locations

The target RATF vector is randomly drawn from the dictionary of RATF vectors. Each RATF vector is associated with a direction and the RATF vectors are drawn from a uniform distribution where the set of possible outcomes is {-90°, -82.5°, ..., 90°}. Hence, the target is located in the frontal half-plane as the HA user in realistic situations is likely to be facing the target speaker [15, 26]. The RATF vector for the interfering speech is randomly selected to be from the directions 75° or 225° and with this choice, the target speech and interfering speech are allowed to overlap in direction, when both the target speaker and interfering speaker are arriving from 75°.

4.2.3 Simulation settings

The sampling frequency of the simulation is 16 kHz. We used (1) to simulate the noisy microphone signals. The STFT and inverse STFT are used to transform the microphone signals into the time-frequency domain. A square-root Hanning window with a window size of 256 samples is used as analysis and synthesis windows. The window overlap is 128 samples. All algorithms in the evaluation have access to an oracle generic VAD that is able to perfectly detect regions with speech absence (i.e. frames with neither own voice, target, nor interfering speech). Since the generic VAD does not require to distinguish between own voice, target, nor interfering speech, this significantly simplifies the task of designing a robust VAD. The generic VAD is used to initialize Γ_v from noise-only region before any speech activity. Furthermore, an oracle OVAD is used in the evaluation for the proposed algorithms. The OVAD can detect the presence of own voice per frame but not per TF-tile. When own voice is detected absent, the proposed algorithms assume the presence of target speech. The duration of an acoustic scene is 5 seconds and Γ_v is initialized in a no-speech region before the beginning of the acoustic scene. The own voice is active in the first 2.5 seconds, followed by 2.5 seconds of own voice absence where the target is active to simulate a conversation. The interfering speaker is active during the whole 5 second simulation. Each reported performance score is an average over 40 acoustic scenes.

The HA user may occasionally rotate the head during conversations [33]. However, such head rotation were not implemented in our simulations. In practice, one might use other sensors e.g. accelerometers on board the HAD to detect or measure head rotations. After such detection, the noise reduction system may then compensate for the head-rotations or resort to a simpler baseline algorithm such as the one presented in [14] to increase robustness. Moreover, the target and interferer locations are fixed during the simulations. The proposed algorithms can in principle handle situations with moving targets, since the target RATF vectors are estimated for each TF-tile independently. Similarly, the proposed algorithms can handle moving interferers, but only during own voice regions. Moving interferers during own voice absence can potentially cause issues, but robustness against such situations can be increased by hypothesis testing. Specifically, if the noisy observations during own voice absence poorly match the interference-plus-noise CPSD matrix estimated from own voice presence (due to a moving interferer), hypothesis testing can help detecting these and resort to a simpler signal model to increase robustness e.g. (D.49).

A summary of the different acoustic settings for the experiments is given in Table D.1 with references to the figures where the results are reported.

Number of mics	4	4	4
Noise type	Canteen	Car noise	Canteen
AIR mismatch	No	No	Yes
Figure	Fig. D.4	Fig. D.5	Fig. D.7

Table D.1: Simulation settings used in the evaluation.

4.3 Baseline noise reduction systems

We compare the proposed system variants to recent state-of-the-art methods used for beamforming in HADs. These methods solve the problem of enhancing a single-target in noise using an MWF beamformer [10, 14]. More advanced techniques presented in [23, 51] can handle multiple speakers but require additional information about the target location or target speech activity. We do not assume that the noise reduction systems have access to such information and therefore these methods were not included in the evaluation. The state-of-the-art methods we have included in the evaluation are:

4.3.1 MWF beamformer with ML PSD estimation assuming frontal target

In the context of HADs, the target speaker is often assumed to be frontal with respect to the HA user [15, 26]. The MWF beamforming scheme presented in [14, 15] is used as a baseline method for MWF beamformers that assume frontal targets. For this particular method, the noisy CPSD matrix is modeled as

$$\mathbf{C}_x = \lambda_t \boldsymbol{d}_t \boldsymbol{d}_t^H + \lambda_v \boldsymbol{\Gamma}_v, \tag{D.49}$$

where $\boldsymbol{d}_t$ is the RATF vector associated with the frontal direction. The PSDs λ_t and λ_v are replaced by ML estimates and used to implement an MWF beamformer as in Fig. D.3 during the remaining 2.5 seconds of an acoustic scene with target presence. The method is referred to as ML-FRONTAL.

4.3.2 MWF beamformer with ML PSD and target RATF estimation

The method proposed in [10] generalizes the method in [14, 15] by including ML estimation of the target RATF vector. The noisy CPSD matrix is modeled as in (D.49), but the frontal RATF vector $\boldsymbol{d}_t$ is replaced by an estimated RATF vector. The log-likelihood function is denoted as $\log f(\boldsymbol{X}; \lambda_t, \lambda_v, \boldsymbol{d}_t)$ and is parameterized by λ_t, λ_v, and $\boldsymbol{d}_t$. These parameters are estimated by solving:

$$\underset{\lambda_t, \lambda_v, \boldsymbol{d}_t \in \mathcal{D}}{\arg\max} \; \log f(\boldsymbol{X}; \lambda_t, \lambda_v, \boldsymbol{d}_t). \tag{D.50}$$

The estimated λ_t, λ_v, $\boldsymbol{d}_t$ are similarly used to implement an MWF beamformer. We refer to this method to as ML-DOA. The baseline methods have access to a generic VAD to detect speech absence where neither own voice, interfering, nor target speech are present. The generic VAD is used to initialize Γ_v. In contrast to the propose methods, the baseline methods do not exploit the own voice to assume the target absence during own voice presence.

As a reference for upper bound performance, an "oracle" MWF beamformer, where all parameters for the MWF beamformer are known, is included.

4.4 Simulation results for canteen and car noise

Beamforming performance is evaluated in terms of estimated speech intelligibility using ESTOI [52] and in terms of speech quality using PESQ [53]. Performance is reported as a function of SINR to compare the robustness towards different noise level and as a function of INR to compare the robustness against the presence of interfering speech. When evaluating the performance as a function of INR, the SINR is fixed to 0 dB to simulate a reasonable noisy acoustic scene. Similarly, when evaluating the performance as a function of SINR, the INR is chosen to be fixed at 6 dB to maintain the presence of a fairly strong interfering speaker.

The beamforming performance in canteen noise is shown in Fig. D.4 and performance for car noise is shown in Fig. D.5. By visual inspection, we see that the proposed methods i.e. ML-BD and UML perform well in the presence of an interfering speaker. At high INRs, the proposed methods outperform both ML-FRONTAL and ML-DOA significantly. This observation indicates that the proposed methods are more efficient at identifying and suppressing interfering speech due to the use of the user's own voice. Sample spectrograms of the beamformer outputs are also shown in Fig. D.6 for a visual comparison between ML-FRONTAL and UML. We only show the baseline method ML-FRONTAL as the target is located at 0° which is the best case scenario for ML-FRONTAL. ML-BD is also omitted from Fig. D.6 as it shows very similar patterns to UML. The spectrograms show that the proposed algorithms suppress the interfering speaker more efficiently than the

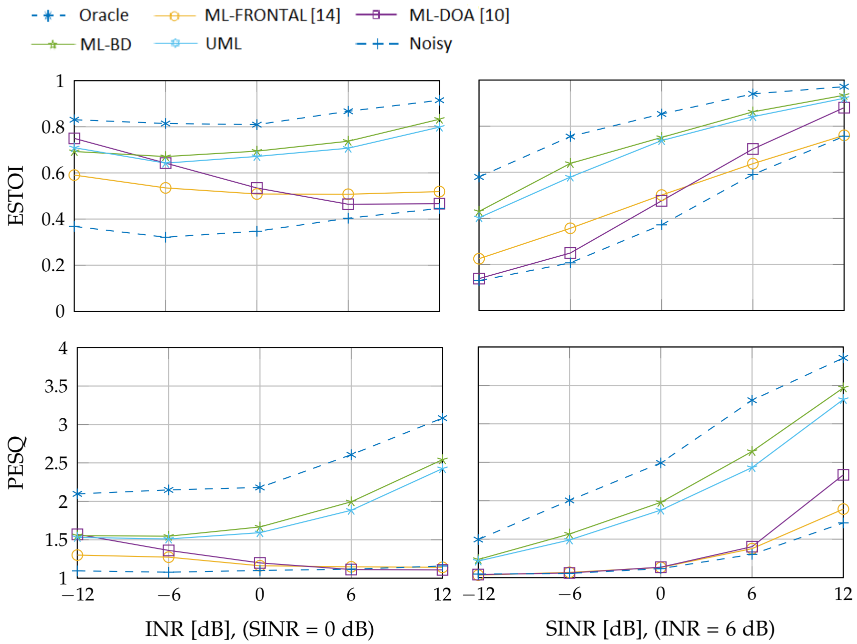


Fig. D.4: Beamforming performance in *canteen noise*.

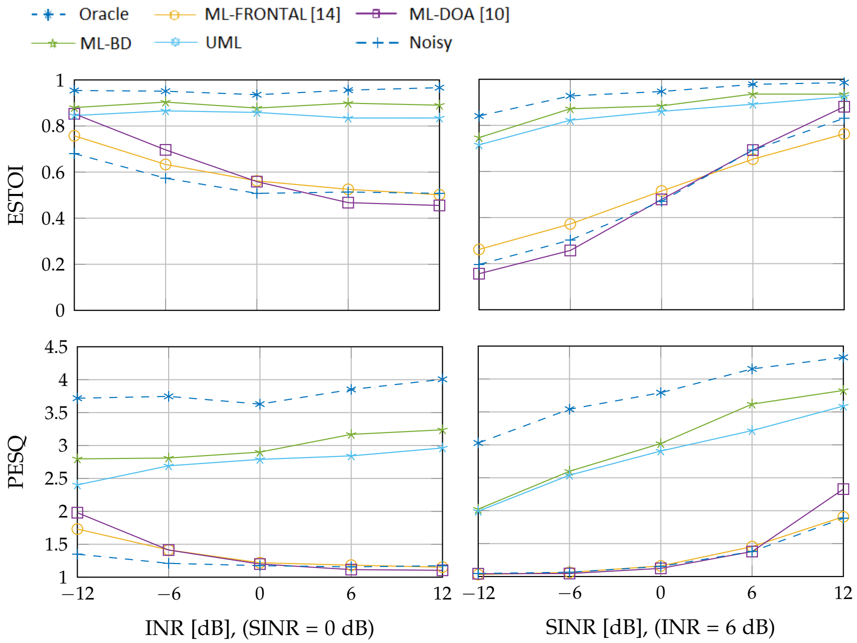
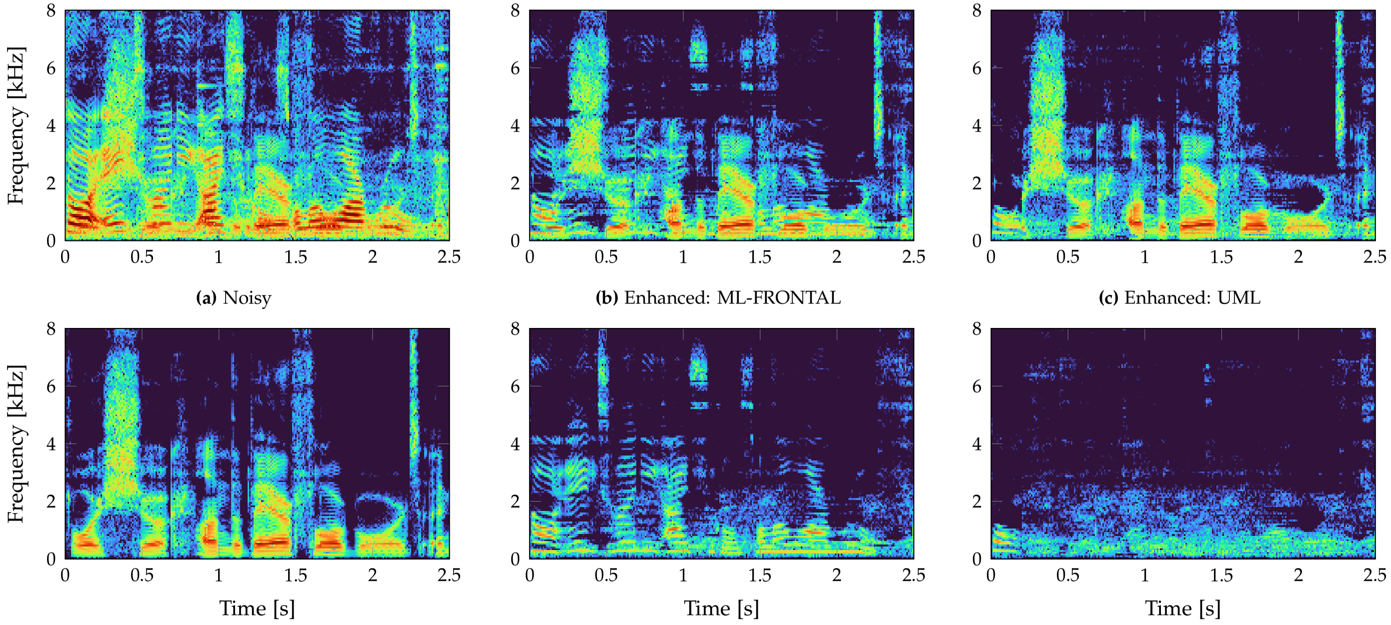


Fig. D.5: Beamforming performance in *car noise*.



(a) Noisy

(b) Enhanced: ML-FRONTAL

(c) Enhanced: UML

(d) Clean target speech

(e) Residual noise: ML-FRONTAL

(f) Residual noise: UML

Fig. D.6: Spectrograms of the noisy, clean target, and processed signals from a single realization of the experiment in Sec. 4.4. The interferer is a competing speaker and the background noise is canteen noise. The INR is set to 12 dB, and the SINR is 0 dB. The target location is in the front (0°), and the interfering speaker is located at 75°. The figures show the spectrograms of the last 2.5 seconds of an acoustic scene with target presence. Fig. D.6a and Fig. D.6d show the noisy and clean target signals at the reference microphone, respectively. Fig. D.6b and Fig. D.6e show the output of the MWF beamformer using ML-FRONTAL, where Fig. D.6b shows the processed signal and Fig. D.6e shows the processed interference-plus-noise components only (i.e. without the target). Fig. D.6c and Fig. D.6f show the output of the MWF beamformer using UML, where Fig. D.6c shows processed signal and Fig. D.6f shows the processed interference-plus-noise components only.

baseline methods while preserving the target speech. This can be seen in Fig. D.6e and Fig. D.6f where a comparison reveals that the interfering speaker is almost completely canceled using UML in contrast to ML-FRONTAL.

Another notable observation in Fig. D.4 and Fig. D.5 is that, the ML-DOA method return a very poor ESTOI and PESQ score when the INR is high. This is due to large amounts of target speech distortion as a consequence of the interfering speech mistakenly being identified as the target speech. In severe situations, e.g. when the INR is 12 dB, the performance of ML-DOA approaches the performance of the noisy signal.

ESTOI and PESQ scores of the proposed methods and the state-of-the-art methods, ML-DOA, are close at low INRs (see left panels in Fig. D.4 and Fig. D.5). To analyze these performance differences, we conduct pairwise t-tests [54] with Bonferroni corrected significance levels. The null-hypothesis is that the mean ESTOI score between two selected methods is identical for a given INR. We choose a significance level of $\alpha = 0.05$ before Bonferroni correction.

For canteen noise in Fig. D.4, we compare ML-DOA with ML-BD and UML. The pairwise t-tests reveal no significant difference at -12 dB and -6 dB INR for ESTOI and -12 dB dB INR for PESQ. For car noise in Fig. D.5, no significant difference is observed at -12 dB INR for ESTOI when comparing ML-DOA with ML-BD and UML. In terms of PESQ, the comparisons reveal that all pairwise comparisons for ML-DOA with ML-BD and UML are significant. The results for car noise, suggest that the proposed methods perform much better in comparison to the state-of-the-art methods, when the noise is approximately isotropic and stationary. A possible explanation is that detection and suppression of weak interferers at low INRs, is substantially easier in these noise fields for the proposed noise reduction systems.

We also examined the performance in situation where the target location is fixed to the front of the user (0°) and known to the beamforming systems. These situations may occur when the user steers the beamformer by rotating the head. However, we did not include these results as they lead to similar conclusions to the experiments with unknown target locations. This is due to the proposed algorithms being able to identify and suppress the interfering speaker more efficiently than the baseline methods.

In summary, the evaluation and statistical tests suggest that both proposed noise reduction systems, i.e. ML-BD and UML, have a significantly advantage over state-of-the-art methods in situations where a strong interferer is present. Additionally, we may conclude that the proposed systems, also perform on pair with state-of-the art in situations with weak interferers.

4.5 Simulation results with AIR mismatch and reverberation

In real-world scenarios, the AIRs of the RATF dictionary may not match the actual AIRs, and the microphone signals may be contaminated by reverbera-

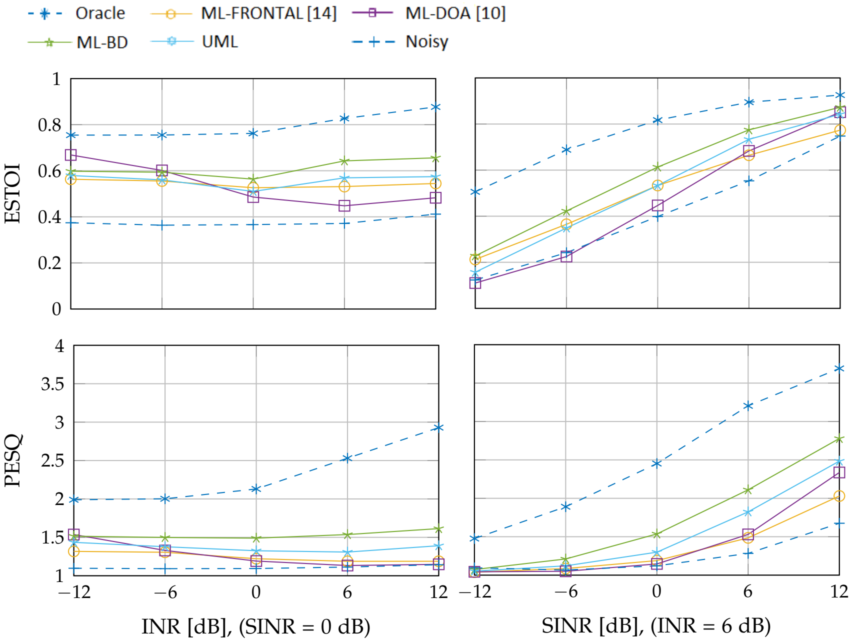


Fig. D.7: Performance in *canteen noise* and *AIR mismatch*.

tion in addition to noise. Both phenomena can potentially have a degrading impact on the beamforming performance of the proposed algorithms. To examine the robustness of the proposed algorithms, we therefore evaluate them against AIR mismatch and reverberation. We perform two experiments where the first experiment examines the robustness against AIR mismatch, and the second experiment examines the robustness against reverberation.

