Informed Sound Source Localization for Hearing Aid Applications

Farmani, Mojtaba

DOI (link to publication from Publisher):
10.5278/vbn.phd.tech.00009

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Farmani, M. (2017). *Informed Sound Source Localization for Hearing Aid Applications*. Aalborg Universitetsforlag. https://doi.org/10.5278/vbn.phd.tech.00009

INFORMED SOUND SOURCE LOCALIZATION FOR HEARING AID APPLICATIONS

BY
MOJTABA FARMANI

DISSERTATION SUBMITTED 2017

AALBORG UNIVERSITY
DENMARK

Informed Sound Source Localization for Hearing Aid Applications

Ph.D. Dissertation

Mojtaba Farmani

Dissertation submitted March 31, 2017

Dissertation submitted:	March 31, 2017
PhD supervisor:	Prof. Jesper Jensen Aalborg University
Assistant PhD supervisors:	Assoc. Prof. Zhen-Hua Tan Aalborg University Dr. Michael Syskind Pedersen Oticon A/S
PhD committee:	Associate Professor Tobias L. Jensen (chairman) Aalborg University Professor James Kates University of Colorado Reader Patrick Naylor Imperial College
PhD Series:	Technical Faculty of IT and Design, Aalborg University

ISSN (online): 2446-1628
ISBN (online): 978-87-7112-936-6

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2017

Curriculum Vitae

Mojtaba Farmani

Mojtaba Farmani received the B.Sc. and the M.Sc. degrees in electrical and computer engineering from the University of Tehran, Tehran, Iran, in 2009 and 2012, respectively. He is currently working towards a Ph.D. degree in electrical engineering at the Aalborg University, Aalborg, Denmark. He was a Research Assistant at the Technical University of Eindhoven, The Netherlands, a Visiting Researcher at Delft University of Technology, The Netherlands, and at University of Rostock, Germany. His research interests include localization, tracking, statistical signal processing, and audio and speech processing.

Abstract

Hearing impaired listeners often face difficulties in understanding speech, especially in noisy situations. A highly effective solution to solve this problem is to employ a hearing aid system (HAS), which can connect to a *wireless microphone* worn by the talker of interest. The wireless microphone allows the HAS to access an essentially noise-free version of the target signals that can be presented to the user. However, despite the increase in intelligibility, some users do not feel comfortable with this solution, because it does not provide the correct spatial cues of the target sound, so that the user cannot localize the target talker. This can reduce a user's sense of immersion and can cause the user to feel detached from the surroundings. Further, in situations where several talkers are simultaneously present and each of them are wearing a wireless microphone, lack of spatial cues can degrade the speech intelligibility of target signals, especially when some of the talkers are talking concurrently.

One solution to address these problems is to impose the correct spatial cues on the wirelessly received signals, before rendering them to the HAS user. To do so, one could solve the *informed* sound source localization (SSL) problem, i.e estimate the location of the target talker(s) based on the knowledge of the noise-free version of the target signal(s). Despite the fact that the informed SSL problem is mainly relevant in acoustically noisy situations, and that HAS microphones are typically located behind/in the users' ears, existing informed SSL algorithms often ignore ambient noise characteristics and effects of the user's head on the received signals. In this thesis, we propose a *maximum likelihood (ML) framework* for solving the informed SSL problem, which allows to take both ambient noise characteristics and effects of a user's head into account. Ambient noise characteristics can be relatively easily estimated based on the wirelessly available noise-free target signal and the noisy target signals captured by the HAS microphones. To model effects of the head, we employ four different head models, which include generic models, which do not depend on a specific user, and individualizable models, which allow to take user-specific details into account. For each of the head models, we propose an informed SSL algorithm using the ML framework. Eventhough the computational complexity of the proposed methods differ, each of the proposed algorithms has been formulated with computational efficiency in mind. Some of the proposed methods are flexible in the sense that they do not depend on a particular microphone array configuration. For these methods, we study how the microphone array geometry affects their performance. This is important because some microphone configurations (e.g. binaural) may require higher implementation costs than others (e.g. monaural). Finally, we assess the performance of the proposed methods in different noisy and reverberant conditions to demonstrate and to compare their effectiveness.

Resumé

Hørehæmmede lyttere har ofte problemer med at forstå tale, specielt i støjfyldte situationer. En særdeles effektiv løsning på dette problem er, at anvende et høreapparat som kan forbindes til en trådløs mikrofon båret af den taler, man ønsker at lytte til. Den trådløse mikrofon giver høreapparatet adgang til en essentielt set støjfri version af talen, som så kan præsenteres direkte til brugeren. Til trods for en forbedret taleforståelse, er mange brugere dog ikke komfortable med denne løsning, da den ikke viderebringer de egenskaber ved lyden, som normalt tillader mennesker at lokalisere taleren. Dette kan reducere brugerens grad af indlevelse og kan få brugeren til at føle sig afkoblet fra omgivelserne. Yderligere, i situationer hvor flere talere bærer mikrofoner på samme tid, kan den manglende evne til at lokalisere talerne føre til tab af taleforståelse, især hvis talerne taler på samme tid. En løsning på disse problemer er, at påtrykke de korrekte egenskaber på lyden fra den trådløse mikrofon, før denne præsenteres til høreapparatsbrugeren. For at gøre dette, kan man gøre brug af informeret lydkildelokalisering (LKL), dvs. estimere placeringen af taleren baseret på kendskab til det støjfrie signal. På trods af det faktum, at LKL oftest er relevant i støjfyldte situationer, og at mikrofonerne på et høreapparat typisk er placeret bag eller i brugerens øre, ignorerer eksisterende LKL-algoritmer ofte både støjkarakteristika og den akustiske effekt, som brugerens hoved har på de modtagede signaler. I denne afhandling foreslår vi et maximum-likelihood-baseret framework til at løse LKL-problemet, som tillader os at tage højde for både baggrundsstøjens karakteristika og den akustiske effekt af brugerens hoved. Baggrundsstøjens karakteristik er relativt enkel at estimere, baseret på det, trådløst tilgængelige, støjfrie talesignal og det støjfyldte signal opfanget af høreapparatets mikrofoner. For at modellere hovedets indflydelse anvender vi forskellige hovedmodeller, hvilke inkluderer generiske modeller, som ikke afhænger af den specifikke bruger, samt individualiserbare modeller som tager højde for brugerspecifikke detaljer. For hver af hovedmodellerne foreslår vi en informeret LKL-algoritme baseret på det nævnte maximum-likelihood-framework. Til trods for, at der er forskelle på beregningskompleksiteten af de forskellige metoder, er alle de udviklede algoritmer designet med beregningseffektivitet i sinde. Visse af de udviklede metoder er fleksible i den forstand, at de ikke afhænger af en bestemt fysisk konfiguration af mikrofoner. For disse metoder undersøger vi relationen mellem mikrofonkonfiguration og opnået ydeevne. Dette er væsentligt, da visse konfigurationer (f.eks. binaurale) er mere omkostningstunge at implementere end andre (f.eks. monaurale). Endeligt vurderer vi de udviklede metoders ydeevne i forskellige situationer med baggrundsstøj og rumklang for at sammenligne deres effektivitet.

Contents

Preface

This thesis is submitted to the *Technical Doctoral School of IT and Design* at Aalborg University, Denmark, in a partial fulfilment of the requirements for the Ph.D. degree.

The work was carried out in the period from April 2014 to April 2017, jointly at *Signal and Information Processing* (SIP) group of the Department of Electronic Systems, Aalborg University, and *Audiological Design* & *Signal Processing* group at Oticon A/S, Denmark. Some of the work was conducted in collaboration with the *Circuits and Systems* (group) of the Department of Microelectronics, Delft University of Technology, The Netherlands.

This thesis is composed of two parts, the introduction and the main body. In the introduction part, we provide an overview of how humans localize sound sources (Chapter 1) , and how hearing loss (Chapter 2) and hearing aids (Chapter 3) can influence the localization performance of humans. Moreover, we review existing sound source localization algorithms, and discuss their advantages and disadvantages for hearing aid applications (Chapter 4). Finally, we motivate our research and provide a summary of contributions made in this thesis (Chapter 5). The main body of the thesis constitutes of eight research papers, which have been published in or submitted to peer-reviewed journals or conferences.

List of Papers

The main body of this thesis consist of the following papers.

[A] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Informed TDoA-based direction of arrival estimation for hearing aid applications," in *Proceedings of IEEE Global Conference on Signal and Information Processing*, 2015, pp. 953–957.

[B] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Informed direction of arrival estimation using a spherical-head model for hearing aid applications," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2016, pp. 360–364.

[C] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Maximum likelihood approach to "informed" sound source localization for hearing aid applications," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2015, pp. 16–20.

[D] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "On the influence of microphone array geometry on HRTF-based sound source localization," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2015, pp. 439-443.

[E] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Informed sound source localization using relative transfer functions for hearing aid applications," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 611–623, March 2017.

[F] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Bias-compensated informed sound source localization using relative transfer functions," submitted to *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

[G] M. Farmani, R. Heusdens, M. S. Pedersen, and J. Jensen, "TDOA-based self-calibration of dual-microphone arrays," in *Proceedings of European Signal Processing Conference*, 2016, pp. 617-621.

[H] M. Farmani, R. Heusdens, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Concurrent localization of sound sources and dual-microphone sub-arrays using TOFs," in *Proceedings of International Conference on Information Fusion*, 2016, pp. 1931-1936.

In addition to the papers, the following patents have been filed.

[I] J. Jensen, M. S. Pedersen, M. Farmani, and P. Minnaar, "Hearing system," European Patent Application, EP14189708.2, October 2014.

[J] J. Jensen, M. Farmani, and M. S. Pedersen, "A hearing device and a hearing system configured to localize a sound source," European Patent Application, EP15189339.3, October 2015.

[J] M. Farmani, M. S. Pedersen, and J. Jensen, "A binaural hearing system configured to localize a sound source," European Patent Application, EP16182987.4, August 2016.

[K] M. Farmani, M. S. Pedersen, J. Jensen, S. O. Petersen, and A. Thule, "A method of localizing a sound source, a hearing device, and a hearing system," European Patent Application, EP17160114.9 , March 2017.

Acknowledgment

First and foremost, I would like to express my appreciation to my main supervisor, *Prof. Jesper Jensen*, without whom this thesis would not have been possible. His deep insights, constructive advice, and precise guidance helped me at various stages of my research. Secondly, I would like to express my gratitude to my industrial supervisor, *Dr. Michael Syskind Pedersen*, for his great support, valuable ideas and enlightening discussions through this project. I am also grateful to my co-supervisor, *Prof. Zheng-Hua Tan*, for his friendly and helpful support in conduction of my research.

I would like to thank my colleagues at the *Signal and Information Processing* (SIP) group of the Department of Electronic Systems, Aalborg University, as well as the people at the *Audiological Design* & *Signal Processing* group at Oticon A/S. It has been a pleasure to work with all these nice people. Particularly, I would like to thank my dear friend *Asger Heidemann Andersen* for translating the abstract of this thesis into Danish.

I am grateful to *Prof. Richard Heusdens*, whom I was visiting at the Delft University of Technology (TUDelft), The Netherlands. I would also like to thank the people at the *Circuits and Systems* (CAS) group at TUDelft for making my visit very pleasant.

My special deepest thanks to my *family*, especially my parents, for their continuous support and encouragement; and finally, I would like to sincerely thank my wife, *Zahra*, for her love, tolerance, and constant support.

Mojtaba Farmani
Aalborg University, March 31, 2017

List of Acronyms

AE	Absolute Error
ANOVA	ANalysis Of VAriance
ASA	Auditory Scene Analysis
BRIR	Binaural Room Impulse Response
BTE	Behind The Ear
CIC	Completely In the Canal
CPSD	Cross Power Spectral Density
DFT	Discrete Fourier Transform
DoA	Direction of Arrival
DRH	Dynamic Range of Hearing
DRR	Direct to Reverberant Ratio
DS	Delay and Sum
EEG	ElectroEncephaloGram
FHSS	Frequency Hopping Spread Spectrum
FM	Frequency Modulation
GCC	Generalized Cross Correlation
HA	Hearing Aid
HAS	Hearing Aid System
HATS	Head And Torso Simulator
HRIR	Head Related Impulse Response
HRSE	High Resolution Spectral Estimation
HRTF	Head Related Transfer Function
IDFT	Inverse Discrete Fourier Transform
IF	Information Fusion
IID	Interaural Intensity Difference
ILD	Interaural Level Difference
IMLD	Inter Microphone Level Difference
IMTD	Inter Microphone Time Difference
ITC	In The Canal
ITD	Interaural Time Difference
ITE	In The Ear

JND	Just Noticeable Difference
LS	Least Squares
LSD	Log-Spectral Distance
MAE	Mean Absolute Error
ML	Maximum Likelihood
MDS	Multi Dimensional Scaling
MLSSL	Maximum Likelihood Sound Source Localization
MMSE	Minimum Mean Squared Error
MTF	Multiplicative Transfer Function
NFMI	Near Field Magnetic Induction
PHAT	PHAse Transform
RF	Radio Frequency
RSS	Received Signal Strength
RTF	Relative Transfer Function
SNR	Signal to Noise Ratio
SRP	Steered Response Power
SSL	Sound Source Localization
STFT	Short Time Fourier Transform
SVD	Singular Value Decomposition
TDE	Time Delay Estimation
TDoA	Time Difference of Arrival
ToA	Time of Arrival
ToF	Time of Flight
UCL	UnComfortable Level
VAD	Voice Activity Detector
WDRC	Wide Dynamic Range Compression

Part I

Introduction

Introduction

1 Human Sound Source Localization

Sound source localization (SSL) refers to the process of identifying the location of a sound source, mostly in terms of direction and distance, relative to the head of a listener [10, 100]. SSL intuitively plays an important role in our perception of and interaction with the environment around us, especially when sound sources are not in our field of vision.

Realistic acoustic scenes generally consist of several spatially distributed sound sources, and sound signals reaching our ears are often complex mixtures of signals originating from different sources. Amazingly, our auditory system can, to a great extent, analyze the mixture of signals reaching our two ears, perceptually decompose the mixture into the constituent signals, and build a picture and a perception of the acoustic scene around us [18]. This ability is known as *auditory scene analysis* (ASA) [18]. The information, which we gain from ASA, generally includes the number of sounds that are present, and the timbre, pitch, loudness, clarity, and location of each [18]. Therefore, we can consider SSL as an essential part of ASA.

The most well-known phenomenon used to illustrate the ASA ability in humans is the so-called *cocktail party effect* [20, 112], which describes the capability of humans to focus intentionally on a conversation with one talker, while other talkers are speaking simultaneously. It has been shown that localization cues of a target talker assist the auditory system to extract the target speech signal from a complex mixture of signals of different talkers, in a way that it is more intelligible [18, 53, 64, 65, 146]. In other words, SSL helps to improve speech intelligibility.

In this chapter, we review the spatial cues, which the human auditory system can exploit to localize a sound source. We do so, because—as we shall see—the SSL algorithms proposed in this thesis exploit these cues, too. Further, we review the SSL performance of the human auditory system, as it defines the accuracy, which the SSL algorithms proposed in this thesis must provide.

Before the review, let us define some terms, which will be used. The coor-

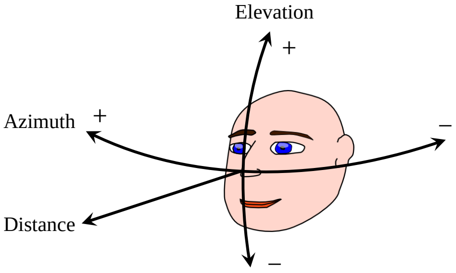


Fig. 1: The coordinate system used to define the position of a sound source relative to the head.

dinate system employed to define the position of a sound source is illustrated in Fig. 1. The direction of a sound source is defined in terms of its *azimuth* and its *elevation* relative to the listener's head (cf. Fig. 1). Sound signals of a source located at an azimuth of 0° and at an elevation of 0° originate exactly from the front of the listener's head. The *horizontal plane* is defined as the locations, whose elevations are 0° [10, 100]. The *median plane* is defined as the locations, whose azimuths are 0° and are at an equal distance from the two ears of the listener [10, 100]. Unless otherwise stated, in this study, we use the term "localization" for identifying only the *direction* of a source. Further, we generally assume that point sound sources are in the far-field[1] with respect to the listener. Generally speaking, the far-field assumption is valid when the sound source distance is 5-10 times larger than the size of the head[2] [128]; therefore, in everyday life, many of the encountered sound sources can be considered in the far-field.

1.1 Localization cues

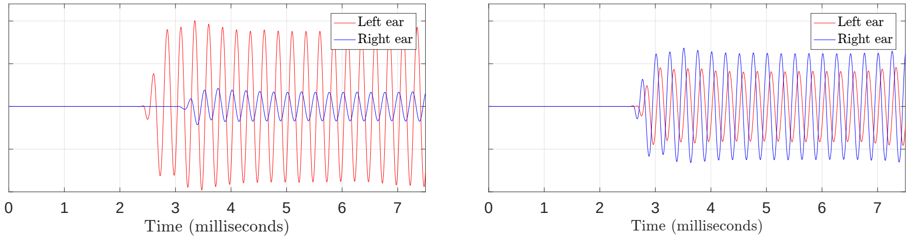
The cues used by the human auditory system to localize a sound source has been investigated widely, e.g [10, 24, 51, 55, 115, 116, 120, 146, 153, 154]. Generally, localization cues can be categorized as *binaural cues* or *monaural cues*.

Binaural cues are based on the differences between the signals received at the left and the right ear [100, 111]. Fig. 2[3] shows exemplary signals received at the left and the right ear of a listener, from a source generating a ramped

[1] Far-field is formally defined as "a region in free space, distant from a sound source, where the sound pressure level obeys the inverse square law (the sound pressure level decreases 6 dB with each doubling of distance from the source). Also, in this region the sound particle velocity is in phase with the sound pressure" [56]. In signal processing algorithms using microphone arrays, a far-field situation implies that the wavefronts impinging on the array can be considered as plane waves [13, 93].

[2] To be more precise, the sound source distance should be 5-10 times larger than the distance between the ears ("microphones").

[3] To generate the figure, the head related impulse responses measured in the ear and provided by [78] have been used.



(a) The source is located at an azimuth of -90° and an elevation of 0°, i.e. on the left-hand side of the listener.

(b) The source is located at an azimuth of $+10^\circ$ and an elevation of 0°, i.e. to the front-right of the listener.

Fig. 2: Exemplary signals received at the ears of a listener from a source placed 0.8 meters away generating a 4-kHz sinusoidal signal.

4-kHz sinusoidal signal. As can be seen, depending on the location of the source, one of the received signals is delayed and attenuated with respect to the other one. More precisely, when the source is located on the left-hand side of the listener, the signal received at the right ear is delayed and attenuated with respect to the signal received at the left ear, and vice versa. These differences are due to the presence of the head and the difference of the distances between the source location and each of the ears. The binaural cues obtained from the relative delays and attenuations of the received signals are, respectively, referred to as *interaural time differences* (ITDs) and *interaural level differences* (ILDs). In addition to binaural cues, several studies, e.g. [10, 24, 142, 153], have shown that humans are still able to localize a sound source, to some extent, when the sound signals are presented to only one of the ears. The cues used by the auditory system to localize a sound source in monaural situations are called monaural cues.

Interaural time difference (ITD)

As demonstrated in Fig. 2, the time of arrivals (ToAs) of the sound signals of a source received at the ears of a listener depend on the relative location of the source and the listener. The ToAs differ because the signals travel different distances and paths to arrive at the ears [111]. The ITD, which we denote by ΔT, is defined as the difference between the ToAs [111], i.e. $\Delta T = t_{\text{right}} - t_{\text{left}}$, where t_{right} and t_{left} represent the ToAs of the sound signal at the left and the right ear, respectively.

In the literature, both physical and perceptual aspects of the ITD have been investigated. Some of the main physical aspects of ITDs include:

- The magnitude of the ITD in adults varies in the range of 0 to approximately 690 μs (the exact range depends on the size of the listener's head) [51, 100]. The maximum magnitude of the ITD occurs when the sound source is located at the sides of the head, i.e. azimuths of ±90°, and the ITD of 0 occurs when the source is on the median plane.

- The relation between ITD ΔT and azimuth θ of a source in the horizontal plane can be approximated by [10]

$$\Delta T \approx \frac{a}{c}(\theta + \sin(\theta)), \tag{1}$$

where a is the radius of the head, and c is the speed of sound. This relation has been derived based on a rigid spherical head model [10, 86].

- In principle, the ITD not only depends on the location of the source, but also on the frequency of the source signal [2, 10, 51, 86, 100, 111]. The frequency-dependency of the ITD is due to the diffraction of sound waves imposed by the head of the listener [43, 86]. As a rule of thumb, the ITD measured at low frequencies ($<$ 500 Hz) is 50% greater than the ITD measured at high frequencies ($\gtrsim$ 2000 Hz) [42, 86].

The main perceptual aspects of the ITD can be summarized as:

- The *smallest detectable ITD* of single tones by human listeners is generally frequency dependent [22, 81]. The smallest threshold is when the frequency of the signal is in the range of 700 to 1000 Hz [22]. At lower frequencies, i.e. in the range of 250 to 700 Hz, the threshold is inversely proportional to the frequency of the signal, while at higher frequencies, more than 1000 Hz, the threshold is increasing rapidly (faster than exponentially) [22]. For a broadband noise in the horizontal plane, the smallest detectable ITD is approximately 10 μs [81].

- The smallest perceivable *change* in an ITD by a human listener is referred to as the *just noticeable difference* (JND) of the ITD [111]. The JND of an ITD directly relates to the magnitude of the ITD [60], i.e. the greater the magnitude of the ITD, the greater the JND of the ITD [60]. In other words, the smallest JNDs of ITD occur when the sound source is in the median plane, while the greatest JNDs of ITD occur when the sound source is at the sides. This intuitively implies that the smallest detectable change in the azimuth of a source directly depends on the current azimuth of the source. In other words, humans have the best spatial resolution when the source is in the front (humans can detect changes of approximately 1° in the angle [111]), while they have the worst spatial resolution when the source is at the sides (they can detect changes of approximately 20° in the angle [111]).

- When the sound signal includes low frequency components ($<$ 1500 Hz), the ITD is the dominant and the main cue used by the auditory system to identify the direction of the sound source [111, 154]. Hence, for most natural sounds, localization is dominated by ITD cues [111].

- Besides the *fine waveform* of a sound, humans can exploit the fluctuations in the sound temporal *envelope*, i.e. "the slower variations in the peak amplitude of the waveform" [111], to estimate ITD, particularly at high frequencies [111, 137].

Interaural level difference (ILD)

To define ILD, let us first define the intensity and the level of a sound signal. The intensity of a sound signal is defined as the energy of the sound waves passing through a unit area [111]. Expressing the intensity in decibels with respect to a reference intensity is called the level (l) of the sound signal. In other words,

$$l = 10 \log_{10}(\frac{I}{I_0}), \tag{2}$$

where I is the sound intensity, and I_0 is a reference intensity [111]. The conventional value of I_0 is $10^{-12} \frac{\mathrm{W}}{\mathrm{m}^2}$ [111], and a sound level, which is expressed with respect to this I_0, is called a sound pressure level (SPL).

The ILD, which we denote by ΔL, is defined as the difference in the level of a sound signal received at the two ears [111], i.e. $\Delta L = l_{\text{right}} - l_{\text{left}}$, where l_{left} and l_{right} are the levels of the sounds received at the left and right ear, respectively. As suggested by Fig. 2, similarly to the ITD, the ILD also depends on the relative location of the source and the listener.

The main cause of the ILD is the *"shadowing effect"* of the listener's head on the received sound [10, 111]. In other words, the head blocks some of the energy of the sound from reaching the contralateral ear [111], and this leads to the ILD. As for the ITD, the ILD is also frequency dependent [10, 61, 100], and it generally has a direct relationship with the frequency of the signal [51, 100], i.e. the higher the frequency, the greater the ILD.

The ILD is smaller at lower frequencies because of the diffraction of the sound signals [100]. In other words, at low frequencies, the wavelength is larger than the size of the listener's head; therefore, the waves can bend around the head, and only a small "shadow" is made by the head. For frequencies less than 500 Hz, the ILD is very small and negligible in far-field situations; however, when the sound source is very close to the listener's head, notable ILDs may occur, even at low frequencies [23, 100]. At high frequencies, the wavelength is smaller than the size of the head, and this prevents diffraction, i.e. a strong "shadow" can occur [100].

Interestingly, the *smallest detectable ILD* by human listeners in a frequency range from 200 to 5000 Hz is roughly frequency-independent [61, 156, 157]. The smallest detectable ILD in this frequency range is approximately 0.5 to

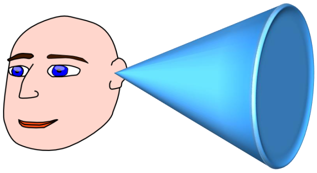


Fig. 3: An exemplary cone of confusion. Every locations on the surface of the cone produce similar ITDs and ILDs.

1 dB [61, 156]. However, the smallest detectable ILD seems to decrease for higher frequencies [52]. Moreover, as with ITD, the greater the magnitude of an ILD, the greater the JND of the ILD [149].

Monaural cues

Even though the binaural cues play the most important role in human SSL, especially for sound sources in the horizontal plane, the binaural cues alone are not sufficient to identify the exact location of a sound source in a three-dimensional space [111]. This is because there is no one-to-one mapping between the binaural cues and the possible locations. For example, an ITD of 0 seconds and an ILD of 0 dB ($\Delta L \approx 0$ and $\Delta T \approx 0$) can essentially be produced from all locations on the median plane [10]. More generally, for any particular ITD, there are an infinity of locations, which can produce that particular ITD. These locations form the surface of a cone, which is known as a *cone of confusion* (cf. Fig. 3) [111]. ILDs produced from locations on a cone of confusion are also similar [111]. Therefore, the ITD and the ILD are not sufficient to determine the exact location of the sound source. Further, as mentioned earlier, when sound signals are presented to just one of the ears, i.e. when binaural cues are not available, humans can still to some extent localize sound sources [10, 100, 111, 142, 153]. Therefore, besides binaural cues, our auditory system exploits other spatial cues, which we refer to as monaural cues.

The monaural cues are based on spectral changes imposed by the head and torso and, especially, by the pinna of the listener on incoming sounds [10, 100, 111]. These spectral changes depend on the direction of arrival (DoA) of the sound signal, and are prominent at higher frequencies [10]. Moreover, these spectral changes are often measured and represented in terms of *head related transfer functions* (HRTFs) [27], which are formally defined as "a specific individuals left or right ear far-field frequency response, as measured from a specific point in the free field to a specific point in the ear canal" [27].

The monaural cues are specifically important for determining the elevation [61, 116], and for SSL at high frequencies, above approximately 4000 Hz, because the wavelength of the signal is comparable to the size of the pinna, head and torso at high frequencies [61, 100, 111]. Moreover, HRTFs differ across different individuals, especially above 6000 Hz, because the size and the shape of their heads and pinnae are different [100]. Several studies have shown that individualized monaural cues (HRTFs) play a main role in resolving front-back and up-down confusions, or more generally, localization along a particular cone of confusion [67, 150].

It should be noted that besides the monaural cues, humans exploit their head movements to resolve confusions and to facilitate the localization of sound sources [72, 102, 108, 152]. Movements of the head relative to a sound source change the monaural and binaural cues and reveal additional evidence of the source position, which humans integrate to resolve ambiguities.

1.2 Distance estimation

So far, we have focused on the spatial cues used by the auditory system to localize a sound source in terms of its direction. Another important location information is the sound source distance, which humans use to interact with the environment, e.g. to avoid a vehicle approaching from behind [159].

The primary cue to distance, especially for familiar sounds, is believed to be *sound level* [100, 111, 159]. Physically, sound level has an inverse relationship with distance. In free-field and far-field situations, this relationship obeys an inverse-square law, i.e. every doubling in distance implies a 6 dB reduction in sound level [111, 159]. In situations, where sound level is the only cue to distance, it has been demonstrated that the perceived distance is independent of the actual physical distance [54, 159]. However, the sound level *change*, e.g. when a listener walks towards a sound source, can be an absolute cue to distance [6, 100]. Moreover, when multiple sound sources are present in the acoustic scene, sound level seems to be the most beneficial cue for distinguishing the relative distances of the sound sources [100].

Besides sound level, reverberation is another factor which contributes to humans' estimate of distance. Several studies have shown that perceived distances are more accurate in reverberant environments than in anechoic situations [21, 97, 159]. Moreover, it has been shown that the perceived distance is inversely related to the direct-to-reverberant ratio (DRR) of the sound, i.e. the lower the DRR, the greater the perceived distance. [100, 111, 159]. Further, this cue seems to be beneficial for distance estimation, even when the sound is not familiar [111]. However, because the ability of humans in detecting changes in the DRR is limited [111, 158]—we can detect changes in the DRR when the sound distance more than doubles—it has been suggested that the DRR provides a primary cue to absolute distance rather than a cue for detecting relative distances [158, 159].

Over long distances, longer than 15 meters, the spectrum of the sound may be changed by the absorbing properties of the air, especially at high frequencies [100, 111, 159]. This change affects the *spectral balance* of the sound [111], and is another cue to distance, especially when the sound is familiar [100]. However, this effect is relatively small, e.g. at 4 kHz, the power loss is 3 to 4 dB per 100 meters [159].

1.3 General remarks

Here, we briefly review some of the important points about the SSL performance of humans:

- Humans generally perform much more accurately in estimating the direction than estimating the distance [10, 159].
- Humans generally perform better in estimating the azimuths of sound sources than their elevations [10].
- Humans perform best in estimating the source direction, when the source is in front of the listener (azimuth of 0°), while perform worst, when the source is to the sides (azimuth of ±90°) [10].
- Humans usually underestimate the distance to faraway sound sources, while they often overestimate the distance to close sound sources (generally, when the distance is less than one meter) [159].
- SSL performance of humans improves when sound signals of the source are familiar [10].
- Non-acoustical cues, such as visual cues, obviously also contribute to SSL in humans [111, 159], e.g. to resolve front-back confusions.

2 Hearing loss

Hearing loss is one of the most common physical disabilities in the world [111]. Statistics show that over 5% of the world's population (around 360 million people) have *disabling* hearing loss that markedly disturbs their daily life [1].

Broadly speaking, hearing loss can be defined as any impairment to humans' ability to receive, process and perceive sounds, which are normally audible. More formally, hearing loss is defined as an increase in the *hearing threshold*, which is the minimum sound pressure level required to perceive a pure tone[4] [121]. Our auditory system is sensitive to frequencies from approximately 20 Hz to 20000 Hz [100]. Fig. 4 shows the hearing threshold of

[4] Note that some auditory deficits do not change the pure tone hearing threshold, but they still limit the perception of some sounds that are above the normal hearing threshold. Such a hearing impairment is called *hidden hearing loss* [110].

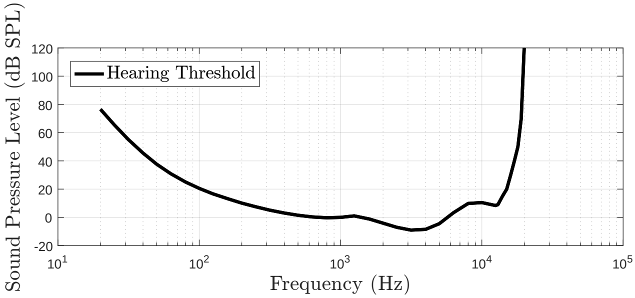


Fig. 4: The hearing threshold of humans with normal hearing [71].

humans with normal hearing at these frequencies [71]. The hearing loss of a listener is recognized by the *difference* between the individual hearing threshold and the threshold shown in Fig. 4 [121]. It should be noted that the hearing threshold generally largely varies among individuals, and Fig. 4 shows an average of hearing thresholds, which are considered to be normal. Any individual hearing threshold up to 15 dB higher than the threshold shown in Fig. 4 is considered normal [121].

To describe details and consequences of hearing loss, it is necessary to briefly review the peripheral auditory system and its different parts.

2.1 The peripheral auditory system

As depicted in Fig. 5, the peripheral auditory system consists of three main parts: *outer ear*, *middle ear*, and *inner ear* [111]. The outer ear, which consists of the pinna and the ear canal, collects sounds from the environment, passively amplify some components of the sounds, and conducts the sounds to the middle ear [121]. The *eardrum* (tympanic membrane) separates the outer ear from the middle ear. The middle ear consists of *ossicles*, which are three tiny bones called *malleus*, *incus* and *stapes* [111, 121]. The main responsibility of the middle ear is to efficiently transmit the pressure vibrations made by the sound waves to the inner ear [111, 121]. The inner ear—*cochlea*—is the, perhaps, most important part of the peripheral auditory system. The main task of the cochlea is to *transduce* the mechanical vibrations into electrical nerve impulses, which are transferred to the brain for processing via the auditory nerve [111, 121].

Structure of the cochlea

The cochlea is a spiral tube with a shape similar to a snail shell [109, 111]. Fig. 6 depicts a cross-section of the cochlea and its components. Two membranes, *Reissner's membrane* and the *basilar membrane*, divide the cochlea along

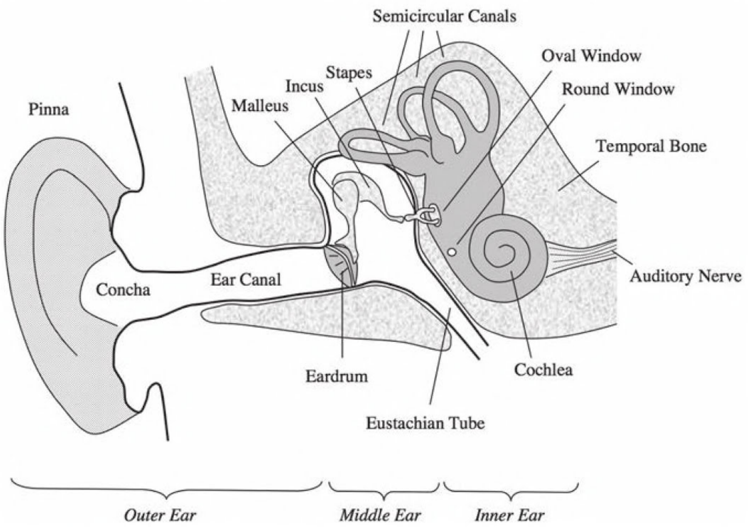


Fig. 5: The peripheral auditory system (copied with permission from [111]).

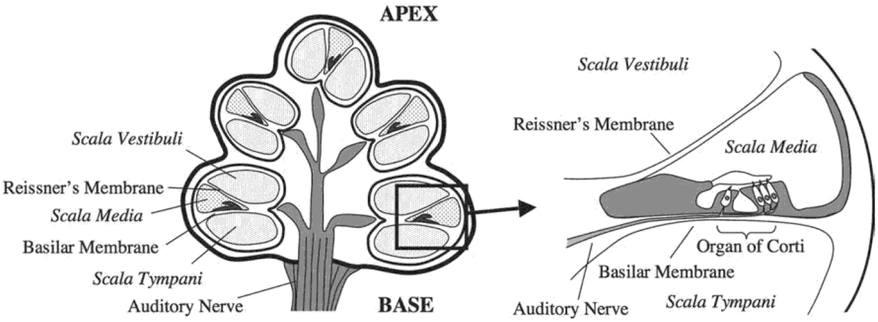


Fig. 6: A cross-section of the cochlea (copied with permission from [111]).

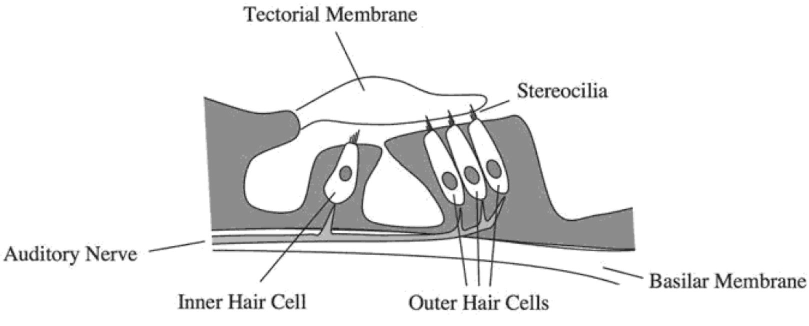


Fig. 7: The tectorial membrane and the organ of Corti (copied with permission from [111]).

its length into three fluid-filled compartments: the *scala vestibuli*, the *scala media*, and the *scala tympani* [100, 111]. The scala vestibuli and the scala tympani are connected to each other via a small opening (*helictrema*) at the apex of cochlea, while the scala media is a completely separate compartment [109, 111].

In the scala media and above the basilar membrane, a gelatinous structure called the *tectorial membrane* is located [109, 111]. Between the tectorial membrane and the basilar membrane is the *organ of Corti* [100, 109, 111]. Fig. 7 depicts more details of this structure. The organ of Corti contains rows of *hair cells*, which can be divided into *inner hair cells* and *outer hair cells* [109, 111]. In the cochlea of humans, there are often one row of inner hair cells and up to five rows of outer hair cells [100, 111]. On top of the hair cells, there are protein filaments called *stereocilia* [100, 109]. In contrast to inner hair cells, the tallest tips of the stereocilia in outer hair cells are implanted in the tectorial membrane [111]. The main task of the inner hair cells is to transduce the vibrations of the basilar membrane into neural activities, while outer hair cells are believed to actively help to amplify the vibrations of the basilar membrane [100, 111]. The amplification of the basilar membrane vibrations by outer hair cells is greatest for low sound levels, whereas it decreases to zeros for sound levels above around 90 dB SPL [111].

Functionality of the cochlea

Sounds are conveyed to the cochlea via the *oval window*, which is covered by a membrane [100, 111]. More precisely, to convey the sounds, the stapes move the membrane of the oval window according to the vibrations of the sounds [111]. Due to the incompressibility of the fluid in the cochlea, the movements of the oval window lead to vibrations of the basilar membrane [100, 111].

The basilar membrane plays a crucial role in hearing. The main effect of the basilar membrane is to break down a sound into its frequency components [109, 111]. The stiffness of the basilar membrane varies continuously along its length; therefore, different parts of the basilar membrane are sensitive to different frequency components of a sound [100, 111]. This implies that different frequency components of a sound excite different parts of the basilar membrane [100, 111]. High frequency components are generally processed at places near to the oval window, while low frequency components are generally processed at places near to the other end of cochlea (apex) [109, 111]. In other words, the basilar membrane acts as a *bank of overlapping band-pass filters* called *auditory filters* [100, 111]. This *tonotopic* (frequency-to-place) representation of frequencies is preserved at almost all levels of auditory processing [100, 111, 146].

When the basilar membrane moves, it leads to electrochemical activities in the pertinent inner hair cells and causes *neurotransmitters* to be released [109, 111]. The neurotransmitters activate the relevant neuron in the auditory nerve and cause the neuron to generate neural *spikes* [109, 111]. The magnitude of the basilar membrane vibrations, at the place where the neuron is linked, is directly correlated with the firing rate of the neuron [109, 111]. Moreover, it has been shown that the neurons in the auditory nerve tend to fire spikes at a particular phase in the vibration of the basilar membrane [100, 111]. This mechanism is referred to as *phase locking* [100, 111]. It should be noted that because the neural activities in inner hair cells are limited, it is believed that the phase locking occurs for frequencies up to 5000 Hz [74, 111]. Above 5000 Hz, the neurons seem to be phase locked to the sound envelope, i.e. "the slower variations in the peak amplitude of the waveform" [111].

Phase locking of neural spikes plays an important role in SSL of humans, as it allows the *central auditory system* (the auditory centers in the brainstem and cerebral cortex) to derive the ToAs of sounds at the two ears [58, 100, 111]. To extract ITDs, it is believed that the central auditory system exploits an array of neurons, where each neuron in the array is most sensitive to a specific delay between the inputs from the ears [73, 118, 146]. To extract ILDs, the firing rate of the neurons plays the most important role. To be more precise, because the firing rate of the neurons linked to the inner hair cells depends on the intensity of sound [111, 146], ILDs are believed to be estimated by the central auditory system from the differences between the firing rates of these neurons at the two ears [146].

2.2 Types of hearing loss

The type of hearing loss is identified by which part of the auditory system that does not function normally [37, 111, 121]. Hearing loss can be categorized into the following types [121]:

1. *Conductive hearing loss:* Any disorder in the outer or middle ear can cause a *conductive hearing loss*, which degrades the efficiency of the conduction of the sound to the inner ear [37, 121]. Conductive hearing loss constitute approximately 10% of all hearing losses [121] and are often treated by surgery or medical treatments. Sometimes, the effects of a conductive hearing loss can be mitigated by bone-anchored hearing aids, which bypass the middle ear by transmitting the sound through vibration of the skull bone in accordance with the sound [121, 123, 131].

2. *Sensorineural hearing loss:* This type of hearing loss occurs when the cochlea or possibly the auditory nerve is damaged [37, 111, 121]. Sensorineural hearing loss is the most common type of hearing loss—it constitutes around 90% of all hearing losses [121]. The underlying cause of most sensorineural hearing is damage of the hair cells in the cochlea [111]. Damage of the inner hair cells diminishes the sensitivity

Table 1: Degree of hearing loss [31, 132].

Degree	Hearing loss range [dB]
Normal	-10 to 15
Minimal	16 to 25
Mild	26 to 40
Moderate	41 to 55
Moderately severe	56 to 70
Severe	71 to 90
Profound	>90

of the ear to the vibrations of the basilar membrane, whereas damage of the outer hair cells weakens the vibrations of the basilar membrane [111]. Nonetheless, both types of damage lead to elevation of the hearing threshold [111]. Currently, sensorineural hearing loss cannot be treated by surgery or medical treatments [37, 121]. However, amplifying the acoustic signals by conventional hearing aids can often mitigate the effects of this type of hearing loss [37, 121].

3. *Mixed hearing loss:* When a conductive hearing loss and a sensorineural hearing loss occur simultaneously, it is called a mixed hearing loss [121].

4. *Central hearing loss (or hidden hearing loss):* This type of hearing loss is due to disorders in the central auditory nerve system [121]. A central hearing loss usually barely changes the hearing thresholds, but it degrades the word or speech recognitions abilities. Central hearing loss is uncommon, and currently does not have any treatment [121].

2.3 Degree of hearing loss

The degree of hearing loss describes the severity of the loss [132]. Based on the average hearing loss at frequencies 500 Hz, 1000 Hz, and 2000 Hz (these frequencies are important for understanding speech, and are known as *speech frequencies* [121]), the degree of a hearing loss has been classified into seven categories shown in Table 1 [31, 132].

2.4 Configuration of hearing loss

The configuration of a hearing loss defines general characteristics of the hearing loss [121, 132]. One of the main characteristics of a hearing loss is its spectral shape [121, 132]. The shape of hearing loss across frequency has generally been categorized into seven groups[5]:

[5] This is an extended version of the categorization proposed in [121].

1. *Flat:* The differences between the hearing loss at different frequencies is less than 20 dB.
2. *Rising:* The hearing loss at low frequencies is at least 20 dB larger than the hearing loss at high frequencies.
3. *Sloping (ski slope):* The hearing loss at high frequencies is at least 20 dB larger than the hearing loss at low frequencies.
4. *Low-frequency:* The hearing loss affects only low frequencies.
5. *High-frequency:* The hearing loss affects only high frequencies.
6. *Cookie bite:* The hearing loss affects only mid frequencies [98].
7. *Precipitous:* The hearing loss is increasing steeply towards high frequencies (at least 20 dB per octave.) [121].

Other characteristics of a hearing loss defined by its configuration can be summarized as [121, 132]

- *Unilateral vs. bilateral:* If a hearing loss affects only one of the ears, it is called a unilateral hearing loss; otherwise, it is a bilateral hearing loss.
- *Symmetrical vs. asymmetrical:* In a bilateral hearing loss, if the shape and the degree of the hearing loss at both ears are similar, it is called a symmetrical hearing loss; otherwise, it is a asymmetrical hearing loss.
- *Sudden vs. progressive:* If a hearing loss has occurred suddenly, it is called a sudden hearing loss. On the other hand, if the hearing loss has evolved over time, it is called a progressive hearing loss.
- *Stable vs. fluctuating:* If a hearing loss improves at times and deteriorates again, it is a fluctuating hearing loss; but if the hearing loss does not change over time, it is a stable hearing loss.

2.5 Effects of hearing loss

Some of the main impacts of a hearing loss, particularly a sensorineural hearing loss, are [37]:

i) *Reduction in audibility:* Due to elevation of the hearing threshold, hearing impaired listeners usually cannot detect quiet sounds. An important consequence of this is difficulties in understanding speech [37].

ii) *Reduction in dynamic range:* The *dynamic range of hearing* (DRH) is the range of sound levels, where our auditory system works effectively [111]. For the frequency range of 1000 Hz to 6000 Hz, an intact DRH is about

120 dB, while at lower and higher frequencies, the DRH decreases [111]. The DRH is characterized by its lower limit, which is the hearing threshold, and its upper limit, which is the pain threshold[6] (threshold of loudness discomfort) [37]. Even though a sensorineural hearing loss elevates the hearing threshold, it usually does not change the pain threshold. In other words, hearing-impaired listeners cannot hear soft sounds, whose levels are below their hearing thresholds, while they can hear intense sounds as loud as a normal-hearing listener [37, 121]. This implies that the DRH of a hearing-impaired person is often less than the DRH of a normal-hearing person [37, 111, 121]. It also implies that each increase of sound level leads to a bigger *loudness*—the perceived sound level—increase for a hearing impaired listener than for a normal hearing listener. The latter implication is referred to as *loudness recruitment* [37].

iii) *Reduction in frequency selectivity:* The frequency selectivity refers to the ability of the auditory system to resolve and separate the frequency components of sounds [100, 111]. This ability allows us to hear sounds with different frequencies that arrive at our ears simultaneously. The frequency selectivity stems from the fact that different frequencies are processed by different places in cochlea [100, 111]. Several studies, e.g. [11, 99, 101], have shown that sensorineural hearing loss reduces the frequency selectivity ability of hearing-impaired people. In other words, sensorineural hearing loss broadens the bandwidths or the *critical bands* of the auditory filters and degrades the ability of hearing-impaired people in detecting a target sound in noisy situations [100, 111].

iv) *Reduction in temporal resolution:* The temporal resolution refers to how fast our auditory system can detect and process *changes* in the characteristics of sounds over the time [100, 111]. This is one of the most important aspects of peripheral auditory system, because information in the auditory domain are primarily encoded in these changes [111]. The temporal resolution plays a crucial rule in understanding speech, especially in a noisy background [37, 100, 111]. For example, it allows us to take advantage of *dip listening*—catching brief moments of speech, when ambient noise levels momentarily drop-off [57]. Several studies, e.g. [68, 99, 126], have shown that hearing loss deteriorates the temporal resolution, which, in turn, leads to degradation in the ability of hearing-impaired people to understand speech [57].

2.6 Sound source localization by hearing-impaired listeners

As explained, hearing loss generally degrades transmission of the received sound information from the peripheral auditory system to the central audi-

[6] also known as uncomfortable loudness level (UCL).

tory system. This obviously can effect human SSL performance. It has been reported that unaided hearing-impaired listeners generally perform worse than normal-hearing listeners in SSL tasks [3, 8, 44, 62, 85, 90, 106, 107]. Degradations of SSL performance in hearing-impaired listeners generally relates to the type and configuration of the hearing loss of listeners, the spatial origin of the sound source, and the signal level [44, 62, 106]. In the following, we briefly describe important points of these relations.

In the horizontal plane, SSL performance of listeners with conductive hearing loss is usually markedly poorer than the performance of listeners with sensorineural hearing loss, when both type of listeners have similar degrees of hearing loss [44, 62, 106]. This is most likely because estimation of ITDs at low frequencies is usually severely disrupted for listeners with conductive hearing losses [106]—as mentioned above, ITDs at low frequencies are normally the dominant cues for SSL in the horizontal plane [100, 111].

In the frontal horizontal plane, SSL performance of listeners with *high-frequency* sensorineural hearing loss is comparable to normal-hearing listeners, as long as the hearing loss is not severe at either low (250 to 1000 Hz) or midrange (2000 to 4000 Hz) frequencies [106]. However, listeners with *high-frequency* hearing loss, generally suffer more from front-back confusions, compared with normal-hearing listeners, because high-frequency components, which assist the auditory system to resolve front-back confusions, have been affected by the hearing loss [106].

In the median plane, performance of listeners with a conductive hearing loss is still generally worse than the performance of listeners with sensorineural hearing loss [106]. However, compared with the horizontal plane, the difference between the performance of these two type of listeners is smaller [106]. This is partly because SSL in the median plane is associated with sensitivity of the auditory system to high frequencies [106], and listeners with conductive hearing loss, to some extent, can discriminate ILDs at high frequencies [62]. It has been shown that listeners with *sloping* hearing loss perform poorly in SSL in the median plane, because high-frequency components of sound signals, which are important for estimating elevation in the median plane, cannot be exploited by the auditory system [106].

3 Hearing aids

A hearing aid is a miniature sound system, which amplifies sounds to reduce the impact of hearing loss [75]. Often, hearing aids are the only solution to mitigate the effects of a sensorineural hearing loss, which is the most common type of hearing loss. Therefore, hearing aids play an important role in daily life of many people. In what follows, we generally restrict our attention to sensorineural hearing loss. Hence, whenever we use the term “hearing loss” alone, it refers to a sensorineural hearing loss.

3.1 Types of hearing aids

The type or style of a hearing aid system (HAS) determines its physical size and to where it should be worn [37, 75, 121]. The common types of hearing aids are: *behind the ear* (BTE), *in the ear* (ITE), *in the canal* (ITC), and *completely in the canal* (CIC) [75, 121]. The BTE is by far the most common type of HAS, and as its name implies, it is placed behind the outer ear (pinna) of a user [37]. The ITE is smaller than the BTE, and it is placed in the concha of the outer ear [37]. The ITC is sufficiently small such that it only occupies a small part of the cavity of concha up to the opening of the ear canal [37]. The CIC is the smallest conventional type of hearing aids and fits the size and shape of the individual ear canal [75].

Hearing aid users usually prefer a hearing aid to be as small as possible so that it is less visible; however, the smaller the hearing aid, the less space for the hearing aid's components (including battery and extra microphones), and hence, the less computational power [37]. Generally, each type of HAS has its own advantages and disadvantages, and the requirements of a user determines the appropriate type [37].

3.2 Hearing aid components

The first industrial HASs were analog devices, which consisted of a microphone, amplifier, and receiver packed in a case [75]. Nowadays, almost all HASs are digital devices, which manipulate the sound using digital signal processing algorithms—such as multichannel dynamic range compression, feedback cancellation and noise reduction—to enhance the effectiveness of the hearing aids [75, 121].

Fig. 8 shows the most basic hardware architecture of a digital hearing aid [37, 75]. It consists of a

- microphone(s): to convert sound waves into analog electrical voltages.
- analog-to-digital converter (ADC): to *digitize* the analog electrical voltages.
- digital signal processor (DSP): to manipulate the digital signal in order to amplify the sound and to improve its quality.
- memory: to store the data, parameters and instructions of signal processing algorithms.
- digital-to-analog converter (DAC): to convert the processed digital signal into electrical voltages, which can be converted to sound waves.

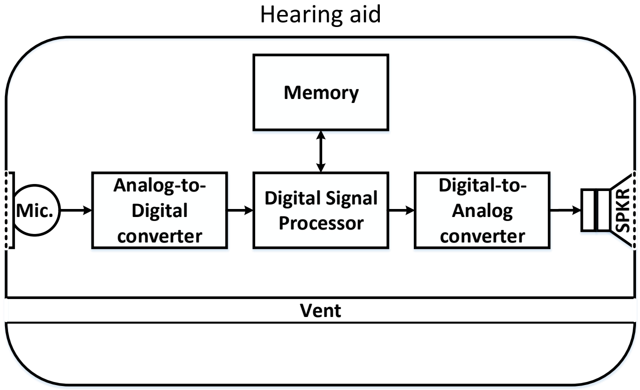


Fig. 8: Basic hardware architecture of a digital hearing aid [37].

- receiver (or speaker): to generate sound waves presented to the user based on the obtained electrical voltages.
- vent[7]: to decrease the *occlusion effect*[8].

HASs have limited processing power and memory capacity, far less than a smart phone. Further, they must operate in real time, i.e. the processing delay must be low [37, 75]. As a rule of thumb, the processing delay of a HAS generally should be less than 10 milliseconds [95]. This is because sounds are reaching the eardrum of a hearing aid user via two different paths: 1) an acoustic path: sounds reach the eardrum by passing around the hearing aid and through the vent, 2) a hearing aid path: sounds are picked up by the HAS microphones, are processed, and are then presented to the eardrum via the HAS receiver. In comparison with the time delay of the hearing aid path, the time delay of the acoustic pass is negligible. Therefore, the processing delay of a HAS can cause spectral ripples (referred to as *comb filtering effect* [37]), which can be audible and disturbing to the user, especially when the processing delay is longer than roughly 10 milliseconds. Hence, signal processing algorithms developed for HASs have to consider these limitations. Moreover, because a hearing loss generally affects different frequencies of a sound in a different manner, processing algorithms of HASs most often operate in the frequency domain [37, 75, 121].

[7] "any opening between the inner part of the ear canal and the free air outside the ear will be called a vent" [37].

[8] "the occlusion effect refers to the increase in low-frequency sound pressure within the blocked ear canal when the hearing-aid user is talking" [75]. In other words, when an object blocks the ear canal of a person, he/she may perceive his/her voice as "hollow" or "booming" [103]

3.3 Signal processing algorithms

Digital HASs are equipped with various signal processing algorithms, e.g. for amplifying or filtering the incoming sound, so that it is audible and intelligible to the hearing aid user. The three main processing algorithms, which are usually available in most digital HASs, are *wide dynamic range compression* (WDRC), *feedback cancellation* and *noise reduction*. Here, we very briefly explain the roles and functionalities of these algorithms.

Wide dynamic range compression (WDRC)

As explained in Sec. 2.5, the DRH of a hearing-impaired listener is smaller than that of a normal-hearing listener, because hearing loss elevates the hearing threshold, while generally does not change the pain threshold (the upper limit of the DRH). Due to the smaller DRH, it is not feasible to amplify all sounds by a fixed gain, in order to make soft sounds audible for a hearing-impaired listener. This is because the fixed gain would amplify intense sounds to levels louder than the pain threshold. Apart from instant discomfort, this could significantly harm the auditory system of the user [37, 121]. Instead, most existing HASs use a WDRC algorithm to reduce the range of sound levels in the environment to fit into the DRH of a hearing-impaired listener. WDRC algorithms usually vary the gain needed to amplify sounds with respect to the input sound level. Two important aspects in designing a WDRC algorithm are [121]: 1) Static aspects: defining the amount of gain as a function of input sound level. 2) Dynamic aspects: defining how fast the algorithm should change the gain with respect to the changes in the input sound level.

Feedback cancellation

Because the microphone and the receiver of a hearing aid are placed close to each other, it is likely that the sound generated by the receiver leaks back to the microphone [59, 75]. This problem is known as "acoustic feedback problem", which can degrade the sound quality of HASs and may lead to an unstable system, which, in turn, may cause the hearing aid to howl [59, 75]. To solve this issue, HASs usually use a *feedback cancellation system* to detect the feedback problem when it occurs, and to cancel its effects [59, 75].

Noise reduction

As mentioned, in noisy situations, hearing-impaired listeners face more difficulties in detecting a target sound and understanding speech compared with normal-hearing listeners. Therefore, HASs usually employ a *noise reduction system* to analyze the input sounds, and to estimate and to reduce the amount of noise in the received sounds, in a way that the target sound is more intelligible or has higher sound quality [66, 121]. Moreover, instead of using one microphone, most existing HASs incorporate an array of microphones—typically two—which allows for spatial filtering, i.e. to amplify the sounds originating from a desired direction while attenuating the interferer sounds coming from other directions. We refer to this feature as *(adaptive) directional microphones* or *beamforming* [37, 75, 121].

Generally, performance of different processing algorithms in a HAS depends on each other. Therefore, another challenge in designing a HAS is to tune the parameters of algorithms, so that the overall performance of a HAS is improved [37].

3.4 Effects of hearing aids on sound source localization

As seen in Sec. 2.5, hearing loss degrades the SSL performance of humans. In this section, we explain the effects of HASs and hearing aid signal processing algorithms on SSL performance of hearing-impaired listeners. In general, several studies, e.g. [104, 133, 136, 140, 141], have shown that most existing HASs not only do not improve the SSL performance of hearing-impaired listeners, but sometimes even deteriorate the SSL performance.

The effects of WDRC systems on SSL performance have been investigated in [79, 124, 133, 151]. These studies showed that WDRC systems generally distort ILDs, and thereby can affect the localization performance, especially for high frequency sounds. However, if low-frequency binaural cues are intact and available, the negative effects of WDRC on SSL performance are reduced [151].

Generally, noise reduction systems can significantly affect the SSL performance [79, 136, 139, 140]. Particularly, most existing directional noise reduction systems can completely distort binaural cues related to interferers, because essentially, these systems are not designed in a way that preserves the localization cues [41, 84]. Moreover, when bilateral hearing aids operate completely independently of each other, mismatches between their noise reduction systems can markedly distort binaural cues, especially ITDs, thereby degrading the localization performance in the horizontal plane [136, 139, 140]. However, more recent noise reduction systems aim at tackling this issue, e.g. [84, 92].

The hearing aids type or more precisely, the location of hearing aids' microphones can also play a role in the SSL performance of users [9, 133, 138]. CIC, ITC and ITE hearing aids generally preserve the monaural cues better than BTE hearing aids [9, 138], and to some extent, can assist their users in resolving front-back confusions [9]. Moreover, it has been shown that the position of the microphones can also affect binaural cues, especially ILDs [133].

BTE hearing aids can distort ILDs up to 30 dB in the frequency range of 6 to 8 kHz [133].

4 Sound Source Localization Algorithms

Taking the spatial information of sound sources into account potentially allows HASs to improve spatial hearing of hearing aid users [84, 130]. To do so, HASs generally need to "know" the location of the target sound source. Most existing HASs assume the target sound source is always in the front of the user, because the user usually looks towards the target to allow the use of visual cues, such as lipreading [127]. However, in practice, the target talker might not be in the front, either because it can be socially awkward to keep the target talker always in the front, or because the physical situation, e.g. a car cabin, does not allow it. In this case, HASs would benefit from being able to localize the sound source. In this section, we review the main existing SSL algorithms proposed in different applications.

As mentioned earlier, most HASs include a microphone array. SSL using a microphone array has been investigated widely over decades, e.g. [7, 13, 32, 36, 69, 70, 91, 94, 105, 113, 114, 122, 135, 146, 147, 160, 162]. In practice, performance of SSL algorithms is limited [13, 69, 128], because :

- Sound signals, particularly speech, are generally wideband, and their spectral contents are changing across time [69].
- Ambient noise, including other interfering sound sources, are typically present, especially in hearing aid applications. Further, typical noise sources are often time varying (nonstationary) [13, 69, 128].
- Reverberation and reflections degenerate the target sound, and make the SSL problem more challenging [13, 69, 128].

In our review of existing SSL algorithms, we categorize them into four type of approaches [50]: 1) Time-difference-of-arrival (TDoA)-based methods, 2) Steered-response-power (SRP)-based methods, 3) High-resolution-spectral-estimation (HRSE)-based methods, and 4) HRTF-based methods. In the following, we explain each of these approaches.

4.1 Time-difference-of-arrival (TDoA)-based methods

TDoA-based algorithms generally consist of two steps [69]. In the first step, a set of TDoAs of the target signal arriving at each pair of microphones in the array are estimated. In the next step, the location of the target source is determined based on the estimated TDoAs and the known geometry of the array [13, 69].

The accuracy of the estimated TDoAs plays a crucial role in the SSL performance of these algorithms [13]. To estimate the TDoAs, several different approaches, which are summarized in [13, 69], have been proposed. The most well-known and computationally efficient approach is based on the *generalized-cross-correlation* (GCC) function [82]. In this approach, the TDoA estimate of signals received by any two microphones is the *time index* or the *time lag* which maximizes their related GCC function [82]. To cope with the effects of non-stationarity, ambient noise and reverberation, variants of the GCC function have been proposed, e.g. [15, 25, 26, 82, 119, 147]. Among the variants of the GCC function, the *GCC-phase-transform* (GCC-PHAT) method is widely used, as it is computationally efficient and operates optimally, in a maximum likelihood sense, in low-noise, highly-reverberant situations [161].

Another main approach to estimate TDoAs is the eigenvalue-decomposition-based method proposed in [7], which performs well for speech signals in reverberant situations. This method uses an iterative solution and has a higher computational complexity than GCC methods [13]. This implies that the eigenvalue-decomposition-based approach is less suitable for low-complexity applications, like hearing aids.

Given the estimated TDoAs, the next step is to determine the location of the sound source. To do so, one often needs to solve a set of nonlinear equations [13]. To solve the equations efficiently in a closed-form manner, several different approaches have been proposed, e.g. [14, 125]. These closed-form solutions are often suboptimal, but their detriment in performance is small, and their computational requirements are relatively low [13].

For hearing aid applications, TDoA-based methods are generally computationally desirable; however, they have two fundamental drawbacks:

1. In the discrete time domain, true TDoAs must be expressed as fractional sample shifts, which are most often ignored in the estimated TDoAs [93].

2. These methods only take the delay of the signals into account for SSL. In other words, head shadowing effects are not considered.

Ignoring fractional delays degrades the SSL performance, especially when the microphones in the array are close to each other, or when the sampling rate is low [19]. To address the fractional delay problem, interpolation methods, e.g. [143], can be used that increase the computational complexity of TDoA-based SSL. To take head shadowing effects into account, HRTF-based methods, which will be explained later in this section, can be used.

4.2 Steered-response-power (SRP)-based methods

The basic idea of SRP-based methods is to steer an adaptive directional microphones, also called a beamformer, towards several candidate locations and look for the candidate location that maximizes the output power [13, 93].

In principle, any type of beamformers can be used for SRP-based SSL. The simplest type of beamformers is a delay-and-sum (DS) beamformer. The output of a DS beamformer is the sum of all the microphone signals, with their ToAs aligned according to the candidate location and the geometry of the array [13, 93, 128]. One of the main disadvantages of a simple DS beamformer for SSL is that its performance significantly degrades, when ambient noise or reverberation is present [13]. To reduce the effects of noise and reverberation, more sophisticated beamformers, which filter the microphone signals as well as align their ToAs, can be used [13, 87]. In practice, the well-known SRP-PHAT method [13], which weight the frequency components of the received signals according to the PHAT weighting, is often used.

At the cost of higher computational complexity, SRP-based methods generally perform better than TDoA-based methods. To decrease the computational load of SRP-based methods, several studies, e.g. [32, 38, 39], propose to replace the exhaustive search among the candidate locations with a more intelligent search strategy for finding the best candidate location.

4.3 High-resolution-spectral-estimation-based methods

High-resolution-spectral-estimation (HRSE)-based methods (also called subspace methods [93]) are based on the spatiospectral correlation matrix derived from the microphones signals [13]. One of the most well-known HRSE-based methods is the multiple signal classification algorithm (known as the MUSIC algorithm), which was originally proposed to localize multiple narrowband uncorrelated sources [93, 122]. Assuming a far-field situation, the MUSIC algorithm exploits an eigenvalue decomposition technique to estimate the source locations from a lower-dimensional vector subspace embedded within the signal space spanned by the columns of the correlation matrix of the noisy microphone signals.

Even-though the MUSIC algorithm has been proposed for narrowband uncorrelated sources, it can be extended to wideband coherent signals, at the cost of higher computational complexity [148, 155]. In practice, the spatiospectral correlation matrix of the noisy microphone signals is unknown, and it is estimated via averaging over a time interval, in which the target signals and the noise are assumed to be statistically stationary and their locations are assumed to be fixed [13]. For speech sound sources, satisfying these conditions over sufficiently long time intervals can be challenging [13].

One of the main drawbacks of HRSE-based methods is that deviations from signal modeling assumptions generally degrade their performance more than the performance of SRP-based methods [13, 144].

4.4 HRTF-based methods

In applications like hearing aids, where the microphone array is mounted close to the head and torso of a user, the shadowing effect and the spectral changes imposed by the head and torso on the received signals can be used to localize the sound source [17, 70, 80, 89, 91, 114, 145, 163]. HRTF-based methods generally can be categorized into two groups:

1. *Model-based*: These methods resort to mathematical models of the head and torso effects on the received signals as a function of the sound source locations [89, 114, 163]. To localize a sound source, these methods estimate the head and torso effects on the current received signals, and use the mathematical models to associate the estimated effects with a location.

2. *Dictionary-based:* In these methods, the spectral changes imposed by the head and torso on the received signal are measured from different locations in advance, and are stored in a database or *dictionary* [17, 80, 91]. Each entry in the dictionary has been labeled by its corresponding location. To localize a sound source, all the entries in the dictionary are evaluated based on the received signals and a *utility (cost) function*. The label of the entry, which maximizes (minimizes) the utility (cost) function, is the estimate of the sound source location.

At the cost of higher computational and storage complexity, dictionary-based methods can potentially perform better than model-based methods.

5 Informed Sound Source Localization

Most existing SSL algorithms have been proposed for applications, where the noise-free target sound is not available, i.e. they are "uninformed" about the noise-free content of the target signal.