4.5.1 Simulation with AIR mismatch

In the first experiment, AIR mismatches can arise due to non-personalized RATF dictionaries. For example, the RATF dictionary may be measured on a different head than the HA user. To simulate such AIR mismatch, we use two sets of AIR databases fitted and measured on two arbitrarily chosen human heads. One is used to simulate the acoustic scene, and the other is used as a non-personalized RATF dictionary for the parameter estimation and MWF beamformer in the noise reduction systems.

Fig. D.7 and Fig. D.8 show the results for canteen noise and car noise, respectively. As expected, mismatches in the RATF dictionary cause performance degradations for all methods. Generally, the difference in mean ESTOI score between methods is smaller than the experiments in Sec. 4.4.

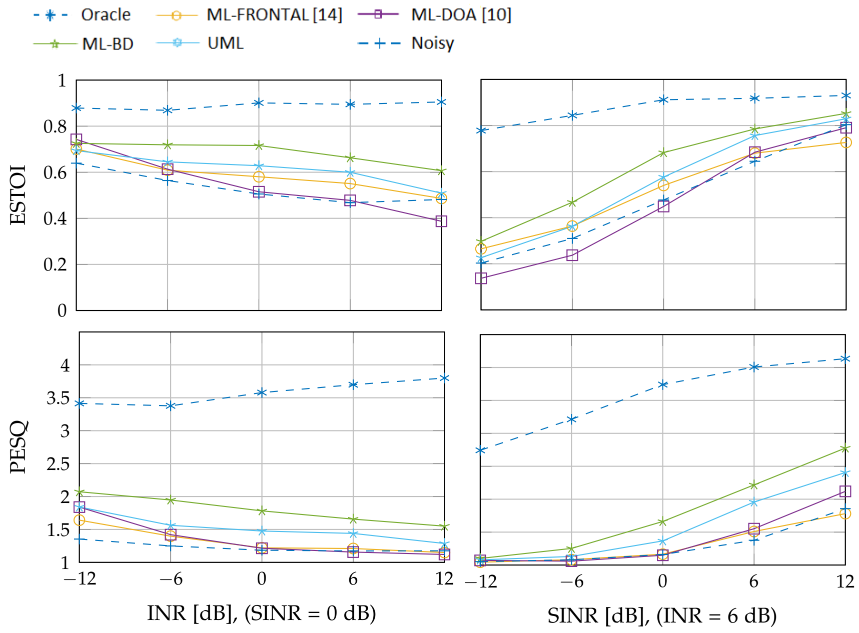


Fig. D.8: Performance in *car noise* and *AIR mismatch*.

As in Sec. 4.4, we perform pairwise t-tests with Bonferroni corrected significance levels. We compare the proposed algorithms with ML-FRONTAL and ML-DOA. Statistically significant differences are primarily observed at high INRs, where ML-BD performs better than any of the baseline methods. At lower INRs, there are no strong indications that any of the noise reduction systems perform differently than the other.

4.5.2 Simulation with reverberation

In the second experiment, we use reverberant AIRs to simulate reverberation on the target and interference sources. The AIRs are measured in a listening room with physical dimensions L x W x H = 7.9 m x 6.0 m x 3.5 m. The reverberation time in the room is approximately $T_{60} = 150$ ms. The clean target and clean interference signals are convolved with reverberant AIRs to simulate the room reverberation. The canteen and car noise already contain natural reverberation from the environment they were measured in, hence we did not convolve the reverberant AIRs with the noise. We did not have access to a reverberant own voice transfer function, and therefore used the dry own voice transfer function in the simulation. We used an RATF dictionary obtained from the dry AIRs for the noise reduction systems.

Figs. D.9 and D.10 show performance results for canteen noise and car

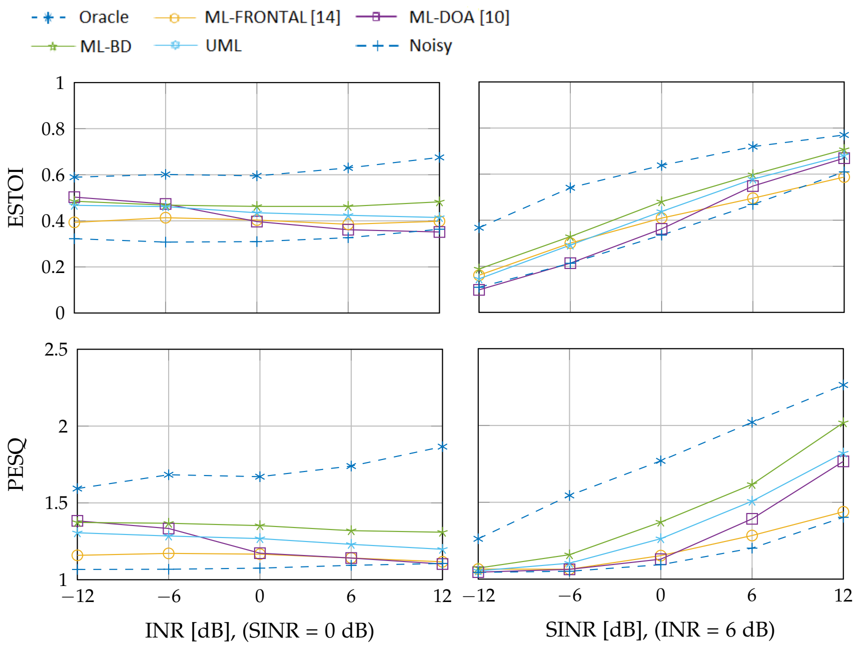


Fig. D.9: Beamforming performance in *canteen noise* and *reverb*.

noise, respectively. Generally, all noise reduction systems suffer substantial performance degradation when the target and interference signals are reverberant. In canteen noise, we measure no strong statistical difference between the noise reduction systems except for ML-BD which performs better than ML-FRONTAL and ML-DOA at 12 dB INR. In terms of PESQ, the results are more decisive and seem to favor ML-BD which is statistically significant better than the baseline methods for INRs between 0 to 12 dB. In car noise, ML-BD performs significantly better than the baseline methods between -6 dB to 12 dB INR both in terms of ESTOI and PESQ. However, UML did not perform statistically significantly differently than any of the baseline methods.

Our evaluations with AIR mismatch and reverberation seem to favor ML-BD over UML and the baseline methods. However, despite reduced performance in these situations compared to the results in Sec. 4.4, it is worth pointing out that the overall conclusion remains: The proposed methods, in particular ML-BD, perform significantly better than the baseline methods in situations with a prominent interfering speaker and perform on par with the baseline methods in the absence of an interfering speaker.

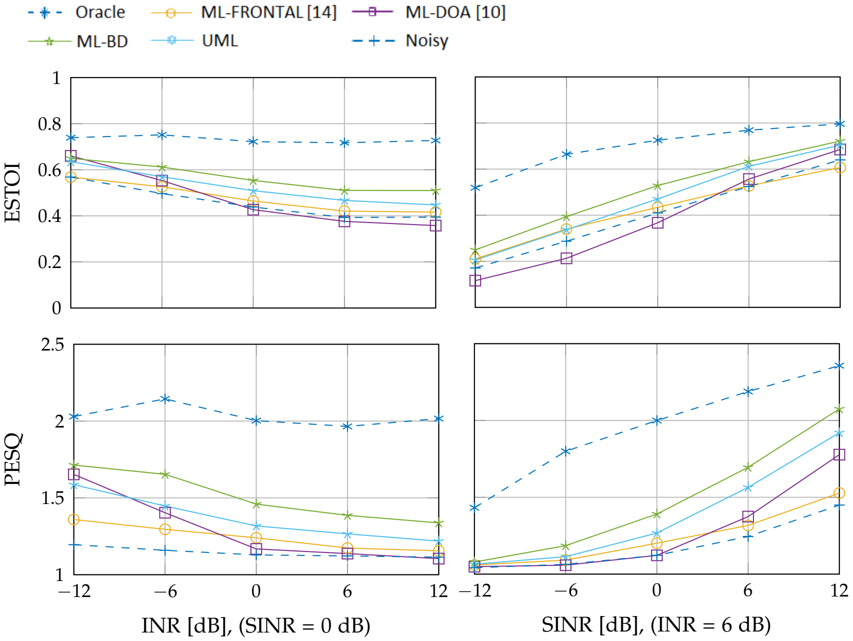


Fig. D.10: Beamforming performance in car noise and *reverb*.

5 Conclusion

In this paper, we propose multichannel noise reduction systems for hearing assistive devices (HADs). The proposed noise reduction systems can solve the problem of enhancing a target speech contaminated by noise and strong interfering speech, which is often considered difficult to solve for existing systems. We rely on the HAD user's own voice to identify interfering speech during own voice presence, but target absence. Furthermore, the multichannel Wiener filter (MWF) is used to retrieve the target speech and we propose three maximum likelihood estimation methods to estimate the target, interfering speech, and noise statistics needed for the MWF beamformer.

The proposed noise reduction systems are compared to state-of-the-art methods in terms ESTOI and PESQ to examine estimated speech intelligibility and speech quality. Simulation results indicate that the proposed noise reduction systems are able to outperform the state-of-the-art methods particularly in situations with a prominent interfering speaker.

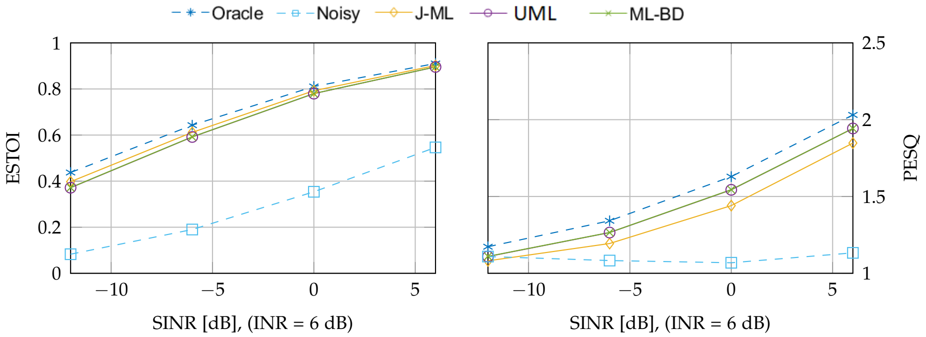


Fig. D.11: Beamforming performance in situations with target, interferer, and noise with known target RATF vector.

A Speech Enhancement Performance Evaluation of the Maximum Likelihood Estimators

In this appendix, we evaluate the speech enhancement performance of the MLEs presented in Sec. 3. The purpose of this evaluation is to show that the performances of ML-BD (Algorithm 3, and UML (Algorithm 4, see Sec. 3.2) are close to identical to the joint MLEs in 3.1 (J-ML). This result is of particular interest as the MLEs for ML-BD and UML only require a one-dimensional search over $\boldsymbol{d}_1$ and hence potentially a much lower computational complexity.

For simplicity, the target RATF vectors are assumed known and the evaluation is focused on speech enhancement performance. Although the target RATF vectors are assumed known for this evaluation, results can still give a sufficient indication on how ML-BD, UML, and J-ML compares. The comparison is made in terms of ESTOI and PESQ as a function of signal-to-interference-plus-noise ratio (SINR).

For this experiment, the setup is similar to the one presented in Sec. 4 but with the target and interferer RATF vector known to the noise reduction systems.

Fig. D.11 shows performance in terms of for ESTOI and PESQ as a function of SINR with the INR fixed to 6 dB. The unprocessed signal and the output of an oracle MWF with known target and noise statistics are also evaluated to indicate lower and upper performance bounds. Each performance score per SINR is averaged over 50 realization of acoustic scenes.

From Fig. D.11, we see that J-ML, ML-BD, and UML perform almost identically without large differences. Furthermore, they perform close to the oracle MWF. We observe that J-ML performs slightly better than UML and ML-BD in terms of ESTOI but slightly worse in terms of PESQ. This is possibly due to UML and ML-BD having a slightly more aggressive noise suppression than J-ML which translates to marginally higher PESQ score at the cost of speech distortion and lower ESTOI score.

An evaluation of the runtime of the algorithms, showed that the real-time factors were 9.46 for UML, 4.98 for ML-BD, and 153.46 for J-ML. We see that J-ML has a significantly higher real-time factor compared to UML and ML-BD. This is partly due to the choice of the grid resolution used in J-ML as a high grid resolution will increase the real-time factor. Lowering the grid resolution will decrease the real-time factor but can result in performance degradation. It should also be noted that the implementation of the algorithms are not code optimized, but the real-time factors can still give a rough indication of computational complexity when comparing the proposed methods.

Because of the insignificant difference between J-ML, ML-BD and UML, we omit J-ML in the evaluation in Sec. 4 due to its high computational complexity. However, if J-ML was chosen to be included in the evaluation in Sec. 4, similar performance to ML-BD and UML would be expected.

References

[1] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer handbook of speech processing*. Berlin ; London: Springer, 2008, oCLC: ocn190966783.

[2] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*, ser. Springer topics in signal processing. Berlin: Springer, 2008, no. 1.

[3] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, 2001.

[4] 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*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[5] X. Zhang and Y. Jia, "A Soft Decision Based Noise Cross Power Spectral Density Estimation for Two-Microphone Speech Enhancement Systems," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. 813–816.

[6] M. Souden, Jingdong Chen, J. Benesty, and S. Affes, "Gaussian Model-Based Multichannel Speech Presence Probability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1072–1077, Jul. 2010.

[7] M. Souden, J. Chen, J. Benesty, and S. Affes, "An Integrated Solution for Online Multichannel Noise Tracking and Reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2159–2169, Sep. 2011.

[8] M. Taseska and E. A. P. Habets, "Non-Stationary Noise PSD Matrix Estimation for Multichannel Blind Speech Extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2017.

[9] ——, *MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator*, International Workshop on Acoustic Signal Enhancement, Ed. Piscataway, N.J.: IEEE, 2012, oCLC: 835582169.

[10] 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, Mar. 2018.

[11] P. Hoang, Z.-H. Tan, J. M. de Haan, and J. Jensen, "Joint maximum likelihood estimation of power spectral densities and relative acoustic transfer functions for acoustic beamforming," in *2021 IEEE ICASSP.* IEEE, 2021.

[12] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[13] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sep. 2003. [Online]. Available: http://ieeexplore.ieee.org/document/1223596/

[14] 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)*, Aug 2012, pp. 295–299.

[15] J. Jensen and M. S. Pedersen, "Analysis of beamformer directed single-channel noise reduction system for hearing aid applications," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 5728–5732.

[16] A. Kuklasinski, S. Doclo, T. Gerkmann, S. Holdt Jensen, and J. Jensen, "Multichannel PSD estimators for speech dereverberation - A theoretical and experimental comparison," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 91–95.

[17] A. Kuklasinski, 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, Sep. 2016.

[18] S. Braun, A. Kuklasinski, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and Comparison of Late Reverberation Power Spectral Density Estimators," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1056–1071, Jun. 2018.

[19] Hao 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, Apr. 1995.

[20] O. Schwartz, S. Gannot, and E. A. P. Habets, "Joint maximum likelihood estimation of late reverberant and speech power spectral density in noisy environments," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Shanghai: IEEE, Mar. 2016, pp. 151–155.

[21] S. Braun and E. P. Habets, "A multichannel diffuse power estimator for dereverberation in the presence of multiple sources," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–14, 2015.

[22] 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 *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2020, p. xx.

[23] O. Thiergart, M. Taseska, and E. A. P. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," vol. 22, no. 12, pp. 2182–2196.

[24] S. Chakrabarty and E. A. P. 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, no. 1, pp. 145–160, Jan. 2018.

[25] 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)*. Ottawa, ON, Canada: IEEE, Nov. 2019, pp. 1–5.

[26] A. Kuklasinski and J. Jensen, "Multichannel Wiener Filters in Binaural and Bilateral Hearing Aids — Speech Intelligibility Improvement and Robustness to DoA Errors," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, pp. 8–16, Feb. 2017.

[27] S. Markovich, S. Gannot, and I. Cohen, "Multichannel Eigenspace Beamforming in a Reverberant Noisy Environment With Multiple Interfering Speech Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[28] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 544–548.

[29] A. J. Sørensen, M. Fereczkowski, and E. N. MacDonald, "Task Dialog By Native-Danish Talkers In Danish And English In Both Quiet And Noise," Mar. 2018, publisher: Zenodo. [Online]. Available: https://zenodo.org/record/1204951

[30] N. Das, A. Bertrand, and T. Francart, "EEG-based auditory attention detection: boundary conditions for background noise and speaker positions," *Journal of Neural Engineering*, vol. 15, no. 6, p. 066017, Dec. 2018.

[31] V. Best, E. Roverud, T. Streeter, C. R. Mason, and G. Kidd, "The Benefit of a Visually Guided Beamformer in a Dynamic Speech Task," *Trends in Hearing*, vol. 21, p. 233121651772230, Dec. 2017.

[32] 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, p. 233121651881438, Jan. 2018.

[33] L. V. Hadley, W. O. Brimijoin, and W. M. Whitmer, “Speech, movement, and gaze behaviours during dyadic conversation in noise,” *Scientific Reports*, vol. 9, no. 1, p. 10451, Dec. 2019.

[34] H. Sacks, E. A. Schegloff, and G. Jefferson, “A Simplest Systematics for the Organization of Turn-Taking for Conversation,” *Language*, vol. 50, no. 4, p. 696, Dec. 1974.

[35] A. J. Munch Sørensen, M. Fereczkowski, and E. N. MacDonald, “Effects of noise and l2 on the timing of turn taking in conversation,” in *7th International Symposium on Auditory and Audiological Research (ISAAR)*, Aug 2019.

[36] P. C. Loizou, *Speech enhancement: theory and practice*. Boca Raton, Fla.: CRC Press, 2013, oCLC: 958799095.

[37] Y. Avargel and I. Cohen, “On multiplicative transfer function approximation in the short-time fourier transform domain,” vol. 14, no. 5, pp. 337–340. [Online]. Available: http://ieeexplore.ieee.org/document/4154719/

[38] S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[39] A. J. M. Sørensen, T. Lunner, and E. N. MacDonald, “Timing of turn taking between normal-hearing and hearing-impaired interlocutors,” in *7th International Symposium on Auditory and Audiological Research (ISAAR)*, Aug 2019.