In this study, on the other hand, we consider an "informed" SSL problem, in which an essentially noise-free target signal is available to the SSL algorithm. To be more precise, we consider a situation (e.g. a classroom situation), where the target talker (e.g. a teacher) is wearing a wireless microphone, and hearing-impaired listeners are wearing HASs that can connect to the wireless microphone. In such a situation, the wireless microphone transmits the essentially noise-free version of the target signal to the HAS, whose goal is to find the relative location of the target talker (the DoA of the target sound) with respect to the head of the HAS user.

Fig. 9 depicts an exemplar scenario of the "informed" SSL problem considered in this study. In this scenario, the HAS consists of two hearing aids, which are mounted behind each ear of the user, and which are connected to

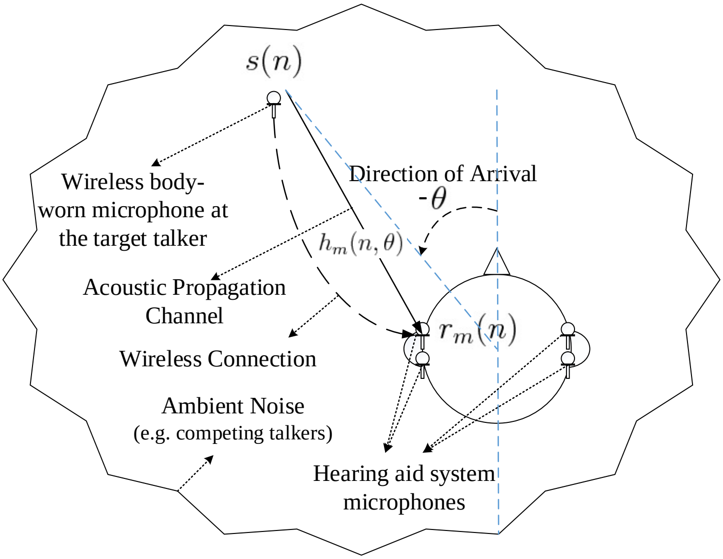


Fig. 9: An informed SSL scenario for hearing aid applications [49].

each other wirelessly. The wireless link between the hearing aids allows them to exchange the signals received by their microphones. Target signal $s(n)$ produced by the target talker, propagates through the acoustic channel $h_m(n,\theta)$, and arrives at microphone m of the HAS from angle θ; n is the discrete-time index. Signal $r_m(n)$ received by microphone m of the HAS is a noisy-version of the target signal, because it has been contaminated by the ambient noise, (e.g. in the classroom situation mentioned above, ambient noise may include irrelevant conversation of students in the class, fan noise from a ventilation system, microphone self-noise, and etc.). Moreover, in this "informed" scenario, the noise-free target signal, i.e. $s(n)$, is also transmitted to the HAS via a wireless connection between the wireless microphone and the HAS. The goal is to estimate the DoA of the target sound, i.e. θ.

To motivate our interest in estimating the DoA in an informed situation, let us first review the existing wireless microphone systems and the advantages, which they provide for HAS users.

5.1 Wireless microphone systems

As mentioned in Sec. 2, hearing-impaired listeners have difficulties in understanding speech in noisy and reverberant situations. In Sec. 3, we mentioned that HASs employ noise reduction systems to decrease the effects of noise and reverberation on the target speech, and to increase intelligibility of the target speech. Even though noise reduction systems can be effective in certain situations (particularly for non-(or slowly-)time-varying noise fields, or when the target position is frontal with respect to the HAS user), there exist many everyday situations, where their performance is limited [40]. Further, the performance of "traditional" noise reduction systems is limited, in principle, due to physical constraints. For example, HAS microphones are mounted close to each other and behind (or in) the ears of a HAS user; therefore, increasing the distance of the target source will generally decrease the performance of noise reduction systems in reverberant, noisy situations [88].

A highly effective solution proposed to avoid negative effects of noise and reverberation on target speech is to capture the speech where it is most powerful with respect to the background noise, e.g. next to the talker's mouth, and transmit this clean speech to a HAS via an electromagnetic signal or a magnetic field, rather than an acoustic signal [37, 88]. Given that HASs have the required receiver to convert the received electromagnetic signal or magnetic field into an acoustic signal, with this strategy, the clean speech is available at the HAS and can be delivered to its user [37].

Existing technologies

Wireless transmission technologies used with HASs can be categorized into [35, 37, 96]: 1) *infrared* systems, 2) *induction loops,* 3) *near-field magnetic induction* (NFMI) systems, and 4) *radio-frequency* (RF) systems.

In infrared systems, the target signal is transmitted electromagnetically (at frequencies of about 10^{14} Hz [37]). Infrared signals are fragile in the sense that they can be reflected easily by flat light-colored surfaces, can be blocked by opaque obstacles, and can be disturbed by direct sunlight [37]. Hence, the use of infrared systems are generally marginal for hearing aid applications [35, 37, 77].

Induction loop systems transmit audio signals by converting them into magnetic fields [37]. If a HAS is equipped with a *telecoil* ("a small coil of wire that produces a voltage when an alternating magnetic field flows through it" [37]), the magnetic field *induces* an electrical voltage in the telecoil [37], and the induced voltage can then be converted back into a sound wave via the HAS receiver [37]. The cost of installing an induction loop system is relatively high [77]. Therefore, induction loop systems are suitable for applications, like auditoriums or theaters, where the number of users is relatively high [77].

An NFMI system can be considered as a "personal induction loop system" that allows for low-power, short-range, wireless communication [96]. NFMI systems are reasonably resistant to interference [96]; however, their practical transmission range is limited (in the range of $1 - 1.5\,\mathrm{m}$) [96]. Therefore, these systems are generally better suited for wireless transmission between hearing aids of a binaural HAS than wireless transmission between a wireless microphone and a HAS.

RF systems offer a portable way to transmit a clean signal from a target talker to a listener [37]. An RF system consists of a *transmitter* along with a microphone worn by the target talker, and a receiver placed at the HAS worn by the user. To transmit the audio signals, RF systems *modulate* an electromagnetic *carrier* signal, which is *demodulated* by the receiver to extract the audio signals [37]. In principle, any modulation technique can be used for transmission, but the two most commonly used for short-range transmissions are *frequency modulation* (FM) and *frequency-hopping spread-spectrum modulation* (FHSS) [37, 96].

FM systems used to be the most common RF system for hearing aid applications [35, 96]. However, new HASs often exploit variants of FHSS technology for RF transmission [37, 96]. FM systems generally suffer from interference when two (or more) FM transmitters are transmitting at the same carrier frequency [37]. In contrast, FHSS systems are less prone to interference and are more suitable for transmission of digital data [37].

The most well-known FHSS system is the Bluetooth protocol, which allows multiple transmitters to work together without interfering with each other [37, 96]. The primary Bluetooth standard was not appropriate for hearing aid applications due to its long delay and high power consumption [37, 96]. However, a new version of the Bluetooth protocol, called "Bluetooth Smart" or "Bluetooth low energy", has been published recently that allows low-power and low-latency wireless transmissions within a range of up to 50 m [96].

Advantages of wireless microphone systems

The benefits of the availability of an almost clean target signal by existing wireless microphone systems have been investigated by several studies, e.g. [12, 29, 63, 77, 129]. Advantages of induction loop systems have been discussed in [77]. Induction loop systems generally seem to improve speech intelligibility, reduce listening effort, and enhance sound quality in a way that even normal-hearing listeners would benefit from them in many situations [35, 77]. In noisy situations, it has been shown that FM systems generally improve speech recognition significantly for hearing-impaired listeners as well as for normal-hearing listeners [12, 63, 88].

Combination of wireless microphone signals and local microphone signals

In practice, hearing-impaired listeners sometimes need to hear more than one talker, e.g. when they are working in a small group [37]. Hence, it is not sufficient for the user to hear only signals of the target talker, who is wearing a wireless microphone, especially when this target talker is far away from the other talkers. One way to solve this issue is to combine the signal received from the FM transmitter and the signal picked up by the HAS microphone (referred to as the *local microphone*). In this way, the HAS user can hear both the target talker and the nearby talkers; however, ambient noise and reverberation picked up by the local microphone can potentially degrade most of the benefits provided by the FM system [37, 63]. To mitigate the effects of noise and reverberation, HASs usually amplify the output of the FM system before combining it with the local microphone output [37]. The difference between the level of the local microphone output and the level of the FM system plus local microphone output is referred to as the *FM advantage* [129]. The American speech-language-hearing association (ASHA) suggests a 10 dB FM advantage for optimum speech recognition in noise [5, 129].

Even though the output of the FM system alone provides a higher speech intelligibility than the output of the FM system plus local microphone [33, 37, 63], it has been shown that hearing-impaired people, specifically children, generally prefer the FM system plus local microphone output [33, 35, 37]. This is most likely because listening only to the output of the FM system causes the users to feel detached from the environment around them [37].

Another commonly-used solution to combine FM systems output and local microphones output is a *dynamic FM* or *adaptive FM* system [37, 129]. In dynamic FM systems, the FM advantage is automatically adapted based on the background noise level [37, 129]. To be more precise, to increase the intelligibility of the target sound when the background noise level is high, the FM advantage is increased. On the other hand, when the background noise level is low, the FM advantage is decreased to develop a feeling of connectedness to the environment for the user [37].

Overall, although RF systems, particularly FM systems, generally provide a great help in noisy situations, it should be noted that the successful use of RF systems in daily life by HAS users requires several counseling, instruction and coaching sessions [12, 29].

5.2 Motivation of research

Most existing HASs, which are using a wireless microphone system, render the wirelessly received signal of a target talker in a monaural or diotic way (the same signal is presented at both ears [100]). These ways of rendering of the target signal obviously remove all spatial cues about the target talker location, which can degrade the sense of immersion and causes the user to feel detached from the environment [35]. Moreover, as mentioned in Sec. 1, lack of spatial cues can degrade the intelligibility of the target speech, in situations where several simultaneous talkers are present, especially, when each of the talkers are wearing a wireless microphone, e.g. in a conference [35]. To be more precise, in these multi-talker multi-microphone situations, if two of the talkers talk simultaneously, rendering of both target signals in a diotic way would deteriorate the intelligibility of both signals (as mentioned in Sec. 1, spatial cues assist humans to "decompose" the mixture of signals in way that they are more intelligible).

One solution to overcome this problem and to improve the sense of immersion is to impose the corresponding spatial cues on the wirelessly received signals, before presenting them to the HAS user [35]. To do so, HASs must know the location of the talker. This leads to the informed SSL problem considered in this study.

5.3 Topics of the thesis

This thesis—apart from this Introduction—consists of a collection of papers, contributing to solving the informed SSL problem introduced earlier in this section. Here, we aim to summarize the scientific contributions of each paper and discuss the relations between the papers.

The informed SSL problem for binaural HASs is an essentially unexplored problem. It was first introduced and addressed in [34]. The method proposed in [34] is a TDoA-based approach, which uses a cross-correlation technique and the wirelessly received clean target to estimate the ToAs of the acoustic target signal received at the left and the right hearing aids, thereby estimating the TDoA. Afterwards, the method resorts to a sine law to map the estimated TDoA to a DoA estimate [34]. The computational complexity of the method proposed in [34] is low enough to nominate this method as a candidate for practical new-future implementation in HAS. However, it does not take the head shadowing effect and ambient noise characteristics into account. This negligence can degrade the estimation performance markedly [46–48, 50].

The main contribution of this thesis, in a few words, is to solve the informed SSL problem using a maximum likelihood (ML) framework, which *does* allow to take ambient noise characteristics and head shadowing effects into account. The general strategy used by the ML framework to localize the target sound source bear similarities to the SRP-based methods discussed in Sec. 4. To be more precise, to localize the target sound source, a proposed likelihood function will be evaluated for a discrete set of candidate locations. The ML estimate of the target sound source location is the candidate location that maximizes the likelihood function. In the following, we briefly explain the scientific contribution of each paper.

Paper A – Informed TDoA-based direction of arrival estimation for hearing aid applications

Situations, where the background noise levels are high, are challenging for informed SSL algorithms. This paper presents a methodology, where easy access to noise statistics may be used to improve informed SSL performance over algorithms that do not take noise characteristics into account, e.g. [34]. To do so, a likelihood function has been proposed, which assumes that the noise signals received at the HAS microphones follows a zero-mean circularly-symmetric complex Gaussian distribution. The proposed likelihood function is based on the noise cross power spectral density (CPSD) matrix, which can be relatively easily estimated in an informed SSL scenario. Moreover, the likelihood function is formulated in a way that can be evaluated efficiently using an inverse discrete Fourier transform (IDFT). Simulation results show that taking the ambient noise characteristics into account can significantly improve the estimation performance, in comparison with the method proposed in [34]. The method proposed in Paper A assumes a free-field and far-field situation, i.e. does not take the presence of the head into account. Therefore, we refer to this method as the *free-field-far-field-model-based* method.

Paper B – Informed direction of arrival estimation using a spherical-head model for hearing aid applications

Microphones of existing HASs are most often placed at/in the ears of the HAS user. This implies that any sound signal picked up by the HAS microphones is affected by the head presence. Hence, Paper B modifies the Paper A's solution to take head shadowing effects, in addition to the noise characteristics, into account. To do so, the method proposed in Paper B resorts to a spherical head model to amend the free-field-far-field-model-based method proposed in Paper A. The proposed spherical head model is generic and does not depend on physical features of any specific user. Similar to the free-field-far-field-model-based method, the likelihood function of the proposed method, which we refer to as the *spherical-head-model-based* method, can be evaluated computationally efficiently using an IDFT. Simulation results show that the spherical-head-model-based method can improve the estimation performance, specifically when the target sound source is at the sides of the user, i.e. where the head has the strongest shadowing effect on the received signals.

Paper C – Maximum likelihood approach to "informed" sound source localization for hearing aid applications

Paper C explores to which extent informed SSL performance can be improved in situations where very detailed and accurate person-specific head-and-torso information is available to the SSL algorithm. The proposed method, called *MLSSL* (maximum likelihood sound source localization), is a dictionary-based method, which exploits the ML framework together with a database (dictionary) of HRTFs measured for a specific user. The HRTF database allows to individually model the presence of the head. MLSSL is highly flexible in the sense that it does not depend on any particular microphone array configuration, and it can work even with one single microphone (the free-field-far-field-model-based method and the spherical-head-model-based method consider a binaural configuration using two microphones—one microphone in each hearing aid). Simulation results show that MLSSL is highly effective under severely noisy conditions, as long as the user-specific HRTF database accurately reflects the real world.

Paper D – On the influence of microphone array geometry on HRTF-based sound source localization

Since MLSSL does not depend on any specific microphone array configuration, Paper D studies how the microphone array geometry can affect the performance of MLSSL. This research question is important because some microphone configurations (e.g. binaural) may require higher implementation costs than others (e.g. monaural). This paper shows that MLSSL performance depends on the location of the target talker and the configuration of the microphone array. It shows that binaural configurations (one microphone in each hearing aid) provide better performance for situations, where the talker is in the front, while monaural configurations provide better performance for situations, where the talker is at the sides.

Paper E – Informed sound source localization using relative transfer functions for hearing aid applications

Even-though MLSSL is highly flexible and effective, its computational load is relatively high. This is because HRTFs generally depend both on the distance and the direction of the target sound source. Therefore, the HRTF database searched by MLSSL should ideally contain a large number of HRTF entries to cover all possible directions and distances; otherwise, MLSSL performance is degraded significantly.

To decrease the MLSSL computational load, the method proposed in Paper E uses a database of *relative transfer functions* (RTFs) rather than the HRTF used in MLSSL. RTFs, in contrast to HRTFs, are relatively distance-independent, especially in far-field situations. Hence, the proposed method, which we refer to as the *measured-RTF-based* method, uses in an RTF database, which has substantially fewer entries than the HRTF database. Similar to the free-field-far-field-model-based method and the spherical-head-model-based method, the likelihood function of the measured-RTF-based method can be evaluated computationally efficiently using an IDFT.

Paper E shows that the measured-RTF-based method is more robust to mismatches between the dictionary elements and characteristics of a specific user than MLSSL. This behavior of the measured-RTF-based method is particularly important in practical situations, where a user-specific database is not available, but a more generic database (e.g. an HRTF database measured for a head-and-torso-simulator (HATS)) is available.

Finally, Paper E provides a unified presentation of the free-field-far-field-model-based, the spherical-head-model-based, and the measured-RTF-based methods, and assesses their performance more extensively.

Paper F – Bias-compensated informed sound source localization using relative transfer functions

The measured-RTF-based method is not as flexible as MLSSL in the sense that it considers only a binaural configuration using two microphones—one microphone in each hearing aid—similarly to the methods proposed in Paper A and Paper B.

In Paper F, the measured-RTF-based method is extended to work in both monaural and binaural configurations. Moreover, the likelihood function of the proposed method is modified, so that it can be evaluated using a sum over frequency components (with a computational complexity of $O(N)$[9,10]) instead of computing an IDFT (with a computational complexity of $O(N \log N)$). Further, a closed-form expression is derived for the bias in the proposed likelihood function, and Paper F proposes a method to analytically compensate for the bias. Finally, to reduce the number of parameters required to be wirelessly exchanged between the hearing aids in binaural configurations, an information fusion strategy is proposed that avoids transmitting microphone signals between the hearing aids. We refer to the method proposed in this paper as the *bias-compensated-measured-RTF-based* method.

Paper G – TDOA-based self-calibration of dual-microphone arrays

This paper does not directly solve the informed SSL problem, but propose an algorithm which supports the free-field-far-field-model-based and spherical-head-model-based methods. In these methods, the relative locations of the microphones must be known. In binaural configurations, the exact relative locations of the microphones is variable in practice, because of different heads radii and varying shapes of pinnae of users. Instead of measuring the relative locations of the microphones manually, Paper G offers an automatic solution, which has lower computational complexity than other existing methods, while showing comparable performance. The proposed method is based on TDoAs of signals received from different sound sources.

[9] Big O notation is used to describe the time, memory or computational complexity of an algorithm [83]. A computational complexity of $O(f(q))$ means the number of operations is lower than $C * f(q)$ for all $q > q_0$, where $f(.)$ is a function defined on some subset of the real numbers, C is a constant, and q_0 is a positive real number.

[10] N is the number of frequency components.

Paper H – Concurrent localization of sound sources and dual-microphone sub-arrays using TOFs

The method proposed in Paper G for estimating the relative locations of the microphones is "uninformed", i.e. it does not exploit the fact that a clean target signal is available in informed scenarios. Hence, Paper H offers an alternative "informed" solution, which uses this extra information. The proposed method is based on the target signals time of flights (TOFs), which can be relatively easily estimated in informed scenarios.

Comparison of the proposed informed SSL methods

To summarize this section, we compare key aspects of the proposed informed SSL methods in Table 2.

Finally, it should be noted that situations, where several talkers are simultaneously present and each of them are wearing a wireless microphone, have not been addressed directly in this thesis. However, all the proposed methods can be extended easily to handle these multi-talker multi-microphone situations by simply executing an instance of the algorithms for each wireless microphone. To be more precise, to estimate the location of a particular talker in these situations, the signals received from other talkers are considered as noise. Therefore, we must separately estimate the noise CPSD matrix for each talker, who is wearing a wireless microphone. Knowing the noise CPSD matrices associated with a particular talker allows us to estimate her/his position using the methods proposed in this thesis.

6 Conclusion

This thesis addresses the problem of localizing a target talker for hearing aid applications. We consider situations where the noise-free content of the target speech emitted at the target talker location is available at the hearing aid system (HAS) via a wireless microphone worn by the target talker.

To solve the problem, a maximum likelihood (ML) framework has been developed. This framework is based on the access to the ambient noise characteristics and the noise-free target signal, and allows to take the shadowing effect of the user's head into account. To localize the target talker, the proposed ML framework evaluates a likelihood function for various candidate locations. The candidate location which has the highest likelihood is the ML estimate of the target talker's location.

To consider the effects of the user's head and torso on the signals received by the HAS microphones, four different models have been employed:

1. *Free-field-far-field model:* This model assumes a free-field situations, i.e. it simply ignores the presence of the head.

Table 2: Comparison of the proposed informed DoA estimators

Method	Required number of microphones and configuration	Computational complexity	Individualizable	Sensitivity to non-individualized databases[*]	Does estimate distance?[†]
Free-field-far-field-model-based (Paper A)	2-mic., binaural	$O(I \times N \times \log N)$[‡]	No	N/A	Yes
Spherical-head-model-based (Paper B)	2-mic., binaural	$O(I \times N \times \log N)$	No	N/A	Yes
MLSSL (Paper C)	M-mic. ($M \geq 1$), monaural or binaural	$O(I \times Q$[††]$\times N)$	Yes	High	Yes
Measured-RTF-based (Paper E)	2-mic., binaural	$O(I \times N \times \log N)$	Yes	Low	Yes
Bias-compensated-measured-RTF-based (Paper F)	M-mic. ($M \geq 2$), monaural or binaural	$O(I \times N)$	Yes	Low	No

[*] In practice, a user-specific database might not always be available. However, a more generic database (e.g. an HRTF database measured for a head-and-torso-simulator (HATS)) can be measured in advance and be available. This column shows how sensitive the performance of proposed informed SSL methods are to the mismatches between a generic database and characteristics of a specific user.

[†] Some of the proposed DoA estimators provide, as a by-product, an ML estimate of the target signal propagation time between the target talker and the user. This propagation time can be straightforwardly converted to a distance estimate.

[‡] I represents the number of candidate DoAs, and N represents the order of the employed discrete Fourier transform.

[††] Q represents the number of candidate distances covered in the HRTF database.

2. *Spherical-head model:* This model considers the head as a rigid sphere.
3. *Head-related-transfer-function (HRTF)-database model:* This model uses a database of HRTFs measured from various candidate locations.
4. *Relative-transfer-function (RTF)-database model:* This model uses a database of RTFs measured from various candidate locations.

These models allow different degrees of individualization. The free-field-far-field model and the spherical-head-model are generic models, i.e. do not depend on any specific user. On the other hand, the HRTF-database and the RTF-database models allow person-specific details to be taken into account.

This thesis shows that

i) Incorporating ambient noise characteristics and head shadowing effects can markedly improve the localization performance.

ii) Individualized head models generally lead to better localization performance than generic head models.

iii) The proposed informed SSL method that relies on person-specific HRTF database tend to be sensitive to model mismatches.

iv) The proposed informed SSL methods based on the RTF-database are more robust to model mismatches, and have a lower computational complexity than the method based on the HRTF-database model.

6.1 Directions of future research

The methods proposed in this study rely on spatiospectral characteristics of signals, such as noise CPSD matrices, and assume these characteristics to be invariant across a short time duration (in the range of milliseconds). To integrate information across longer time durations, as a topic of future research, one can extend the localization methods to take temporal characteristics of the acoustic scene into account. This might be done by noting that physical objects, such as target talker and the user's head (and hence, the HAS microphones), can only move with finite velocity with respect to each other. Therefore, modeling and tracking of the target source and user's head movements could improve SSL performance further [117, 134].

In practice, informed SSL scenarios often occurs in reverberant environments. However, the signal model used in this study does not directly consider and model reverberation. Another topic for future research is to explicitly take the reverberation into account, e.g. by modeling the reverberation as a highly time varying isotropic noise field, e.g. [16, 87].

Extra sensors in hearing aids, such as accelerometers [28, 163], or technologies, such as *electroencephalogram* (EEG) [30, 45], eye tracking and head movement detection [4], generally provide additional non-acoustic information about how an acoustic scene is changing over time. It is a topic for future research to investigate how to incorporate this additional information to solve the informed SSL problem.

Finally, practical implementation of the informed SSL algorithms in HASs and investigation of how to render the acoustic scene to the HAS user based on the outputs of the proposed algorithms is another direction of future research [35, 76].

References

[1] "Deafness and hearing loss," World Health Organization, http://www.who.int/mediacentre/factsheets/fs300/en/, accessed: 2017-02-09.

[2] L. A. Abbagnaro, B. B. Bauer, and E. L. Torick, "Measurements of diffraction and interaural delay of a progressive sound wave caused by the human head. II," *The Journal of the Acoustical Society of America*, vol. 58, no. 3, pp. 693–700, 1975.

[3] M. A. Akeroyd and F. H. Guy, "The effect of hearing impairment on localization dominance for single-word stimuli," *The Journal of the Acoustical Society of America*, vol. 130, no. 1, pp. 312–323, 2011.

[4] A. Al-Rahayfeh and M. Faezipour, "Eye tracking and head movement detection: A state-of-art survey," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 1, no. 2100212, 2013.

[5] "Guidelines for fitting and monitoring FM systems," American Speech-Language-Hearing Association, 2002.

[6] D. H. Ashmead, D. L. Davis, and A. Northington, "Contribution of listeners' approaching motion to auditory distance perception." *Journal of experimental psychology: Human perception and performance*, vol. 21, no. 2, pp. 239–256, 1995.

[7] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[8] V. Best, S. Carlile, N. Kopco, and A. van Schaik, "Localization in speech mixtures by listeners with hearing loss," *The Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. EL210–EL215, 2011.

[9] V. Best, S. Kalluri, S. McLachlan, S. Valentine, B. Edwards, and S. Carlile, "A comparison of CIC and BTE hearing aids for three-dimensional localization of speech," *International Journal of Audiology*, vol. 49, no. 10, pp. 723–732, 2010.

[10] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, 1997.

[11] P. Bonding, "Frequency selectivity and speech discrimination in sensorineural hearing loss," *Scandinavian Audiology*, vol. 8, no. 4, pp. 205–215, 1979.

[12] A. Boothroyd, "Hearing aid accessories for adults: The remote FM microphone," *Ear and Hearing*, vol. 25, no. 1, pp. 22–33, 2004.

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

[14] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 1, pp. 45–50, Jan 1997.

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

[16] S. Braun and E. A. 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. 34, pp. 1–14, 2015.

[17] S. Braun, W. Zhou, and E. A. P. Habets, “Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2015, pp. 1–5.

[18] A. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT press, 1994.

[19] B. V. D. Broeck, A. Bertrand, P. Karsmakers, B. Vanrumste, H. V. hamme, and M. Moonen, “Time-domain generalized cross correlation phase transform sound source localization for small microphone arrays,” in *Proceedings of European DSP Education and Research Conference*, Sept 2012, pp. 76–80.

[20] A. W. Bronkhorst, “The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions,” *Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[21] A. W. Bronkhorst and T. Houtgast, “Auditory distance perception in rooms,” *Nature*, vol. 397, no. 6719, pp. 517–520, 1999.

[22] A. Brughera, L. Dunai, and W. M. Hartmann, “Human interaural time difference thresholds for sine tones: The high-frequency limit,” *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2839–2855, 2013.

[23] D. S. Brungart and W. M. Rabinowitz, “Auditory localization of nearby sources. head-related transfer functions,” *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1465–1479, 1999.

[24] R. A. Butler, R. A. Humanski, and A. D. Musicant, “Binaural and monaural localization of sound in two-dimensional space,” *Perception*, vol. 19, no. 2, pp. 241–256, 1990.

[25] G. C. Carter, A. H. Nuttall, and P. G. Cable, “The smoothed coherence transform,” *Proceedings of the IEEE*, vol. 61, no. 10, pp. 1497–1498, 1973.

[26] J. Chen, J. Benesty, and Y. Huang, “Time delay estimation in room acoustic environments: an overview,” *EURASIP Journal on applied signal processing*, vol. 2006, no. 26503, pp. 1–19, 2006.

[27] C. I. Cheng and G. H. Wakefield, “Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space,” in *Audio Engineering Society Convention 107*, 1999, paper 5026.

[28] Y. Chisaki and S. Tanaka, “Improvement in estimation accuracy of a sound source direction by a frequency domain binaural model with information on listener’s head movement in a conversation,” in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Dec 2014, pp. 1–6.

[29] T. H. Chisolm, C. M. Noe, R. McArdle, and H. Abrams, “Evidence for the use of hearing assistive technology by adults: The role of the FM system,” *Trends in Amplification*, vol. 11, no. 2, pp. 73–89, 2007.

[30] C. J. Chun, S. H. Jeong, J. W. Shin, H. K. Kim, and J. A. Kang, "Feasibility study for objective measurement on sound localization using auditory evoked potential," in *Proceedings of International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Aug 2014, pp. 610–613.

[31] J. G. Clark, "Uses and abuses of hearing loss classification," *Journal of the American Speech-Language-Hearing Association*, vol. 23, no. 7, pp. 493–500, 1981.

[32] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Processing Letters*, vol. 18, no. 1, pp. 71–74, 2011.

[33] S. E. Cotton, "Evaluation of FM fittings," Master's thesis, Macquarie University, Sydney, 1988.

[34] G. Courtois, P. Marmaroli, M. Lindberg, Y. Oesch, and W. Balande, "Implementation of a binaural localization algorithm in hearing aids: specifications and achievable solutions," in *Audio Engineering Society Convention 136*, April 2014, paper 9034.

[35] G. A. Courtois, "Spatial hearing rendering in wireless microphone systems for binaural hearing aids," Ph.D. dissertation, Swiss Federal Institute of Technology in Lausanne (EPFL), 2016.

[36] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, 2000.

[37] H. Dillon, *Hearing Aids*, 2nd ed. Thieme, 2012.

[38] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2510–2526, 2007.

[39] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction(src) on a large-aperture microphone array," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, April 2007, pp. I–121–I–124.

[40] 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, March 2015.

[41] S. Doclo, T. J. Klasen, T. Van den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural cue preservation using multi-channel wiener filtering and interaural transfer functions," in *Proceedings of International Workshop on Acoustic Echo and Noise Control*, 2006.

[42] R. O. Duda, "Modeling head related transfer functions," in *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*. IEEE, 1993, pp. 996–1000.

[43] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *The Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048–3058, 1998.

[44] N. I. Durlach, C. L. Thompson, and H. S. Colburn, "Binaural interaction in impaired listeners: A review of past research," *Audiology*, vol. 20, no. 3, pp. 181–211, 1981.

[45] M. Ebisawa, M. Kogure, S. h. Yano, S. i. Matsuzaki, and Y. Wada, "Estimation of direction of attention using eeg and out-of-head sound localization," in *Proceedings of International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2011, pp. 7417–7420.

[46] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Informed TDoA-based direction of arrival estimation for hearing aid applications," in *Proceedings of IEEE Global Conference on Signal and Information Processing*, 2015, pp. 953–957.

[47] ——, "Maximum likelihood approach to "informed" sound source localization for hearing aid applications," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2015, pp. 16–20.

[48] ——, "Informed direction of arrival estimation using a spherical-head model for hearing aid applications," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2016, pp. 360–364.

[49] ——, "Bias-compensated informed sound source localization using relative transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017, manuscript submitted for publication.

[50] ——, "Informed sound source localization using relative transfer functions for hearing aid applications," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 611–623, March 2017.

[51] W. E. Feddersen, T. T. Sandel, D. C. Teas, and L. A. Jeffress, "Localization of high frequency tones," *The Journal of the Acoustical Society of America*, vol. 29, no. 9, pp. 988–991, 1957.

[52] T. Francart and J. Wouters, "Perception of across-frequency interaural level differences," *The Journal of the Acoustical Society of America*, vol. 122, no. 5, pp. 2826–2831, 2007.

[53] R. L. Freyman, K. S. Helfer, D. D. McCall, and R. K. Clifton, "The role of perceived spatial separation in the unmasking of speech," *The Journal of the Acoustical Society of America*, vol. 106, no. 6, pp. 3578–3588, 1999.

[54] M. B. Gardner, "Distance estimation of 0° or apparent 0°-oriented speech signals in anechoic space," *The Journal of the Acoustical Society of America*, vol. 45, no. 1, pp. 47–53, 1969.

[55] ——, "Some monaural and binaural facets of median plane localization," *The Journal of the Acoustical Society of America*, vol. 54, no. 6, pp. 1489–1495, 1973.

[56] "Acoustic Glossary," Gracey & Associates, http://www.acoustic-glossary.co.uk/sound-fields.htm, [Accessed 27-02-2017].

[57] S. Greenberg and W. Ainsworth, *Listening to Speech: An Auditory Perspective*. Taylor & Francis, 2012.

[58] B. Grothe, M. Pecka, and D. McAlpine, "Mechanisms of sound localization in mammals," *Physiological reviews*, vol. 90, no. 3, pp. 983–1012, 2010.

[59] M. Guo, “Analysis, design, and evaluation of acoustic feedback cancellation systems for hearing aids,” Ph.D. dissertation, Aalborg University, 2013.

[60] K. Hale and K. Stanney, *Handbook of Virtual Environments: Design, Implementation, and Applications, Second Edition*, ser. Human Factors and Ergonomics. Taylor & Francis, 2014.

[61] W. M. Hartmann, “How we localize sound,” *Physics today*, vol. 52, no. 11, pp. 24–29, 1999.

[62] R. Häusler, S. Colburn, and E. Marr, “Sound localization in subjects with impaired hearing,” *Acta Oto-Laryngologica*, vol. 96, no. sup400, pp. 1–62, 1983.

[63] D. B. Hawkins, “Comparisons of speech recognition in noise by mildly-to-moderately hearing-impaired children using hearing aids and FM systems,” *Journal of Speech and Hearing Disorders*, vol. 49, no. 4, pp. 409–418, 1984.

[64] M. L. Hawley, R. Y. Litovsky, and J. F. Culling, “The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer,” *The Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 833–843, 2004.

[65] M. L. Hawley, R. Y. Litovsky, and H. S. Colburn, “Speech intelligibility and localization in a multi-source environment,” *The Journal of the Acoustical Society of America*, vol. 105, no. 6, pp. 3436–3448, 1999.

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

[67] P. M. Hofman, J. G. Van Riswick, and A. J. Van Opstal, “Relearning sound localization with new ears,” *Nature neuroscience*, vol. 1, no. 5, pp. 417–421, 1998.

[68] K. Hopkins and B. C. J. Moore, “The effects of age and cochlear hearing loss on temporal fine structure sensitivity, frequency selectivity, and speech reception in noise,” *The Journal of the Acoustical Society of America*, vol. 130, no. 1, pp. 334–349, 2011.

[69] Y. Huang, J. Benesty, and J. Chen, *Time Delay Estimation and Source Localization*. Springer Berlin Heidelberg, 2008, pp. 1043–1063.

[70] S. Hwang, Y. Park, and Y. Park, “Sound source localization using HRTF database,” in *Proceedings of International Conference on Control, Automation, and Systems*, 2005, pp. 751–755.

[71] *Acoustics – Normal equal-loudness-level contours*, ISO Std. 226:2003, 2003.

[72] Y. Iwaya, Y. Suzuki, and D. Kimura, “Effects of head movement on front-back error in sound localization,” *Acoustical Science and Technology*, vol. 24, no. 5, pp. 322–324, 2003.

[73] L. A. Jeffress, “A place theory of sound localization.” *Journal of comparative and physiological psychology*, vol. 41, no. 1, pp. 35–39, 1948.

[74] D. H. Johnson, “The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones,” *The Journal of the Acoustical Society of America*, vol. 68, no. 4, pp. 1115–1122, 1980.

[75] J. M. Kates, *Digital Hearing Aids*. Plural Publishing, Incorporated, 2008.

[76] J. M. Kates, K. H. Arehart, and R. K. Muralimanohar, "Improving externalization in remote microphone systems," in *Poster presented at International Hearing Aid Research Conference*, Aug 2016.

[77] T. Kaufmann, J. Sterkens, and J. M. Woodgate, "Hearing loops, the preferred assistive listening technology," *Journal of Audio Engineering Society*, vol. 63, no. 4, pp. 298–302, 2015.

[78] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 298605, pp. 1–10, 2009.

[79] G. Keidser, K. Rohrseitz, H. Dillon, V. Hamacher, L. Carter, U. Rass, and E. Convery, "The effect of multi-channel wide dynamic range compression, noise reduction, and the directional microphone on horizontal localization performance in hearing aid wearers," *International Journal of Audiology*, vol. 45, no. 10, pp. 563–579, 2006.

[80] F. Keyrouz, Y. Naous, and K. Diepold, "A new method for binaural 3-D localization based on HRTFs," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, 2006, pp. V 341–V 344.

[81] R. G. Klumpp and H. R. Eady, "Some measurements of interaural time difference thresholds," *The Journal of the Acoustical Society of America*, vol. 28, no. 5, pp. 859–860, 1956.

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

[83] T. Koshy, *Discrete mathematics with applications*. Academic Press, 2004.

[84] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Relaxed binaural LCMV beamforming," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 137–152, Jan 2017.

[85] T. Kubo, T. Sakashita, M. Kusuki, K. Kyunai, K. Ueno, C. Hikawa, T. Wada, T. Shibata, and Y. Nakai, "Sound lateralization and speech discrimination in patients with sensorineural hearing loss," *Acta Oto-Laryngologica*, vol. 118, no. 543, pp. 63–69, 1998.

[86] G. F. Kuhn, "Model for the interaural time differences in the azimuthal plane," *The Journal of the Acoustical Society of America*, vol. 62, no. 1, pp. 157–167, 1977.

[87] A. Kuklasinski, "Multi-channel dereverberation for speech intelligibility improvement in hearing aid applications," Ph.D. dissertation, Aalborg University, 2016.

[88] M. S. Lewis, C. C. Crandell, M. Valente, and J. E. Horn, "Speech perception in noise: Directional microphones versus frequency modulation (FM) systems," *Journal of the American Academy of Audiology*, vol. 15, no. 6, pp. 426–439, 2004.

[89] C. Lim and R. Duda, "Estimating the azimuth and elevation of a sound source from the output of a cochlear model," in *Proceedings of the Twenty-eighth Annual Asilomer Conference on Signals, Systems, and Computers*, vol. 1, 1994, pp. 399–403.

[90] C. Lorenzi, S. Gatehouse, and C. Lever, "Sound localization in noise in hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 105, no. 6, pp. 3454–3463, 1999.

[91] J. A. MacDonald, "A localization algorithm based on head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4290–4296, 2008.

[92] D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, "Theoretical analysis of linearly constrained multi-channel wiener filtering algorithms for combined noise reduction and binaural cue preservation in binaural hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2384–2397, Dec 2015.

[93] R. Martin, U. Heute, and C. Antweiler, *Advances in Digital Speech Transmission*. Wiley, 2008.

[94] T. May, S. van de Par, and A. Kohlrausch, *Binaural Localization and Detection of Speakers in Complex Acoustic Scenes*. Springer, 2013, pp. 397–425.

[95] M. F. McKinney, J. R. Burwinkel, and T. Zhang, "Maximum acceptable delay in hearing aids under noisy conditions," in *Poster presented at the Annual Scientific and Technology Conference of the American Auditory Society*, Scottsdale, Arizona, 2015.

[96] J. Mecklenburger and T. Groth, "Wireless technologies and hearing aid connectivity," in *Hearing Aids*. Springer, 2016, pp. 131–149.

[97] D. H. Mershon and L. E. King, "Intensity and reverberation as factors in the auditory perception of egocentric distance," *Attention, Perception, & Psychophysics*, vol. 18, no. 6, pp. 409–415, 1975.

[98] A. R. Møller, *Hearing: Its Physiology and Pathophysiology*. Academic Press, 2000.

[99] B. C. J. Moore, *Cochlear hearing loss: physiological, psychological and technical issues*. John Wiley & Sons, 2007.

[100] ——, *An Introduction to the Psychology of Hearing*. BRILL, 2012.

[101] B. C. J. Moore, B. R. Glasberg, and K. Hopkins, "Frequency discrimination of complex tones by hearing-impaired subjects: Evidence for loss of ability to use temporal fine structure," *Hearing research*, vol. 222, no. 1, pp. 16–27, 2006.

[102] D. Morikawa, Y. Toyoda, and T. Hirahara, "Head movement during horizontal and median sound localization experiments in which head-rotation is allowed," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3510–3510, 2013.

[103] H. G. Mueller, "There's less talking in barrels, but the occlusion effect is still with us," *The Hearing Journal*, vol. 56, no. 8, pp. 10–16, 2003.

[104] M. F. Mueller, A. Kegel, S. M. Schimmel, N. Dillier, and M. Hofbauer, "Localization of virtual sound sources with bilateral hearing aids in realistic acoustical scenes," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4732–4742, 2012.

[105] H. Nakashima and T. Mukai, “3D sound source localization system based on learning of binaural hearing,” in *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, vol. 4, Oct 2005, pp. 3534–3539.

[106] W. Noble, D. Byrne, and B. Lepage, “Effects on sound localization of configuration and type of hearing impairment,” *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 992–1005, 1994.

[107] W. Noble, D. Byrne, and K. Ter-Horst, “Auditory localization, detection of spatial separateness, and speech hearing in noise by hearing impaired listeners,” *The Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2343–2352, 1997.

[108] S. Perrett and W. Noble, “The effect of head rotations on vertical plane sound localization,” *The Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2325–2332, 1997.

[109] J. Pickles, *An Introduction to the Physiology of Hearing*. Brill, 2013.

[110] C. J. Plack, D. Barker, and G. Prendergast, “Perceptual consequences of “hidden” hearing loss,” *Trends in hearing*, vol. 18, pp. 1–11, 2014.

[111] C. Plack, *The Sense of Hearing: Second Edition*. Taylor & Francis, 2013.

[112] I. Pollack and J. M. Pickett, “Cocktail party effect,” *The Journal of the Acoustical Society of America*, vol. 29, no. 11, pp. 1262–1262, 1957.

[113] A. Pourmohammad and S. M. Ahadi, “Real time high accuracy 3-D PHAT-based sound source localization using a simple 4-microphone arrangement,” *IEEE Systems Journal*, vol. 6, no. 3, pp. 455–468, Sept 2012.

[114] M. Raspaud, H. Viste, and G. Evangelista, “Binaural source localization by joint estimation of ILD and ITD,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68–77, Jan 2010.

[115] L. Rayleigh, “XII. on our perception of sound direction,” *Philosophical Magazine Series 6*, vol. 13, no. 74, pp. 214–232, 1907.

[116] S. K. Roffler and R. A. Butler, “Factors that influence the localization of sound in the vertical plane,” *The Journal of the Acoustical Society of America*, vol. 43, no. 6, pp. 1255–1259, 1968.

[117] N. Roman and D. Wang, “Binaural tracking of multiple moving sources,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 728–739, May 2008.

[118] J. E. Rose, N. B. Gross, C. D. Geisler, and J. E. Hind, “Some neural mechanisms in the inferior colliculus of the cat which may be relevant to localization of a sound source.” *Journal of Neurophysiology*, vol. 29, no. 2, pp. 288–314, 1966.

[119] P. R. Roth, “Effective measurements using digital signal analysis,” *IEEE spectrum*, vol. 8, no. 4, pp. 62–70, 1971.

[120] T. T. Sandel, D. C. Teas, W. E. Feddersen, and L. A. Jeffress, “Localization of sound from single and paired sources,” *The Journal of the Acoustical Society of America*, vol. 27, no. 5, pp. 842–852, 1955.

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

[122] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[123] D. R. Schumaier, "Bone conduction hearing aid," US Patent 6 643 378 B2, 2003.

[124] A. H. Schwartz, "Effect of dynamic range compression on attending to sounds based on spatial location," Ph.D. dissertation, Massachusetts Institute of Technology, 2013.

[125] J. Smith and J. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 12, pp. 1661–1669, 1987.

[126] O. Strelcyk and T. Dau, "Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3328–3345, 2009.

[127] Q. Summerfield, "Lipreading and audio-visual speech perception," *Philosophical Transactions: Biological Sciences*, vol. 335, no. 1273, pp. 71–78, 1992.

[128] I. Tashev, *Sound Capture and Processing: Practical Approaches*. Wiley, 2009.

[129] L. Thibodeau, "Benefits of adaptive FM systems on speech recognition in noise for listeners who use hearing aids," *American Journal of Audiology*, vol. 19, no. 1, pp. 36–45, 2010.

[130] J. Thiemann, M. Müller, D. Marquardt, S. Doclo, and S. van de Par, "Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 12, pp. 1–11, 2016.

[131] A. Tjellström and B. Håkansson, "The bone-anchored hearing aid. design principles, indications, and long-term clinical results." *Otolaryngologic clinics of north America*, vol. 28, no. 1, pp. 53–72, 1995.

[132] "Type, degree, and configuration of hearing loss," *Audiology Information Series*, American Speech-Language-Hearing Association, 2015.

[133] J. Udesen, T. Piechowiak, F. Gran, and A. B. Dittberner, "Degradation of spatial sound by the hearing aid," in *Proceedings of the International Symposium on Auditory and Audiological Research*, vol. 4, 2013, pp. 271–278.

[134] M. Usman, F. Keyrouz, and K. Diepold, "Real time humanoid sound source localization and tracking in a highly reverberant environment," in *Proceedings of International Conference on Signal Processing*, Oct 2008, pp. 2661–2664.

[135] J. M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 2, 2003, pp. 1228–1233.

[136] T. Van de Bogaert, J. Wouters, T. J. Klasen, and M. Moonen, "Distortion of interaural time cues by directional noise reduction systems in modern digital hearing aids," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, pp. 57–60.

[137] S. van de Par and A. Kohlrausch, "A new approach to comparing binaural masking level differences at low and high frequencies," *The Journal of the Acoustical Society of America*, vol. 101, no. 3, pp. 1671–1680, 1997.

[138] T. Van den Bogaert, E. Carette, and J. Wouters, "Sound source localization using hearing aids with microphones placed behind-the-ear, in-the-canal, and in-the-pinna," *International Journal of Audiology*, vol. 50, no. 3, pp. 164–176, 2011.

[139] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 484–497, 2008.

[140] T. Van den Bogaert, T. J. Klasen, M. Moonen, L. Van Deun, and J. Wouters, "Horizontal localization with bilateral hearing aids: Without is better than with," *The Journal of the Acoustical Society of America*, vol. 119, no. 1, pp. 515–526, 2006.

[141] T. Van den Bogaertl, E. Carettel, and J. WoutersI, "Sound localization with and without hearing aids," in *NAG-DAGA International Conference on Acoustics*, 2009, pp. 1314–1317.

[142] M. M. Van Wanrooij and A. J. Van Opstal, "Contribution of head shadow and pinna cues to chronic monaural sound localization," *Journal of Neuroscience*, vol. 24, no. 17, pp. 4163–4171, 2004.

[143] F. Viola and W. F. Walker, "A spline-based algorithm for continuous time-delay estimation using sampled data," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 52, no. 1, pp. 80–93, 2005.

[144] A. M. Vural, "Effects of perturbations on the performance of optimum/adaptive arrays," *IEEE Transactions on Aerospace and Electronic Systems*, no. 1, pp. 76–87, 1979.

[145] X. Wan and J. L., "Robust and low complexity localization algorithm based on head-related impulse responses and interaural time difference," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. EL40–EL46, 2013.

[146] D. Wang and G. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley, 2006.

[147] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1997, pp. 187–190.

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

[149] H. Wang, C. Zhang, and Y. Wu, "Just noticeable difference of interaural level difference to frequency and interaural level difference," in *Audio Engineering Society Convention 140*, May 2016, paper 9511.

[150] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, 1993.

[151] I. M. Wiggins and B. U. Seeber, "Dynamic-range compression affects the lateral position of sounds," *The Journal of the Acoustical Society of America*, vol. 130, no. 6, pp. 3939–3953, 2011.

[152] F. Wightman, D. Kistler, and K. Andersen, "Reassessment of the role of head movements in human sound localization," *The Journal of the Acoustical Society of America*, vol. 95, no. 5, pp. 3003–3004, 1994.

[153] F. L. Wightman and D. J. Kistler, "Monaural sound localization revisited," *The Journal of the Acoustical Society of America*, vol. 101, no. 2, pp. 1050–1063, 1997.

[154] ——, "The dominant role of low-frequency interaural time differences in sound localization," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1648–1661, 1992.

[155] Y. S. Yoon, L. M. Kaplan, and J. H. McClellan, "Tops: New doa estimator for wideband signals," *IEEE Transactions on Signal processing*, vol. 54, no. 6, pp. 1977–1989, 2006.

[156] W. A. Yost, "Lateral position of sinusoids presented with interaural intensive and temporal differences," *The Journal of the Acoustical Society of America*, vol. 70, no. 2, pp. 397–409, 1981.

[157] W. A. Yost and R. H. Dye J., "Discrimination of interaural differences of level as a function of frequency," *The Journal of the Acoustical Society of America*, vol. 83, no. 5, pp. 1846–1851, 1988.

[158] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1832–1846, 2002.

[159] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica united with Acustica*, vol. 91, no. 3, pp. 409–420, 2005.

[160] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[161] C. Zhang, D. Florêncio, and Z. Zhang, "Why does phat work well in lownoise, reverberative environments?" in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008, pp. 2565–2568.

[162] C. Zhang, Z. Zhang, and D. Florêncio, "Maximum likelihood sound source localization for multiple directional microphones," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2007, pp. I–125–I–128.

[163] M. Zohourian and R. Martin, "Binaural speaker localization and separation based on a joint ITD/ILD model and head movement tracking," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2016, pp. 430–434.

Part II

Papers

Paper A

Informed TDoA-based direction of arrival estimation for hearing aid applications

Mojtaba Farmani, Michael Syskind Pedersen, Zheng-Hua Tan, and Jesper Jensen

The paper has been published in the
Proceedings of IEEE Global Conference on Signal and Information Processing,
pp. 953–957, 2015.

Abstract

This paper deals with estimation of the target sound direction of arrival (DoA) for a hearing aid system (HAS) which can connect to a wireless microphone worn by a target talker. In this setup, the HAS is "informed" about the almost noise-free content of the target sound via the wireless microphone and can use this information for the DoA estimation. Here, we propose an "informed" DoA estimator based on the time difference of arrival (TDoA) of the target sound at two microphones mounted on the ears of the HAS user—one microphone on each ear. To estimate the TDoA and the DoA, we propose a maximum likelihood framework relying on the noise-free target sound and estimation of the ambient noise characteristics. We show how the proposed ML framework allows us to estimate the TDoA and the DoA jointly or consecutively. Further, to evaluate the likelihood function efficiently, we resort to an inverse discrete Fourier transform (IDFT) technique. To study the performance of the proposed algorithms, we run simulations for various DoAs, signal to noise ratios (SNRs), and distances in large crowd noise situations. In these situations, the proposed estimator improves the estimation performance markedly over a recently proposed "informed" TDoA-based DoA estimator.

1 Introduction

Estimation of the target sound direction of arrival (DoA) enables hearing aid systems (HASs) to improve the spatial hearing of their users by maintaining or accentuating the spatial cues of the target sound [1, 2]. State-of-the-art HASs usually consist of a pair of wirelessly connected hearing aids which enables them to utilize binaural signals in their speech enhancement and DoA estimation algorithms. The DoA estimation problem has been investigated with different approaches, e.g. [1–7]. Most of them have been proposed for applications which do not have any access to the noise-free target sound, e.g. [2–6]; in other words, they are "uninformed" about the content of the target sound. However, recent advances in wireless technology allow new HASs—where the target talker is wearing a wireless microphone—to have access to an essentially noise-free version of the target signal [1, 7]. This information turns the "uninformed" DoA estimation problem into the "informed" DoA estimation problem (Fig. A.1) considered in this paper.

In previous work [7], we proposed an "informed" maximum likelihood DoA estimation algorithm, named MLSSL (maximum likelihood sound source localization), relying on the noise-free target signal, a database of head related transfer functions (HRTFs) of the specific HAS user, and estimation of the ambient noise characteristics. MLSSL is markedly effective under severely noisy conditions when the individual HRTFs are available [7].

In some situations, measuring HRTFs for each HAS user is impractical,

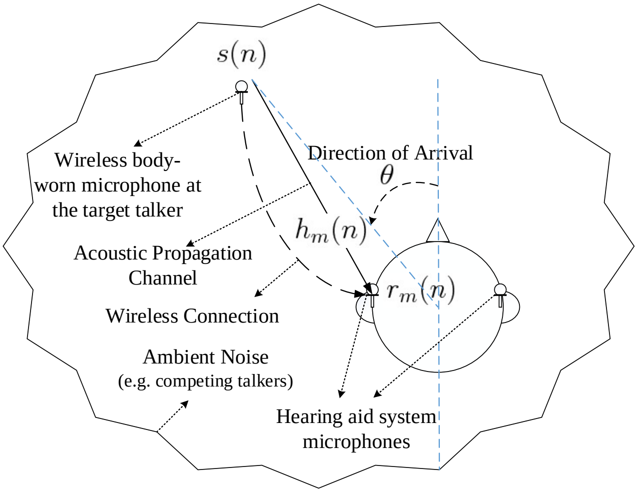

$s(n)$
Wireless body-worn microphone at the target talker
Direction of Arrival
θ
$h_m(n)$
Acoustic Propagation Channel
$r_m(n)$
Wireless Connection
Ambient Noise
(e.g. competing talkers)
Hearing aid system microphones

Fig. A.1: An "informed" binaural DoA estimation scenario for a hearing aid system using a wireless microphone. $r_m(n)$, $s(n)$ and $h_m(n)$ are the noisy received sound at microphone m, the noise-free target sound and the acoustic channel impulse response between the target talker and microphone m, respectively. $s(n)$ is available at the hearing aid via wireless connection to the wireless microphone at the target talker. The goal is to estimate θ.

and alternative methods which do not depend on user-specific HRTFs are of interest. In this paper, we propose an "informed" binaural DoA estimator that relies on a minimal number of user-related prior assumptions. The proposed DoA estimator depends on the Time Difference of Arrival (TDoA) of the target sound at two microphones placed on the ears of the HAS user—one microphone on each ear. As a signal model for the DoA estimation problem, we disregard the "shadowing effect" of the HAS user's head and consider a far field and a free field model. To estimate the TDoA and the DoA, we propose a maximum likelihood (ML) framework which performs well despite the crude modeling assumptions. Further, this framework allows us to estimate the TDoA and the DoA jointly or consecutively. We show that the joint estimation of the TDoA and the DoA improves the accuracy of the estimations in the cost of higher computation.

The "informed" TDoA-based DoA estimation problem was first studied in [1]. This method estimates the TDoA and the DoA consecutively by resorting to a cross-correlation technique and a sine law. The proposed method in this paper is different from [1], because it takes into account the background noise characteristics, which are relatively readily available, to improve the estimation performance. Moreover, the proposed method allows to estimate the TDoA and the DoA both jointly and consecutively, where the "joint approach" enhance the estimation performance.

2 Signal Model

The noisy signal r_m received at microphone m in Fig.A.1 is given by:

$$r_m(n) = s(n) \star h_m(n) + v_m(n), \qquad m = 1, 2; \tag{A.1}$$

where $s(n)$, $h_m(n)$ and $v_m(n)$ are the noise-free target signal emitted at the target talker's position, the acoustic channel impulse response between the target talker and microphone m, and an additive noise component, respectively. Furthermore, n is the discrete time index, and $\star$ is the convolution operator.

In a free field and far field situation, the acoustic channel can be modeled as a function that delays and attenuates its input signals uniformly across frequencies. This allows us to model $h_m(n)$ as:

$$H_m(k) = \sum_{n=0}^{N-1} h_m(n) \mathrm{e}^{-\frac{j2\pi kn}{N}} = \alpha_m e^{\frac{-j2\pi k}{N} D_m}, \tag{A.2}$$

where $H_m(k)$ denotes the discrete Fourier transform (DFT) of $h_m(n)$, α_m is a real number and denotes the attenuation factor due to propagation effects, D_m is the propagation time from the target sound source to microphone m, and the DFT order N is greater or equal to the duration of $h_m(n)$

Most state-of-the-art HASs operate in the short time Fourier transform (STFT) domain because of frequency dependent processing, computational efficiency and the ability to adapt to the changing conditions. Therefore, let $R_m(l,k)$, $S(l,k)$ and $V_m(l,k)$ denote the STFT of $r_m(n)$, $s(n)$ and $v_m(n)$, respectively. Specifically,

$$R_m(l,k) = \sum_n r_m(n) w(n - lA) \mathrm{e}^{-\frac{j2\pi k}{N}(n-lA)}, \tag{A.3}$$

where l and k are frame and frequency bin indexes, respectively, N is the frame length, A is the decimation factor, $w(n)$ is the windowing function, and $j = \sqrt{-1}$ is the imaginary unit. We define $S(l,k)$ and $V_m(l,k)$ similarly. Eq.(A.1) can be approximated in the STFT domain as:

$$R_m(l,k) = S(l,k) H_m(k) + V_m(l,k). \tag{A.4}$$

The accuracy of this approximation depends on the length and smoothness of $w(n)$; the longer and the smoother the support of $w(n)$, the more accurate the approximation [8].

3 Problem Statement

Let us consider a free field scenario shown in Fig. A.2; and let d_m, c, θ, and a denote the distance between the target sound source and microphone m,

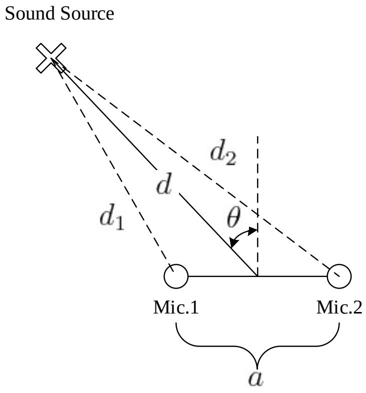


Fig. A.2: A free field scenario.

the sound speed, the target sound DoA, and the distance between the microphones, respectively. The propagation time of the target signal received at each microphone is given by:

$$D_m = \frac{d_m}{c}, \qquad m = 1, 2; \tag{A.5}$$

and the inter-microphone TDoA is:

$$D_1 - D_2 = \frac{a}{c}\sin(\theta), \tag{A.6}$$

which leads to

$$\theta = \arcsin\left((D_1 - D_2)\frac{c}{a}\right). \tag{A.7}$$

We consider two different approaches to find θ:

1. Independent delays: D_1 and D_2 are estimated independently, and the estimated D_1 and D_2 are substituted into (A.7) to estimate θ, i.e. the TDoA and the DoA are estimated consecutively.
2. Dependent delays: from (A.6), D_1 and D_2 relate to each other via θ, which allows us to use this fact and estimate the TDoA and the DoA jointly.

The proposed ML framework encompasses as special cases both these approaches for estimation of θ.

4 Maximum Likelihood Framework

In this section, we define the likelihood function for a general case, where we consider M received microphone signals ($1 \leq M \leq$ number of HAS microphones). Afterwards, to estimate θ and to maximize the likelihood function efficiently, we will reformulate the likelihood function for two special cases where $M = 1$ and $M = 2$, which correspond to the two different approaches for estimating θ outlined in section 3.

4.1 Likelihood function

To define the likelihood function, we model the additive noise $\mathbf{V}(l,k) = [V_1(l,k)\ V_2(l,k)\ ...\ V_M(l,k)]^{\mathrm{T}}$ as a zero-mean circularly-symmetric complex Gaussian vector:

$$\mathbf{V}(l,k) \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_v(l,k)), \tag{A.8}$$

where $\mathbf{C}_v(l,k) = E\{\mathbf{V}(l,k)\mathbf{V}^{\mathrm{H}}(l,k)\}$, and where $E\{.\}$ and the superscript H represent the expectation and Hermitian transpose operators, respectively. Since the noise-free signal S is available at the HAS, we can relatively easily determine the time-frequency regions in the noisy microphones signals where the target speech is essentially absent; therefore, we adaptively estimate $\mathbf{C}_v$ using exponential smoothing over the time-frequency regions where the noise is dominant. Moreover, we assume the noisy observations are independent across frequencies and frames; therefore, the likelihood function for each frame is defined by:

$$\begin{aligned} & p(\underline{\underline{\mathbf{R}}}(l)|\mathbf{S}(l), \underline{\underline{\mathbf{H}}}, \underline{\underline{\mathbf{C}_v}}(l)) = \\ & \prod_{k=1}^{N} \frac{1}{\pi^M |\mathbf{C}_v(l,k)|} e^{\{-(\mathbf{Z}(l,k))^{\mathrm{H}}\mathbf{C}_v^{-1}(l,k)(\mathbf{Z}(l,k))\}}, \end{aligned} \tag{A.9}$$

where $|.|$ denotes the matrix determinant, N is the number of frequency indexes and

$$\begin{aligned} \underline{\underline{\mathbf{R}}}(l) &= [\mathbf{R}(l,1)\ \mathbf{R}(l,2)\ \cdots\ \mathbf{R}(l,N)], \\ \mathbf{R}(l,k) &= [R_1(l,k)\ R_2(l,k)\ ...\ R_M(l,k)]^{\mathrm{T}},\ 1 \le k \le N, \\ \mathbf{S}(l) &= [S(l,1)\ S(l,2)\ \cdots S(l,N)], \\ \underline{\underline{\mathbf{C}_v}}(l) &= [\mathbf{C}_v(l,1)\ \mathbf{C}_v(l,2)\ \cdots\ \mathbf{C}_v(l,N)]^{\mathrm{T}}, \\ \mathbf{Z}(l,k) &= \mathbf{R}(l,k) - S(l,k)\mathbf{H}(k), \\ \underline{\underline{\mathbf{H}}} &= [\mathbf{H}(1)\ \mathbf{H}(2)\ \cdots\ \mathbf{H}(N)], \\ \mathbf{H}(k) &= [H_1(k)\ H_2(k)\ ...\ H_M(k)]^{\mathrm{T}},\ 1 \le k \le N, \\ &= [\alpha_1 e^{-j2\pi\frac{k}{N}D_1}\ ...\ \alpha_M e^{-j2\pi\frac{k}{N}D_M}]^{\mathrm{T}}. \end{aligned}$$

The corresponding reduced log-likelihood function, with terms independent of D_m and θ omitted, is given by:

$$\hat{\mathcal{L}} = \sum_{k=1}^{N} \{-(\mathbf{Z}(l,k))^{\mathrm{H}}\mathbf{C}_v^{-1}(l,k)(\mathbf{Z}(l,k))\}. \tag{A.10}$$

4.2 Independent delays estimation ($M = 1$)

If we consider the received signal of each microphone independently, i.e. $M = 1$, the reduced log-likelihood function $\hat{\mathcal{L}}_m$ for the m^{th} microphone can be written as

$$\hat{\mathcal{L}}_m(\alpha_m, D_m) = -\sum_{k=1}^{N} \frac{Z_m^*(l,k)Z_m(l,k)}{C_v(l,k)}, \tag{A.11}$$

where $Z_m(l,k) = R_m(l,k) - \alpha_m S(l,k)e^{-j2\pi\frac{K}{N}D_m}$, and $*$ represents the complex conjugate operator.

We aim to find the maximum likelihood estimate (MLE) of D_m. To make $\hat{\mathcal{L}}_m$ independent of α_m, we replace the MLE of α_m into (A.11). To find the MLE of α_m, we solve $\frac{\partial\hat{\mathcal{L}}_m}{\partial\alpha_m} = 0$. This leads to:

$$\hat{\alpha}_m = \sum_{k=1}^{N} \frac{R_m(l,k)}{S(l,k)} e^{j2\pi\frac{k}{N}D_m}. \tag{A.12}$$

By inserting (A.12) into (A.11) and making simplifications, we have:

$$\tilde{\mathcal{L}}_m(D_m) = \sum_{k=1}^{N} \frac{1}{C_v(l,k)} S^*(l,k)R_m(l,k)e^{j2\pi\frac{k}{N}D_m}, \tag{A.13}$$

which must be maximized for D_m. This can be done efficiently because (A.13) is an inverse discrete Fourier transform (IDFT) with respect to D_m. Equation (A.13) can also be interpreted as a Generalized Cross Correlation (GCC) relation [9] with a weighting function of $\psi(k) = \frac{1}{C_v(l,k)}$. The MLE of D_m equals:

$$\hat{D}_m = \arg\max_{D_m} \tilde{\mathcal{L}}_m(D_m), \qquad m = 1, 2\,, \tag{A.14}$$

which when inserted in (A.7) leads to an estimate of θ:

$$\hat{\theta} = \arcsin\left((\hat{D}_1 - \hat{D}_2)\frac{c}{a}\right). \tag{A.15}$$

4.3 Dependent delays estimation ($M = 2$)

In the previous subsection, we estimated D_1 and D_2 independently. However, D_1 and D_2 depends on each other via θ. In this subsection, we consider (A.6) and the received signals of $M = 2$ microphones together to estimate D_m and θ jointly.

In the following, we find the MLE of θ for two different cases of $\mathbf{C}_v(l,k)$. We first consider the general case of $\mathbf{C}_v(l,k)$ without any constraints. Afterwards, we assume that V_1 and V_2 are uncorrelated, and we model $\mathbf{C}_v(l,k)$ as a diagonal matrix to decrease the computation overhead.