[40] H. Dillon, *Hearing aids*, 2nd ed. Sydney: Boomerang Press [u.a.], 2012.

[41] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. de Ruiter, K.-E. Yoon, and S. C. Levinson, “Universals and cultural variation in turn-taking in conversation,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10 587–10 592, Jun. 2009.

[42] H. Akaike, “A new look at the statistical model identification,” vol. 19, no. 6, pp. 716–723.

[43] P. Stoica and Y. Selen, “Model-order selection,” vol. 21, no. 4, pp. 36–47. [Online]. Available: http://ieeexplore.ieee.org/document/1311138/

[44] Y. Laufer, B. Laufer-Goldshtein, and S. Gannot, “ML Estimation and CRBs for Reverberation, Speech, and Noise PSDs in Rank-Deficient Noise Field,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 619–634, 2020.

[45] A. Moore, J. M. de Haan, M. Pedersen, D. Brookes, P. Naylor, and J. Jensen, “Personalized signal-independent beamforming for binaural hearing aids,” *Journal of the Acoustical Society of America*, vol. 145, pp. 2971–2981, 2019.

[46] D. Marquardt, S. Doclo, V. Hohmann, and R. Martin, *Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques*, 1st ed. München: Verlag Dr. Hut, 2016.

[47] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, “Multichannel Signal Enhancement Algorithms for Assisted Listening Devices: Exploiting spatial diversity using multiple microphones,” *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[48] Pei Chee Yong, S. Nordholm, and Hai Huyen Dam, "Effective Binaural Multi-Channel Processing Algorithm for Improved Environmental Presence," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2012–2024, Dec. 2014.

[49] "TIMIT: acoustic-phonetic continuous speech corpus." Philadelphia, Pa., 1993.

[50] P. Minnaar, S. F. Albeck, C. Simonsen, B. Søndersted, S. Oakley, and J. Bennedbæk, "Reproducing real-life listening situations in the laboratory for testing hearing aids," *Journal of The Audio Engineering Society*, 2013.

[51] E. Hadad, S. Doclo, and S. Gannot, "The binaural LCMV beamformer and its performance analysis," vol. 24, no. 3, pp. 543–558.

[52] 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, Nov. 2016.

[53] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2. Salt Lake City, UT, USA: IEEE, 2001, pp. 749–752.

[54] S. M. Ross, *Introduction to probability and statistics for engineers and scientists*, 4th ed. Amsterdam ; Boston: Academic Press/Elsevier, 2009, oCLC: ocn255903275.

Paper E

The Minimum Overlap-Gap Algorithm for Speech Enhancement

Poul Hoang, Zheng-Hua Tan, Jan Mark de Haan, and Jesper Jensen

The paper has been published in
IEEE Access, vol. 10, pp. 14698-14716, 2022.

Abstract

In this paper, we propose a novel speech enhancement paradigm which can effectively solve the problem of retrieving a desired speech signal in a multi-talker environment. The proposed speech enhancement paradigm involves a three-step procedure consisting of separation, ranking, and enhancement. First, a speech separation system – which could be a conventional spatial filter bank or more advanced separation systems – separates mixtures of speech signals captured by microphones into speech signals from candidate speakers. Next, novel ranking algorithms – proposed in this paper – are applied to determine the talker-of-interest amongst the separated speech signals. Finally, the speech signal of the talker-of-interest is estimated as a linear combination of the separated signals, whose weights are determined by the ranking algorithms. We propose ranking algorithms, which exploit turn-taking patterns between conversational partners in order to determine the talker-of-interest amongst competing speakers. Unlike some existing solutions, our ranking algorithms do not require access to additional sensors, e.g., EEG electrodes, cameras, etc., but only rely on microphone signals. Specifically, the proposed algorithms rank the separated speech signals based on the probability of speech overlaps and gaps with the user's own voice. The speech signal with highest ranking is the talker with minimum probability of speech overlap and gap with the user's own voice. The proposed ranking algorithms are shown highly effective at determining the talker-of-interest, since conversational partners, i.e., the user and the talker-of-interest, behaviorally avoid speech overlaps and gaps. We evaluate the proposed speech enhancement paradigm in two practical hearing aid related applications, where the objective is to enhance a speech signal of a conversational partner in a multi-talker environment. The results of the evaluation demonstrate that the proposed speech enhancement systems in both applications significantly outperform conventional speech enhancement systems.

1 Introduction

The cocktail party problem is often regarded as one of the most difficult situations any speech enhancement system may encounter. The complexity in the acoustic environment is vast and its composition may include multiple competing speakers, music, reverberation, and noise. Solving the cocktail party problem, i.e., the speech signal(s)-of-interest, i.e. the *target signal(s)*, is commonly the goal for speech enhancement systems in applications such as hearing assistive devices (HADs) and speaker-phone systems. The enhancement system in these applications is crucial for many humans as they rely on the aid to communicate more efficiently in noisy environments, particularly when competing speech and noise become dominant. However, achieving effective suppression of loud competing speech and noise remains a remarkably difficult problem to solve even with the most recent state-of-the-art speech enhancement systems.

The problem of interest in this paper is to enhance a conversational partner, i.e., the talker-of-interest, in the presence of multiple competing speakers and noise. The competing speakers are obviously undesired and can potentially be louder than the conversational partner. In order to be able to enhance the conversational partner in such multi-speaker situations, any enhancement system faces the question: *"Who is the user listening and talking to?"*. The traditional speech enhancement paradigm for single-microphone systems involves estimation of temporal statistics of the conversational partner and noise for implementation of linear filters. For the multiple microphone case, beamformers are often implemented and typically require estimation of the direction-of-arrival (DOA) and/or spatio-temporal statistics of the conversational partner and noise [1–5]. However, the presence of multiple speakers poses a great estimation challenge, since the conversational partner and competing speakers are often indistinguishable from an acoustic perspective. In worst case scenarios, speech enhancement algorithms might in fact suppress the conversational partner and enhance competing speech. For example, current DOA estimators such as SRP-PHAT [6], maximum likelihood [7, 8], and deep learning-based DOA estimators [9], are not able to robustly handle a conversational partner in a multi-speaker environment, without additional a priori information on the conversational partner's location or voice activity. Consequently, these DOA estimators will indecisively switch between the candidate speakers as being the conversational partner leading to an enhanced signal of unacceptable intelligibility and quality.

In this paper, we propose a speech enhancement paradigm that can efficiently identify the conversational partner in a multi-speaker environment and retrieve the desired speech signal. The paradigm is described through the three step-procedure as shown in Fig. E.1.

In the first step, the noisy microphone signals are fed into a speech separation system to separate mixtures of speech signals into individual source signals/components, which we refer to as candidate speakers. Example of speech separation systems include beamforming systems which separate speech using beams steered in different directions, or deep neural network (DNN) based separation algorithm e.g. uPit [10, 11] and TasNet [12, 13]. Some applications allow microphones to be placed physically on the candidate talker, in which case the separation is trivial.

In the second step, the separated candidate speakers are ranked according to their likelihood of being the conversational partner. Existing ranking strategies may involve additional sensor signals and prior knowledge to support the decision of estimating the conversational partner channel after speech separation. As an example, beamforming systems in HADs often rank, or simply assume, the frontal speaker as the most likely conversational partner [2]. However, unfortunately, the user may not always face the con-

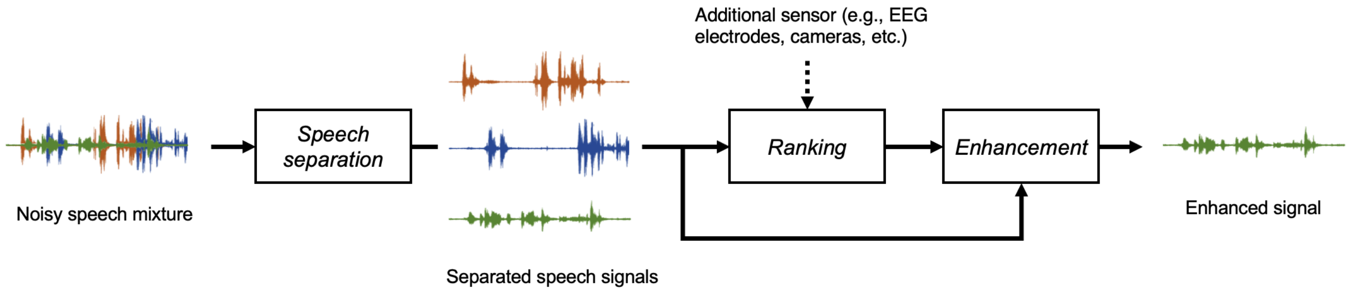


Fig. E.1: Speech enhancement paradigm for enhancement of a conversational partner in multi-talker situations.

versational partner in all situations, which leads to a loss of performance. Alternatively, estimated candidate speakers may be ranked using EEG-signals, retrieved from EEG-electrodes placed on the scalp of the user, to detect the user's attention on conversational partner, EOG-signals to estimate eye-gaze from in-ear electrodes, and cameras to track eye-movements and estimate eye-gaze [14–18]. While these signals have the potential to support the decision of determining the talker-of-interest, they require additional sensors which increase equipment cost, increase wearing inconvenience, and likely also increase computational cost and power consumption. These trade-offs make acquisition of EEG, EOG, and visual signals impractical for small devices such as HADs where power consumption and wearing inconvenience matters for the end user.

Finally, the last step involves enhancement of the conversational partner signal. The enhanced signal is formed as a linear combination of the separated speech signals where the weights are determined from the speaker ranking algorithm.

Additionally, we propose a method to the ranking problem in Fig. E.1, which does not require additional sensors apart from microphones. A microphone-only system is highly desirable from a practical perspective, both due to the cost of additional sensors and from a algorithm complexity perspective. Our method is based on exploiting the conversational behavior between the user and the conversational partner. We use the so-called *turn-taking* behavior between two conversational partners [19–23] to rank the candidate speakers according to the talker which is most likely the user's conversational partner. Specifically, the method analyses the speech overlaps and gaps between the user and a candidate speaker to quantify turn-taking, and then selects the speaker with minimum probability of speech overlap and gap with the user as the talker-of-interest.

This paper is organized as follows. Sec. II introduces the basics of conversation and turn-taking behavior and its potential use in ranking the candidate speakers and determining the talker-of-interest. In Sec. III, we derive our minimum overlap-gap (MOG) method and propose statistical models of speech overlap and gap behavior between a user and a conversational partner. Based on the statistical model, we propose an extension, namely, the Bayesian MOG (BMOG) algorithm. In Sec. IV, we describe the estimation of the parameters for the proposed statistical models of turn-taking from datasets of real conversations. We use the statistical models to derive the theoretical performance of the (B)MOG algorithm. Finally in Sec. V, we evaluate the performance of the proposed speech enhancement paradigm and (B)MOG algorithms in two speech enhancement applications.

2 Speech interaction in conversations

Determining the talker-of-interest and ranking the candidate speakers are needed for the proposed speech enhancement paradigm and can be an extremely difficult problem to solve. We propose to rank the candidate talker using the turn-taking model presented in [19]. Human interaction is a group of behavioral mechanisms that are taught since childhood to use when engaged in conversations to structurize exchange of information [20]. *Addressing* and *turn-taking* mechanisms found in conversations are examples of interaction management between conversational partners [20].

Addressing is used by the addressee, i.e., the talking person, to indicate whom the speech is directed to. For example, humans may use gaze, gestures, and speech to indicate the conversational partner. Strong indicators are typically head pose and eye-gaze which potentially could be utilized by speech enhancement systems to determine the talker-of-interest [20]. However, measuring the head pose and eye-gaze would usually require additional sensors such as accelerometers, electrodes, or cameras for applications such as HADs.

The turn-taking mechanism is another type of interaction management and is universal across cultures and languages. Turn-taking is used to structurize conversations. Turn-taking is used to coordinate who should speak next and when, to ensure that only one speaker is talking at a time, while others remain silent. Conversational partners may occasionally overlap and gap in conversations, but these are often of short duration such as when the listener responds the talker by saying "yes" or "uh hm" [19, 24]. In order to maintain rapid turn-taking, listeners also try to predict the end of a speech utterance of their conversational partner to minimize speech overlap and gap.

We use the turn-taking model in [19] to model a) the conversational behavior between the user and the conversational partner, and b) the voice activity pattern between the user and a competing speaker. We may describe a) and b) in terms of four voice activity states.

$\mathcal{S}_1$: Conversational partner/competing speaker speaks while user is silent.

$\mathcal{S}_2$: User speaks while conversational partner/competing speaker is silent.

$\mathcal{S}_3$: Conversational partner/competing speaker and user are both silent.

$\mathcal{S}_4$: Conversational partner/competing speaker and user are both speaking.

State $\mathcal{S}_1$ and $\mathcal{S}_2$ are *turns* of the conversational partner/competing speaker and user, respectively, while state $\mathcal{S}_3$ is referred to as *gaps* or pauses and state $\mathcal{S}_4$ is referred to as *overlaps* [19, 24, 25]. In [24] it was found that 77% of all recorded conversations between a user and a conversational partner were in state $\mathcal{S}_1$ or $\mathcal{S}_2$, 19.2 % belonged to state $\mathcal{S}_3$, and 3.8% were in state $\mathcal{S}_4$. For a user and a competing speaker, the proportion of time spent in each state, may be argued to be significantly different compared to a user and a conversational partner. Specifically, a larger proportion of speech overlaps and gaps would be expected between a user and a competing speaker, since the turn-taking mechanisms would not exist. In addition, when the conversational partners are exposed to noisy environments, the proportion of time spend in each state changes, with overlaps becoming more common as the noise level increases. In [26] it was found that in very noisy environments the proportion of time spent in state $\mathcal{S}_1$ or $\mathcal{S}_2$ decreased from 70% at a noise level of 54 dB SPL to 50% at 78 dB SPL, $\mathcal{S}_3$ increased from 8% at 54 dB SPL to 24% at 78 dB SPL, and for $\mathcal{S}_4$ from approximately 22% at 54 dB SPL to 26% at 78 dB SPL, where normal conversation breaks down. A possible reason for these observations is that conversational partners insist on maintaining rapid turn-taking during conversations, resulting in poorer timing and prediction of their partners end of a turn, hence increasing the proportion of overlaps and gaps.

These results indicate that humans rely significantly on turn-taking to maintain normal conversations even in very noisy environments as conversations otherwise would break down. Although speech overlaps and gaps become more frequent in noisy environments, these conversational patterns remain robust in noisy condition and the turn-taking patterns between a user and a conversational partner would presumably still be significantly different than the voice activity patterns between a user and a competing speaker. Hence, in the following we propose a method that exploits these turn-taking patterns to determine the talker-of-interest in a multi-talker environment.

3 The Minimum Overlap-Gap Algorithm

In this section, we derive the proposed algorithm for ranking the candidate speakers using expected turn-taking patterns. Our primary focus in this section is the task of ranking the speakers by their likelihood of being the conversational partner, i.e. the *Ranking* block in Fig. E.1.

First, the speech separation system separates mixtures of speech signals into individual discrete time-sequences $s_i(n)$, $i = 0, 1, ..., I$, where $s_0(n)$ is the user's own voice, and the remaining $s_i(n)$, $i = 1, ..., I$ are the I candidate speech signals. For each speech signal $s_i(n)$, a binary output $\alpha_i(n)$ of voice activity detector (VAD) is defined as

$$\alpha_i(n) = \begin{cases} 1, \text{ if } s_i(n) \text{ contains speech at time } n \\ 0, \text{ if } s_i(n) \text{ contains no speech at time } n. \end{cases} \tag{E.1}$$

where $\alpha_0(n)$ is the user's own voice VAD (OVAD). We assume that $\alpha_i(n)$ represents the actual speech activity of the various speech sources – as we demonstrate in Sec. 5.5, the proposed ranking and enhancement system work well, even when $\alpha_i(n)$ are estimated from sources separated with a practical beamforming system. Fig. E.2 shows an example of VAD outputs a real conversation between the user and the conversational partner in addition to two competing speakers. The outputs of the VADs are used to determine the voice activity state between the user and a candidate speaker i i.e.

$\mathcal{S}_1$: if $\alpha_0(n) = 0$ and $\alpha_i(n) = 1$.

$\mathcal{S}_2$: if $\alpha_0(n) = 1$ and $\alpha_i(n) = 0$.

$\mathcal{S}_3$: if $\alpha_0(n) = 0$ and $\alpha_i(n) = 0$.

$\mathcal{S}_4$: if $\alpha_0(n) = 1$ and $\alpha_i(n) = 1$.

As discussed in Sec. 2, conversational partners use turn-taking when engaged in a conversation. A consequence of the turn-taking mechanism is that conversational partners avoid speech overlaps and gaps, i.e., they minimize the proportion of time spent in state $\mathcal{S}_3$ and $\mathcal{S}_4$. In the following, we use this observation to propose an algorithm exploiting this to determine the talker-of-interest.