General $\mathbf{C}_v(l,k)$

Let us denote $\mathbf{C}_v^{-1}(l,k)$ for $M = 2$ as

$$\mathbf{C}_v^{-1}(l,k) = \begin{bmatrix} C_{11}(l,k) & C_{12}(l,k) \\ C_{21}(l,k) & C_{22}(l,k) \end{bmatrix}. \tag{A.16}$$

Further, in a far field and a free field situation, we have that $\alpha_1 = \alpha_2 = \alpha$. Using this assumption, we expand (A.10) for $M = 2$ and note that $D_2 = D_1 - \sin(\theta)\frac{a}{c}$. The obtained expansion $\hat{\mathcal{L}}(\theta, \alpha, D_1)$ is a function of θ, α, and D_1, and we aim to find the MLE of θ and D_1. To eliminate the dependency on α, we substitute the MLE of α in $\hat{\mathcal{L}}(\theta, \alpha, D_1)$. It can be shown that the MLE of α is:

$$\hat{\alpha} = \frac{f(\theta, D_1)}{g(\theta)}, \tag{A.17}$$

where

$$\begin{aligned} f(\theta, D_1) \quad = \quad & \sum_{k=1}^{N} \Big(C_{11}(l,k)R_1(l,k) + C_{12}(l,k)R_2(l,k) \\ & \quad + \big(C_{21}(l,k)R_1(l,k) + C_{22}(l,k)R_2(l,k)\big) \\ & \quad e^{j2\pi \frac{k}{N}[-\sin(\theta)\frac{a}{c}]}\Big) S^*(l,k) e^{j2\pi \frac{k}{N} D_1}, \end{aligned} \tag{A.18}$$

and

$$\begin{aligned} g(\theta) \quad = \quad & \sum_{k=1}^{N} \Big(C_{11}(l,k) + 2C_{21}(l,k) e^{j2\pi \frac{k}{N}[-\sin(\theta)\frac{a}{c}]} + \\ & \qquad C_{22}(l,k)\Big) |S(l,k)|^2. \end{aligned} \tag{A.19}$$

where [.] rounds to nearest integer. Inserting $\hat{\alpha}$ into $\hat{\mathcal{L}}(\theta, \alpha, D_1)$ gives us:

$$\tilde{\mathcal{L}}(\theta, D_1) = \frac{f^2(\theta, D_1)}{g(\theta)}. \tag{A.20}$$

From (A.18), it can be seen that $f(\theta, D_1)$ is an IDFT, which can be evaluated efficiently, with respect to D_1; therefore, for a given θ, computing $\tilde{\mathcal{L}}(\theta, D_1)$ results in a discrete-time sequence, where the MLE of D_1 is the time index of the maximum of the sequence. Since θ is unknown, we consider a discrete set Θ of different θs, and compute $\tilde{\mathcal{L}}(\theta, D_1)$ for each $\theta \in \Theta$. The MLEs of D_1 and θ are then found from the global maximum:

$$[\hat{\theta}, \hat{D_1}] = \arg\, \max_{\theta \in \Theta, D_1} \tilde{\mathcal{L}}(\theta, D_1). \tag{A.21}$$

Diagonal $\mathbf{C}_v(l,k)$

To decrease the computation overhead and to simplify the solution, let us assume $V_1(l,k)$ and $V_2(l,k)$ are uncorrelated, so that the noise covariance matrix is diagonal:

$$\mathbf{C}_v^{-1}(l,k) = \begin{bmatrix} C_{11}(l,k) & 0 \\ 0 & C_{22}(l,k) \end{bmatrix}. \tag{A.22}$$

Following a similar procedure as in the previous section leads to a reduced log-likelihood function

$$\tilde{\mathcal{L}}(\theta, D_1) = \sum_{k=1}^{N} \Big(p(l,k) + q(l,k,\theta) \Big) S^*(l,k) e^{j2\pi \frac{k}{N} D_1}, \tag{A.23}$$

where

$$\begin{aligned} p(l,k) &= C_{11}(l,k) R_1(l,k), && \text{(A.24)} \\ q(l,k,\theta) &= C_{22}(l,k) R_2(l,k) e^{j2\pi \frac{k}{N}[-\sin(\theta)\frac{a}{c}]}. && \text{(A.25)} \end{aligned}$$

As before, (A.23) can be evaluated using an IDFT with respect to D_1; however, due to its simpler structure, it is computationally cheaper than (A.20). The MLEs of D_1 and θ are found as in (A.21).

5 Related Work

We compare the proposed methods with the method proposed in [1], which belongs to the "independent delays" class of approaches and which uses conventional cross correlation to find D_1 and D_2. In general, any method based on Generalized Cross Correlation (GCC) method [9] can be used to estimate D_1 and D_2 independently:

$$\hat{D}_m = \arg\max_{D_m} \mathcal{R}^{\mathrm{GCC}}_{S,R_m}(D_m), \qquad m = 1, 2; \tag{A.26}$$

$$\mathcal{R}^{\mathrm{GCC}}_{S,R_m}(D_m) = \sum_{k=1}^{N} \psi(k) S^*(l,k) R_m(l,k) e^{j2\pi \frac{k}{N} D_m}. \tag{A.27}$$

The method proposed in [1] uses $\psi(k) = 1$. Regarding the PHAT weighting function [9], we propose an "informed" PHAT weighting function as $\psi(k) = \frac{1}{|S^*(l,k) R_m(l,k)|}$ for comparison.

6 Simulation Experiments

In this section, we evaluate estimation performance in simulation experiments. Specifically, we study the effects of the target sound DoA θ, the

Signal-to-Noise ratio (SNR) and the distance d between the target source and the user.

6.1 Setup

To simulate a real world situation, we use a set of head related impulse responses (HRIRs) measured with behind-the-ear (BTE) hearing aids which are mounted behind each pinna of a head-and-torso-simulator (HATS) in an anechoic chamber. The HRIRs were measured for 35 positions uniformly spaced on a semicircle in the front-horizontal plane with radius 1.2 m centered at the HATS, i.e. $\theta \in \{-85°, -80°, \cdots, 85°\}$. To simulate a signal from a position, we convolve the signal with the corresponding HRIR.

We consider a 20-second sample of the ISTS signal [10] composed of 21 female voices in 6 different languages as the target speech signal. To approximate a large-crowd noise field, we synthesize different speech signals originating from each of the 35 positions simultaneously. The TSP database [11], which consists of different male and female voices, is used as noise sound sources. The global SNR of a given simulation experiment is expressed relative to the left-ear microphone signals. The other simulation parameters are as follows: the sampling frequency is 20 kHz, $N = 2048$, $A = 1024$, $w(n)$ is a hamming window, and $\Theta = \{-85°, -80°, \cdots, 85°\}$. Due to the presence of the head, a must be chosen greater than the distance between the microphones in a free field situation to model the curved path around the head. Therefore, a has been calibrated separately to maximize the performance for each method. As a performance metric, we use the mean absolute error (MAE) given by:

$$\sigma = \frac{1}{L} \sum_{j=1}^{L} |\theta - \hat{\theta}_j|, \tag{A.28}$$

where $\hat{\theta}_j$ is the estimated DoA for the j^{th} frame of the signal.

6.2 Results and discussion

Fig. A.3 shows the MAE of the DoA estimators as a function of θ at an SNR of 0 dB. As can be seen, the proposed ML-based methods perform better than the Cross-Correlation-based method [1] and the proposed "informed" PHAT method. Among the ML-based methods, the ones which consider dependent delays estimate θ more accurately, at a higher computation cost. However, using a non-diagonal $\mathbf{C}_v$ does not provide considerable improvement compared with modeling $\mathbf{C}_v$ as diagonal. The estimators perform worse for θs towards the sides of the head because the far field and free field assumption (i.e. $\alpha_1 = \alpha_2$) is less valid for these θs.

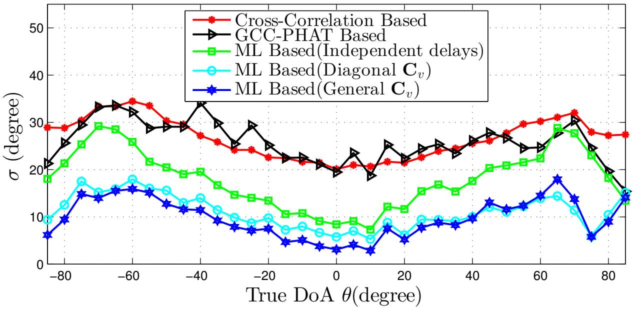


Fig. A.3: Performance as a function of θ at SNR $= 0\,$dB.

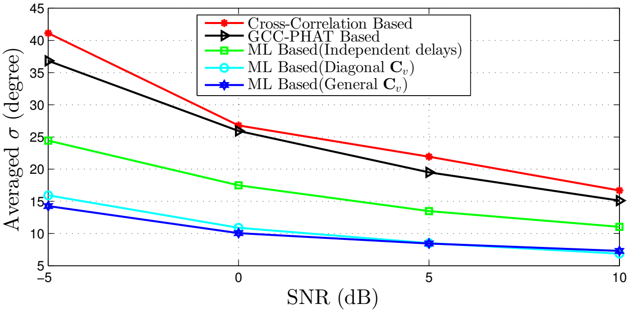


Fig. A.4: Performance as a function of SNR, averaged across θs.

Fig. A.4 shows the MAE averaged over all θs as a function of SNR. As expected, the higher the SNR, the better the performance. The general performance order of Fig. A.3 remains at different SNRs.

We finally study the performance of the estimators as a function of distance between the target source and the HAS user. Since HRTF measurements for long distances were not available, we use the analytical HRTFs for a spherical head model computed by the model of Duda and Martens in [12]. Fig. A.5 shows the performance for different ds. As can be seen, the ML-based methods are more effective than the other "informed" DoA estimators. The ML-based method which considers a general $\mathbf{C}_v$ is less accurate than the one which considers a diagonal $\mathbf{C}_v$; apparently, using a general $\mathbf{C}_v$ makes the ML framework more sensitive to the violation of the free field assumption. Moreover, it should be noted that the SNR is kept at $0\,$dB for all distances, explaining why the performance hardly degrades with distance.

7 Conclusion

In this paper, we proposed a binaural TDoA-based DoA estimator for a new hearing aid system which is able to connect to a wireless microphone and

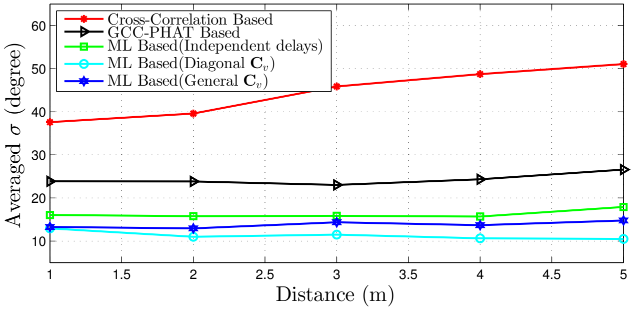


Fig. A.5: Performance as a function of d at SNR $= 0$ dB, averaged across θs.

has access to the noise-free version of the target signal. To rely on minimal number of user-specific prior assumptions, we considered a free field and far field signal model, and we proposed a maximum likelihood framework based on the noise-free target sound and the back-ground noise characteristics to estimate the DoA. We showed that for $M = 1$ and $M = 2$ microphones, the likelihood function can be calculated efficiently via inverse-discrete-Fourier-transform techniques. In simulation experiments with a target speech signal in a large-crowd noise, the proposed ML framework performs better than a recently proposed "informed" TDoA-based DoA estimator [1]. Future work includes investigating the effect of reverberation and more realistic acoustic setups.

References

[1] G. Courtois, P. Marmaroli, M. Lindberg, Y. Oesch, and W. Balande, "Implementation of a binaural localization algorithm in hearing aids: specifications and achievable solutions," in *Audio Engineering Society Convention 136*, April 2014, paper 9034.

[2] W. C. Wu, C. H. Hsieh, H. C. Huang, and O. C. Chen, "Hearing aid system with 3D sound localization," in *IEEE TENCON*, 2007, pp. 1–4.

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

[4] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[5] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68–77, Jan 2010.

[6] F. keyrouz, "Advanced binaural sound localization in 3-D for humanoid robots," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 9, pp. 2098–2107, Sept 2014.

[7] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "On the influence of microphone array geometry on HRTF-based sound source localization," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2015, pp. 439–443.

[8] Y. Avargel, "Linear system identification in the short-time Fourier transform domain," Ph.D. dissertation, Israel Institute of Technology, 2008.

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

[10] European Hearing Industry Manufactures, "International Speech Test Signal," http://www.ehima.com.

[11] P. Kabal, "TSP speech database," Department of Electrical and Computer Engineering, McGill University, Tech. Rep., 2002.

[12] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *The Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048–3058, 1998.

Paper B

Informed direction of arrival estimation using a spherical-head model for hearing aid applications

Mojtaba Farmani, Michael Syskind Pedersen, Zheng-Hua Tan,
and Jesper Jensen

The paper has been published in the
Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 360–364, 2016.

The layout has been revised.

Abstract

In this paper, we propose a direction of arrival (DoA) estimator for a hearing aid system (HAS) which can connect to a wireless microphone worn by a target talker. The wireless microphone "informs" the HAS about the almost noise-free content of the target sound, and the proposed DoA estimator uses the knowledge of the noise-free target sound and the received microphone signals to estimate the DoA via a maximum likelihood approach. Moreover, the proposed DoA estimator resorts to a user-independent spherical-head model to consider the acoustic impacts of the head on the received signals at the HAS. Further, the proposed DoA estimator uses an inverse discrete Fourier transform (IDFT) technique to evaluate the likelihood function computationally efficiently. We assessed the performance of the proposed estimator for various DoAs, signal to noise Ratios (SNRs), and target distances in different noisy and reverberant situations. The proposed estimator improves the performance markedly over other recently proposed "informed" DoA estimators.

1 Introduction

Direction of arrival (DoA) estimation of a target sound has been investigated with different approaches in various applications, such as robotics [1–3], video conferencing [4], surveillance [5], wireless acoustic sensor network [6], and hearing aids [7–10]. In this paper, we propose a DoA estimator for an advanced hearing aid system (HAS) which can connect to a wireless microphone worn by a target talker. Recognizing the target sound DoA allows HASs to enhance the spatial hearing of the HAS user by maintaining or accentuating the spatial cues of the target sound [10–12].

Most DoA estimation algorithms have been proposed for applications which are "uninformed" about the noise-free content of the target sound, e.g. [1–7, 12–15]. However, advances in wireless technology enable new HASs—where the target talker is wearing a wireless microphone—to have access to an essentially noise-free version of the target signal [8–11]. This change introduces the "informed" DoA estimation problem considered in this paper (Fig. B.1).

The "informed" DoA estimation problem was first studied and tackled via a binaural time-difference-of-arrival (TDoA)-based method in [10]. This method estimates the TDoA by resorting to a cross-correlation technique and then maps the estimated TDoA to a DoA estimate through a sine law. This method [10] has a low computational overhead and confines the target locations to the front-horizontal plane.

In previous papers [8, 9], we also dealt with the "informed" DoA estimation problem. Specifically, we proposed a maximum likelihood (ML) framework that utilizes the wirelessly transmitted signal and ambient noise

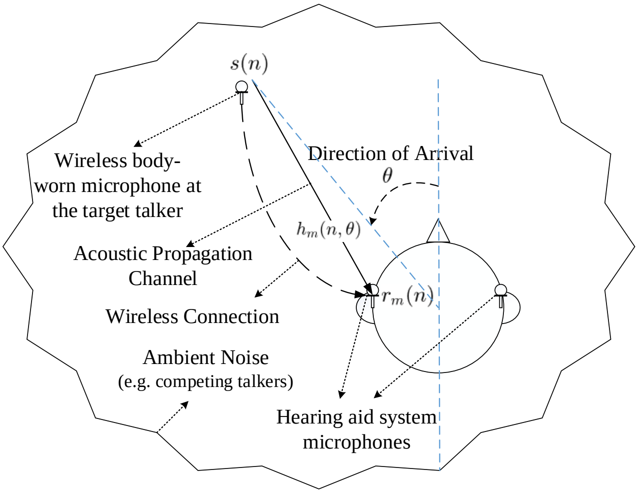


Fig. B.1: An "informed" DoA estimation scenario for a hearing aid system using a wireless microphone. $r_m(n)$, $s(n)$ and $h_m(n,\theta)$ are the noisy received sound at microphone m, the noise-free target sound and the acoustic channel impulse response between the target talker and microphone m, respectively. $s(n)$ is available at the hearing aid via wireless connection, and the goal is to estimate θ.

characteristics for DoA estimation. The algorithm proposed in [8]—called MLSSL (maximum likelihood sound source localization)—uses a database of measured head related transfer functions (HRTFs) of the specific HAS user in order to model the user's head shadowing effect and the acoustic channel. On the other hand, the estimator proposed in [9], which is a TDoA-based DoA estimator, employs a free-field and far-field model to avoid user-related prior assumptions. The signal model in [9] enabled the use of inverse discrete Fourier transform (IDFT) techniques to evaluate the likelihood function computationally efficiently.

MLSSL [8] and the TDoA-based method [9] form a family of ML-based methods for solving the "informed" DoA estimation problem. These two methods are the two extremes in this family regarding modeling of and dependence on the acoustic characteristics of the specific user's head: MLSSL [8] relies on detailed knowledge of the head characteristics of a specific user, while the TDoA-based method [9] totally ignores the acoustic shadowing effect of the head. In general, MLSSL is more accurate than the TDoA-based method at the cost of higher computation and prior knowledge of HRTFs.

In this paper, we propose an intermediate approach to gain advantages of both methods. To improve the accuracy over the TDoA-based method, we propose a simplified spherical-head model which allows to consider the acoustic effects of the head without being user-dependent. Further, we show that the likelihood function in the proposed method can be computed efficiently using IDFTs. The proposed method is different from [10] because it uses a maximum likelihood approach, which considers the background noise characteristics, models the presence of the head, and estimates the DoA and the TDoA jointly.

2 Signal Model

Regarding Fig. B.1, for microphone m of the HAS, we can write:

$$r_m(n) = s(n) * h_m(n,\theta) + v_m(n), \qquad m \in \{\text{left, right}\}\,, \tag{B.1}$$

where r_m, s, h_m and v_m are the noisy signal received at microphone m, the noise-free target signal emitted at the target talker's position, the acoustic channel impulse response between the target talker and microphone m, and an additive noise component, respectively. n is the discrete time index, and $*$ is the convolution operator.

Let $R_m(l,k)$, $S(l,k)$ and $V_m(l,k)$ denote the short time Fourier transform (STFT) of r_m, s and v_m, respectively. Specifically, let

$$R_m(l,k) = \sum_n r_m(n) w(n - lA) \mathrm{e}^{-\frac{j2\pi k}{N}(n-lA)}, \tag{B.2}$$

where l and k are frame and frequency bin indexes, respectively, N is the frame length, A is the decimation factor, $w(n)$ is the windowing function, and $j = \sqrt{-1}$ is the imaginary unit. We define $S(l,k)$ and $V_m(l,k)$ similarly. Moreover, let $H_m(k,\theta)$ denote the discrete Fourier transform (DFT) of h_m:

$$\begin{aligned} H_m(k,\theta) &= \sum_n h_m(n,\theta) \mathrm{e}^{-\frac{j2\pi kn}{N}} && \text{(B.3)} \\ &= \alpha_m(k,\theta) \mathrm{e}^{-\frac{j2\pi k}{N} D_m(k,\theta)}, && \text{(B.4)} \end{aligned}$$

where N is the DFT order, $\alpha_m(k,\theta)$ is a real number and denotes the frequency-dependent attenuation factor due to propagation effects, and $D_m(k,\theta)$ is the frequency-dependent propagation time from the target sound source to microphone m. For simplicity and to decrease computation overhead, we model the acoustic channel as a function that delays and attenuates its input signals uniformly across frequencies [9], i.e.

$$\tilde{H}_m(k,\theta) = \tilde{\alpha}_m(\theta) \mathrm{e}^{-\frac{j2\pi k}{N} \tilde{D}_m(\theta)}, \tag{B.5}$$

where $\tilde{D}_m(\theta)$ and $\tilde{\alpha}_m(\theta)$ are frequency-independent. Now, we can approximate Eq. (B.1) in the STFT domain as:

$$R_m(l,k) = S(l,k) \tilde{H}_m(k,\theta) + V_m(l,k). \tag{B.6}$$

The vector form of Eq. (B.6) is written as:

$$\boldsymbol{R}(l,k) = S(l,k)\tilde{\boldsymbol{H}}(k,\theta) + \boldsymbol{V}(l,k), \tag{B.7}$$

where

$$\begin{aligned}
\boldsymbol{R}(l,k) &= [\ R_{\text{left}}(l,k),\ R_{\text{right}}(l,k)\ ]^{\mathsf{T}}, \\
\tilde{\boldsymbol{H}}(k,\theta) &= [\ \tilde{H}_{\text{left}}(k,\theta),\ \tilde{H}_{\text{right}}(k,\theta)\ ]^{\mathsf{T}}, \\
\boldsymbol{V}(l,k) &= [\ V_{\text{left}}(l,k),\ V_{\text{right}}(l,k)\ ]^{\mathsf{T}},
\end{aligned}$$

and the superscript T is the transpose operator.

3 Maximum Likelihood Framework

To define the likelihood function, we assume the additive noise observed at the microphones is distributed according to a zero-mean circularly-symmetric complex Gaussian distribution, i.e. $\boldsymbol{V}(l,k) \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_v(l,k))$, where $\mathbf{C}_v(l,k) = \mathrm{E}\{\boldsymbol{V}(l,k)\boldsymbol{V}^{\mathrm{H}}(l,k)\}$, and where $\mathrm{E}\{.\}$ and superscript H represent the expectation and Hermitian transpose operators, respectively. Since $S(l,k)$ is available at the HAS, we can relatively easily determine the time-frequency regions in the received noisy microphone signals, where the target speech is essentially absent; therefore, we adaptively estimate $\mathbf{C}_v(l,k)$ using exponential smoothing over these time-frequency regions. Moreover, we assume the noisy observations are independent across frequencies; therefore, the likelihood function for each frame is defined by:

$$\begin{aligned}
&p(\underline{\underline{\mathbf{R}}}(l)|\boldsymbol{S}(l),\underline{\underline{\tilde{\mathbf{H}}}}(\theta),\underline{\underline{\mathbf{C}_v}}(l)) = \\
&\qquad \prod_{k=1}^{N} \frac{1}{\pi^M |\mathbf{C}_v(l,k)|} e^{\{-(\mathbf{Z}(l,k))^{\mathrm{H}}\mathbf{C}_v^{-1}(l,k)(\mathbf{Z}(l,k))\}},
\end{aligned} \tag{B.8}$$

where $|.|$ denotes the matrix determinant, N is the number of frequency indexes and

$$\begin{aligned}
\underline{\underline{\mathbf{R}}}(l) &= [\ \boldsymbol{R}(l,1),\ \boldsymbol{R}(l,2),\ \cdots,\ \boldsymbol{R}(l,N)\ ], \\
\boldsymbol{R}(l,k) &= [\ R_{\text{left}}(l,k),\ R_{\text{right}}(l,k)\ ]^{\mathsf{T}},\ 1 \le k \le N, \\
\boldsymbol{S}(l) &= [\ S(l,1),\ S(l,2),\ \cdots,\ S(l,N)\ ]^{\mathsf{T}}, \\
\underline{\underline{\tilde{\mathbf{H}}}}(\theta) &= [\ \tilde{\boldsymbol{H}}(1,\theta),\ \tilde{\boldsymbol{H}}(2,\theta),\ \cdots,\ \tilde{\boldsymbol{H}}(N,\theta)\ ], \\
\tilde{\boldsymbol{H}}(k,\theta) &= [\ \tilde{H}_{\text{left}}(k,\theta),\ \tilde{H}_{\text{right}}(k,\theta)\ ]^{\mathsf{T}} \\
&= \begin{bmatrix} \tilde{\alpha}_{\text{left}}(\theta)e^{-j2\pi\frac{k}{N}\tilde{D}_{\text{left}}(\theta)} \\ \tilde{\alpha}_{\text{right}}(\theta)e^{-j2\pi\frac{k}{N}\tilde{D}_{\text{right}}(\theta)} \end{bmatrix},\ 1 \le k \le N, \\
\underline{\underline{\mathbf{C}_v}}(l) &= [\ \mathbf{C}_v(l,1),\ \mathbf{C}_v(l,2),\ \cdots,\ \mathbf{C}_v(l,N)\ ]^{\mathsf{T}}, \\
\boldsymbol{Z}(l,k) &= \boldsymbol{R}(l,k) - S(l,k)\tilde{\boldsymbol{H}}(k).
\end{aligned}$$

The corresponding reduced log-likelihood function, with terms independent of θ omitted, is given by:

$$\tilde{\mathcal{L}} = \sum_{k=1}^{N} \{-(\mathbf{Z}(l,k))^{\mathrm{H}} \mathbf{C}_v^{-1}(l,k)(\mathbf{Z}(l,k))\}. \tag{B.9}$$

4 DoA Estimation using a Head Model

In this section, we aim to find the MLE of θ. The first step is to describe the acoustic model of the head.

4.1 Spherical-head model

To describe the acoustic characteristics of a head, we use the "inter-microphone time difference" (IMTD) and the "inter-microphone level difference" (IMLD), which are defined as follows:

$$\begin{aligned} \text{IMTD} : \Delta T(\theta) &= \tilde{D}_{\text{left}}(\theta) - \tilde{D}_{\text{right}}(\theta), && \text{(B.10)} \\ \text{IMLD} : \Delta L(\theta) &= 20 \log_{10} \left(\frac{\tilde{\alpha}_{\text{left}}(\theta)}{\tilde{\alpha}_{\text{right}}(\theta)} \right), && \text{(B.11)} \end{aligned}$$

where $\tilde{D}_m$ and $\tilde{\alpha}_m$ are defined in Eq. (B.5).

In general, IMTD and IMLD are frequency-dependent; however, to compute the likelihood function computationally efficiently using IDFTs, we assume they are frequency-independent. Despite this crude assumption, we show in our simulation experiments that this leads to performance improvements. For a rigid spherical head, the IMTD can be approximated by [16]:

$$\text{IMTD} : \tilde{D}_{\text{left}}(\theta) - \tilde{D}_{\text{right}}(\theta) = \frac{b}{c} \left(\sin(\theta) + \theta \right), \tag{B.12}$$

where b is the sphere radius and c is the speed of sound. To model the IMLD, we use the following relation inspired by the work in [15]:

$$\text{IMLD} : 20 \log_{10} \left(\frac{\tilde{\alpha}_{\text{left}}(\theta)}{\tilde{\alpha}_{\text{right}}(\theta)} \right) = \gamma \sin(\theta). \tag{B.13}$$

In [15], γ is a frequency-dependent scaling factor, which is generally smaller at lower frequencies and larger at higher frequencies; however, to be able to apply IDFTs, we assume γ to be frequency-independent. We describe how to determine this value in sec. 4.3.

4.2 DoA estimator

To find the MLE of θ, we expand Eq. (B.9). Let us denote

$$\mathbf{C}_v^{-1}(l,k) \equiv \begin{bmatrix} C_{11}(l,k) & C_{12}(l,k) \\ C_{21}(l,k) & C_{22}(l,k) \end{bmatrix}. \tag{B.14}$$

From Eqs. (B.12) and (B.13), $\tilde{D}_{\text{right}}$ and $\tilde{\alpha}_{\text{right}}$ can be expressed in terms of $\tilde{D}_{\text{left}}$ and $\tilde{\alpha}_{\text{left}}$, respectively. Inserting these expressions in Eq. (B.9), we arrive at $\tilde{\mathcal{L}}(\theta, \tilde{D}_{\text{left}}, \tilde{\alpha}_{\text{left}})$ which is independent of $\tilde{D}_{\text{right}}$ and $\tilde{\alpha}_{\text{right}}$. To eliminate the dependency on $\tilde{\alpha}_{\text{left}}$, we insert the MLE of $\tilde{\alpha}_{\text{left}}$ in $\tilde{\mathcal{L}}$. It can be shown that the MLE of $\tilde{\alpha}_{\text{left}}$ is:

$$\hat{\alpha}_{\text{left}} = \frac{f(\theta, D_{\text{left}})}{g(\theta)}, \tag{B.15}$$

where

$$\begin{aligned} f(\theta, D_{\text{left}}) = \sum_{k=1}^{N} \Bigg(& C_{11}(l,k) R_{\text{left}}(l,k) + \\ & C_{12}(l,k) R_{\text{right}}(l,k) + 10^{\frac{\gamma \sin(\theta)}{20}} \Big(C_{21}(l,k) R_{\text{left}}(l,k) + \\ & C_{22}(l,k) R_{\text{right}}(l,k) \Big) e^{j2\pi \frac{k}{N} [-\frac{b}{c}(\sin(\theta)+\theta)]} \Bigg) \times \\ & S^*(l,k) e^{j2\pi \frac{k}{N} D_{\text{left}}(\theta)}, \end{aligned} \tag{B.16}$$

$$\begin{aligned} g(\theta) = \sum_{k=1}^{N} \Big(& C_{11}(l,k) + \\ & 2 \times 10^{\frac{\gamma \sin(\theta)}{20}} C_{21} e^{j2\pi \frac{k}{N} [-\frac{b}{c}(sin(\theta)+\theta)]} + \\ & 10^{\frac{\gamma \sin(\theta)}{10}} C_{22}(l,k) \Big) |S(l,k)|^2, \end{aligned} \tag{B.17}$$

where [.] rounds to nearest integer. Inserting $\hat{\alpha}_{\text{left}}$ into $\tilde{\mathcal{L}}$ gives us:

$$\tilde{\mathcal{L}}(\theta, D_{\text{left}}) = \frac{f^2(\theta, D_{\text{left}})}{g(\theta)}. \tag{B.18}$$

Note that $f(\theta, D_{\text{left}})$ in Eq. (B.16) has a structure of an IDFT, which can be evaluated computationally efficiently, with respect to D_{left}; therefore, for a given θ, computing $\tilde{\mathcal{L}}(\theta, D_{\text{left}})$ results in a discrete-time sequence, where the MLE of D_{left} is the time index of the maximum of the sequence. Since θ is unknown, we consider a discrete set Θ of different θs, and evaluate $\tilde{\mathcal{L}}(\theta, D_{\text{left}})$ using an IDFT for each $\theta \in \Theta$. The MLEs of D_{left} and θ are then given by the global maximum:

$$[\hat{\theta}, \hat{D}_{\text{left}}] = \text{arg max}_{\theta \in \Theta, D_{\text{left}}} \tilde{\mathcal{L}}(\theta, D_{\text{left}}). \tag{B.19}$$

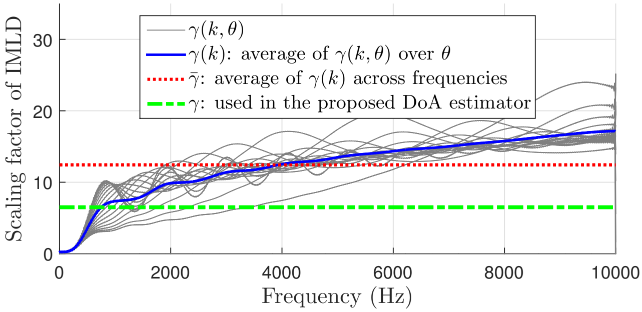


Fig. B.2: Scaling factor γ of IMLD (Eq. (B.13)) for a spherical head using theoretical HRTFs [17].

4.3 Scaling factor γ

The only remaining issue is the value of γ, which should be inserted in Eqs. (B.16) and (B.17) to evaluate Eq. (B.18). As shown in Fig. B.2, ideally, the scaling factor is frequency- and DoA-dependent ($\gamma(k,\theta)$). To find a frequency- and DoA-independent γ, one could consider averaging over DoAs and frequencies, which leads to $\bar{\gamma} \approx 12.4$. However, in the considered application, the target signal is speech, which is a relatively low-pass signal. Therefore, we expect that low-frequency components should play a larger role in finding γ.

To find the appropriate value of γ, we run simulations for numerous acoustic setups and different $\gamma \in \Gamma = \{1, 1.5, 2, ..., 20\}$ and select the γ leading to the best DoA estimation performance. We evaluate the performance in terms of Mean Absolute Error (MAE):

$$\sigma = \frac{1}{\mathrm{L}} \sum_{j=1}^{\mathrm{L}} |\theta - \hat{\theta}_j|, \tag{B.20}$$

where $\hat{\theta}_j$ is the estimated DoA for the j^{th} frame of the signal, and L is the number of target-active frames. We use the value of γ which minimizes the MAE over the considered conditions.

To simulate a rigid spherical-head, we use theoretical HRTFs proposed in [17]. We run simulations for 72 different configurations: four different target sources (two males and two females), three different distances (1 m, 5 m and 10 m), three different SNRs (-10 dB, 0 dB and 10 dB) and two different noise types (large-crowd noise and bottling-factory-hall noise). The signal duration for each configuration is 60 s, and we use the speech database provided by [18] for the target signals. For each configuration, the target source is placed at 35 different angles at the front-horizontal plane, i.e. $\theta \in \{-85°, -80°, \cdots, 85°\}$. The other simulation parameters are as follows: the sampling frequency is 20 kHz, $N = 2048$, $A = 1024$, and $w(n)$ is a Ham-

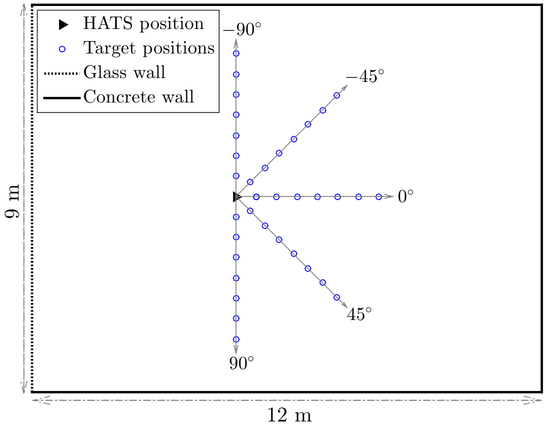


Fig. B.3: The map of the room used for HRIRs measurements.

ming window.

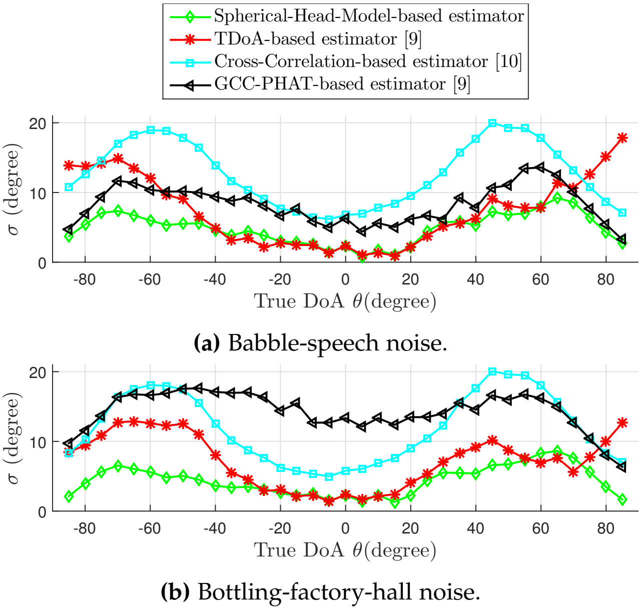
From the simulation results, we find that $\gamma = 6.5$ provides minimum MAE averaged over all considered configurations and θs. As expected, the obtained value of $\gamma = 6.5$ is less than the result of a simple averaging of the scaling factor over the frequencies for the considered spherical head, i.e. $\bar{\gamma} \approx 12.4$ (Fig. B.2).

5 Simulation Results

In this section, we evaluate the proposed estimator under realistic conditions which were not used in the simulation experiments to find γ. Here, we study the impacts of the true DoA, noise type, SNR, reverberation level, and the target distance on the performance of the proposed estimator. In the following, the proposed estimator is referred as "Spherical-Head-Model-based DoA estimator".

5.1 Setup

To simulate real world scenarios, we use two different sets of head related impulse responses (HRIRs) measured with behind-the-ear (BTE) hearing aids mounted behind each pinna of a head-and-torso-simulator (HATS). The first set of HRIRs was measured in an anechoic chamber for 35 positions uniformly spaced on a semicircle in the front-horizontal plane with radius 1 m centered at the HATS, i.e. $\theta \in \{-85°, -80°, ..., 85°\}$. The second set was measured in a reverberant room shown in Fig. B.3. These HRIRs were measured for 35 positions: five DoAs $\theta \in \{-90°, -45°, 0°, 45°, 90°\}$ versus seven dis-



(a) Babble-speech noise.

(b) Bottling-factory-hall noise.

Fig. B.4: Performance as a function of θ at SNR = 0 dB in an anechoic room.

tances $d \in \{0.5\,\mathrm{m}, 1\,\mathrm{m}, 1.5\,\mathrm{m}, ..., 3.5\,\mathrm{m}\}$. To simulate a signal from a position, the signal is convolved with its related HRIR.

As target signal, we consider a four-minute signal composed of two male and two female speech signals [18]. We consider two different noise-types: speech-babble and bottling-factory-hall noise. Speech-babble is synthesized by playing back different speech signals from each θ simultaneously. The TSP database [18], which consists of different male and female voices, is used as noise sources. The wide-band SNR in each simulation experiments is expressed relative to the left-ear microphone signals. The other simulation parameters are as follows: the sampling frequency is 20 kHz, $N = 2048$, $A = 1024$, $w(n)$ is a Hamming window, the length of $w(n)$ and the DFT order are the same, and $\Theta = \{-90^\circ, -85^\circ, \cdots, 90^\circ\}$. We use the MAE (Eq. (B.20)) as performance metric.

5.2 Results and discussion

Fig. B.4 shows the MAE of various “informed” DoA estimators as a function of θ at an SNR of 0 dB for two different noise-types in an anechoic room. Clearly, the proposed spherical-head-model-based estimator performs better than existing “informed” DoA estimators, and appears robust against the noise types. In contrast, the performance of the “informed” GCC-PHAT-based estimator, introduced in [9], is quite dependent on the noise types.

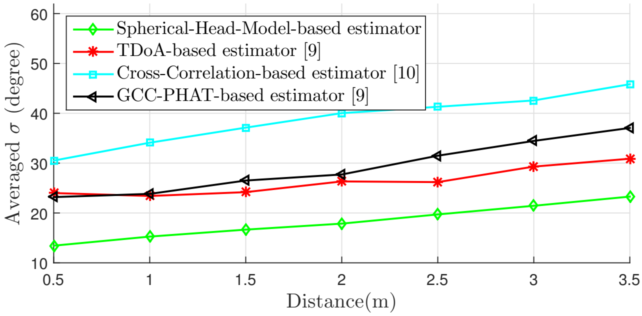


Fig. B.5: Performance as a function of distance in a reverberant room shown in Fig. B.3 at SNR = 0 dB, and σ is averaged over all θs.

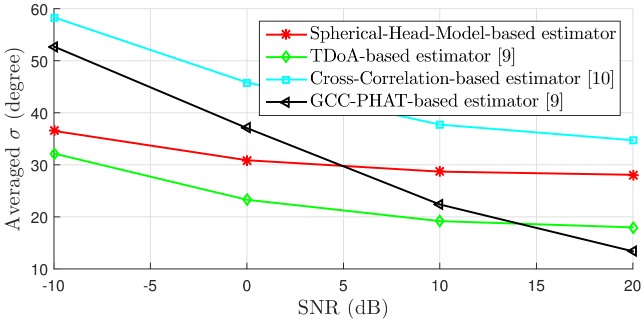


Fig. B.6: Performance as a function of SNR in a reverberant room shown in Fig. B.3. $d = 3.5$m, and σ is averaged over all θs.

As mentioned before, the TDoA-based estimator [9] relies on a free-field assumption, which is more valid for $\theta \approx 0^\circ$ and less valid for $\theta \approx \pm 90^\circ$. The influence of the free-field assumption is clearly visible in the results of the TDoA-based estimator. On the other hand, because the proposed estimator simulates the presence of the head, it improves the performance of the DoA estimation compared with the TDoA-based estimator for $\theta \in [-90, -50]$ or $\theta \in [60, 90]$.

Fig. B.5 shows the MAE of the estimators averaged across the noise types and θs as a function of target distance in a reverberant room (Fig. B.3). In general, increasing the distance will decrease the direct-to-reverberant energy ratio [19], i.e. reverberation will degrade the received signals more at larger distances. However, the proposed estimator still shows consistent improvement.

Fig. B.6 shows the MAE of the estimators averaged across the noise types and θs as a function of SNR in a reverberant situation. As expected, the higher the SNR, the better the performance. The excellent performance of the GCC-PHAT-based estimator at high SNRs may be explained by the fact that the PHAT algorithm is almost ML optimal in low-noise reverberant environments [20]. While the proposed method already performs decently in this situation, we expect that a signal model which directly takes the reverberation into account, e.g. [4, 21], would improve the performance further.

6 Conclusion

In this paper, we proposed a DoA estimator for a hearing aid system which has access to the noise-free target signal via a wireless microphone. We employed a spherical-head model and proposed a maximum likelihood approach to estimate the DoA. We showed that the considered signal model allowed the likelihood function to be calculated efficiently via inverse-discrete-Fourier-transform techniques. In simulation experiments, we studied the effects of the true DoA, noise type, SNR, reverberation level, and target distance on the performance of the proposed algorithm. The proposed method improves the estimation performance over recently proposed "informed" DoA estimators, especially, when the target is at the sides of the head, where the influence of a head model is largest.

References

[1] J. A. MacDonald, "A localization algorithm based on head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4290–4296, 2008.

[2] C. Vina, S. Argentieri, and M. Rébillat, "A spherical cross-channel algorithm for binaural sound localization," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 2921–2926.

[3] F. keyrouz, "Advanced binaural sound localization in 3-D for humanoid robots," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 9, pp. 2098–2107, Sept 2014.

[4] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[5] J. Kotus, K. Lopatka, and A. Czyzewski, "Detection and localization of selected acoustic events in acoustic field for smart surveillance applications," *Multimedia Tools and Applications*, vol. 68, no. 1, pp. 5–21, 2014.

[6] A. Hassani, A. Bertrand, and M. Moonen, "Cooperative integrated noise reduction and node-specific direction-of-arrival estimation in a fully

connected wireless acoustic sensor network," *Signal Processing*, vol. 107, pp. 68–81, Feb. 2015.

[7] S. Goetze, T. Rohdenburg, V. Hohmann, B. Kollmeier, and K.-D. Kammeyer, "Direction of arrival estimation based on the dual delay line approach for binaural hearing aid microphone arrays," in *International Symposium on Intelligent Signal Processing and Communication Systems*, Nov 2007, pp. 84–87.

[8] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Maximum likelihood approach to "informed" sound source localization for hearing aid applications," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2015, pp. 16–20.

[9] ——, "Informed TDoA-based direction of arrival estimation for hearing aid applications," in *Proceedings of IEEE Global Conference on Signal and Information Processing*, 2015, pp. 953–957.

[10] G. Courtois, P. Marmaroli, M. Lindberg, Y. Oesch, and W. Balande, "Implementation of a binaural localization algorithm in hearing aids: specifications and achievable solutions," in *Audio Engineering Society Convention 136*, April 2014, paper 9034.

[11] G. Courtois, P. Marmaroli, H. Lissek, Y. Oesch, and W. Balande, "Binaural hearing aids with wireless microphone systems including speaker localization and spatialization," in *Audio Engineering Society Convention 138*, May 2015, paper 9242.

[12] W. C. Wu, C. H. Hsieh, H. C. Huang, and O. C. Chen, "Hearing aid system with 3D sound localization," in *IEEE TENCON*, 2007, pp. 1–4.

[13] A. Pourmohammad and S. M. Ahadi, "Real time high accuracy 3-D PHAT-based sound source localization using a simple 4-microphone arrangement," *IEEE Systems Journal*, vol. 6, no. 3, pp. 455–468, Sept 2012.

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

[15] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68–77, Jan 2010.

[16] V. R. Algazi, C. Avendano, and R. O. Duda, "Estimation of a spherical-head model from anthropometry," *Journal of Audio Engineering Society*, vol. 49, no. 6, pp. 472–479, 2001.

[17] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *The Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048–3058, 1998.

[18] P. Kabal, "TSP speech database," Department of Electrical and Computer Engineering, McGill University, Tech. Rep., 2002.

[19] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio using d/r spatial correlation matrix model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2374–2384, Nov 2011.

[20] C. Zhang, D. Florêncio, and Z. Zhang, "Why does phat work well in lownoise, reverberative environments?" in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008, pp. 2565–2568.

[21] 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 *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2015, pp. 91–95.

Paper C

Maximum likelihood approach to “informed” sound source localization for hearing aid applications

Mojtaba Farmani, Michael Syskind Pedersen, Zheng-Hua Tan, and Jesper Jensen

The paper has been published in the
Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing , pp. 16–20, 2015.

Abstract

Most state-of-the-art sound source localization (SSL) algorithms have been proposed for applications which are "uninformed" about the target sound content; however, utilizing a wireless microphone worn by a target talker, enables recent Hearing Aid Systems (HASs) to access to an almost noise-free sound signal of the target talker at the HAS via the wireless connection. Therefore, in this paper, we propose a maximum likelihood (ML) approach, which we call MLSSL, to estimate the direction of arrival (DoA) of the target signal given access to the target signal content. Compared with other "informed" SSL algorithms which use binaural microphones for localization, MLSSL performs better using signals of one or more microphones placed on just one ear, thereby reducing the wireless transmission overhead of binaural hearing aids. More specifically, when the target location confined to the front-horizontal plane, MLSSL shows an average absolute DoA estimation error of 5 degrees at SNR of -5 *dB in a large-crowd noise and non-reverberant situation. Moreover, MLSSL suffers less from front-back confusions compared with the recent approaches.*

1 Introduction

Sound source localization (SSL) has been investigated in many applications, such as robotics [1–3], video conferencing [4], and hearing aids [5]. In a sense, SSL is a primitive task which would improve performance of higher level tasks. For example, in a hearing aid system (HAS), knowing the location of the target sound may improve noise reduction algorithms [6, 7], leading to better speech enhancement performance.

In general, different acoustic localization strategies using microphone arrays have been investigated [8, ch. 8]:

- Steered-beamformer-based location estimators: these methods steer the beam to the potential sound source locations and search for a maximum in output power (termed focalization) [9].

- High-resolution-spectral-estimation-based location Estimators: these methods exploit the spatiospectral correlation matrix obtained from the microphones signals. Under certain assumptions, the sound source locations can be derived from a lower-dimensional vector subspace embedded within the signal space spanned by the columns of the correlation matrix [8, ch. 8].

- Time-difference-of-arrival (TDoA)-Based Location Estimators: these methods use a set of TDoA estimations of the signals reaching each pair of microphones to estimate the sound source location [8, ch. 8] [10].

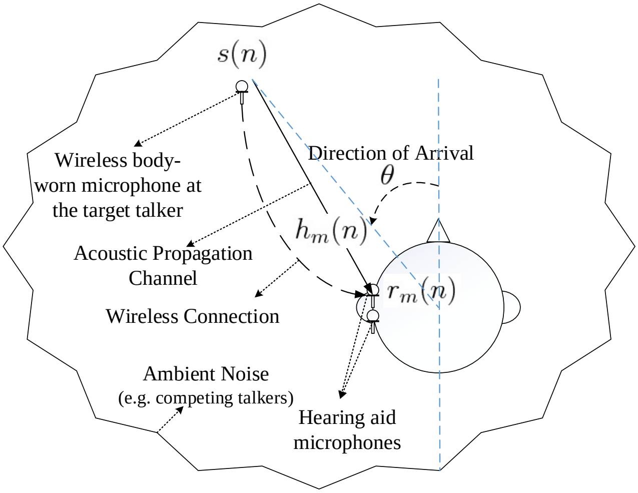

$s(n)$
Wireless body-worn microphone at the target talker
Direction of Arrival
θ
$h_m(n)$
Acoustic Propagation Channel
$r_m(n)$
Wireless Connection
Ambient Noise (e.g. competing talkers)
Hearing aid microphones

Fig. C.1: SSL scenario for a hearing aid system using a wireless microphone: $r_m(n)$, $s(n)$ and $h_m(n)$ are the noisy received sound, the noise-free target sound and the corresponding HRIR for microphone m, respectively. $s(n)$ is available at the hearing aid via wireless connection to the wireless microphone at the target talker. Estimating the direction of arrival θ is the goal in this scenario.

When the microphone array is located next to the ears, like in HASs or humanoid robots, bio-inspired binaural cues, such as interaural time difference (ITD), interaural intensity difference (IID) and monaural cues represented by head related transfer functions (HRTFs) [called head related impulse responses (HRIRs) in the time domain] are often used for SSL [11]. Roughly, humans are thought to use ITDs for low frequency components, up to approximately 1500 Hz, and IIDs for higher frequency components [12]. For monaural spatial hearing, humans are believed to utilize the spectral filtering of the incoming sound at the head, torso and pinnae [11], i.e. filtering of the incoming sound through HRTFs.

Most current SSL algorithms have been proposed for applications which are "uninformed" about the target source signal content [1, 3, 4], i.e. they do not have any access to the noise-free target signal content. However, recent advances in wireless technology enables new HASs, where the target talker is wearing a wireless microphone, to have access to an essentially noise-free version of the target signal [5]. This turns the "uninformed" SSL problem into the "informed" SSL problem considered in this paper.

Fig. C.1 depicts the system considered in this paper. The target signal $s(n)$ is transmitted through the acoustic channel $h_m(n)$ and reaches the m^{th} microphone of the HAS. Due to additive environmental noise, a noisy signal $r_m(n)$ is received at the m^{th} microphone. Moreover, the noise-free target signal $s(n)$ is also transmitted to the HAS via the wireless connection. We aim at estimating the target signal direction of arrival (DoA) θ based on these signals.

In HASs, since microphones are located at the ears, the acoustic shadowing effects of the user's head and torso cause $h_m(m)$ to depend on θ [11]. However, for simplicity, many SSL algorithms, e.g. [5, 10], assume a free field situation and disregard the user's head and torso acoustic shadowing effect, causing the location estimation performance to be reduced. In this paper, we propose a method which does take the head presence into account to distinguish directions, thereby improving localization performance. The proposed method is a maximum likelihood approach; therefore, we call it maximum likelihood sound source localization (MLSSL).

2 Signal Model

Fig. C.1 shows the situation at hand: the noisy received sound signal $r_m(n)$ at microphone m is a result of the convolution of the target signal $s(n)$ with the acoustic channel impulse response $h_m(n)$ from the target talker to microphone m, and is contaminated by additive noise $v_m(n)$. For each microphone of the HAS, we can write:

$$\begin{aligned} r_m(n) &= d_m(n) + v_m(n), \qquad m = 1, \cdots, M\,, && \text{(C.1)} \\ d_m(n) &= s(n) * h_m(n), && \text{(C.2)} \end{aligned}$$

where $M \geq 1$ is the number of available microphones, n is the discrete time index, and $*$ is the convolution operator.

Most state-of-the-art HASs operate in the short time Fourier transform (STFT) domain because it allows frequency dependent processing, computational efficiency and low latency algorithm implementation. Therefore, let

$$\begin{aligned} S(l,k) &= \sum_n s(n) w(n - lA) \mathrm{e}^{-\frac{j2\pi k}{N}(n-lA)}, && \text{(C.3)} \\ D_m(l,k) &= \sum_n \sum_t h_m(t) s(n-t) \times \\ &\qquad w(n-lA)\mathrm{e}^{-\frac{j2\pi k}{N}(n-lA)} \\ &= \sum_n s(n) \sum_t h_m(t) \times \\ &\qquad w(n+t-lA)\mathrm{e}^{-\frac{j2\pi k}{N}(n+t-lA)} && \text{(C.4)} \end{aligned}$$

denote the STFT representations of $s(n)$ and $d_m(n)$, respectively, where l and k are frame and frequency bin indices, respectively, N is the frame length, A is the decimation factor, $w(n)$ is the windowing function, and $j = \sqrt{-1}$ is the imaginary unit. Moreover, let

$$H_m(k) = \sum_t h_m(t) \mathrm{e}^{-\frac{j2\pi kt}{N}} \tag{C.5}$$

denote the discrete Fourier transform of $h_m(n)$, where N is greater or equal to the duration of $h_m(n)$. Eq. (C.4) implies that $D_m(l,k) \neq S(l,k)H_m(k)$. However, if the support of $w(n)$ is smoothly long enough compared with the duration of $h_m(n)$, then $w(n-t)h_m(t) \approx w(n)h_m(t)$ [13]; in this case, we find:

$$\begin{aligned} D_m(l,k) &\approx \sum_n s(n)w(n-lA)\mathrm{e}^{-\frac{j2\pi k}{N}(n-lA)} \times \\ &\quad \sum_t h_m(t)\mathrm{e}^{-\frac{j2\pi k}{N}(t)} \qquad \text{(C.6)} \\ &= S(l,k)H_m(k), \qquad \text{(C.7)} \end{aligned}$$

i.e. $D_m(l,k)$ can be approximated as a point-wise multiplication of $S(l,k)$ and $H_m(l,k)$ [13]. With this approximation, Eq. (C.1) can be approximated in the STFT domain as:

$$R_m(l,k) = S(l,k)H_m(k) + V_m(l,k), \tag{C.8}$$

where $R_m(l,k)$ and $V_m(l,k)$ are STFT coefficients of the received signal and noise signal for the m^{th} microphone, respectively, and are defined analogously to $S(l,k)$ in Eq. (C.3).

Collecting the M microphone equations (Eq. (C.8)) in a column vector gives rise to the following signal model:

$$\boldsymbol{R}(l,k) = S(l,k)\boldsymbol{H}(k) + \boldsymbol{V}(l,k), \tag{C.9}$$

where

$$\begin{aligned} \boldsymbol{R}(l,k) &= [R_1(l,k), R_2(l,k), \cdots, R_M(l,k)]^{\mathrm{T}}, \qquad \text{(C.10)} \\ \boldsymbol{H}(k) &= [H_1(k), H_2(k), \cdots, H_M(k)]^{\mathrm{T}}, \qquad \text{(C.11)} \\ \boldsymbol{V}(l,k) &= [V_1(l,k), V_2(l,k), \cdots, V_M(l,k)]^{\mathrm{T}}. \qquad \text{(C.12)} \end{aligned}$$

3 Maximum Likelihood Estimation of DoA

The acoustic shadowing effects of the head and torso cause $\boldsymbol{H}(k)$ to depend on θ [11]; therefore, if we possess a prestored database $\mathcal{H} = \{\boldsymbol{H}_1, \boldsymbol{H}_2, \cdots, \boldsymbol{H}_I\}$, which consists of I sets of HRTFs labelled by their corresponding θ, the target θ may be estimated by finding the best candidate in $\mathcal{H}$. In fact, $\mathcal{H}$ is a discrete model of the continuous space of HRTFs. To find the best $\boldsymbol{H}_i$ in $\mathcal{H}$ based on the received signals, we introduce a maximum likelihood strategy.

Let us assume that $\boldsymbol{V}(l,k)$ in Eq. (C.9) is a zero-mean, circularly-symmetric complex Gaussian random vector, i.e. $\boldsymbol{V}(l,k) \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_V(l,k))$, where $\mathbf{C}_V(l,k)$ is the inter-microphone noise covariance matrix. Since we assume the target signal is picked up without any noise by the wireless microphone, $S(l,k)$ is available at the HAS, and we consider it as deterministic and known. $\boldsymbol{H}(k)$ is also considered deterministic but unknown ($\boldsymbol{H} \in \mathcal{H}$). Hence, from Eq. (C.9) follows:

$$\boldsymbol{R}(l,k) \sim \mathcal{N}(S(l,k)\boldsymbol{H}(k), \mathbf{C}_V(l,k)). \tag{C.13}$$

Since $S(l,k)$ is available at the HAS, we can relatively easily determine the time-frequency regions in the noisy microphones signals where the target speech is essentially absent; therefore, we adaptively estimate $\mathbf{C}_V(l,k)$ using exponential smoothing over the frames where the noise is dominant. Furthermore, for mathematical convenience, we assume that the noisy observations are independent over time and frequency. Therefore, the likelihood function of each $\boldsymbol{H}_i \in \mathcal{H}$ regarding the received signals at frame l is defined as:

$$\begin{aligned} &f_l(\boldsymbol{R}, S; \boldsymbol{H}_i) = \\ &\prod_{j=l-\mathrm{D}+1}^{l} \prod_{k=1}^{\mathrm{K}} \frac{1}{\pi^M |\mathbf{C}_V(j,k)|} \mathrm{e}^{\{-\mathbf{Z}_i^{\mathrm{H}}(j,k)\mathbf{C}_V^{-1}(j,k)\mathbf{Z}_i(j,k)\}}, \end{aligned} \tag{C.14}$$

where $\boldsymbol{Z}_i(j,k) = \boldsymbol{R}(j,k) - S(j,k)\boldsymbol{H}_i(k)$, and $|.|$ and $^{\mathrm{H}}$ denotes the matrix determinant and Hermitian transpose operator, respectively. D is the number of frames and K is the number of frequency indices used to compute the likelihood. It should be noted that we assume that the target source location is fixed across D frames. The corresponding log-likelihood function is given by:

$$\begin{aligned} \mathcal{L}_l(\boldsymbol{H}_i) \quad = \quad & -M\mathrm{DK}\log\pi - \sum_{j=l-\mathrm{D}+1}^{l} \sum_{k=1}^{\mathrm{K}} \log|\mathbf{C}_V(j,k)| - \\ & \sum_{j=l-\mathrm{D}+1}^{l} \sum_{k=1}^{\mathrm{K}} \mathbf{Z}_i^{\mathrm{H}}(j,k)\mathbf{C}_V^{-1}(j,k)\mathbf{Z}_i(j,k), \end{aligned} \tag{C.15}$$

leading to the maximum likelihood estimation of the HRTF:

$$\boldsymbol{H}_{\mathrm{ML}} = \arg\max_{\boldsymbol{H}_i \in \mathcal{H}} \mathcal{L}_l(\boldsymbol{H}_i), \tag{C.16}$$

from which the corresponding DoA estimate $\hat{\theta}$ follows. We solve Eq. (C.16) via an exhaustive search in $\mathcal{H}$.

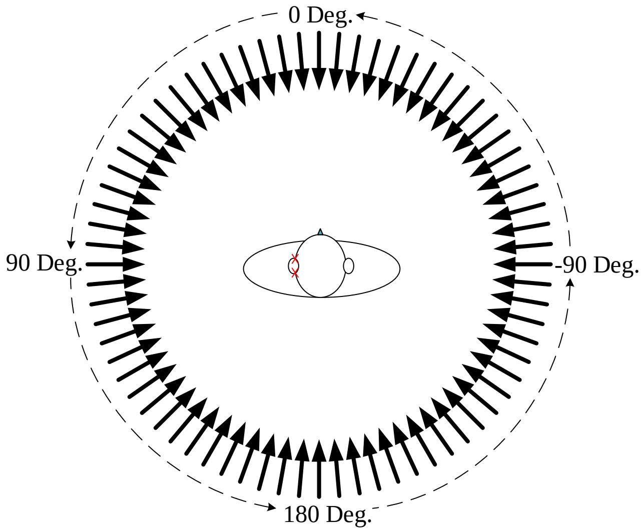


Fig. C.2: Experiment setup. In an anechoic chamber, 72 loudspeakers, represented by arrows, are placed on a circle with radius 1.5 m in the horizontal plane centered at the HATS. Microphones locations are represented by × behind the left ear of the HATS (i.e. around 90°).

4 Simulations Results

4.1 Experiment setup

Fig. C.2 shows the situation considered for assessing the algorithm. The target source is assumed to be placed at one of 72 uniformly spaced possible positions, i.e. with a 5 degrees resolution, on a circle in the horizontal plane with radius 1.5 m centered at a head-and-torso-simulator (HATS). Behind the left pinna of the HATS a two-microphone behind-the-ear (BTE) hearing aid is placed. The distance between front and rear microphones is 12 mm, and the sampling frequency of the microphone signals is 20 kHz. The other simulation parameters are as follows: $N = 2048$ samples, $A = 1024$ samples, and D = 2. $\mathcal{H}$ consists of $I = 72$ sets of HRTFs, measured from each loudspeaker to microphones, and the target speech signal is a 10-second sample of the ISTS signal [14] composed of 21 female voices in 6 different languages. To approximate a practical large-crowd noise field, we play back different speech signals from each of the $I = 72$ target positions simultaneously. The database provided by [15], which consists of different male and female voices, is used as noise sound sources.

When the power of the noise sources is fixed, then the signal-to-noise-ratio (SNR) observed at each of the microphones is a function of θ since the target signal is filtered by the head and torso of the HAS user. Specifically, the SNR is generally reduced when the microphone is in the "shadow" part of the head compared with the case where the microphone is in the "sunny" part. Moreover, for the same θ, "sunny" part microphones have higher SNRs than "shadow" part microphones; therefore, the reference SNRs of the simulations are expressed relative to the left-front microphone and $\theta = 0°$.

As performance metrics, we define the percentage of the DoA correct detection and the DoA estimation mean absolute error (MAE) in the following. Let Q_θ denote the number of frames for which $\hat{\theta} = \theta$. The percentage of the DoA correct detections is:

$$P_\theta = \frac{Q_\theta}{\mathrm{L}} \times 100, \tag{C.17}$$

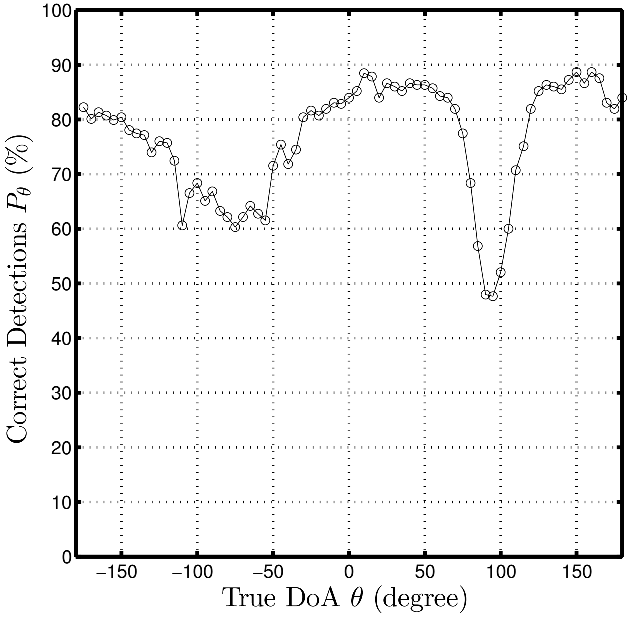
where L is the total number of frames of the received signals. Moreover, the mean absolute error (MAE) of the DoA estimation is given by:

$$\sigma_{\hat{\theta}} = \frac{1}{\mathrm{L}} \sum_{j=1}^{\mathrm{L}} |\theta - \hat{\theta}_j|, \tag{C.18}$$

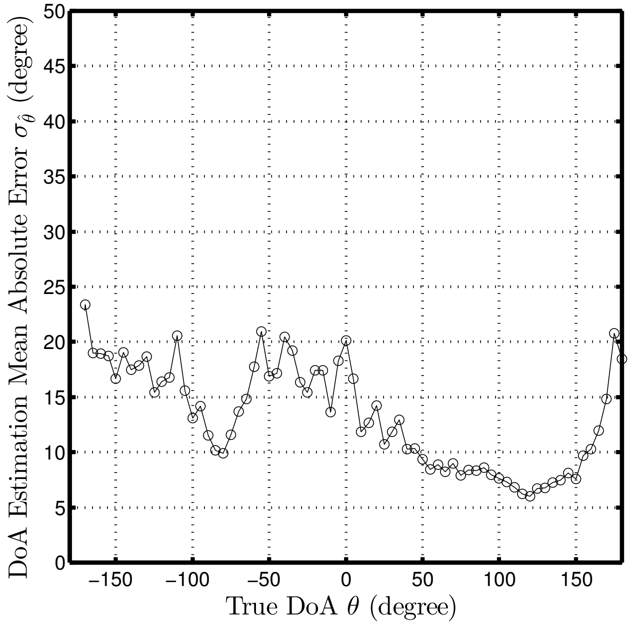
where $\hat{\theta}_j$ is the estimated DoA for the j^{th} frame of the signal.

4.2 MLSSL using one microphone

In contrast to other SSL algorithms which often use two microphones, MLSSL allows us to estimate θ with just one microphone. Figs. C.3a and C.3b show the MLSSL performance in terms of P_θ and $\sigma_{\hat{\theta}}$ at a reference SNR of 0 dB for the full-band signal using $M = 1$ microphone signal (the Left-Front microphone). As can be seen, P_θ drops when the target is located at the sides of the HATS (i.e. $\theta \approx -90°$ and $\theta \approx 90°$), compared with when the target is in front ($\theta \approx 0°$) or behind ($\theta \approx 180°$). On the other hand, $\sigma_{\hat{\theta}}$ in Fig. C.3b shows that even though MLSSL has lower P_θ for θ close to $-90°$ or $90°$, the MAE is less than the cases where θ is close to $0°$ or $180°$. To explain these behaviours, we plot the MLSSL confusion matrix shown in Fig. C.4. Each column of the matrix relates to a θ, and represents the normalized histogram of $\hat{\theta}$s for that particular θ. The almost red diagonal of the matrix shows that MLSSL is generally successful in estimating the θ. However, the two parallel anti-diagonal lines show that when MLSSL fails in detecting the correct θ, then the most probable cause of errors is a front-back confusion. Front-back confusions result in larger estimation errors for the θs in the front or back of the HATS than the left or right sides θs and explain the higher $\sigma_{\hat{\theta}}$ around $\theta = 0°$ or $180°$. As mentioned before, the SNR is a function θ and is almost higher for $\theta \approx 90°$ when the microphones are on the left ear, but since influences of the head and torso are small for $\theta \approx 90°$, their HRTFs are locally very similar and cause local errors and relatively low P_θ. Finally, as can be seen in Fig. C.3b, $\sigma_{\hat{\theta}}$ is generally higher when $\theta \in [-180°, 0°]$ since the microphone is in the head shadow region, and the SNR is lower.