3.1 Minimum probability of speech overlap and gap

The paradigm presented in Fig. E.1 ranks the candidate speakers using their voice activity patterns, prior to the enhancement. The proposed algorithm selects the speaker with minimum probability of speech overlap and gap related to the user's own voice as the talker-of-interest. We refer to this method as the Minimum Overlap-Gap (MOG) algorithm. Let $A_i(n)$, $i = 0, 1, ..., I$ be Bernoulli random variables of the VADs and let $\alpha_i(n)$, $i = 0, 1, ..., I$ be their corresponding realizations. The probability of a speech overlap and speech gap between the user's own voice and candidate speaker i, is denoted as $P_{A_0 A_i}(\alpha_0(n) = 1, \alpha_i(n) = 1)$ and $P_{A_0 A_i}(\alpha_0(n) = 0, \alpha_i(n) = 0)$, respectively. The MOG algorithm selects the speaker with minimum probability of overlaps and gaps:

$$\hat{i}_{\text{MOG}}(n) = \arg\min_{i \in \{1,...,I\}} \sum_{k=0}^{1} P_{A_0 A_i}(\alpha_0(n) {=} k, \alpha_i(n) {=} k), \tag{E.2}$$

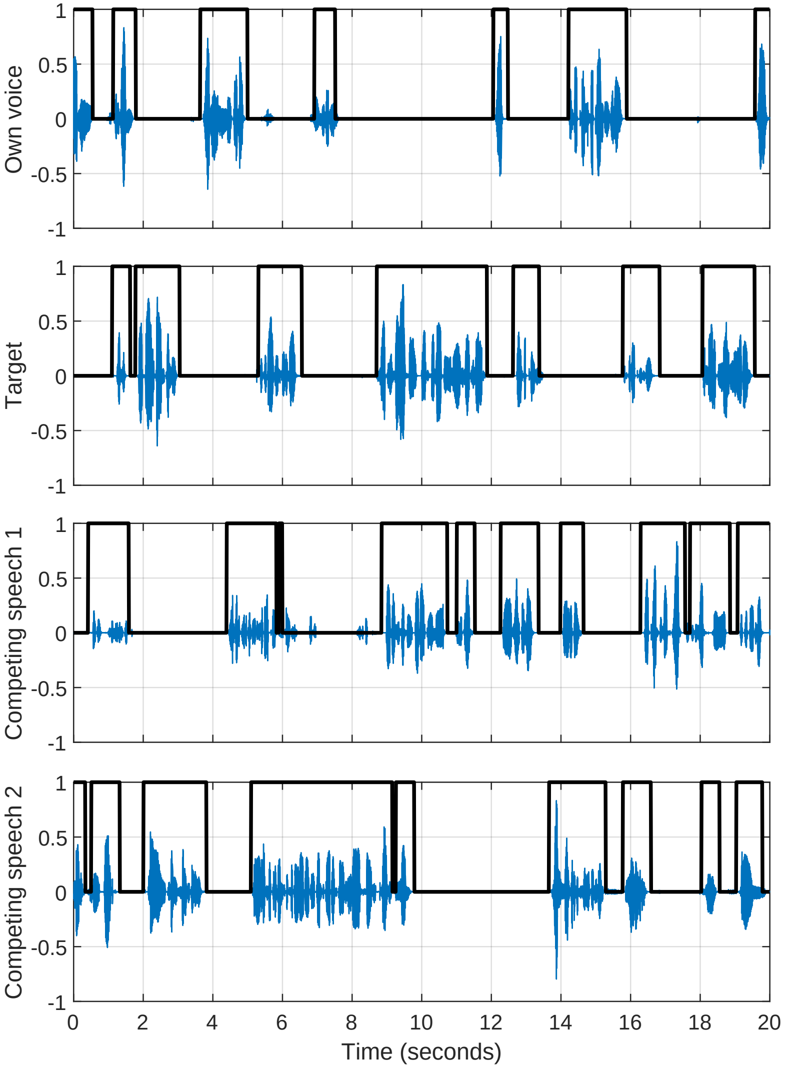


Fig. E.2: Speech signals with VAD outputs plotted on top of real conversations between the own voice and the conversational partner. The two top plots are conversations between a user and a conversational partner (target). The two bottom plots are competing speakers unrelated to the conversation between the user and the conversational partner.

where $\hat{\imath}_{\text{MOG}}(n)$ is the estimated conversational partner channel index and minimizing the cost in (E.2) is equivalent to minimizing the occurrences of the states $\mathcal{S}_3(n,i)$ and $\mathcal{S}_4(n,i)$, i.e. gaps and overlaps, respectively. Alternatively, the optimization problem may also be formulated as maximizing the probability of mutual exclusion between the binary sequences $\alpha_0(n)$ and $\alpha_i(n)$ (see Appendix A.1) i.e.

$$\hat{\imath}_{\text{MOG}}(n) = \underset{i\in\{1,\ldots,I\}}{\arg\max} \sum_{k=0}^{1} P_{A_0 A_i}(\alpha_0(n) = k, \alpha_i(n) = 1-k). \tag{E.3}$$

Furthermore, as shown in Appendix A.2, solving (E.3) is also equivalent to finding the candidate speaker index, which maximizes the mean-square-error (MSE) between the user own-voice VAD (OVAD) and candidate speaker's VAD, i.e.,

$$\hat{\imath}_{\text{MOG}}(n) = \underset{i\in\{1,\ldots,I\}}{\arg\max} \mathbb{E}\left[(A_0(n) - A_i(n))^2\right]. \tag{E.4}$$

Note that the optimization problem is bounded in $[0,1]$ as $A_0(n)$ and $A_i(n)$ are binary values. The definition of the MOG algorithm in (E.4) is a maximization of the MSE between two binary sequences and is thus computationally simple.

3.2 Bayesian MOG for probability-based speaker ranking

Probability-based ranking of the candidate speakers can provide additional insights compared to the MOG algorithm in (E.4) which only identifies a single talker-of-interest. In this approach, a posterior probability is estimated for each candidate speaker which quantifies the uncertainty of a candidate speaker being the talker-of-interest. This information can be particularly useful for a speech enhancement system, for example, to adjust the level of noise suppression.

3.2.1 Statistical models of the sum of squared error

One approach to derive posterior probabilities for each candidate speaker, is to statistically model the distribution of overlaps and gaps between 1) a user and a conversational partner, and 2) a user and a competing speaker, and then use Bayes theorem to estimate the probabilities. To model the statistical distribution of overlaps and gaps, we introduce the random variable $Z_i(n)$, which represents the squared error between the own voice VAD and the candidate speaker VAD:

$$Z_i(n) = (A_0(n) - A_i(n))^2, \tag{E.5}$$

where $Z_i(n)$, $A_0(n)$, and $A_i(n)$ are Bernoulli random variables. The random variable $Z_i(n)$ quantifies if $A_0(n)$ and $A_i(n)$ are overlapping or gapping, i.e., when $Z_i(n) = 0$, or not. We define the sum of squared errors (SSEs) as

$$\Phi_i(n) = \sum_{k=n-N+1}^{n} Z_i(k), \tag{E.6}$$

where N is the number of past observations of $Z_i(n)$ upon which the decision will be based. The SSE quantifies the total amount of observed overlaps and gaps within N observations. Low SSEs indicate large amounts of overlaps and gaps between $A_0(n)$ and $A_i(n)$, whereas high SSEs indicate small amounts of overlaps and gaps. It is also worth noting that N is related to the *integration time*, which we define as

$$T_{\text{int}} = N \cdot f_{\text{s,vad}}, \tag{E.7}$$

where $f_{\text{s,vad}}$ is the sampling frequency of the VADs. The integration time T_{int}, is easier interpreted than N as it also accounts for the sampling frequency of the VADs.

In order to model the distribution of $\Phi_i(n)$, we use that $\Phi_i(n)$ is a sum of N Bernoulli distributed random variables. For independently and identically distributed $Z_i(n)$, then $\Phi_i(n)$ follows a binomial distribution. However, preliminary experiments with natural conversations have shown that observations of $\Phi_i(n)$ have a higher dispersion than a binomial distribution, hence the binomial distribution is too restrictive to explain the observations. Instead, we have found that a beta-binomial distribution provides a significantly better fit than the binomial distribution. The beta-binomial distribution is parameterized by N and two shaping parameters γ and β and its probability mass function (PMF) is given as

$$p_{\Phi_i}(\phi_i; \gamma, \beta, N) = \binom{N}{\phi_i} \frac{B(\phi_i + \gamma, N - \phi_i + \beta)}{B(\gamma, \beta)}. \tag{E.8}$$

$B(\cdot, \cdot)$ is the Beta-function parameterized by γ and β, and

$$\binom{N}{\phi_i} = \frac{N!}{\phi_i!(N - \phi_i)!}, \tag{E.9}$$

denotes the binomial coefficient. In the remaining part of the paper, we use the PMF notation $p_{\Phi_i}(\phi_i; \gamma, \beta, N) \triangleq p(\Phi_i = \phi_i; \gamma, \beta, N)$ for brevity. First, we statistical model Φ_i when the user is engaged in a conversation and afterwards model Φ_i for the interaction between the user and a competing speaker. Hence, the first statistical distribution $p_{\Phi_i}(\phi_i; \gamma_t, \beta_t, N)$ is fitted to observations of SSEs between a user and conversational partner engaged in a conversation, where the subscript t denotes that the shaping parameters are related to the true conversational partner. The second distribution

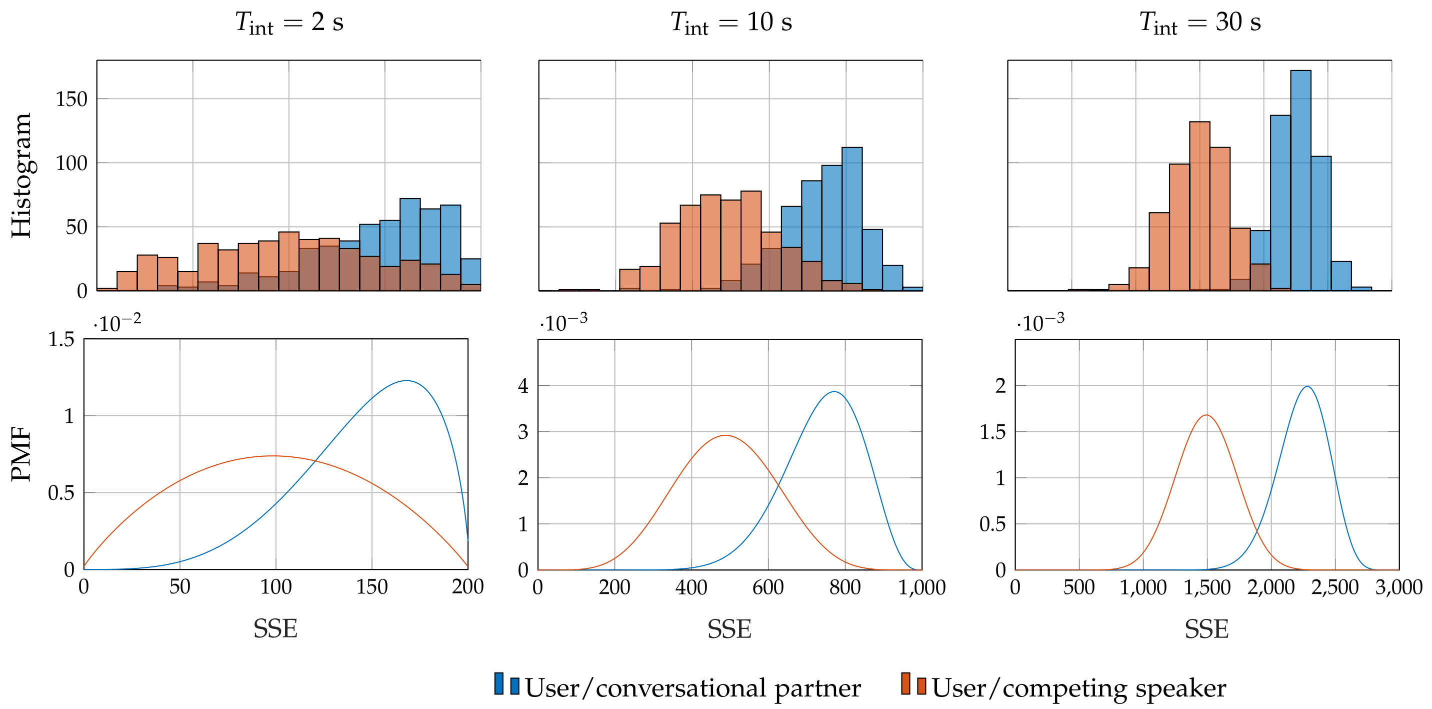


Fig. E.3: Histograms of $K = 500$ observations of SSEs from real conversations are shown in the upper plots. The bottom plots show beta-binomial distributions where the parameters are found from the observations. In all plots the blue color denotes the conversational partner and the red color denotes the competing speaker.

$p_{\Phi_j}(\phi_j;\gamma_v,\beta_v,N)$ is fitted to observations of SSEs between a user and competing speakers, where the user and competing speaker are engaged in different conversations.

3.2.2 Hypothesis testing

In order to estimate probabilities for each candidate speaker, we define I hypotheses

> $\mathcal{H}_i$: Candidate speaker i is the conversational partner, and the remaining $I-1$ speakers are competing speakers for $i=1,...,I$.

Under $\mathcal{H}_i$, it follows that Φ_i is distributed according to $p_{\Phi_i}(\phi_i;\gamma_t,\beta_t,N)$ and Φ_j for $j\neq i$, is distributed according to $p_{\Phi_j}(\phi_j;\gamma_v,\beta_v,N)$, i.e.

$$\begin{aligned}\Phi_i &\sim p_{\Phi_i}(\phi_i|\mathcal{H}_i)\triangleq p_{\Phi_i}(\phi_i;\gamma_t,\beta_t,N)\\ \Phi_j &\sim p_{\Phi_j}(\phi_j|\mathcal{H}_i)\triangleq p_{\Phi_j}(\phi_j;\gamma_v,\beta_v,N)\text{ for } j\neq i.\end{aligned}\tag{E.10}$$

For each time n, we observe realizations, ϕ_k, of Φ_k for all $k=1,...,I$. Assuming that Φ_k, are statistically independent, the likelihood function conditioned on $\mathcal{H}_i$ is given by

$$\begin{aligned}p_{\Phi_1,...,\Phi_I}(\phi_1,...,\phi_I|\mathcal{H}_i) &= \prod_{k=1}^{I} p_{\Phi_k}(\phi_k|\mathcal{H}_i)\\ &= p_{\Phi_i}(\phi_i|\mathcal{H}_i)\prod_{j\in\mathcal{I}\backslash i} p_{\Phi_j}(\phi_j|\mathcal{H}_i)\\ &= p_{\Phi_i}(\phi_i;\gamma_t,\beta_t,N)\prod_{j\in\mathcal{I}\backslash i} p_{\Phi_j}(\phi_j;\gamma_v,\beta_v,N),\end{aligned}\tag{E.11}$$

where $\mathcal{I}=\{1,...,I\}$ is the set of candidate speaker indices, and $\mathcal{I}\backslash i$ denotes the set of competing speakers under hypothesis $\mathcal{H}_i$, i.e. $\mathcal{I}$ excluding the element i. Using Bayes theorem, the posterior probability of $\mathcal{H}_i$ is given by

$$\begin{aligned}P(\mathcal{H}_i|\phi_1,...,\phi_I) &= \frac{P(\mathcal{H}_i)p_{\Phi_1,...,\Phi_I}(\phi_1,...,\phi_I|\mathcal{H}_i)}{p_{\Phi_1,...,\Phi_I}(\phi_1,...,\phi_I)}\\ &= \frac{P(\mathcal{H}_i)p_{\Phi_i}(\phi_i;\gamma_t,\beta_t,N)\prod_{j\in\mathcal{I}\backslash i} p_{\Phi_j}(\phi_j;\gamma_v,\beta_v,N)}{\sum_{k=1}^{I} P(\mathcal{H}_k)p_{\Phi_k}(\phi_k;\gamma_t,\beta_t,N)\prod_{l\in\mathcal{I}\backslash k} p_{\Phi_l}(\phi_l;\gamma_v,\beta_v,N)},\end{aligned}\tag{E.12}$$

where $P(\mathcal{H}_i)$ is the prior probability of the conversational partner being channel i. This method of estimating the posterior probability is referred to as the Bayesian MOG algorithm.

4 Parameter Estimation from Conversational Speech Database

To implement the Bayesian MOG (BMOG) algorithm in (E.12), the shaping parameters γ_t, β_t, γ_v, and β_v for the statistical models $p_\Phi(\phi;\gamma_t,\beta_t,N)$ and $p_\Phi(\phi;\gamma_v,\beta_v,N)$ are estimated from speech databases containing real conversations. Next, using the estimated statistical models, we analyze the theoretical speaker ranking performance of the MOG algorithm in terms of misclassification rate.

4.1 Setup and speech database

4.1.1 Conversational speech database

In order to estimate the shaping parameters γ_t, β_t, γ_v, and β_v, we use the speech database in [27] which contains dialogues between 19 pairs of native-Danish talkers recorded during a task dialog experiment. The participants had normal hearing and were coupled into pairs to collaborate solving DiapixUK tasks [28]. DiapixUK is *spot-the-difference* tasks where partners were given two almost identical cartoon pictures with a few differences. The participants were not allowed to view each others pictures, but had to solve the DiapixUK task by exchanging descriptions of their picture through verbal communication. The partners were placed in different sound booths and communicated through headphones and head-worn microphones. The experiment had four test conditions: 1) native language (Danish) and no noise, 2) native language (Danish) and babble noise, 3) second language (English) and no noise, and 4) native language (English) and babble noise.

4.1.2 Voice activity detection

The presence of speech in the signal $s_j(n)$ is determined by a binary VAD which produces an output sequence $\alpha_j(n) = \{0, 1\}$ for either of the speakers in the dialogue. For voice activity detection, we used the robust voice activity detector (rVAD) proposed in [29] applied to the essentially noise-free dialogue recordings. The input to rVAD is L consecutive samples of $s_i(l)$ with sampling frequency f_s. The output of rVAD is a sequence of N voice activity decisions $\alpha_i(n)$ at sampling frequency $f_{\text{s,vad}} = 100$ Hz. Version rVAD2.0 was used in this paper and can be found in [30].

4.2 Parameter Estimation for the Beta-Binomial Distribution

We used the speech data set recorded in a quiet condition and in Danish language for parameter estimation. The speech signals are sampled at 22.05

kHz but downsampled to 16 kHz for compatibility with rVAD. In order to collect observations of the SSEs for a user and a conversational partner, we used the following procedure:

1. Select an integration time T_{int}, e.g. $T_{\text{int}} = 10$ seconds, where the integration time is related to N by $N = \frac{T_{\text{int}}}{f_{\text{s,vad}}}$.
2. Divide the speech signals into non-overlapping segments with length T_{int}.
3. Apply the rVAD on the speech signals of conversational partners.
4. Compute the SSE from the VAD outputs using (E.6).

To gather observations of the SSE between the user and a competing speaker, we perform a similar procedure, but instead of choosing a matching conversational pair, we randomly choose two non-conversational speakers to form a pair and compute the SSE. Histograms and fitted beta-binomial distribution of SSEs between a user and a conversational partner, as well as a user and a competing speaker are shown in Fig. E.3 for different integration times. Clearly and as expected, the separability between $p_\Phi(\phi; \gamma_t, \beta_t, N)$ and $p_\Phi(\phi; \gamma_v, \beta_v, N)$ becomes greater as T_{int} becomes larger. The dispersion of SSE becomes smaller for both distributions as T_{int} increases. The shaping parameters γ_t, β_t, γ_v, and β_v are functions of T_{int}.