(a) Correct detection percentage.



(b) Mean absolute error.

Fig. C.3: MLSSL simulation results for the left-front microphone at 0 dB SNR.

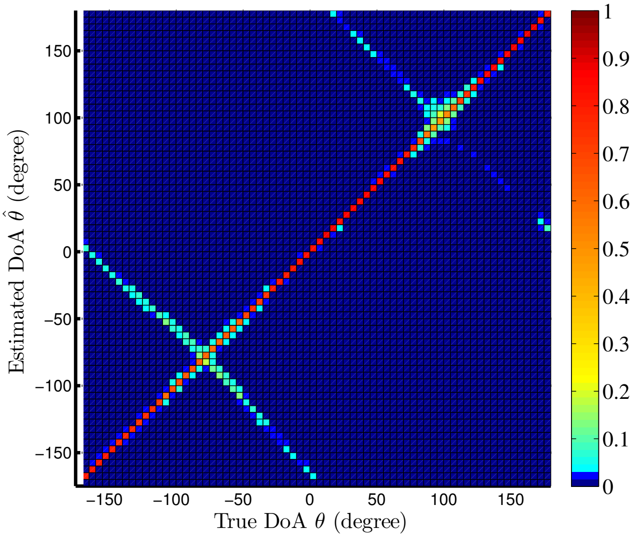


Fig. C.4: Confusion matrix of MLSSL for the left-front microphone at 0 dB SNR.

4.3 Comparison with the state-of-the-art

Courtois et al. in [5] recently introduced the informed SSL problem and proposed a solution based on ITD and binaural signals. They use the wirelessly received noise-free target signal as a time reference to estimate the ITD, and then to estimate θ, they resort to a "sine law" [5]. This causes their method to be unable to differentiate between front and back angles, e.g. $\theta = 45^\circ$ and $\theta = 135^\circ$. Although MLSSL does not have this limitation, for comparison, we consider the frontal horizontal plane only.

Fig. C.5 shows $\sigma_{\hat{\theta}}$ for MLSSL using one or two microphones placed behind the left ear, compared with the Courtois et al. method [5]. As can be seen, MLSSL performs significantly better for all θs. The Courtois et al. results are symmetric with respect to $\theta = 0$ since they use binaural signals, but MLSSL results are asymmetric because microphones are located at one ear only, and the head shadow influences signals which are coming from left and right differently. Furthermore, comparing Figs. C.5 and C.3b shows that knowing a priori that θ is in the frontal plane, improves the MLSSL $\sigma_{\hat{\theta}}$ significantly by eliminating front-back confusions.

In practice, since θ is a continuous variable, it may be represented exactly by none of the HRTFs in $\mathcal{H}$. To assess MLSSL performance in this situation, we made a reduced database $\mathcal{H}'$ by eliminating every other HRTF from $\mathcal{H}$, i.e. there is no HRTF in $\mathcal{H}'$ for half of the considered θs. Fig. C.6 shows MAE of the methods averaged over all the frontal θs as a function of SNR.

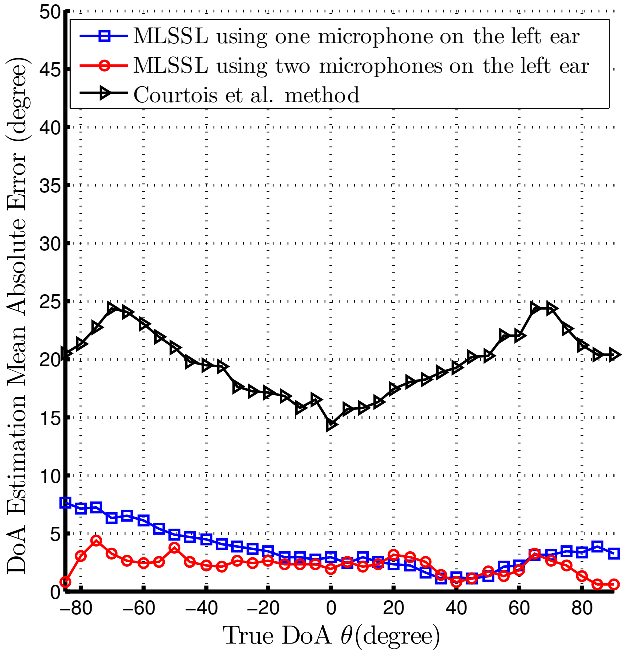


Fig. C.5: Performance comparison of MLSSL with the Courtois et al. method for the frontal plane DoAs at the reference SNR of 0 dB.

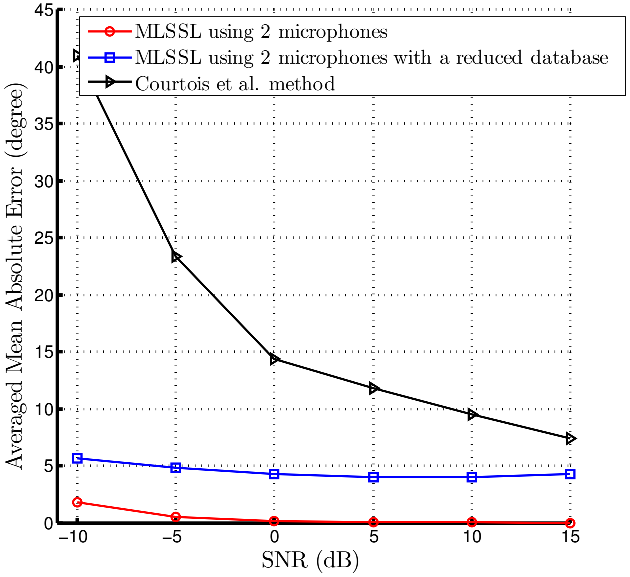


Fig. C.6: Mean absolute error averaged over all the DoAs in the front-horizontal plane as a function of SNR.

As expected, MLSSL has the best performance when θ is represented in the database. But when θ is not in $\mathcal{H}'$, MLSSL mostly finds the nearest DoA in the database, which means that the resolution of the database is a key factor that influences DoA estimation performance using MLSSL.

5 Conclusion and future work

In this paper, we formulated a target sound DoA estimation problem for a new infrastructure of hearing aid systems, which employs a wireless microphone worn by a sound source of interest. To solve the problem, we considered a maximum likelihood strategy which exploits the noise-free target sound and pre-stored HRTFs. In simulations, MLSSL showed better performance than a recent binaural method proposed by Courtois et al. in [5] even when MLSSL uses only a single microphone. The proposed framework is flexible and easily scalable to any number of microphones. Considering an intelligent search instead of an exhaustive search in the HRTFs database would decrease the computation overhead, and moreover, considering elevation and range in addition to the azimuth will generalize the method. Furthermore, robustness to reverberation is an important issue for SSL. These topics will be investigated in future work.

References

[1] J. A. MacDonald, “A localization algorithm based on head-related transfer functions,” *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4290–4296, 2008.

[2] C. Vina, S. Argentieri, and M. Rébillat, “A spherical cross-channel algorithm for binaural sound localization,” in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 2921–2926.

[3] F. keyrouz, “Advanced binaural sound localization in 3-D for humanoid robots,” *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 9, pp. 2098–2107, Sept 2014.

[4] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, “Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings,” *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[5] G. Courtois, P. Marmaroli, M. Lindberg, Y. Oesch, and W. Balande, “Implementation of a binaural localization algorithm in hearing aids: specifications and achievable solutions,” in *Audio Engineering Society Convention 136*, April 2014, paper 9034.

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

[7] O. Thiergart and E. A. P. Habets, "An informed LCMV filter based on multiple instantaneous direction-of-arrival estimates," in *IEEE International Conference on Acoustics Speech and Signal Processing*, 2013, pp. 659–663.

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

[9] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction(src) on a large-aperture microphone array," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, April 2007, pp. I–121–I–124.

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

[11] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space," in *Audio Engineering Society Convention 107*, 1999, paper 5026.

[12] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, 1997.

[13] Y. Avargel, "Linear system identification in the short-time Fourier transform domain," Ph.D. dissertation, Israel Institute of Technology, 2008.

[14] European Hearing Industry Manufactures, "International Speech Test Signal," http://www.ehima.com.

[15] P. Kabal, "TSP speech database," Department of Electrical and Computer Engineering, McGill University, Tech. Rep., 2002.

Paper D

On the influence of microphone array geometry on HRTF-based sound source localization

Mojtaba Farmani, Michael Syskind Pedersen, Zheng-Hua Tan, and Jesper Jensen

The paper has been published in the
Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing , pp. 439–443, 2015.

The layout has been revised.

Abstract

The direction dependence of head related transfer functions (HRTFs) forms the basis for HRTF-based sound source localization (SSL) algorithms. In this paper, we show how spectral similarities of the HRTFs of different directions in the horizontal plane influence performance of HRTF-based SSL algorithms; the more similar the HRTFs of different angles to the HRTF of the target angle, the worse the performance. However, we also show how the microphone array geometry can assist in differentiating between the HRTFs of the different angles, thereby improving performance of HRTF-based SSL algorithms. Furthermore, to demonstrate the analysis results, we show the impact of HRTFs similarities and microphone array geometry on an exemplary HRTF-based SSL algorithm, called MLSSL. This algorithm is well-suited for this purpose as it allows to estimate the direction of arrival (DoA) of the target sound using any number of microphones and any geometries of the microphone array around the head.

1 Introduction

Sound source localization (SSL) using a microphone array has been studied in different applications, such as robotics [1–3], video conferencing [4], and hearing aids [5].

Bio-inspired spatial cues, like interaural time difference (ITD), interaural intensity difference (IID) and the monaural spectral cues in head related transfer functions (HRTFs) [called head related impulse responses (HRIRs) in the time domain] are often used for SSL when the microphone array is located next to the ears[1], such as in hearing aid systems (HASs).

Acoustic shadowing effects of the head and torso of a HAS user or a humanoid robot cause the HRTFs to depend on the target sound direction of arrival (DoA) θ [6]. HRTF-based SSL algorithms use this fact and often exploit a dictionary of HRTFs, labelled by their corresponding θ, to estimate the target sound DoA by finding the best HRTF match in the dictionary [1, 3].

The SSL scenario, which is considered in this paper, is shown in Fig. D.1. Because of recent advances in wireless technology for HASs, the depicted scenario is of practical interest. The target signal $s(n)$ is transmitted through the acoustic channel $h_m(n)$ and is "polluted" by environmental noise to generate the noisy signal $r_m(n)$ at microphone m of the HAS. Moreover, we assume that the noise-free target signal $s(n)$ is also available at the HAS via a wireless connection. We aim at estimating θ in this scenario.

[1] While formally, an HRTF is defined to be "a specific individual's left or right ear far-field frequency response, as measured from a specific point in the free field to a specific point in the ear canal" [6], in this paper we use the term HRTF to describe the frequency response from a target source to a microphone of a hearing aid system.

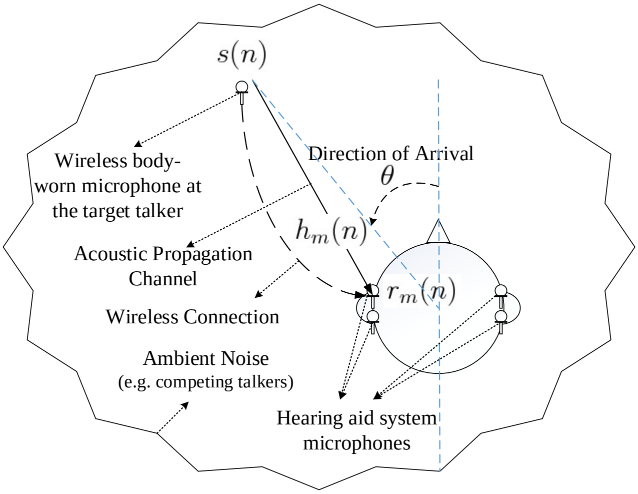

$s(n)$
Wireless body-worn microphone at the target talker
Direction of Arrival
θ
Acoustic Propagation Channel
$h_m(n)$
$r_m(n)$
Wireless Connection
Ambient Noise (e.g. competing talkers)
Hearing aid system microphones

Fig. D.1: SSL scenario for a HAS using a wireless microphone: $r_m(n)$, $s(n)$ and $h_m(n)$ are the noisy received sound, the clean target signal and the correspondent HRIR for microphone m, respectively. $s(n)$ is available at the HAS via wireless connection to the wireless microphone. We aim at estimating θ in this scenario.

In general, spectral similarities of the HRTFs and microphone array geometries affect HRTF-based SSL performance. The spectral similarities of the HRTFs in the dictionary may complicate finding the best candidate in the dictionary and reduce the SSL performance. However, the microphone array geometry may assist to improve the SSL performance. Different microphone array geometries impose different amounts of computation and wireless transmission overhead; for example, generally, two different microphone array configurations are conceivable for a HAS: a) a binaural configuration which allows usage of microphones from wirelessly connected hearing aids, but impose wireless transmission overhead, and b) a monaural configuration, which is restricted to use microphones of one hearing aid only, but which does not impose any transmission overhead. The goal of this paper is to compare different microphone array configurations in terms of performance for HRTF-based SSL. Specifically, we wish to study to which extent the need for wireless data transmission in binaural configuration is justified in terms of performance improvements over a monaural configuration.

To demonstrate our investigation results about HRTFs spectral similarities and microphone array geometry, we consider an exemplary SSL algorithm, called maximum likelihood sound source localization (MLSSL) [7], that uses the noisy microphone signals, the noise-free target signal and a maximum likelihood (ML) strategy to find the best HRTF match in the dictionary to estimate θ. MLSSL is well-suited for the purpose of this paper since it is scalable to any number of microphones and any array geometry around the head.

2 Signal Model and MLSSL

In this section, we briefly review the MLSSL algorithm [7]. Regarding Fig. D.1, for microphone m of the HAS, we can write:

$$r_m(n) = s(n) * h_m(n) + v_m(n), \qquad m = 1, \cdots, M, \tag{D.1}$$

where $r_m(n)$, $s(n)$, $h_m(n)$ and $v_m(n)$ are the noisy microphone signal, the noise-free target signal, the HRIR between the target source and microphone m, and the noise signal, respectively. $M \geq 1$ is the number of available HAS microphones, n is the discrete time index, and $*$ represents the convolution operator. It can be shown that Eq. (D.1) can be approximated in the short-time Fourier-transform (STFT) domain as [7, 8]:

$$R_m(l,k) = S(l,k)H_m(k) + V_m(l,k), \tag{D.2}$$

where $R_m(l,k)$, $S(l,k)$ and $V_m(l,k)$ are STFT coefficients of the noisy microphone signal, target signal and noise signal for the m^{th} microphone, respectively. $H_m(k)$ is the corresponding HRTF, and l and k are frame and frequency bin indices, respectively.

Collecting expressions for the received microphone signals in a column vector leads to:

$$\boldsymbol{R}(l,k) = S(l,k)\boldsymbol{H}(k) + \boldsymbol{V}(l,k), \tag{D.3}$$

where

$$\begin{aligned} \boldsymbol{R}(l,k) &= [R_1(l,k), R_2(l,k), \cdots, R_M(l,k)]^{\mathrm{T}}, & \text{(D.4)} \\ \boldsymbol{H}(k) &= [H_1(k), H_2(k), \cdots, H_M(k)]^{\mathrm{T}}, & \text{(D.5)} \\ \boldsymbol{V}(l,k) &= [V_1(l,k), V_2(l,k), \cdots, V_M(l,k)]^{\mathrm{T}}. & \text{(D.6)} \end{aligned}$$

Assume we possess a dictionary $\mathcal{H} = \{\boldsymbol{H}_1, \boldsymbol{H}_2, \cdots, \boldsymbol{H}_I\}$ of I sets of HRTFs labelled by their corresponding θs, then MLSSL aims at finding the $\boldsymbol{H}_i$ in $\mathcal{H}$ that fits best the observed signals, and in this way estimate the target θ.

Let us assume that $\boldsymbol{V}(l,k)$ in Eq. (D.3) is a zero-mean, circularly-symmetric complex Gaussian random vector, i.e. $\boldsymbol{V}(l,k) \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_V(l,k))$, where $\mathbf{C}_V(l,k)$ is the inter-microphone noise covariance matrix. Since we assume the noise-free $S(l,k)$ is available at the HAS, it is considered as known and deterministic. $\boldsymbol{H}(k)$ is also considered as deterministic but unknown ($\boldsymbol{H} \in \mathcal{H}$). Therefore, $\boldsymbol{R}(l,k)$ in Eq. (D.3) obeys a Gaussian distribution according to:

$$\boldsymbol{R}(l,k) \sim \mathcal{N}(S(l,k)\boldsymbol{H}(k), \mathbf{C}_V(l,k)). \tag{D.7}$$

Because $S(l,k)$ is available at the HAS, it is easy to determine the time-frequency regions in the noisy microphones signals where the target speech is essentially absent, and therefore, adaptively estimate $\mathbf{C}_V(l,k)$ over the frames where the noise is dominant. Moreover, for mathematical convenience, the noisy observations are considered to be independent over time and frequencies. Therefore, the likelihood function of $\boldsymbol{H}_i \in \mathcal{H}$ at frame l, regarding the received signals is given by:

$$f_l(\boldsymbol{R}, S; \boldsymbol{H}_i) = \prod_{j=l-\mathrm{D}+1}^{l} \prod_{k=1}^{\mathrm{K}} \frac{1}{\pi^M |\mathbf{C}_V(j,k)|} e^{\{-\boldsymbol{Z}_i^{\mathrm{H}}(j,k)\mathbf{C}_V^{-1}(j,k)\boldsymbol{Z}_i(j,k)\}}, \tag{D.8}$$

where $\boldsymbol{Z}_i(j,k) = \boldsymbol{R}(j,k) - S(j,k)\boldsymbol{H}_i(k)$, and $|.|$ and $^{\mathrm{H}}$ denotes the matrix determinant and Hermitian transpose operator, respectively. D and K are the number of frames and frequency indices, respectively, used for calculating f_l. We assume the target sound source location is fixed during the D frames. The corresponding log-likelihood function is:

$$\begin{aligned} \mathcal{L}_l(\boldsymbol{H}_i) \quad = \quad & -M\mathrm{DK}\log\pi - \sum_{j=l-\mathrm{D}+1}^{l} \sum_{k=1}^{\mathrm{K}} \log|\mathbf{C}_V(j,k)| - \\ & \sum_{j=l-\mathrm{D}+1}^{l} \sum_{k=1}^{\mathrm{K}} \boldsymbol{Z}_i^{\mathrm{H}}(j,k)\mathbf{C}_V^{-1}(j,k)\boldsymbol{Z}_i(j,k), \end{aligned} \tag{D.9}$$

leading to the maximum likelihood estimation of the HRTF:

$$\boldsymbol{H}_{\mathrm{ML}} = \underset{\boldsymbol{H}_i \in \mathcal{H}}{\arg\max}\, \mathcal{L}_l(\boldsymbol{H}_i) \tag{D.10}$$

from which the corresponding DoA estimate $\hat{\theta}$ follows. For implementation of Eq. (D.10), we use an exhaustive search in $\mathcal{H}$.

3 Performance Analysis

3.1 Acoustic setup and experiment configurations

For investigating effects of different factors on SSL algorithm performance, an anechoic chamber environment is considered (Fig. D.2). The target source can be located at one of 72 uniformly spaced positions, i.e. with 5 degrees resolution, on a horizontal circle with radius 1.5 m centered at a head-and-torso simulator (HATS). Behind-the-ear (BTE) hearing aids are mounted behind each ear of the HATS. The microphone signals of each hearing aid can be wirelessly exchanged such that a maximum of four microphones can be used

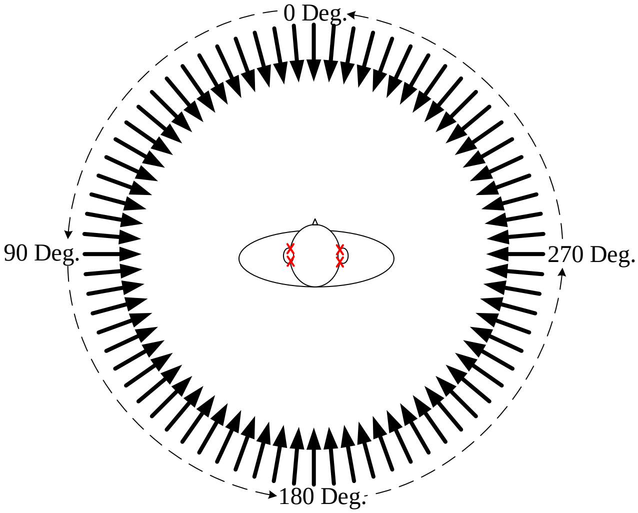


Fig. D.2: Acoustic setup. In an anechoic chamber, 72 loudspeakers are placed on a circle with radius of 1.5 m in a horizontal plane centered at the HATS. Possible microphones locations are represented by × behind the HATS' pinnae.

to perform SSL. We assume this exchange to be instantaneous and error-free. The distance between front and rear microphones in each hearing aid is 12 mm, and the sampling frequency of the microphone signals is 20 kHz. The STFT uses a frame length of 2048 samples, and a decimation factor of 1024 samples. We use a number of D = 2 frames and the dictionary $\mathcal{H}$ consists of $I = 72$ sets of microphones HRTFs, measured from each loudspeaker to the microphones. The target speech signal is a 10-seconds sample of the ISTS V1.0 [9] which is composed of 21 female voices in 6 different languages.

To generate a realistic and difficult situation, we approximate a cylindrically isotropic large-crowd noise field [10], which is simulated by a number of speech sources that are uniformly spaced on the considered circle. The large-crowd speech signals are from the TSP speech database [11] which consists of different male and female voices. The power of the noise sources is constant for all θs. Therefore, the acoustic shadowing of the HATS causes the effective signal-to-noise-ratios (SNRs) observed at each microphone to be a function of target direction θ. For this reason, the simulation SNRs are expressed relative to the left-front microphone and $\theta = 0^\circ$.

To quantify SSL performance, we define the percentage of the DoA correct detection and the DoA estimation mean absolute error (MAE) as following. Let Q_θ denote the number of frames for which $\hat{\theta} = \theta$. The percentage of the DoA correct detections is:

$$P_{_\theta} = \frac{Q_\theta}{\mathrm{L}} \times 100, \tag{D.11}$$

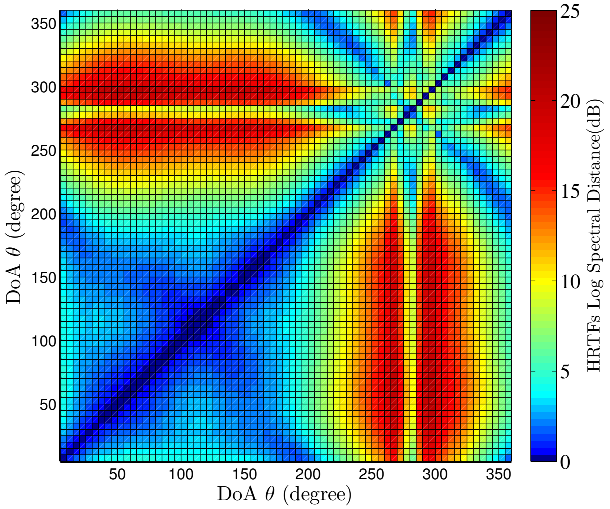


Fig. D.3: Log-spectral distances of the HRTFs of θs for the front microphone of the left hearing aid.

where L is the total frames of the target signal. Furthermore, the mean absolute error (MAE) of the DoA estimation is given by:

$$\sigma_{\hat{\theta}} = \frac{1}{\mathrm{L}} \sum_{j=1}^{\mathrm{L}} |\theta - \hat{\theta}_j|, \tag{D.12}$$

where $\hat{\theta}_j$ is the estimated DoA for the j^{th} frame of the signal.

3.2 HRTF similarities

In this section, we study spectral similarities of HRTFs to be able to identify general challenges that any HRTF-based SSL algorithm faces. Intuitively, we expect that HRTF similarities reduce performance of HRTF-based SSL algorithms. To quantify the similarity between two HRTFs H_i and H_j in $\mathcal{H}$, we use the log-spectral distance (LSD) measure [12]:

$$\mathrm{LSD}(H_i, H_j) = \sqrt{\frac{1}{\mathrm{K}} \sum_{k=1}^{\mathrm{K}} \left(20 \log_{10} \frac{|H_i(k)|}{|H_j(k)|} \right)^2}, \tag{D.13}$$

where $|.|$ denotes the absolute value, and K is the number of frequency bin indices.

Fig. D.3 depicts the LSDs of pairs of HRTFs in $\mathcal{H}$ for the front microphone of the left hearing aid. As can be seen, for θs which are on the left side

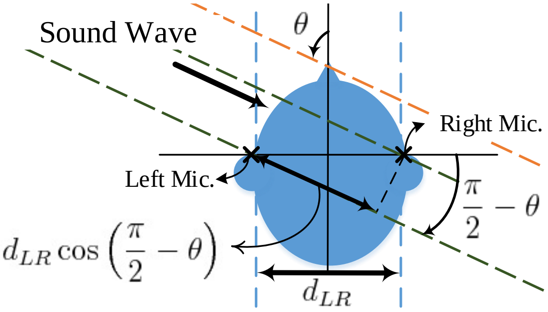


Fig. D.4: Left-Right microphone axis.

of the head ($\theta \in [0°; 180°]$), i.e. the same side of the hearing aid, their corresponding HRTFs are more similar to each other than to the HRTFs of θs of the other side. This fact helps HRTF-based SSL algorithms to decrease right-left confusions. On the other hand, HRTFs corresponding to angles which are almost symmetric relative to the axis between the left and right ears have similar HRTFs, represented by almost two anti-diagonal blue lines in Fig. D.3. These two anti-diagonal blue lines represent the projection of the 3D cone-of-confusion [13] onto the 2D horizontal plane. These similarities cause front-back confusions, which result in larger estimation errors for the θs in the front or back of the HATS than the left or right sides θs.

3.3 Microphone array configurations

To analyze the impact of microphone array geometry, let us first focus on the two-microphone ($M = 2$) situation. In a HAS context, two configurations are of interest: Left-Right axis (Fig. D.4) and Front-Rear axis (Fig. D.5). Left-Right axis is a binaural configuration and uses the front microphones of the left and right hearing aids. On the other hand, Front-Rear axis is a monaural configuration and uses the front and rear microphone of a single hearing aid. Without loss of generality, we assume the Front-Rear microphone axis is placed on the left ear. The Left-Right axis needs wireless communication between the hearing aids while the Front-Rear axis does not.

To explain the influence of different configurations on HRTF-based SSL, we analyze the inter-microphone time differences of arrival (TDoA). Since HRTFs can be treated as a minimum phase FIR filter [6], inter-microphone TDoAs are “encoded” in the HRTFs and implicitly affect HRTF-based SSL. To simplify the analysis, we consider a free field and far field situation (ignoring the head and torso filtering effects and assuming a planar wavefront). Let d_{LR} and d_{FR} denote the distance between left and right microphones (Fig. D.4), and front and rear microphones (Fig. D.5), respectively, and ‘c’ the sound velocity. The inter-microphone TDoAs for the Left-Right and Front-

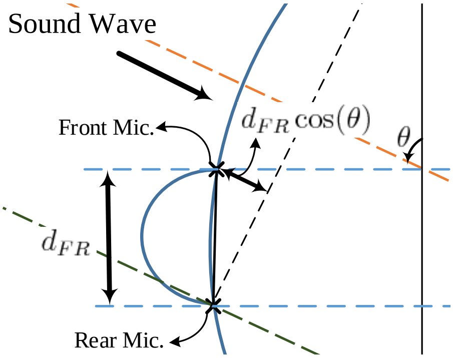


Fig. D.5: Front-Rear microphone axis.

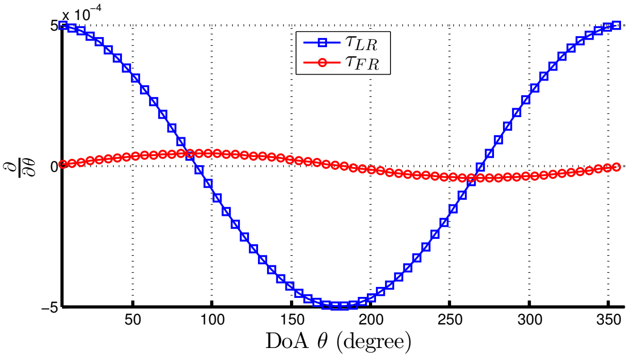


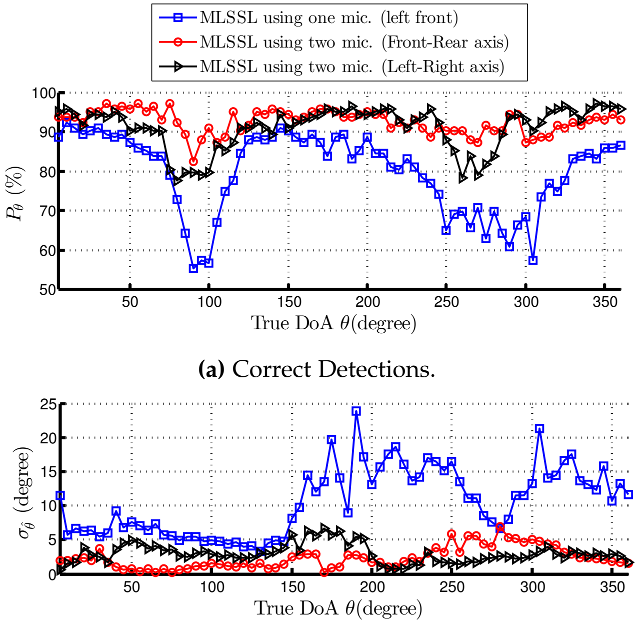
Fig. D.6: Inter-Microphone TDoA derivation for different configurations of two microphones.

Rear microphone axes are given by:

$$\tau_{LR} = \frac{d_{LR}\sin\theta}{c}, \ \tau_{FR} = \frac{d_{FR}\cos\theta}{c}, \text{respectively.}$$

The different microphone axes provide different sensitivities to changes in θ; the higher the change in TDoA with respect to the change in θ, the better SSL performance. To measure the sensitivity of TDoA to θ changes, the derivatives of τ_{LR} and τ_{FR} with respect to θ are shown in Fig. D.6. Clearly, the Front-Rear microphone axis is more sensitive to θ changes when θ is around 90° and 270° while the Left-Right microphone axis is more sensitive to the changes of θ when θ is around 0° and 180°. Moreover, the sensitivity to the DoA changes is a function of the microphone distance; the larger the distance, the higher the sensitivity to θ changes.

Regarding Fig. D.1, increasing the number of microphones to $M = 3$ enables us to take advantage of both the Left-Right axis and the Front-Rear axis at the cost of computation and wireless transmission overhead. Increasing M



(a) Correct Detections.

(b) Mean Absolute Error.

Fig. D.7: The MLSSL performance using one and two microphones with different axes at 0 dB SNR.

further to $M = 4$ will add another Front-Rear axis in the horizontal plane. However, we would not expect this extra Front-Rear axis to provide significant information for SSL in a horizontal plane because the plane is already spanned by the existing microphone axes.

3.4 MLSSL performance

To validate and demonstrate the above analysis, we show the performance of the MLSSL algorithm. Fig. D.7 shows P_θ and $\sigma_{\hat{\theta}}$ using one and two microphones signals in different configurations as a function of θ at 0 dB reference SNR. As can be seen in Fig. D.7a, P_θ generally falls when the target is located at the sides of the HATS (i.e. $\theta \approx 90°$ and $\theta \approx 270°$), compared with when the target is in the front ($\theta \approx 0$) or behind ($\theta \approx 180°$). This is because the HRTFs around 90° and 270° are locally more similar than the HRTFs around 0° and 180° (Sec. 3.2). Moreover, as can be seen in Fig. D.7b, $\sigma_{\hat{\theta}}$ shows different and sometimes opposite behaviour, specifically, for MLSSL using one microphone. This behaviour is because of front-back confusions which cause larger estimation errors for the θs in the front or back of the HATS than the left or right sides θs (Sec. 3.2).

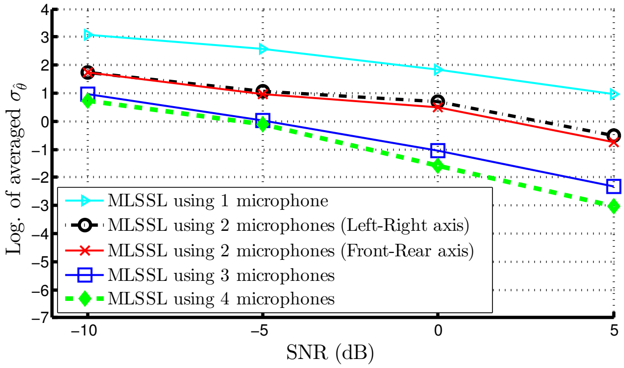


Fig. D.8: The MLSSL performance in terms of logarithm of averaged $\sigma_{\hat{\theta}}$ over considered θs for different number of microphones as a function of reference SNR.