4.2.1 Parameter estimation of γ_t, β_t, γ_v, and β_v given T_{int}

For each T_{int}, the parameters γ_t, β_t, γ_v, and β_v are estimated using observations of the SSEs. The observations of SSEs are denoted as $\phi_t^{(k)}$ and $\phi_v^{(k)}$, $k = 1, ..., K$, respectively, where the subscript t denotes the SSE between the user and conversational partner, v is the SSE between the user and a competing speaker, and K is the total number of observations. Each observation of $\phi_t^{(k)}$ and $\phi_v^{(k)}$ are assumed independent. The parameters are found numerically using maximum likelihood estimation such that

$$\hat{\gamma}_t(T_{\text{int}}), \hat{\beta}_t(T_{\text{int}}) = \arg\max_{\gamma_t, \beta_t} \prod_{k=1}^{K} p_\Phi(\phi_t^{(k)}; \gamma_t, \beta_t, N)$$

and

$$\hat{\gamma}_v(T_{\text{int}}), \hat{\beta}_v(T_{\text{int}}) = \arg\max_{\gamma_v, \beta_v} \prod_{k=1}^{K} p_\Phi(\phi_v^{(k)}; \gamma_v, \beta_v, N),$$

where $T_{\text{int}} = N \cdot f_{\text{s,vad}}$. In order to provide simple models of γ_t, β_t, γ_v, and β_v, scatter plots of estimated shaping parameters for different T_{int} are shown in Fig. E.4. We choose to describe the shaping parameters using a power model. Let $\tilde{h}(T_{\text{int}}; a, b)$ be the general form of a power model with parameters a and b:

$$\tilde{h}(T_{\text{int}}; a, b) = a \cdot T_{\text{int}}^{b}. \tag{E.13}$$

This model can be useful for implementation of the BMOG algorithm for any T_{int}, and to facilitate the theoretical performance evaluation of the MOG algorithm in Sec. 4.3. To estimate the parameters a and b of the power model, we use a non-linear least squares procedure with the general form of

$$\hat{a}, \hat{b} = \arg\min_{a,b} \sum_{j=1}^{J} \left[\tilde{h}\left(T_{\text{int}}^{(j)}; a, b\right) - \hat{h}\left(T_{\text{int}}^{(j)}\right)\right]^2, \tag{E.14}$$

where $\hat{h}\left(T_{\text{int}}^{(j)}\right)$ is an estimated shaping parameter, i.e., either $\hat{\gamma}_t(T_{\text{int}})$, $\hat{\beta}_t(T_{\text{int}})$, $\hat{\gamma}_v(T_{\text{int}})$, or $\hat{\beta}_v(T_{\text{int}})$, and J is the total number of data points for each ML estimated shaping parameter. We minimize (E.14) numerically. The estimated power model parameters are summarized in Table E.1. Fig. E.4 shows that the fitted power models provide an excellent fit to the ML estimated shaping parameters as a function of T_{int}.

	$\tilde{\gamma}_t(\cdot; \hat{a}, \hat{b})$	$\tilde{\beta}_t(\cdot; \hat{a}, \hat{b})$	$\tilde{\gamma}_v(\cdot; \hat{a}, \hat{b})$	$\tilde{\beta}_v(\cdot; \hat{a}, \hat{b})$
$\hat{a}$	2.5091	0.8522	0.7736	0.8057
$\hat{b}$	0.7879	0.7817	0.9727	0.9681

Table E.1: Power model parameters for modeling the estimated shaping parameters of the beta-binomial distributions.

4.3 Theoretical Performance of the MOG algorithm

In this section, we analyze the theoretical performance of the MOG algorithm and compare it with performance achieved through simulations. Two quantities that have a significant impact on the performance of the MOG algorithm, are the number of candidate speakers, I, and the integration time T_{int}. Increasing the number of candidate speakers will increase the solution search space, hence increase the a priori risk of choosing a wrong candidate as the target speaker. Decreasing the integration time T_{int} will lead to higher variance in the estimation of the SSEs in (E.6).

The misclassification rate is used to measure the performance of the MOG algorithm and is defined as the probability of classifying a competing speaker as the conversational partner. We denote the misclassification rate as $P(E = 1; I, T_{\text{int}})$ where $E \in \{0, 1\}$ is a Bernoulli random variable with $E = 1$

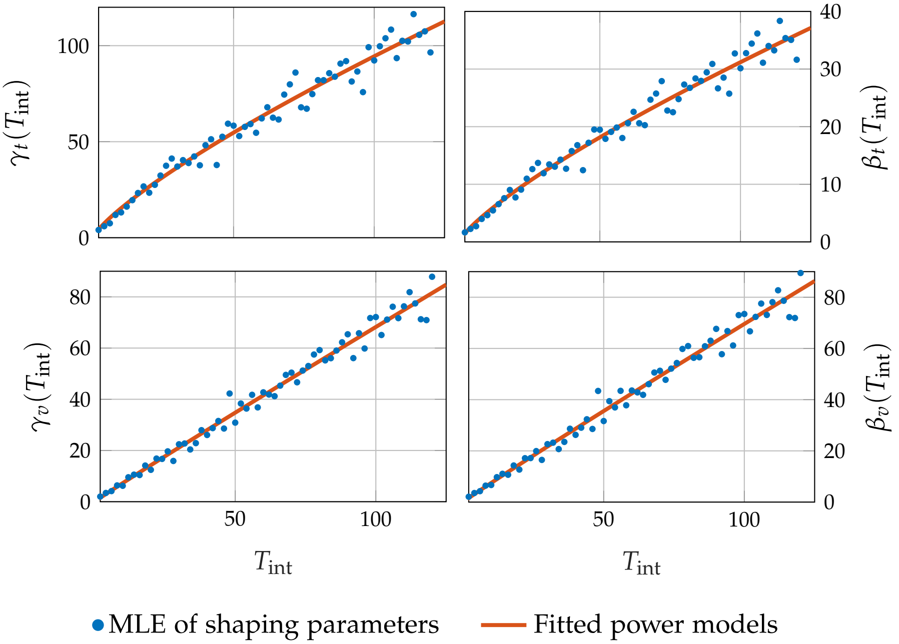


Fig. E.4: Shaping parameters of the beta-binomial distribution as a function of T_{int}. The blue data points are obtained from maximum likelihood estimation for different T_{int}. The red curves are fitted power models on the blue data points.

representing a misclassification. To derive an expression for the misclassification rate, we define $P(\Phi_t > \Phi_{v,1}, ..., \Phi_t > \Phi_{v,I-1})$ as the probability of correct classification, where Φ_t denotes the SSE between the user and conversational partner, and $\Phi_{v,j}$ is the SSE between the user and the j'th competing speaker. The misclassification rate is then given by

$$P(E = 1; I, T_{\text{int}}) = 1 - P(\Phi_t > \Phi_{v,1}, ..., \Phi_t > \Phi_{v,I-1}).$$

In Appendix A.3, we show that the misclassification rate of the MOG algorithm can be expressed as

$$\begin{aligned} &P(E = 1; I, T_{\text{int}}) = \\ &1 - \sum_{\phi=1}^{N} p_\Phi(\phi; \gamma_t, \beta_t, N) P_\Phi^{I-1}(\phi - 1; \gamma_v, \beta_v, N), \end{aligned} \tag{E.15}$$

where $P_\Phi(\phi - 1; \gamma_v, \beta_v, N)$ is the cumulative distribution of $p_\Phi(\phi - 1; \gamma_v, \beta_v, N)$ which is given by

$$P_\Phi(\phi - 1; \gamma_v, \beta_v, N) = \sum_{\kappa=0}^{\phi-1} p_\Phi(\kappa; \gamma_v, \beta_v, N). \tag{E.16}$$

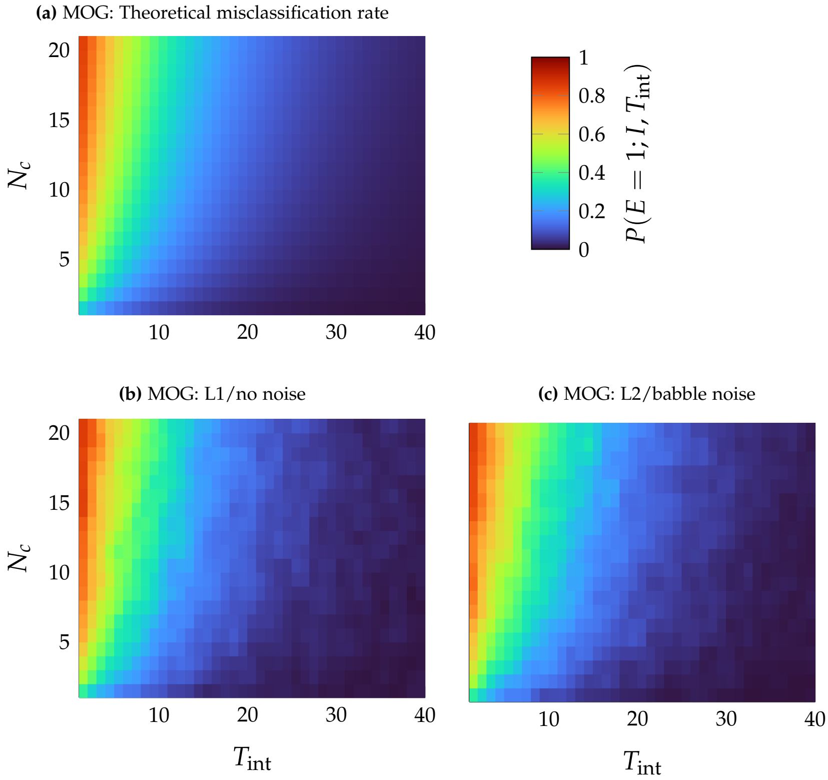


Fig. E.5: Probability of misclassifying the conversational partner as a function of the number of competing speakers N_c and integration time T_{int}. The theoretical performance of the MOG algorithm is shown in Fig. E.5a. The simulated MOG performance using the datasets L1/no noise and L2/babble noise is shown in Fig. E.5b and Fig. E.5c.

For verification, we compare the theoretical misclassification rate given by (E.15) with the misclassification rate achieved with the MOG algorithm in simulations as seen in Fig. E.5. From Fig. E.5b, we clearly see a close match between the theoretical and simulated misclassification rates, where the conversational partners are speaking in Danish without any noise stimuli. Likewise, a close match between the theoretical and simulated misclassification rate can be seen in Fig. E.5b, where the conversational partners are speaking in English (second language) with babble noise as noise stimuli. The close match indicates that the fitted statistical models are able to generalize to unseen conditions.

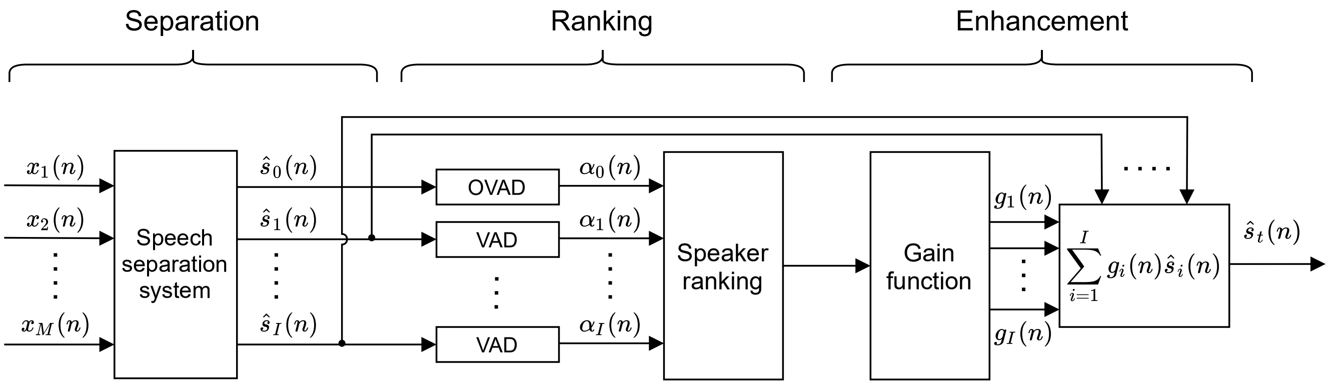


Fig. E.6: The proposed speech enhancement system consists of speech separation, speaker ranking, and enhancement. The output of the speaker ranking algorithm is for example, $\hat{i}_{\mathrm{MOG}}(n)$ or $P(\mathcal{H}_i|\phi_1, ..., \phi_I)$.

5 Evaluation in Speech Enhancement Applications

In this section, we demonstrate the use of MOG and BMOG for solving problem of enhancing a conversational partner in a multi-talker environment, using the speech enhancement paradigm of Fig. E.1. In particular, we use MOG/BMOG to rank the candidate speakers according to how likely they are to be the conversational partner. In Secs. 5.1 and 5.2, we outline the practical implementation of the MOG and BMOG algorithms and in Sec. 5.3, we present the reference/baseline speaker ranking methods that will be used in our experiments. In Secs. 5.4 and 5.5, we demonstrate the use of the proposed speech enhancement systems in two different applications for HADs.

5.1 Speech enhancement system using speaker ranking

Fig. E.6 shows an example of the speech enhancement paradigm of Fig. E.1 employing multiple microphones. In many situations, the microphone signals consist of a mixture of speech signals (including target and potential competing speakers) and noise from the environment. The unprocessed microphone signals are denoted as $x_m(n)$ for $m = 1, ..., M$, where M is the number of microphones and n is the discrete-time index. Let $\boldsymbol{x}(n) = [x_1(n), ..., x_M(n)]^T$ be the noisy microphone signals stacked in a vector, which is processed by a speech separation system. The speech separation system separates the microphone signals into estimated speech signals $\hat{\boldsymbol{s}}(n) = [\hat{s}_0(n), \hat{s}_1(n), ..., \hat{s}_I(n)]^T$. Next, voice activity detection is applied to each of the separated signals, $\hat{s}_i(n)$, $i = 1, ..., I$. A speaker ranking algorithm, e.g., MOG or BMOG, ranks the conversational partner by assigning a ranking score to each candidate speaker. Finally, in the example system in Fig. E.6, the enhancement of the conversational partner is achieved simply as a linear combination of the separated speech signals $\hat{\boldsymbol{s}}(n)$. The weights are found using a gain function which maps the ranking score to a gain value for each separated speech signal. A straightforward gain function for the MOG algorithm, is to set the gain to a value of '1' to the estimated conversational partner channel, and a value of $0 < g_{\min} < 1$ for the remaining channels, i.e.,

$$g_j(n) = \begin{cases} 1, & \text{if } j = \hat{i} \\ g_{\min}, & \text{otherwise.} \end{cases} \tag{E.17}$$

where $\hat{i}$ is the estimated channel of the conversational partner. It might occur that a competing speaker is estimated as being the conversational partner which can lead to severe loss in speech enhancement performance. It can also disrupt an ongoing conversational between a user and a conversational partner if the speaker ranking algorithm suddenly changes the estimated conversational partner. To increase the robustness, a minimum gain $g_{\min}$ can be applied such that a small amount of speech from all candidate speakers are always let through. Likewise, $g_{\min}$ can be made as a function of n, such that $g_{\min} = 1$ in the initial phase of a conversation, and gradually decreases towards a minimum value when the conversation has been established.

Another approach, specifically for the BMOG algorithm, is to use the estimated posterior probabilities as weights for the linear combination such that

$$g_j(n) = \max\left(g_{\min}, P(\mathcal{H}_i|\phi_1, ..., \phi_I)\right). \tag{E.18}$$

A potential advantage of the posterior probability as a gain function is similar to that of introducing $g_{\min} > 0$ in (E.17): It reduces perceptual switching artifacts and limits the effect of target loss in case of misclassification. For both approaches, the estimated conversational partner signal is

$$\hat{s}_t(n) = \sum_{i=1}^{I} g_j(n)\hat{s}_i(n). \tag{E.19}$$

5.2 Implementation of the MOG and BMOG algorithms

In order to implement the MOG algorithm in (E.4), we estimate the MSE as the average square-error between $\alpha_0(n)$ and $\alpha_i(n)$ over integration time T_{int}. The MOG estimate of the conversational partner index then becomes

$$\hat{i} = \arg\max_{i \in \{1,...,I\}} \sum_{k=n-N+1}^{n} \left(\alpha_0(k) - \alpha_i(k)\right)^2. \tag{E.20}$$

Implementation of the BMOG algorithm is a two-step procedure. First, the shaping parameters are computed for beta-binomial distributions given T_{int} using (E.13) and TABLE E.1, which may be done offline. Secondly, the posterior probabilities $P(\mathcal{H}_i|\phi_1, ..., \phi_I)$ are computed. To do so, the likelihood function in (E.11) is computed in the logarithmic domain for numerical stability. For this purpose, we first define the variable ψ_i as:

$$\psi_i \triangleq P(\mathcal{H}_i) p_{\Phi_i}(\phi_i; \gamma_t, \beta_t, N) \prod_{j \in \mathcal{I} \setminus i} p_{\Phi_j}(\phi_j; \gamma_v, \beta_v, N). \tag{E.21}$$

The natural logarithm of ψ_i is

$$\begin{aligned} \ln \psi_i &= \ln P(\mathcal{H}_i) + \ln p_{\Phi_i}(\phi_i; \gamma_t, \beta_t, N) \\ &\quad + \sum_{j \in \mathcal{I} \setminus i} \ln p_{\Phi_j}(\phi_j; \gamma_v, \beta_v, N). \end{aligned} \tag{E.22}$$

Substituting (E.21) into (E.12) gives

$$P(\mathcal{H}_i|\phi_1, ..., \phi_I) = \frac{\psi_i}{\sum_{k=1}^{I} \psi_k}. \tag{E.23}$$

Using the logarithm function on both sides yields

$$\ln P(\mathcal{H}_i|\phi_1,...,\phi_I) = \ln\psi_i - \ln\left(\sum_{k=1}^{I}\psi_k\right), \tag{E.24}$$

where

$$\ln\left(\sum_{k=1}^{I}\psi_k\right) = \ln\psi_1 + \ln\left(1+\sum_{j=2}^{I} e^{(\ln\psi_j - \ln\psi_1)}\right). \tag{E.25}$$

The posterior probability can be found by inserting (E.22), (E.22), and (E.25) into (E.24) and applying the exponential function $\exp(\cdot)$ to (E.24). The implementation of the BMOG algorithm is summarized in Algorithm 5.

Algorithm 5 Implementation of the BMOG algorithm.

```
Input: $\alpha_i(n)$ for $i = 0,1,...,I$. Set the parameters $T_{\text{int}}$, $N = \left\lfloor \frac{T_{\text{int}}}{f_{\text{s,vad}}} \right\rfloor$, and
    $P(\mathcal{H}_i)\forall i$.
Compute the shaping parameters $\hat{\gamma}_t(T_{\text{int}})$, $\hat{\beta}_t(T_{\text{int}})$, $\hat{\gamma}_v(T_{\text{int}})$, $\hat{\beta}_v(T_{\text{int}})$ using (E.13) and TABLE E.1.
Compute the SSEs using (E.6) to obtain $\phi_i(n)$ for all $i$.
Compute the log-likelihoods
    $\ln p_{\Phi_i}(\phi_i(n); \hat{\gamma}_t(T_{\text{int}}), \hat{\beta}_t(T_{\text{int}}), N)$,
    $\ln p_{\Phi_i}(\phi_i(n); \hat{\gamma}_v(T_{\text{int}}), \hat{\beta}_v(T_{\text{int}}), N)$,
    for all $i$ using (E.8).
Compute $\ln\psi_i(n)$ in (E.22) for all $i$.
Compute the log posterior probabilities from (E.24).
Use the exponential function $\exp(\cdot)$ on (E.24) to obtain the posterior probability in (E.23).
```

5.3 State-of-the-art methods for speaker ranking

The idea of using turn-taking to detect conversations between two speakers has been explored in [31–33] but was not used in the context of enhancing a conversational partner of a user as presented in Fig. E.1. In [31], the presence of a conversation between two speakers was quantified using mutual information between the user's and candidate speakers' voice activity sequences. The normalized cross-correlation function was later proposed as a quantifier of conversations in [32]. Both methods can be compared to the MOG/BMOG algorithms in a fair manner, since all methods require access to VAD sequences for each speaker and they return a cost that can be used for ranking the candidate speakers.

5.3.1 Maximum mutual information[31]

The mutual information method is based on finding the candidate speaker that maximizes the mutual information between the user's and candidate speaker's voice activity sequences

$$\hat{\imath}_{\text{MMI}} = \arg\max_{i\in\{1,\dots,I\}} \sum_{k=0}^{1}\sum_{j=0}^{1} P_{A_0A_i}(\alpha_0(n)=k,\alpha_i(n)=j) \\ \times \log \frac{P_{A_0A_i}(\alpha_0(n)=k,\alpha_i(n)=j)}{P_{A_0}(\alpha_0(n)=k)P_{A_i}(\alpha_i(n)=k)},$$

where all joint and marginal probabilities are sample estimates obtained from $\alpha_i(n)$ over integration time T_{int}. One problem with the MMI algorithm is situations where the numerator or denominator of the logarithmic function becomes zero. These situations might occur if the integration time is short, e.g., 2 seconds, as there is a risk that the user or candidate speaker i might be silent within the period of time. In the evaluations, we removed results where the numerator or denominator of the MMI algorithm becomes zero.

5.3.2 Normalized cross-correlation[32]

Similarly, the normalized cross-correlation (NCC) method is here used to detect the presence of a conversational partner. The optimization problem of NCC is formulated as

$$\hat{\imath}_{\text{NCC}} = \arg\max_{i\in\{1,\dots,I\}} \frac{1-\min\limits_{p\in[r1,r2]} R_{0,i}(p)}{2}, \tag{E.26}$$

where $R_{0,i}(p)$ is the normalized cross-correlation between A_0 and A_i at lag p. $r1$ and $r2$ are search region bounds for the lag p. We set p equal to zero in our evaluation.

5.3.3 Speaker ranking performance

We examine the speaker ranking performance between the proposed MOG algorithm against MMI and NCC. The performance is reported in terms of misclassification rate as a function of the number of competing speakers N_{c} and integration time T_{int}. We use speech signals from [27] for the performance evaluation. Specifically, we use the subset of the data set containing

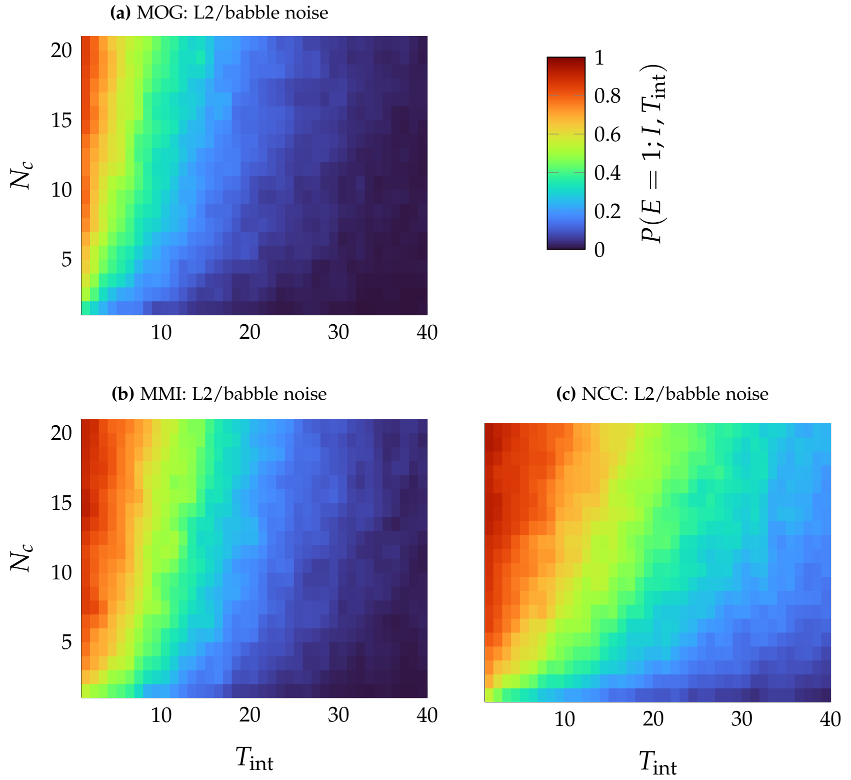


Fig. E.7: Misclassification rate as a function of the number of competing speakers N_c and integration time T_{int} for MOG (proposed), MMI, and NCC. The database used for this evaluation is from [27] with the subset where all speakers were talking in their second language (L2) in babble noise.

2-person conversations in second language English (L2) in babble noise. The speech signals are segmented into segments of length T_{int}. For each T_{int}, one 2-person conversation is randomly selected to constitute the user's own voice and the user's conversational partner. A number of N_c arbitrarily chosen speakers from the data set are selected to constitute the competing speakers. Fig. E.7 shows the misclassification rate $P(E = 1; I, T_{\mathrm{int}})$ as a function of T_{int} and the number of competing speakers $N_c = I - 1$ for each ranking algorithm. A comparison between MOG, MMI, and NCC shows that the misclassification rate is significantly lower for the MOG algorithm compared to the MMI and NCC, particularly, when 1) the integration time is short, and/or 2) there is a large number of competing speakers. At long integration times, e.g. 40 sec, the difference between the algorithms is smaller. However, the MOG algorithm consistently performs better than the MMI and NCC algorithms.

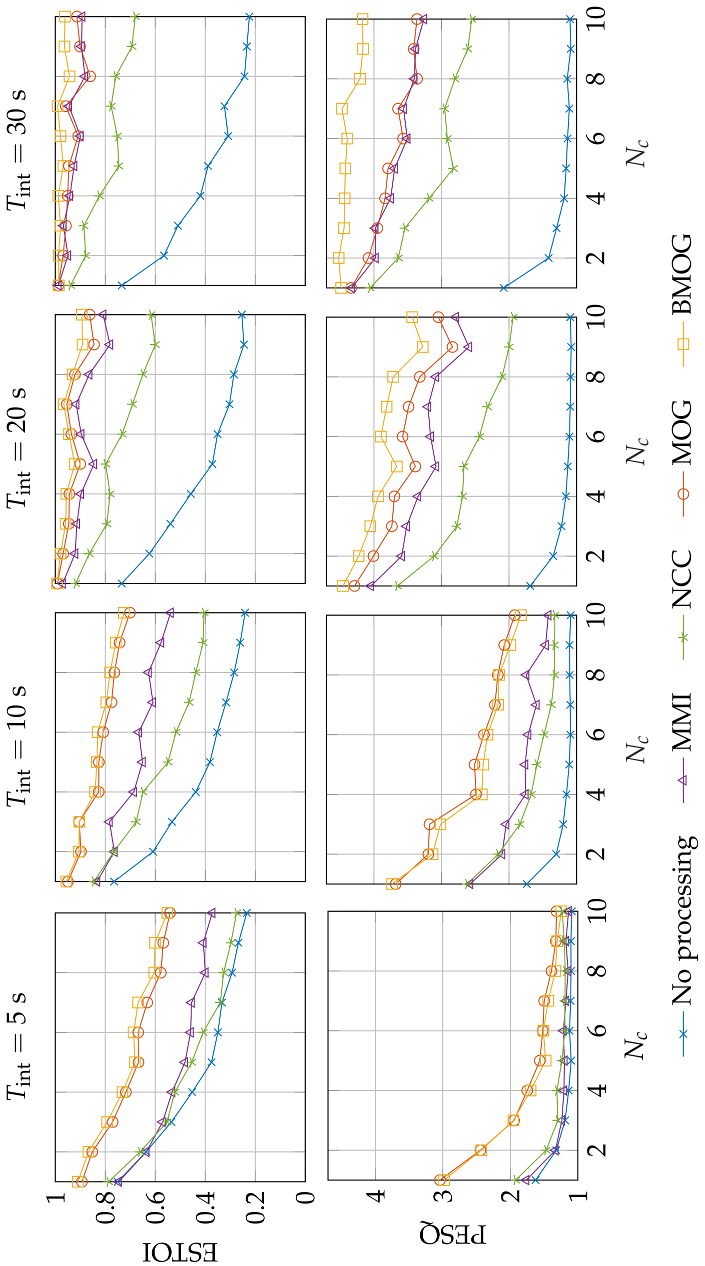


Fig. E.8: Averaged ESTOI and PESQ scores as a function of the number of competing speakers N_c and integration time T_{int}.

5.4 Application 1: Wireless hearing aid network

In this section, we demonstrate the use of the proposed (B)MOG based speech enhancement paradigm, cf. Fig. E.1 in a hearing aid (HA) application, in which the HAs of several users are wirelessly connected. The basic idea is that multiple HA users can distribute their own voice signal to the other users' HAs through a wireless network. This can be useful, e.g., in acoustically challenging social gatherings with multiple HA users. The proposed speech enhancement paradigm can in this situation assist the HA user by first ranking and then enhancing the estimated conversational partner amongst the users. [1]

The signal model of the sound picked-up by the user's HA microphone can be described as

$$x_i(n) = s_i(n), \quad i = 0, ..., I, \tag{E.27}$$

where $s_0(n)$ is the HA user's own voice signal as picked-up by the user's microphone while $s_i(n)$ for $i = 1, ..., I$ are the clean speech signals picked up by the microphones located at the candidate speakers.

5.4.1 Simulation Setup

We reuse the speech database presented in Sec. 5.3.3 for the candidate speakers and own voice signals. We use the data set with conversations in second language and babble noise, which was not used for estimation of the shaping parameters. Two conversational partners are randomly chosen from the data set, where one is randomly chosen as the HA user and the other as the conversational partner for each signal realization. The competing speakers are chosen from the same data set, but are not conversing with the HA user. The HA user's conversational partner is unknown to the speech enhancement systems. We use rVAD 2.0 [29, 30] for voice activity detection and the sampling frequency of the VAD output is $f_{\text{s,vad}} = 100$ Hz. The integration time needed for the speaker ranking algorithms, is implemented as sliding windows with length T_{int} and with a hop size of 1 sample at sampling frequency $f_{\text{s,vad}}$. The speech enhancement systems used in the evaluation are referred to as:

- **No processing**: The speech enhancement system does not apply any speaker ranking algorithms and simply outputs the sum of all candidate speakers.

[1] In this situation, the separation stage in Figs. E.1 and E.6 is obsolete – microphones are located on each candidate speaker and allow direct estimation of their voice activity pattern $\alpha_i(n)$; hence, the separation stage is unnecessary.

- **MMI**, **NCC**, and **MOG**: The MMI, NCC, and the proposed MOG algorithms are used as speaker ranking. The gain function is implemented as in (E.17).
- **BMOG**: Posterior probabilities of the conversational partner are estimated using BMOG and used as a gain function for enhancement, cf. (E.19). The prior probability distribution $P(\mathcal{H}_i)$ in (E.12) is set uniform.

5.4.2 Results: Wireless hearing aid network

We evaluate the speech enhancement performance by comparing the enhanced conversational partner with the clean speech signal of the conversational partner in terms of ESTOI [34, 35], PESQ [36], and segmental SNR [37]. We evaluate the speech enhancement systems for $T_{\text{int}} = \{5, 10, 20, 30\}$ s and the following number of competing speakers $N_c = \{1, 2, ..., 10\}$ [s] . The minimum gain for MMI, NCC, and MOG is set to $g_{\text{min}} = 0.01$. A minimum gain of $g_{\text{min}} > 0$ is necessary for the MMI, NCC, and MOG enhancement systems to avoid rare situations with a complete suppression of the conversational partner. These situations typically arise at low T_{int} and results in undefined PESQ and segmental SNR scores. The minimum gain for BMOG was set to $g_{\text{min}} = 0$ as it did not experience similar problems. The results are shown in Fig. E.8 and each score is averaged over 100 realizations of conversations. Generally, we see a significant improvement in terms of both ESTOI and PESQ when using MOG and BMOG compared to NCC and MMI. The improvement is particularly notably at low integration time such as $T_{\text{int}} = 5$ s and $T_{\text{int}} = 10$ s. At higher integration times, the improvements become less prominent with the exception of NCC, which seems to perform the worst. We note that BMOG seems to perform much better than MOG in terms of PESQ at $T_{\text{int}} = 30$. This is due to the minimum gain which is set to 0.01 for MOG but 0 for BMOG. From our experiments, we have observed that setting the minimum gain to be above 0 can help NCC, MMI, and MOG perform better on average at low integration times, e.g., $T_{\text{int}} = 5$. However, the trade-off is slightly degraded performance at high integration times as shown in the results.

From these results, it is clear that speech enhancement systems that use MOG and BMOG generally outperform the NCC and MMI methods for this particular application.

5.5 Application 2: Beamforming System in Hearing Aids

In this section, we demonstrate the use of the proposed speech enhancement paradigm in another hearing aid application. Modern hearing aids are equipped with multiple microphones which allow for implementation of acoustic beamformers to enhance the speech signal of a conversational partner of a HA user. However, retrieving the speech signal can be particularly difficult in situations with multiple competing speakers, because it is hard to decide who is the conversational partner. Hence, in this application the proposed (B)MOG speech enhancement paradigm is used to efficiently retrieve the speech signal of the conversational partner amongst several competing speakers.

First, we model the received signal at the microphones of the HAs. The user's and candidate speakers' speech signals propagate to the microphones and are simulated using acoustic impulse responses (AIRs). The AIR from the i'th speaker to the m'th microphone is denoted as $h_{i,m}(n)$ where $i = 0, 1, ..., I$ is the speaker index, and $m = 1, ..., M$ is the microphone index. The index value $i = 0$ is used to denote the user's index. The AIRs can be decomposed into $h_{i,m}(n) = h_{i,m'}(n) * d_{i,m}(n)$ where $*$ denotes the convolution operator, $h_{i,m'}(n)$ is the AIR from the i'th speaker to a pre-selected reference microphone $m' \in \{1, ..., M\}$, and $d_{i,m}(n)$ is the impulse response from the reference microphone to the m'th microphone also referred to as the relative impulse response. Let $s'_i(n)$ be the received signal of the i'th speaker at the reference microphone, m', i.e. $s'_i(n) = s_i(n) * h_{i,m'}(n)$. Then the received signal of the i'th speaker at the m'th microphone is $s'_{i,m}(n) = s'_i(n) * d_{i,m}(n)$. We denote $v_m(n)$ as being the noise vector (e.g. ambient noise and microphone self-noise) as received at the m'th microphone. The noisy signal at the m'th microphone is then modeled as

$$x_m(n) = \sum_{i=0}^{I} s'_{i,m}(n) + v_m(n). \tag{E.28}$$

5.5.1 Speech separation using beamformers

The received microphone signal, $x_m(n)$, is a mixture of clean user and candidate speaker signals received at microphone m, $s'_{i,m}(n)$, plus noise $v_m(n)$. Following the speech enhancement paradigm in Fig. E.6, the microphone signals are first separated into user and candidate speaker signals before applying speaker ranking. We use the minimum power distortionless response (MPDR) beamformer to separate the speech signals. The MPDR beamformers are implemented in the time-frequency domain using the short-time Fourier transform (STFT) and are for each time-frequency tile computed as [38]

$$\boldsymbol{W}_i(k,l) = \frac{\mathbf{C}_x^{-1}(k,l)\boldsymbol{D}_i(k)}{\boldsymbol{D}_i^H(k)\mathbf{C}_x^{-1}(k,l)\boldsymbol{D}_i(k)}, \; \boldsymbol{W}_i(k,l) \in \mathbb{C}^M, \tag{E.29}$$

where k and l denote the frequency and frame indices, respectively. $\mathbf{C}_x(k,l) = \mathbb{E}\{\boldsymbol{X}(k,l)\boldsymbol{X}^H(k,l)\}$ is the cross power spectral density (CPSD) matrix of the noisy microphone signals and $\boldsymbol{D}_i(k) = [D_{i,1}(k), ..., D_{i,M}(k)]^T$, $i = 0, 1, ..., I$,

$k = 0, 1, ..., K$ denotes the relative acoustic transfer function (RATF) vector for the i'th speaker and k'th frequency bin [39, 40]. The m'th element of the RATF vector is the frequency domain representation of $d_{i,m}(n)$. Unfortunately, the number of candidate speakers and their RATF vectors are seldomly known in practice. Instead, we use I to denote the number of MPDR beamformers steered towards a set of I unique and fixed directions in the acoustic environment. In other words, the spatial filter bank is implemented using a dictionary of RATF vectors $\mathcal{D}(k) = \{\boldsymbol{D}_0(k), \boldsymbol{D}_1(k), ..., \boldsymbol{D}_I(k)\}$, $k = 0, 1, ..., K$, where we assume that the dictionary is given in advance. Assuming that each beam contains a maximum of one candidate speaker (i.e., that candidate sources are sufficiently spatially separated), each beamformer output, $\hat{s}'_i(n)$, is treated as a candidate speaker signal. The output of each beamformer is

$$\hat{S}'_i(k,l) = \boldsymbol{W}_i^H(k,l)\boldsymbol{X}(k,l), \tag{E.30}$$

where $\hat{S}'_i(k,l)$ is the enhanced signal from direction i, and is treated as a speech signal from a candidate speaker. The beamformer outputs $\hat{S}'_i(k,l)$ are transformed back to the time-domain using the inverse STFT to obtain $\hat{s}'_i(n)$. The remaining part of the speech enhancement system, i.e., ranking and enhancement, follows the same procedure as in application 1 in Sec. 5.4.

5.5.2 Simulation of the acoustic scene

We simulate the acoustic scenes to resemble a cocktail party-like scenario with a HA user engaged in a conversation with a conversational partner. Such a situation involves the presence of speech signals from the HA user, the conversational partner, and competing speakers, and the presence of noise from the environment.

To simulate the received signals at the microphones, we use a database of AIRs measured in a sound studio where room reverberation has been removed [41]. The measurement setup consists of a spherical loudspeaker array with a HA user seated in the center of the array. The HA user is wearing a behind-the-ear (BTE) hearing aid on each ear. Each BTE hearing aid has three microphones where two are placed in a front/rear configuration on the HA and the third is placed in the ear canal. The microphones are used in a binaural HA configuration where we assume wireless, simultaneous, and error-free signal exchange between the left and right HAs. Hence, beamformers are implemented using a the total number of $M = 6$ microphones. The AIRs are measured from uniformly spaced positions in the horizontal plane with respect to the head of the HA user and with a resolution of 7.5° resulting in AIRs for 48 different angles. We define 0° as the frontal direction from the user's point-of-view. The own voice AIRs are measured using a mouth reference microphone placed in front of the HA user's mouth.

We use the conversational speech database in [27], as in Application 1, as speech material in our simulation. Realistic noise measured in a canteen is used in our simulation. The noise is measured using a spherical microphone array to accurately capture the noise field [42]. The noise recordings are transformed and convolved with the AIRs to reproduce the same noise field as would have been experienced by a HA user in the canteen.

Competing speakers are added to the acoustic scenes. The speech material for the competing speaker are from the same speech database as in Sec. 5.4 [27]. The speech of the competing speakers is unrelated to the conversation between the user and conversational partner. We experiment with $N_c = 3$ and $N_c = 5$ competing speakers in our evaluation. Increasing the number of competing speaker to much larger than $N_c = 5$, results in poorer speech separation as the beamformers cannot sufficiently suppress the speakers from other directions. The purpose here is mainly to demonstrate the feasibility of using (B)MOG ranking in a beamforming context and for a larger number of competing speakers, other better performing speech separation systems could be used, e.g., (Conv-)TasNet [12, 13] or Wavesplit [43].

For $N_c = 3$, the conversational partner is placed randomly in the positions $\{0°, 90°, 180°, 270°\}$, and the competing speakers are placed at the remaining 3 positions. Similarly for $N_c = 5$, the conversational partner is placed randomly in the positions $\{0°, 60°, 120°, 180°, 240°, 300°\}$, and the competing speakers are placed at the remaining 5 positions. The positions of the speakers are fixed for the whole duration of a realization of an acoustic scene. We do not simulate head-movements of the HA user but these movement can be compensated with other sensors, e.g., accelerometers in practice. To simulate the received signals of the speech sources at the microphones, we convolve the speech signals with the AIRs associated with the direction. The speech power of each competing speaker is approximately identical to the speech power of the conversational partner before convolving with the AIRs. Canteen noise is added to the acoustic scenes and the SNR is defined as the ratio between the clean speech power of the conversational partner at the source location and the power of the background noise. The SNR is set to 12 dB.

The search region of the beamformers, $\boldsymbol{W}_i(k,l)$, is $0°, 7.5°, ..., 352.5°$ in the azimuth angle. The RATF dictionary is given as $\mathcal{D}(k) = \{\boldsymbol{D}_{0,m}(k), \boldsymbol{D}_{1,m}(k), ..., \boldsymbol{D}_{I,m}(k)\}$ where $\boldsymbol{D}_{0,m}(n)$ is the own voice RATF vector. The elements $\boldsymbol{D}_{i,m}(n)$ $i = 1, ..., I$ are RATF vectors associated with sound sources impinging from direction $\theta = (i-1) \cdot 7.5°$ in the horizontal plane where $\theta = 0°$ is the frontal direction with respect to the HA user.

To implement the OVAD/VAD blocks in Fig. E.6, rVAD 2.0 [30] is used for voice activity detection on the separated speech signals $\hat{s}_i(n)$ and the own voice signal $\hat{s}_0(n)$.

The sampling frequency of the received microphone signals is set to 16 kHz. We use a square-root Hann window with a window size of 256 samples

Algorithm 6 Beamforming system for application 2.

```
Input: $\boldsymbol{x}(n) = [x_1(n), ..., x_M(n)]^T$, $\mathcal{D}_k = \{\boldsymbol{D}_1(k), ..., \boldsymbol{D}_I(k)\}$
Apply STFT to $\boldsymbol{x}(n)$ to obtain $\boldsymbol{X}(k,l)$ for all $k$ and $l$.
for all $i \in \Theta$ do
    for $k = 1, 2, ..., K$ do
        Estimate the noisy CPSD matrix:
            $\hat{\mathbf{C}}_x(k,l) = \frac{1}{L}\underline{\boldsymbol{X}}(k,l)\underline{\boldsymbol{X}}^H(k,l)$
        where $\underline{\boldsymbol{X}}(k,j) = [\boldsymbol{X}(k,l-L+1), ..., \boldsymbol{X}(k,l)]$.
        Compute the MPDR beamformer weights, $\boldsymbol{W}_i(k,l)$, using (E.29).
        Enhance the signal from direction $i$ using (E.30).
    end for
    Inverse STFT $\hat{S}'_i(k,l)$ to obtain $\hat{s}'_i(n)$.
end for
Estimate voice activity of each candidate speaker $\alpha_i(n) = \text{VAD}\left(\hat{s}'_i(n)\right)$.
Use a speaker ranking algorithm e.g. Algorithm 5 and compute the gain
    function to obtain $g_i(n)$.
Enhance the conversational partner using (E.19).
```

for the STFT and inverse STFT. The hop-size is 128 samples.

The beamforming system is summarized in Algorithm 6 in pseudo-code.

5.5.3 Evaluation of the speech enhancement paradigm in beamforming systems

We evaluate the performance in terms of 1) speaker ranking performance in Sec. 5.5.4 and 2) speech enhancement performance in Sec. 5.5.5. First, the speaker ranking in this application is closely related to direction-of-arrival (DOA) estimation. DOA estimation often arises in beamforming applications where the goal is to estimate the direction of the talker-of-interest in order to steer a beamformer. In our context, DOA estimation is related to estimating the channel of the conversational partner. Hence, the MOG algorithm is in fact a DOA estimator in this context. Secondly, the speech enhancement performance will quantify the potential benefit of using the proposed speech enhancement paradigm in a beamforming context for HAs. The reported performance scores are averaged from simulations of 40 realizations of the acoustic scenes for the results in Sec. 5.5.4 and Sec. 5.5.5.

To evaluate the speaker ranking performance, we evaluate the DOA accuracy and the mean-absolute-error (MAE) between the estimated DOA $\hat{\theta}_n$

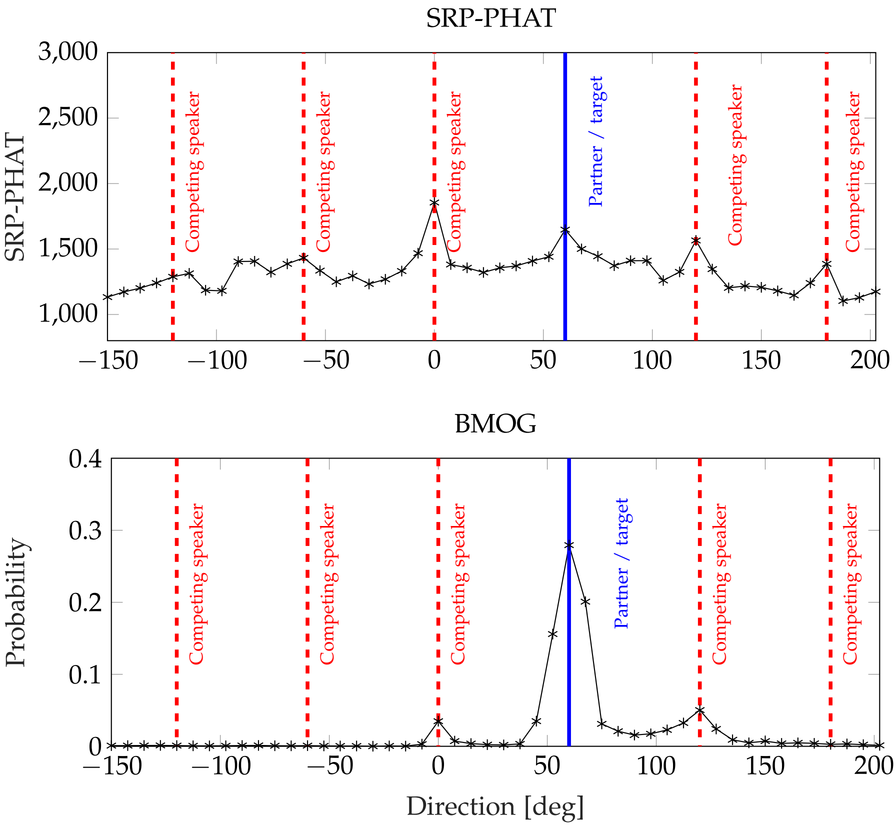


Fig. E.9: Example of the average output over 24 seconds of the SRP-PHAT algorithm and BMOG algorithm as a function of direction. The acoustic scene consist of 1 conversational partner and 5 competing speakers in canteen noise. The SRP-PHAT algorithm is not able to distinguish between the conversational partner and competing speakers, however, the proposed BMOG algorithm ($T_{\text{int}} = 10$ s) is however effective at locating the conversational partner.

MOG	$T_{\text{int}} = 5$	$T_{\text{int}} = 10$	$T_{\text{int}} = 20$	$T_{\text{int}} = 30$
$N_c = 3$	31.60%	46.23%	65.94%	75.03%
$N_c = 5$	35.54%	48.23%	63.83%	70.97%
MMI				
$N_c = 3$	16.91%	32.69%	51.54%	61.26%
$N_c = 5$	20.00%	31.77%	49.66%	58.91%
SRP-PHAT				
$N_c = 3$	19.72%	19.72%	19.72%	19.72%
$N_c = 5$	23.39%	23.39%	23.39%	23.39%

Table E.2: DOA estimation accuracy as a function of integration time T_{int} and number of competing speakers N_c.

MOG	$T_{\text{int}} = 5$	$T_{\text{int}} = 10$	$T_{\text{int}} = 20$	$T_{\text{int}} = 30$
$N_c = 3$	60.55°	45.41°	29.81°	22.30°
$N_c = 5$	58.58°	46.67°	30.81°	22.39°
MMI				
$N_c = 3$	79.68°	60.35°	42.27°	33.77°
$N_c = 5$	72.90°	59.71°	43.16°	33.17°
SRP-PHAT				
$N_c = 3$	79.62°	79.62°	79.62°	79.62°
$N_c = 5$	77.25°	77.25°	77.25°	77.25°

Table E.3: Mean-absolute-error of estimated DOA as a function of integration time T_{int} and number of competing speakers N_c.

and the true DOA θ_n of the conversational partner. The DOA accuracy is the probability of estimating the correct DOA of the conversational partner and the MAE is estimated as the average absolute error:

$$\hat{\text{MAE}} = \frac{1}{N} \sum_{n=1}^{N} \left| \arg \left(\exp(j(\theta_n - \hat{\theta}_n)) \right) \right|, \tag{E.31}$$

where θ_n and $\hat{\theta}_n$ are in radians, $\sqrt{j} = -1$, and $\arg(\cdot)$ is the argument of a complex number. The $\hat{\text{MAE}}$ is averaged over

The speech enhancement performance is reported in terms of ESTOI, PESQ, and segmental SNR scores to estimate the speech intelligibility, speech quality, and noise suppression performance of the proposed speech enhancement paradigm, respectively. The ESTOI, PESQ, and segmental SNR scores are computed using the output of the enhancement system $\hat{s}_t(n)$ and the clean conversational partner speech signal received at the reference microphone index $s'_{t,m'}(n)$.

Our evaluation includes four beamforming systems which are based on

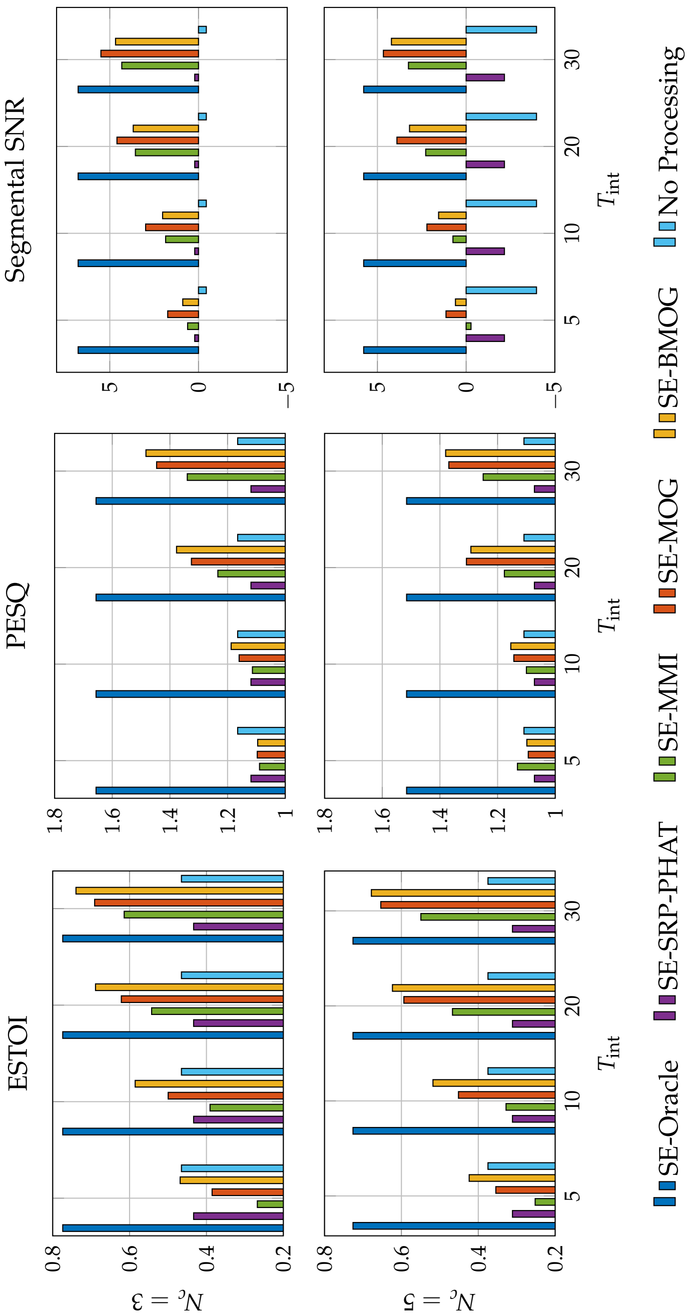


Fig. E.10: Averaged beamforming performance as function of the number of competing speakers N_c and integration time T_{int}.

the speech enhancement paradigm in Fig. E.6. All systems use the same spatial filter bank of MPDR beamformers for speech separation. We use the rVAD 2.0 for voice activity detection for all systems. We refer the beamforming systems to as

- **SE-Oracle**: The beamforming system, SE-Oracle, is used as a reference system to indicate the upper bound performance if the direction of the conversational partner is known in advance.

- **SE-MMI**: The beamforming system, SE-MMI, uses the MMI algorithm to find the direction of the conversational partner. The output at time n of SE-MMI is $\hat{s}_t(n) = \hat{s}'_{\hat{i}_{\text{MMI}}(n)}(n)$ where $\hat{i}_{\text{MMI}}(n)$ is the DOA estimate of the conversational partner at time n.

- **SE-MOG**: The beamforming system, SE-MOG, uses the MOG algorithm to find the direction of the conversational partner. The output at time n of SE-MOG is $\hat{s}_t(n) = \hat{s}'_{\hat{i}_{\text{MOG}}(n)}(n)$ where $\hat{i}_{\text{MOG}}(n)$ is the DOA estimate of the conversational partner at time n.

- **SE-SRP-PHAT**: The beamforming system, SE-SRP-PHAT, uses the well-known SRP-PHAT algorithm [6] to estimate the DOA of the conversational partner. In contrast to the speaker ranking algorithms NCC, MMI, and MOG, the SRP-PHAT algorithm does not utilize turn-taking to the candidate speakers related to conversations but instead searches for the most dominant speaker. The output at time n of SE-SRP-PHAT is $\hat{s}_t(n) = \hat{s}'_{\hat{i}_{\text{SRP-PHAT}}(n)}(n)$ where $\hat{i}_{\text{SRP-PHAT}}(n)$ is the DOA estimate of the conversational partner at time n.

- **SE-BMOG**: The beamforming system, SE-BMOG, uses the BMOG algorithm to compute a posterior probability distribution of the direction of the conversational partner. The output of SE-BMOG at time n is a linear combination of the separated candidate speakers using the posterior probabilities as weights, i.e., $\hat{s}_t(n) = \sum_{i=1}^{I} P(\mathcal{H}_i|\phi_1, ..., \phi_I)\hat{s}'_i(n)$. The prior probability distribution for BMOG was set to be a uniform prior probability distribution.

We did not include a beamforming system with a NCC-based speaker ranking algorithm as it performed significantly poorer than the other algorithms in preliminary experiments.

5.5.4 Results: DOA estimation performance in beamforming systems

This section focuses on speaker ranking/DOA performance and not speech enhancement performance of the complete beamforming system, which is treated in Sec. 5.5.5. Therefore, BMOG is not included since the output of

BMOG is a probability distribution and not an estimate of the conversational partner as the MOG algorithm.

The results for DOA estimation performance in terms of DOA accuracy and MAE are shown in table E.2 and E.3, respectively. Each score in the table is an average over 40 realizations of the acoustic scenes.

The MOG algorithm seems to outperform the MMI algorithm consistently by approximately 15%-points. Similarly, the MAE for the MOG algorithm is lower than the MAE for MMI and SRP-PHAT for all T_{int} and N_c. It is also clear, that the SRP-PHAT algorithm in general struggles in estimating the conversational partner DOA in a multi-speaker situation which is demonstrated in Fig. E.9. Essentially, the SRP-PHAT algorithm constantly switches between the candidate speakers as the estimate of the conversational partner. The MOG algorithm, however, effectively exploits the turn-taking mechanism in conversations and is able to detect the conversational partner.

An interesting observation is that the DOA estimation accuracy is slightly higher for $N_c = 5$ compared to $N_c = 3$ at low integration times, e.g., $T_{\text{int}} = 5$ s. Likewise, the MAE is lower for $N_c = 5$ compared to $N_c = 3$ at low integration times. However, note that the angular distance between the conversational partner and competing speakers becomes larger at $N_c = 3$ compared to $N_c = 5$. That is, for $N_c = 3$ the speakers are located at $\{0°, 90°, 180°, 270°\}$ whereas for $N_c = 5$ the speakers are located at $\{0°, 60°, 120°, 180°, 240°, 300°\}$. Therefore, possible explanations of these observations at low integration times are that 1) the DOA estimates of MMI and MOG become more biased for $N_c = 3$, which results in a lower accuracy and 2) MMI and MOG are more likely to return a higher absolute error for $N_c = 3$ than for $N_c = 5$ in case of a DOA estimation error. However, it is evident from the results, that the MOG algorithm has significantly higher accuracy compared to MMI and SRP-PHAT for all combinations of T_{int} and N_c.

5.5.5 Results: Speech enhancement performance in beamforming systems

The results for beamforming performance are shown in Fig. E.10, which plots performance scores ESTOI, PESQ, and segmental-SNR as a function of integration time T_{int} for different beamforming systems. Cleacly, the MOG algorithm outperforms the MMI and SRP-PHAT algorithms significantly in most situations. The results also indicate that the SRP-PHAT algorithm performs slightly worse than the unprocessed signal in multi-speaker environments unless additional knowledge on the conversational partner is given. The MMI algorithm also performs slightly worse than the unprocessed signal in terms of ESTOI at $T_{\text{int}} = 5$ s and $T_{\text{int}} = 10$ s as the MMI algorithm can erroneously estimate a competing speaker as being the conversational partner for low integration times. The speech enhancement system using the BMOG algorithm, however, performs best on average across all scores, especially in terms of ESTOI and PESQ. This is likely due to a softer gain function based on the estimated posterior probability, which is less aggressive compared to the gain function used in the MOG algorithm. The softer gain function translates to higher ESTOI and PESQ scores, but a slightly lower segmental SNR score. With long integration times, both speech enhancement systems using MOG and BMOG are extremely effective at retrieving a conversational partner in a multi-speaker situation as they perform close to the oracle beamformer. However, long integration times also require that the conversational partner stays within the same beam for longer duration, e.g. in a restaurant where the speakers are seated.

6 Conclusion

In this paper, we have proposed a speech enhancement paradigm using a speaker ranking algorithm which can effectively retrieve a desired speech signal in a multi-talker environment. Specifically, the proposed speech enhancement paradigm exploits turn-taking behavior to determine the conversational partner amongst a set of candidate talker of a user by finding the talker with minimum probability of speech overlaps and gaps. The proposed algorithm only requires access to microphone signals, which is in contrast to existing methods which require additional sensor inputs, e.g. EEG, cameras, etc. We demonstrated the proposed speech enhancement paradigm in two applications, where retrieval of a conversational partner's speech signal in a multi-talker environment, is desired. We compared the proposed systems to current state-of-the-art speech enhancement systems, and results indicate that the proposed systems significantly outperform the state-of-the-art systems.

A Appendix

A.1 Proof of minimizing speech overlap and gap, and maximizing mutual exclusion

In this Appendix, it is shown that minimizing the probability of speech overlap and gap is equivalent to maximizing the probability of mutual speech exclusion between the user's own voice VAD and candidate speaker's VAD.

Specifically, we show that

$$\begin{aligned}&\arg\min_i \sum_{k=0}^{1} P_{A_0A_i}(\alpha_0 = k, \alpha_i = k)\\&= \arg\max_i \sum_{k=0}^{1} P_{A_0A_i}(\alpha_0 = k, \alpha_i = 1-k).\end{aligned} \tag{E.32}$$

Proof

The sum of the support of $P_{A_0A_i}(\alpha_0 = k, \alpha_i = j)$ is equal to one such that

$$\sum_{k=0}^{1}\sum_{j=0}^{1} P_{A_0A_i}(\alpha_0(n) = k, \alpha_i(n) = j) = 1. \tag{E.33}$$

The probabilities are split into the probability of speech overlap and gap, and the probability of mutual speech exclusion

$$\begin{aligned}&\sum_{k=0}^{1} P_{A_0A_i}(\alpha_0(n) = k, \alpha_i(n) = k) =\\&1 - \sum_{j=0}^{1} P_{A_0A_i}(\alpha_0(n) = j, \alpha_i(n) = 1-j).\end{aligned} \tag{E.34}$$

where the left-hand side is the probability of speech overlap and gap and the right-hand side is '1' subtracted by the probability of mutual speech exclusion. Hence, minimizing the probability of speech overlap and gaps is equivalent to:

$$\hat{i}_{\mathrm{MOG}}(n) = \arg\min_i \; 1 - \sum_{k=0}^{1} P_{A_0A_i}(\alpha_0(n) = k, \alpha_i(n) = 1-k), \tag{E.35}$$

or maximizing the probability of mutual speech exclusion:

$$\hat{i}_{\mathrm{MOG}}(n) = \arg\max_i \sum_{k=0}^{1} P_{A_0A_i}(\alpha_0(n) = k, \alpha_i(n) = 1-k), \tag{E.36}$$

hence proving the equivalence in (E.32).

A.2 Proof of minimizing speech overlap and gap, and maximizing mean-square-error

In this Appendix, we show that minimizing the probability of speech overlap and gap is identical to maximizing the mean-square-error between the own voice VAD and the candidate speaker VAD i.e.

$$\begin{aligned}&\arg\min_i \sum_{k=0}^{1} P_{A_0A_i}(\alpha_0(n)=k, \alpha_i(n)=k)\\&=\arg\max_i \mathbb{E}\left[(A_0(n)-A_i(n))^2\right].\end{aligned} \tag{E.37}$$

Proof

The probability of speech overlap and gap is

$$\begin{aligned}\sum_{k=0}^{1} P_{A_0A_i}(\alpha_0=k, \alpha_i=k) &= P_{A_0A_i}(\alpha_0=1, \alpha_i=1)\\&\quad + P_{A_0A_i}(\alpha_0=0, \alpha_i=0).\end{aligned} \tag{E.38}$$

We may then write

$$\begin{aligned}&P_{A_0A_i}(\alpha_0=1, \alpha_i=1) + P_{A_0A_i}(\alpha_0=0, \alpha_i=0)\\&=\sum_{k=0}^{1}\sum_{j=0}^{1} kjP_{A_0A_i}(\alpha_0=k, \alpha_i=j)\\&+\sum_{m=0}^{1}\sum_{n=0}^{1}(1-m)(1-n)P_{A_0A_i}(\alpha_0=m, \alpha_i=n)\end{aligned} \tag{E.39}$$

and using the expectation operator, we have that

$$\begin{aligned}\mathbb{E}\left[A_0A_i\right] &= \sum_{k=0}^{1}\sum_{j=0}^{1} kjP_{A_0A_i}(\alpha_0=k, \alpha_i=j)\\ \mathbb{E}\left[(1-A_0)(1-A_i)\right] &=\\&\sum_{m=0}^{1}\sum_{n=0}^{1}(1-m)(1-n)P_{A_0A_i}(\alpha_0=m, \alpha_i=n).\end{aligned} \tag{E.40}$$

Hence, the probability of speech overlap and gap is

$$\begin{aligned}\sum_{k=0}^{1} P_{A_0A_i}(\alpha_0=k, \alpha_i=k) &= \mathbb{E}\left[A_0A_i\right] + \mathbb{E}\left[(1-A_0)(1-A_i)\right]\\&= 1-\mathbb{E}\left[A_0\right]-\mathbb{E}\left[A_i\right]+2\mathbb{E}\left[A_0A_i\right].\end{aligned} \tag{E.41}$$

Since A_0 and A_i are Bernoulli random variables, then $\mathbb{E}\left[A_0\right] = \mathbb{E}\left[A_0^2\right]$ and $\mathbb{E}\left[A_i\right] = \mathbb{E}\left[A_i^2\right]$, such that the probability of speech overlap and gap is

$$\sum_{k=0}^{1} P_{A_0A_i}(\alpha_0=k, \alpha_i=k) = 1-\mathbb{E}\left[(A_0-A_i)^2\right], \tag{E.42}$$

where $\mathbb{E}\left[(A_0 - A_i)^2\right]$ is the mean-square-error (MSE) between A_0 and A_i. We see that the probability of speech overlap and gap is equivalent to $1 - \mathbb{E}\left[(A_0 - A_i)^2\right]$. Hence, the optimization problem for the MOG algorithm is

$$\begin{aligned} \hat{i}_{\text{MOG}}(n) &= \arg\min_i \; 1 - \mathbb{E}\left[(A_0(n) - A_i(n))^2\right] \\ &= \arg\max_i \; \mathbb{E}\left[(A_0(n) - A_i(n))^2\right], \end{aligned} \tag{E.43}$$

which is a maximization of the MSE between the own voice VAD and a candidate speaker VAD.

A.3 Expected misclassification rate for MOG

The speaker misclassification rate is defined as the probability of classifying a wrong candidate speaker as the conversational partner. Using the MOG algorithm, we consider a misclassification as when Φ_t is equal to or smaller than Φ_v. For a number I of candidate speakers and integration time T_{int}, the misclassification rate $P(E = 1; I, T_{\text{int}})$ is given by

$$P(E = 1; I, T_{\text{int}}) = 1 - P\left(\Phi_t > \Phi_{v,1}, ..., \Phi_t > \Phi_{v,I-1}\right),$$

where

$$\begin{aligned} &P\left(\Phi_t > \Phi_{v,1}, ..., \Phi_t > \Phi_{v,I-1}\right) \\ &= \sum_{\phi=1}^{N} p_\Phi(\phi; \gamma_t, \beta_t, N) P_\Phi^{I-1}(\phi - 1; \gamma_v, \beta_v, N) \end{aligned} \tag{E.44}$$

denotes the correct classification rate, and $P_\Phi(\phi - 1; \gamma_v, \beta_v, N)$ is the cumulative distribution function of $p_\Phi(\phi - 1; \gamma_v, \beta_v, N)$,

$$P_\Phi(\phi - 1; \gamma_v, \beta_v, N) = \sum_{\kappa=0}^{\phi-1} p_\Phi(\kappa; \gamma_v, \beta_v, N). \tag{E.45}$$

Proof

First we consider the probability of correct classification under the assumption that $\Phi_{v,j}$ for all j's are independent:

$$\begin{aligned}&P\left(\Phi_t > \Phi_{v,1}, ..., \Phi_t > \Phi_{v,I-1}\right)\\&= \sum_{\phi=1}^{N} \sum_{\kappa_1=0}^{\phi-1} \cdots \sum_{\kappa_{I-1}=0}^{\phi-1} p_\Phi(\phi; \gamma_t, \beta_t, N) \times\\&p_\Phi(\kappa_1; \gamma_{v,1}, \beta_{v,1}, N) \times \cdots \times p_\Phi(\kappa_{I-1}; \gamma_{v,I-1}, \beta_{v,I-1}, N)\\&= \sum_{\phi=1}^{N} p_\Phi(\phi; \gamma_t, \beta_t, N) \sum_{\kappa_1=0}^{\phi-1} p_\Phi(\kappa_1; \gamma_{v,1}, \beta_{v,1}, N) \times \cdots \times\\&\sum_{\kappa_{I-1}=0}^{\phi-1} p_\Phi(\kappa_{I-1}; \gamma_{v,I-1}, \beta_{v,I-1}, N).\end{aligned} \tag{E.46}$$

To simplify the expression, we define following cumulative distribution function

$$P_\Phi(\phi-1; \gamma_{v,j}, \beta_{v,j}, N) = \sum_{\kappa_j=0}^{\phi-1} p_\Phi(\kappa_j; \gamma_{v,j}, \beta_{v,j}, N), \tag{E.47}$$

for all $j = 1, ..., I-1$. Inserting (E.47) into (E.46), we have

$$\begin{aligned}&P\left(\Phi_t > \Phi_{v,1}, ..., \Phi_t > \Phi_{v,I-1}\right)\\&= \sum_{\phi=1}^{N} p_\Phi(\phi; \gamma_t, \beta_t, N) P_\Phi(\phi-1; \gamma_{v,1}, \beta_{v,1}, N) \times \cdots \times\\&P_\Phi(\phi-1; \gamma_{v,I-1}, \beta_{v,I-1}, N).\end{aligned} \tag{E.48}$$

Assuming that $\Phi_{v,j}$ are independent and identically distributed such that $\gamma_{v,j} = \gamma_v, \beta_{v,j} = \beta_v$ for all j, we may simplify to

$$\begin{aligned}&P\left(\Phi_t > \Phi_{v,1}, ..., \Phi_t \geq \Phi_{v,I-1}\right) =\\&\sum_{\phi=1}^{N} p_\Phi(\phi; \gamma_t, \beta_t, N) P_\Phi^{I-1}(\phi-1; \gamma_v, \beta_v, N).\end{aligned} \tag{E.49}$$

As the misclassification rate is

$$P(E=1; I, T_{\text{int}}) = 1 - P\left(\Phi_t > \Phi_{v,1} ..., \Phi_t > \Phi_{v,I-1}\right), \tag{E.50}$$

then inserting (E.49) into (E.50) yields the derived result

$$\begin{aligned}&P(E=1; I, T_{\text{int}}) =\\&1 - \sum_{\phi=1}^{N} p_\Phi(\phi; \gamma_t, \beta_t, N) P_\Phi^{I-1}(\phi-1; \gamma_v, \beta_v, N).\end{aligned}$$

References

[1] A. Kuklasinski, 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, Sep. 2016.

[2] J. Jensen and U. Kjems, “Maximum Likelihood Based Noise Covariance Matrix Estimation For Multi-Microphone Speech Enhancement,” *EUSIPCO*, Aug. 2012.

[3] J. Jensen and M. S. Pedersen, “Analysis of beamformer directed single-channel noise reduction system for hearing aid applications,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 5728–5732.

[4] S. Braun, A. Kuklasinski, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, “Evaluation and Comparison of Late Reverberation Power Spectral Density Estimators,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1056–1071, Jun. 2018.

[5] O. Schwartz, S. Gannot, and E. A. P. Habets, “Joint maximum likelihood estimation of late reverberant and speech power spectral density in noisy environments,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai: IEEE, Mar. 2016, pp. 151–155.

[6] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, 2001.

[7] 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, Mar. 2018.

[8] 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)*. Ottawa, ON, Canada: IEEE, Nov. 2019, pp. 1–5.

[9] S. Chakrabarty and E. A. P. Habets, “Broadband DOA estimation using Convolutional neural networks trained with noise signals,” *arXiv:1705.00919 [cs, stat]*, May 2017, arXiv: 1705.00919.

[10] D. Yu, M. Kolbaek, Z.-H. Tan, and J. Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA: IEEE, Mar. 2017, pp. 241–245.

[11] M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, “Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[12] Y. Luo and N. Mesgarani, “TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB: IEEE, Apr. 2018, pp. 696–700.

[13] ——, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," vol. 27, no. 8, pp. 1256–1266.

[14] A. A. Nair, A. Reiter, C. Zheng, and S. Nayar, "Audiovisual Zooming: What You See Is What You Hear," in *Proceedings of the 27th ACM International Conference on Multimedia*. Nice France: ACM, Oct. 2019, pp. 1107–1118.

[15] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, "A Tutorial on Auditory Attention Identification Methods," *Frontiers in Neuroscience*, vol. 13, p. 153, Mar. 2019.

[16] 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, p. 233121651881438, Jan. 2018.

[17] N. Das, A. Bertrand, and T. Francart, "EEG-based auditory attention detection: boundary conditions for background noise and speaker positions," *Journal of Neural Engineering*, vol. 15, no. 6, p. 066017, Dec. 2018.

[18] A. Aroudi and S. Doclo, "EEG-based auditory attention decoding: Impact of reverberation, noise and interference reduction," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. Banff, AB: IEEE, Oct. 2017, pp. 3042–3047.

[19] H. Sacks, E. A. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation," *Language*, vol. 50, no. 4, p. 696, Dec. 1974.

[20] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, Nov. 2009.

[21] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard, "Modeling human interaction in meetings," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, vol. 4. Hong Kong, China: IEEE, 2003, pp. IV–748–51.

[22] L. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 305–317, Mar. 2005.

[23] T. Choudhury and S. Basu, "Modeling conversational dynamics as a mixed-memory markov process," in *Advances in Neural Information Processing Systems*, L. Saul, Y. Weiss, and L. Bottou, Eds., vol. 17. MIT Press, 2005.

[24] S. C. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language," *Frontiers in Psychology*, vol. 6, Jun. 2015.

[25] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, Oct. 2010.

[26] L. V. Hadley, W. O. Brimijoin, and W. M. Whitmer, "Speech, movement, and gaze behaviours during dyadic conversation in noise," *Scientific Reports*, vol. 9, no. 1, p. 10451, Dec. 2019.

[27] A. J. Sørensen, M. Fereczkowski, and E. N. MacDonald, "Task Dialog By Native-Danish Talkers In Danish And English In Both Quiet And Noise," Mar. 2018, publisher: Zenodo.

[28] R. Baker and V. Hazan, "DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs," *Behavior Research Methods*, vol. 43, no. 3, pp. 761–770, Sep. 2011.

[29] Z.-H. Tan, A. k. Sarkar, and N. Dehak, "rVAD: An unsupervised segment-based robust voice activity detection method," *Computer Speech & Language*, vol. 59, pp. 1–21, Jan. 2020.

[30] Z.-H. Tan, "rVAD2.0," Jul. 2020. [Online]. Available: https://github.com/zhenghuatan/rVAD

[31] S. Basu, "Conversational Scene Analysis," Ph.D. dissertation, Massachusetts Institute of Technology, Sep. 2002.

[32] A. Harma and K. Pham, "Conversation detection in ambient telephony," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 4641–4644.

[33] Y. O. Yohei Kawaguchi, Masahito Togami, "Turn Taking-Based Conversation Detection by Using DOA Estimation," in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, Makuhari, Chiba, Japan, Sep. 2010.

[34] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[35] 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, Nov. 2016.

[36] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2. Salt Lake City, UT, USA: IEEE, 2001, pp. 749–752.

[37] J. H. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Fifth international conference on spoken language processing*, 1998.

[38] H. L. Van Trees, *Optimum array processing*, ser. Detection, estimation, and modulation theory / Harry L. Van Trees. New York: Wiley, 2002, no. 4.

[39] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[40] 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*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[41] A. Moore, J. M. de Haan, M. Pedersen, D. Brookes, P. Naylor, and J. Jensen, "Personalized signal-independent beamforming for binaural hearing aids," *Journal of the Acoustical Society of America*, vol. 145, pp. 2971–2981, 2019.

[42] P. Minnaar, S. F. Albeck, C. Simonsen, B. Søndersted, S. Oakley, and J. Bennedbæk, "Reproducing real-life listening situations in the laboratory for testing hearing aids," *Journal of The Audio Engineering Society*, 2013.

[43] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," vol. 29, pp. 2840–2849.

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-786-4

AALBORG UNIVERSITY PRESS