Fig. D.7 shows that increasing the number of microphones generally improves the performance. However, as expected, the configuration of the microphones also affect MLSSL performance. From Fig. D.7a, it is clear that MLSSL ($M = 1$) has lower P_θ for θs around 90° and 270°. For $M = 2$, the Front-Rear configuration is preferred (over the Left-Right) because the Front-Rear axis is more sensitive to changes in θ at these angles (Sec. 3.3, Fig. D.6).

Fig. D.8 shows the MLSSL performance in terms of the logarithm of the averaged $\sigma_{\hat{\theta}}$ over the 72 θs for different number of microphones as a function of reference SNR. The performance difference between $M = 1$ and $M = 2$ of the MLSSL is significant due to the fact that two microphones can form a new microphone axis in the plane. It is clear that for $M = 2$ the Left-Right axis, which requires wireless communication capabilities, does not offer any advantage over the Front-Rear axis. Increasing the number of microphones to $M = 3$ or $M = 4$, improve the performance of the MLSSL at the cost of higher computation and communication overhead. The performance difference between $M = 2$ and $M = 3$ is also relatively significant, since three microphones configuration allows the MLSSL to make use of both Right-Left and Front-Rear axes via the wireless connection. The performance differences between using $M = 3$ and $M = 4$ are relatively small since the planar dimensions are already spanned when $M = 3$.

4 Conclusion

In this paper, we analyzed the performance of HRTF-based SSL algorithms in terms of spectral similarities between HRTFs and microphone array geometry. We showed that due to similarities of different HRTFs, the performance of HRTF-based SSL algorithms depends on the DoA of the target signal. Moreover, we showed that even though increasing the number of microphones in the microphone array improves SSL performance, the geometry of the microphone array plays a key role in improving the performance. For example, a binaural wireless configuration does not necessarily improve the SSL performance compared with a monaural configuration. In this paper, we only considered target locations in the horizontal plane and BTE hearing aids; future research includes considering elevation and range in addition to the azimuth. Furthermore, considering other types of hearing aids than BTE will help complement the investigation.

References

[1] J. A. MacDonald, "A localization algorithm based on head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4290–4296, 2008.

[2] C. Vina, S. Argentieri, and M. Rébillat, "A spherical cross-channel algorithm for binaural sound localization," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 2921–2926.

[3] F. keyrouz, "Advanced binaural sound localization in 3-D for humanoid robots," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 9, pp. 2098–2107, Sept 2014.

[4] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[5] G. Courtois, P. Marmaroli, M. Lindberg, Y. Oesch, and W. Balande, "Implementation of a binaural localization algorithm in hearing aids: specifications and achievable solutions," in *Audio Engineering Society Convention 136*, April 2014, paper 9034.

[6] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space," in *Audio Engineering Society Convention 107*, 1999, paper 5026.

[7] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Maximum likelihood approach to "informed" sound source localization for hearing aid applications," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2015, pp. 16–20.

[8] Y. Avargel, "Linear system identification in the short-time Fourier transform domain," Ph.D. dissertation, Israel Institute of Technology, 2008.

[9] European Hearing Industry Manufactures, “International Speech Test Signal,” http://www.ehima.com.

[10] E. A. P. Habets and S. Gannot, “Generating sensor signals in isotropic noise fields,” *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, Dec. 2007.

[11] P. Kabal, “TSP speech database,” Department of Electrical and Computer Engineering, McGill University, Tech. Rep., 2002.

[12] A. Gray and J. Markel, “Distance measures for speech processing,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.

[13] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, 1997.

Paper E

Informed sound source localization using relative transfer functions for hearing aid applications

Mojtaba Farmani, Michael Syskind Pedersen, Zheng-Hua Tan,
and Jesper Jensen

The paper has been published in the
Journal of IEEE/ACM Transactions on Audio, Speech, and Language Processing,
vol. 25, no. 3, pp. 611–623, 2017.

The layout has been revised.

Abstract

Recent hearing aid systems (HASs) can connect to a wireless microphone worn by the talker of interest. This feature gives the HASs access to a noise-free version of the target signal. In this paper, we address the problem of estimating the target sound direction of arrival (DoA) for a binaural HAS given access to the noise-free content of the target signal. To estimate the DoA, we present a maximum likelihood framework which takes the shadowing effect of the user's head on the received signals into account by modeling the relative transfer functions (RTFs) between the HAS's microphones. We propose three different RTF models which have different degrees of accuracy and individualization. Further, we show that the proposed DoA estimators can be formulated in terms of inverse discrete Fourier transforms (IDFTs) to evaluate the likelihood function computationally efficiently. We extensively assess the performance of the proposed DoA estimators for various DoAs, signal to noise ratios (SNRs), and in different noisy and reverberant situations. The results show that the proposed estimators improve the performance markedly over other recently proposed "informed" DoA estimators.

1 Introduction

In realistic acoustic scenes, where several sound sources are present simultaneously, the auditory scene analysis (ASA) ability in humans allows them to focus deliberately on a sound source while suppressing the other irrelevant sound sources [1]. Sensorineural hearing loss degrades this ability [2], and hearing impaired listeners face difficulties in interacting with the environment. Hearing aid systems (HASs) may take some of these ASA responsibilities to restore the normal interactions of the hearing impaired users with the environment.

Sound source localization (SSL) is one of the main tasks in ASA, and different SSL approaches have been proposed for various applications, such as robotics [3, 4], video conferencing [5], surveillance [6], and hearing aids [7].

SSL strategies using microphone arrays can be generally categorized as[1]:

- Steered-beamformer-based (also called steered response power methods): the main idea of these methods is to steer a beamformer towards potential locations and look for a maximum in the output power [8, ch. 8], [9].
- High-resolution-spectral-estimation-based: these methods are based on the spatiospectral correlation matrix obtained from the microphones signals. Under certain assumptions, the sound source locations can be

[1] This is an extended version of the categorization proposed in [8, ch. 8].

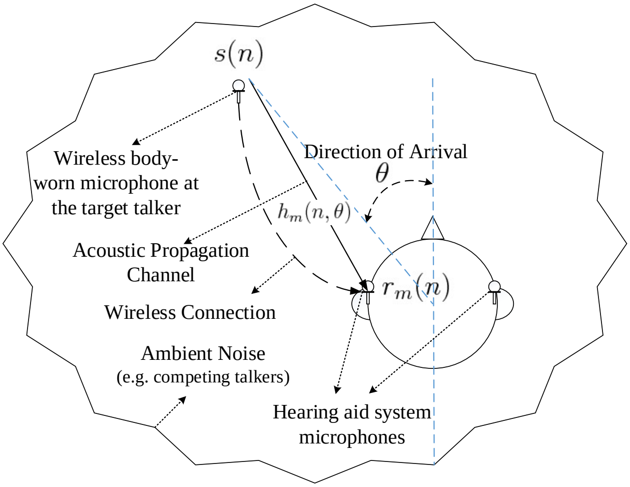


Fig. E.1: An "informed" SSL scenario for a binaural hearing aid system using a wireless microphone. $r_m(n)$ is the noisy received sound at microphone m, $s(n)$ is the noise-free target sound emitted at the target location, and $h_m(n,\theta)$ is the acoustic channel impulse response between the target talker and microphone m. $s(n)$ is available at the HAS via the wireless connection, and the hearing aids are also connected to each other wirelessly. The goal is to estimate θ.

estimated from a lower-dimensional vector subspace embedded within the signal space spanned by the columns of the correlation matrix [10, 11].

- Time-difference-of-arrival (TDoA)-based: these methods first estimate a set of TDoAs of the signals reaching each pair of the microphones in the microphone array, then map the estimated TDoAs to an estimate of the sound source location using a mapping function [12, 13].

- Head-related-transfer-function (HRTF)-based: when the microphone array is mounted at the head and torso of humans or humanoid robots, the filtering effects of the head and torso on the incoming sounds can be used for SSL [4, 14–17].

Most existing SSL algorithms have been proposed for applications which are "uninformed" about the noise-free content of the target sound, e.g. [3–7, 9–16]. However, recent HASs can employ a wireless microphone worn by the target talker to access an essentially noise-free version of the target signal emitted at the target talker's position [17–20]. Using a wireless microphone worn by the target talker introduces the "informed" SSL problem considered in this paper.

Fig. E.1 depicts the situation considered in this paper. The HAS consists of two hearing aids (HAs) connected wirelessly and mounted on each ear of the user, and a wireless microphone worn by the target talker. The target signal $s(n)$ is emitted at the target location, propagates through the acoustic channel $h_m(n, \theta)$, and reaches microphone $m \in \{\text{left}, \text{right}\}$ of the binaural HAS. Due to additive environmental noise, the signal captured by microphone m, denoted by $r_m(n)$, is a noisy version of the target signal impinging on the microphone. The problem considered in this paper is to estimate the target signal direction of arrival (DoA) θ based on the wirelessly available target signal $s(n)$ and the noisy microphone signals $r_m(n)$. Estimating the target sound DoA in this system allows the HAS to enhance the spatial correctness of the acoustic scene presented to the HAS user, e.g. by imposing the corresponding binaural cues on the wirelessly received target sound [21].

The "informed" SSL problem for hearing aid applications was first investigated via a TDoA-based approach in [18]. The method proposed in [18] uses a cross-correlation technique to estimate the TDoA, then uses a sine law to map the estimated TDoA to a DoA estimate. The approach proposed in [18] has relatively low computational load, because it does not take the shadowing effect of the user's head and the ambient noise characteristics into account. Disregarding the head shadowing effect inevitably degrades the DoA estimation performance, especially when the target sound is located at the sides of the user's head, where the head shadowing has the highest impact on the received signals. Moreover, neglecting the ambient noise characteristics causes the estimator performance to be sensitive to the noise type.

In this paper, we present a maximum likelihood (ML) framework for "informed" SSL relying on the noise-free target signal and the ambient noise characteristics. Moreover, to improve the estimation accuracy, we consider the effects of the user's head on the received signals by modeling the direction-dependent relative transfer functions (RTFs) between the left and right microphones of the HAS. More precisely, we present three different RTF models: i) the free-field-far-field model, ii) the spherical-head model, and iii) the measured-RTF model. These models have different degrees of accuracy and individualization. Using the proposed ML framework and based on each of the RTF models, we propose an ML estimator for the target sound DoA. Moreover, besides the DoA, as a by-product, the proposed methods provide an ML estimate of the target signal propagation time between the target talker and the user. The propagation time can be easily converted to a distance estimate, which is an important information about the target location.

The free-field-far-field model and the spherical-head model have been proposed and used for informed DoA estimation in [19] and [20], respectively. In this paper, we introduce the measured-RTF model and its corresponding ML DoA estimator. Moreover, we provide a new unified presentation of all the models and investigate their performances extensively.

The idea of using measured RTFs for "uninformed" DoA estimation was already presented in [22]. The method proposed in [22] considers a narrowband “uniformed” DoA estimation problem and solves it using a minimum mean square error approach. In contrast, our proposed estimator based on the measured-RTF model solves a wide-band “informed” DoA estimation problem using a ML approach. We show that formulating the “informed” DoA estimation problem as wide-band allows us to evaluate the proposed likelihood function in all frequency bins at once using inverse discrete Fourier transforms (IDFTs), which can be computed efficiently.

The general ML framework presented in this paper was first proposed in [17] for the informed SSL, using a database of measured HRTFs. The HRTF database was used to model the acoustic channel and the shadowing effect of a particular user's head. To estimate the DoA, the proposed method in [17], called MLSSL (maximum likelihood sound source localization), looks for the HRTF entry in the database which maximizes the likelihood of the observed microphone signals. MLSSL is markedly effective under severely noisy conditions when the detailed information of the user-specific HRTFs for different directions and different distances is available.

Compared with MLSSL, which is based on HRTFs, the proposed estimators in this paper are based on RTFs. In contrast to HRTFs, which are distance-dependent, RTFs are almost independent of the distance between the target talker and the user, especially in far-field situations [23]. The distance independency decreases the required memory and the computational overhead of the proposed estimators. This is because to estimate the DoA, the proposed estimators must search in a RTF database, which is only a function of DoA, while MLSSL searches in an HRTF database which is a function of both DoA and distance. Further, the proposed estimators in this paper can all be formulated in terms of IDFTs which can be computed efficiently.

The structure of this paper is as follows. In Sections 2 and 3, the signal model and the ML framework are presented, respectively. Afterwards, in Section 4, different RTF models used for modeling the presence of the head are introduced. The proposed DoA estimators using the proposed RTF models and the ML framework are derived in Section V. In Section VI, the performance of the proposed estimators is evaluated and compared using experimental simulations. Lastly, we conclude the paper in Section VII.

2 Signal Model

Regarding Fig. E.1, the noisy signal received at microphone $m \in \{\text{left}, \text{right}\}$ of the HAS is given by:

$$r_m(n) = s(n) * h_m(n, \theta) + v_m(n), \tag{E.1}$$

where $s(n)$, $h_m(n, \theta)$ and $v_m(n)$ are the noise-free target signal emitted at the target talker's position, the acoustic channel impulse response between the target talker and microphone m, and an additive noise component, respectively. Further, n is the discrete time index, and $*$ denotes the convolution operator.

Most state-of-the-art HASs operate in the short time Fourier transform (STFT) domain because it allows frequency dependent processing, computational efficiency and low latency algorithm implementations. Therefore, Let

$$R_m(l,k) = \sum_n r_m(n) w(n - lA) \mathrm{e}^{-\frac{j2\pi k}{N}(n-lA)},$$

denote the STFT of $r_m(n)$, where l and k are frame and frequency bin indexes, respectively, N is the discrete Fourier transform (DFT) order, A is the decimation factor, $w(n)$ is the windowing function, and $j = \sqrt{-1}$ is the imaginary unit. Similarly, let us denote the STFT of $s(n)$ and $v_m(n)$ by $S(l,k)$ and $V_m(l,k)$, respectively, which are defined analogously to $R_m(l,k)$. Moreover, let

$$\begin{aligned} H_m(k,\theta) &= \sum_n h_m(n,\theta) \mathrm{e}^{-\frac{j2\pi kn}{N}} \\ &= \alpha_m(k,\theta) \mathrm{e}^{-\frac{j2\pi k}{N} D_m(k,\theta)}, \end{aligned} \tag{E.2}$$

denote the discrete Fourier transform (DFT) of $h_m(n,\theta)$, where $\alpha_m(k,\theta)$ is a real positive number and denotes the frequency-dependent attenuation factor due to propagation effects, and $D_m(k,\theta)$ is the frequency-dependent propagation time measured in samples, from the target sound source to microphone m. Eq. (E.1) can be approximated in the STFT domain as:

$$R_m(l,k) = S(l,k) H_m(k,\theta) + V_m(l,k). \tag{E.3}$$

This approximation is known as the multiplicative transfer function (MTF) approximation [24], and its accuracy depends on the length and smoothness of the windowing function $w(n)$: the longer and the smoother the analysis window $w(n)$, the more accurate the approximation [24].

3 Maximum Likelihood Framework

To define the likelihood function, let us assume that the additive noise observed at the microphones follows a zero-mean circularly-symmetric complex Gaussian distribution:

$$\boldsymbol{V}(l,k) = \begin{bmatrix} V_{\text{left}}(l,k) \\ V_{\text{right}}(l,k) \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_v(l,k)), \tag{E.4}$$

where $\mathbf{C}_v(l,k)$ is the noise cross power spectral density (CPSD) matrix defined as $\mathbf{C}_v(l,k) = \mathrm{E}\{\boldsymbol{V}(l,k)\boldsymbol{V}^{\mathrm{H}}(l,k)\}$, where $\mathrm{E}\{.\}$ and superscript H represent the expectation and Hermitian transpose operators, respectively. Further, let us assume that the noisy observations are independent across frequencies (strictly speaking, this assumption holds when the correlation time of the signal is short compared with the frame length [25, 26]). Therefore, the likelihood function for frame l is defined by:

$$p(\underline{\underline{\mathbf{R}}}(l);\underline{\underline{\mathbf{H}}}(\theta)) = \prod_{k=0}^{N-1} \frac{1}{\pi^M \det\left[\mathbf{C}_v(l,k)\right]} \mathrm{e}^{\{-(\mathbf{Z}(l,k))^{\mathrm{H}}\mathbf{C}_v^{-1}(l,k)(\mathbf{Z}(l,k))\}}, \tag{E.5}$$

where det[.] denotes the matrix determinant, and

$$\begin{aligned}
\underline{\underline{\mathbf{R}}}(l) &= [\ \boldsymbol{R}(l,0),\ \boldsymbol{R}(l,1),\ \cdots,\ \boldsymbol{R}(l,N-1)\ ],\\
\boldsymbol{R}(l,k) &= [\ R_{\text{left}}(l,k),\ R_{\text{right}}(l,k)\ ]^{\mathrm{T}},\ 0 \le k \le N-1,\\
\underline{\underline{\mathbf{H}}}(\theta) &= [\ \boldsymbol{H}(0,\theta),\ \boldsymbol{H}(1,\theta),\ \cdots,\ \boldsymbol{H}(N-1,\theta)\ ],\\
\boldsymbol{H}(k,\theta) &= [\ H_{\text{left}}(k,\theta),\ H_{\text{right}}(k,\theta)\ ]^{\mathrm{T}}\\
&= \begin{bmatrix} \alpha_{\text{left}}(k,\theta)e^{-j2\pi\frac{k}{N}D_{\text{left}}(k,\theta)} \\ \alpha_{\text{right}}(k,\theta)e^{-j2\pi\frac{k}{N}D_{\text{right}}(k,\theta)} \end{bmatrix},\\
\boldsymbol{Z}(l,k) &= \boldsymbol{R}(l,k) - S(l,k)\boldsymbol{H}(k).
\end{aligned}$$

To reduce the computational overhead, we consider the log-likelihood function and omit the terms independent of θ. Therefore, the reduced log-likelihood function is given by:

$$\mathcal{L}(\underline{\underline{\mathbf{R}}}(l);\underline{\underline{\mathbf{H}}}(\theta)) = \sum_{k=0}^{N-1}\{-(\mathbf{Z}(l,k))^{\mathrm{H}}\mathbf{C}_v^{-1}(l,k)(\mathbf{Z}(l,k))\}. \tag{E.6}$$

The ML estimate of θ is found by maximizing $\mathcal{L}$. However, to maximize $\mathcal{L}$ with respect to θ, we need to model and find the ML estimate of the parameters ($\alpha_{\text{left}}, D_{\text{left}}, \alpha_{\text{right}}$ and D_{right}) in $\underline{\underline{\mathbf{H}}}(\theta)$. Instead of estimating all the parameters separately, in the following, we present three different RTF models, which model and define the relations between the parameters in $\underline{\underline{\mathbf{H}}}(\theta)$ considering the influence of the user's head, and with different degrees of accuracy and individualization. These RTF models allow us to formulate $\mathcal{L}$ depending on the parameters of the transfer function between the target and only one, not both, of the microphones, while it also considers the head presence.

4 Relative Transfer Function (RTF) Models

The RTF between the left and the right microphones represents the filtering effect of the user's head. Moreover, this RTF defines the relation between the acoustic channels' parameters (the attenuations and the delays) corresponding to the left and the right microphones. An RTF is usually defined with respect to a reference microphone. Without loss of generality, let us consider the left microphone as the reference microphone; therefore, considering Eq. (E.2), the RTF at frequency bin k is defined by

$$\begin{aligned}\Psi(k,\theta) &= \frac{H_{\text{right}}(k,\theta)}{H_{\text{left}}(k,\theta)}\\ &= \Gamma(k,\theta)\mathrm{e}^{-j2\pi\frac{k}{N}\Delta D(k,\theta)},\end{aligned}$$

where

$$\begin{aligned}\Gamma(k,\theta) &= \frac{\alpha_{\text{right}}(k,\theta)}{\alpha_{\text{left}}(k,\theta)},\\ \Delta D(k,\theta) &= D_{\text{right}}(k,\theta) - D_{\text{left}}(k,\theta).\end{aligned}$$

We refer to $\Gamma(k,\theta)$ in dB as the inter-microphone level difference (IMLD), and to $\Delta D(k,\theta)$ in discrete time samples as the inter-microphone time difference (IMTD). In the following, three different models are presented for the RTF with different degrees of accuracy.

4.1 The free-field-far-field model

The free-field-far-field model $\boldsymbol{\Psi}_{\text{ff}}(\theta)$ is the simplest and the most straightforward model, which simply ignores the shadowing effect of the user's head and relies on a minimal number of user-related prior assumptions. In a free-field and far-field situation, the delay and the attenuation of an acoustic channel are frequency-independent. Therefore, using basic geometry rules, the IMTD can be formulated as [19]

$$\begin{aligned}\Delta D_{\text{ff}}(\theta) &= D_{\text{right}}(\theta) - D_{\text{left}}(\theta)\\ &= -\frac{a}{c}\sin(\theta), \qquad \text{(E.7)}\end{aligned}$$

where a is the head diameter (or more precisely, the distance between the microphones) and c is the sound speed. It should be noted that $\theta = 0^\circ$ is exactly at the front of the user, and DoAs are defined clockwise with respect to 0°. Moreover, in a free-field and far-field situation, $\alpha_{\text{left}}(\theta) = \alpha_{\text{right}}(\theta)$, i.e.

$$\Gamma_{\text{ff}}(\theta) = \frac{\alpha_{\text{right}}(\theta)}{\alpha_{\text{left}}(\theta)} = 1. \qquad \text{(E.8)}$$

Accordingly, the RTF in a free-field and far-field situation is given by:

$$\boldsymbol{\Psi}_{\text{ff}}(\theta) = [\Psi_{\text{ff}}(0,\theta), \Psi_{\text{ff}}(1,\theta), \cdots, \Psi_{\text{ff}}(N-1,\theta)]^{\text{T}}$$

where

$$\Psi_{\text{ff}}(k,\theta) = \mathrm{e}^{j2\pi\frac{k}{N}(\frac{a}{c}\sin(\theta))}, 0 \leq k \leq N-1.$$

4.2 The spherical-head model

For the spherical-head model $\boldsymbol{\Psi}_{\text{sp}}(\theta)$, we model the user's head as a rigid sphere. Even though the IMTD and the IMLD for a spherical head are generally frequency-dependent, here we assume that the IMTD and the IMLD, or more precisely the delays and the attenuations of the acoustic channels, are frequency-independent. The frequency-independency assumption keeps the model simple and decreases the computational load [20]. Moreover, our preliminary simulation results reveal that a frequency-dependent spherical-head model, which is a more accurate model with more parameters, does not necessarily provide more accurate DoA estimation. This is partly because the frequency-dependent model is over-fitted to the spherical head, while there is a mismatch between the spherical head and an actual head.

For a spherical head, the IMTD can be approximated by the Woodworth model [27, pp. 520–-523]:

$$\Delta D_{\text{sp}}(\theta) = -\frac{a}{2c}\left(\theta + \sin(\theta)\right). \tag{E.9}$$

Moreover, to model the IMLD, we use the following expression inspired by the work in [28]:

$$20\log_{10}\Gamma_{\text{sp}}(\theta) = \gamma\sin(\theta), \tag{E.10}$$

where γ is a frequency-independent scaling factor. In [20], to find the best γ for the DoA estimation, we ran simulation using the theoretical HRTF of the spherical-head model proposed in [23]. The results showed that $\gamma = 6.5$ provides the best DoA estimation performance [20]. Therefore, the RTF for the spherical-head model is given by

$$\boldsymbol{\Psi}_{\text{sp}}(\theta) = \left[\Psi_{\text{sp}}(0,\theta), \Psi_{\text{sp}}(1,\theta), \cdots, \Psi_{\text{sp}}(N-1,\theta)\right]^{\text{T}},$$

where

$$\Psi_{\text{sp}}(k,\theta) = 10^{\frac{6.5\sin(\theta)}{20}}\text{e}^{j2\pi\frac{k}{N}\left(\frac{a}{2c}(\theta+\sin(\theta))\right)}, 0 \leq k \leq N-1.$$

4.3 The measured-RTF model

The measured-RTF model $\boldsymbol{\Psi}_{\text{ms}}(\theta)$ is the most detailed and individualized model. This model uses a database of RTFs for different directions obtained from the corresponding HRTFs measured for the specific user. The measured RTF model is defined as

$$\boldsymbol{\Psi}_{\text{ms}}(\theta) = [\Psi_{\text{ms}}(0,\theta), \Psi_{\text{ms}}(1,\theta), \cdots, \Psi_{\text{ms}}(N-1,\theta)]^{\text{T}},$$

where

$$\Psi_{\text{ms}}(k,\theta) = \Gamma_{\text{ms}}(k,\theta)\text{e}^{j\Phi_{\text{ms}}(k,\theta)}, 0 \leq k \leq N-1,$$

where

$$\Gamma_{\text{ms}}(k,\theta) = \frac{|\tilde{H}_{\text{right}}(k,\theta)|}{|\tilde{H}_{\text{left}}(k,\theta)|}, \tag{E.11}$$

$$\Phi_{\text{ms}}(k,\theta) = \angle \frac{\tilde{H}_{\text{right}}(k,\theta)}{\tilde{H}_{\text{left}}(k,\theta)}, \tag{E.12}$$

where $\tilde{H}_{\text{left}}(k,\theta)$ and $\tilde{H}_{\text{right}}(k,\theta)$ are the measured HRTFs[2] for the left and right microphones, respectively, and $|.|$ and $\angle$ denote the magnitude and the phase angle of a complex number, respectively.

5 Proposed DoA Estimators

In this section, we derive DoA estimators based on each of the proposed RTF models (Section 4) using the ML framework (Section 3). In the derivations, we denote the inverse of the noise CPSD matrix as

$$\mathbf{C}_v^{-1}(l,k) \equiv \begin{bmatrix} C_{11}(l,k) & C_{12}(l,k) \\ C_{21}(l,k) & C_{22}(l,k) \end{bmatrix}. \tag{E.13}$$

To derive the DoA estimators, we expand the reduced log-likelihood function $\mathcal{L}$ presented in Eq. (E.6). Let

$$\begin{aligned} \boldsymbol{\alpha}_{\text{left}}(\theta) &= [\alpha_{\text{left}}(0,\theta), \alpha_{\text{left}}(1,\theta), \cdots, \alpha_{\text{left}}(N-1,\theta)]^{\text{T}}, \\ \boldsymbol{D}_{\text{left}}(\theta) &= [D_{\text{left}}(0,\theta), D_{\text{left}}(1,\theta), \cdots, D_{\text{left}}(N-1,\theta)]^{\text{T}}, \\ \boldsymbol{\alpha}_{\text{right}}(\theta) &= [\alpha_{\text{right}}(0,\theta), \alpha_{\text{right}}(1,\theta), \cdots, \alpha_{\text{right}}(N-1,\theta)]^{\text{T}}, \end{aligned}$$

and

$$\boldsymbol{D}_{\text{right}}(\theta) = [D_{\text{right}}(0,\theta), D_{\text{right}}(1,\theta), \cdots, D_{\text{right}}(N-1,\theta)]^{\text{T}}.$$

[2] Formally, an HRTF is defined as "a specific individuals left or right ear far-field frequency response, as measured from a specific point in the free field to a specific point in the ear canal" [29]. However, in this paper we relax this definition and use the term HRTF to describe the frequency response from a target source to the microphone of a hearing aid system.

The expansion of $\mathcal{L}$ is

$$\begin{aligned}
&\mathcal{L}\left(\underline{\underline{\mathbf{R}}}(l); \boldsymbol{\alpha}_{\text{left}}(\theta), \boldsymbol{D}_{\text{left}}(\theta), \boldsymbol{\alpha}_{\text{right}}(\theta), \boldsymbol{D}_{\text{right}}(\theta)\right) = \\
&\sum_{k=1}^{N} 2\alpha_{\text{left}}(k,\theta) C_{11}(l,k) R_{\text{left}}(l,k) S^*(l,k) \mathrm{e}^{\frac{j2\pi k D_{\text{left}}(k,\theta)}{N}} + \\
&\quad 2\alpha_{\text{left}}(k,\theta) C_{12}(l,k) R_{\text{right}}(l,k) S^*(l,k) \mathrm{e}^{\frac{j2\pi k D_{\text{left}}(k,\theta)}{N}} + \\
&\quad 2\alpha_{\text{right}}(k,\theta) C_{21}(l,k) R_{\text{left}}(l,k) S^*(l,k) \mathrm{e}^{\frac{j2\pi k D_{\text{right}}(k,\theta)}{N}} + \\
&\quad 2\alpha_{\text{right}}(k,\theta) C_{22}(l,k) R_{\text{right}}(l,k) S^*(l,k) \mathrm{e}^{\frac{j2\pi k D_{\text{right}}(k,\theta)}{N}} + \\
&\quad \left(\alpha^2_{\text{left}}(k,\theta) C_{11}(l,k) + \alpha^2_{\text{right}}(k,\theta) C_{22}(l,k)\right) |S(l,k)|^2 + \\
&\quad 2\alpha_{\text{left}}(k,\theta) \alpha_{\text{right}}(k,\theta) C_{21}(l,k) |S(l,k)|^2 \times \\
&\quad \mathrm{e}^{\frac{j2\pi k}{N}\left(D_{\text{right}}(k,\theta) - D_{\text{left}}(k,\theta)\right)}. \qquad \text{(E.14)}
\end{aligned}$$

In the following, we aim to make $\mathcal{L}$ independent of all other parameters except θ, using the proposed RTF models.

5.1 The free-field-far-field model DoA estimator

As mentioned, in a free-field and far-field situation, the delays and the attenuations of acoustic channels are frequency independent. Based on Eqs. (E.7) and (E.8), $D_{\text{right}}(\theta)$ and $\alpha_{\text{right}}(\theta)$ can be written as functions of $D_{\text{left}}(\theta)$ and $\alpha_{\text{left}}(\theta)$, respectively:

$$\begin{aligned}
D_{\text{right}}(\theta) &= \Delta D_{\text{ff}}(\theta) + D_{\text{left}}(\theta) \\
&= -\frac{a}{c}\sin(\theta) + D_{\text{left}}(\theta), \\
\alpha_{\text{right}}(\theta) &= \Gamma_{\text{ff}}(\theta)\alpha_{\text{left}}(\theta) \\
&= \alpha_{\text{left}}(\theta).
\end{aligned}$$

Inserting these relations in Eq. (E.14), we arrive at the reduced log-likelihood function $\mathcal{L}(\underline{\underline{\mathbf{R}}}(l); \boldsymbol{\Psi}_{\text{ff}}(\theta), \alpha_{\text{left}}(\theta), D_{\text{left}}(\theta))$ which is independent of H_{right} parameters (i.e. $D_{\text{right}}(\theta)$ and $\alpha_{\text{right}}(\theta)$). To eliminate the dependency of $\mathcal{L}$ on $\alpha_{\text{left}}(\theta)$, we find the maximum likelihood estimate (MLE) of $\alpha_{\text{left}}(\theta)$ in terms of other parameters, and replace the result into $\mathcal{L}$. To do so, we solve $\frac{\partial \mathcal{L}}{\partial \alpha_{\text{left}}(\theta)} = 0$, which leads to

$$\hat{\alpha}_{\text{left}}(\theta) = \frac{f_{\text{ff}}(\boldsymbol{\Psi}_{\text{ff}}(\theta), D_{\text{left}}(\theta))}{g_{\text{ff}}(\boldsymbol{\Psi}_{\text{ff}}(\theta))}, \qquad \text{(E.15)}$$

where

$$f_{\mathrm{ff}}(\boldsymbol{\Upsilon}_{\mathrm{ff}}(\theta), D_{\mathrm{left}}(\theta)) = \sum_{k=1}^{N} \Big(C_{11}(l,k) R_{\mathrm{left}}(l,k) + \\ C_{12}(l,k) R_{\mathrm{right}}(l,k) + \big(C_{21}(l,k) R_{\mathrm{left}}(l,k) + \\ C_{22}(l,k) R_{\mathrm{right}}(l,k) \big) \Upsilon^{*}_{\mathrm{ff}}(k,\theta) \Big) \times \\ S^{*}(l,k) e^{\frac{j2\pi k D_{\mathrm{left}}(\theta)}{N}}, \tag{E.16}$$

and

$$g_{\mathrm{ff}}(\boldsymbol{\Upsilon}_{\mathrm{ff}}(\theta)) = \sum_{k=1}^{N} \Big(C_{11}(l,k) + 2C_{21}(l,k) \Upsilon^{*}_{\mathrm{ff}}(k,\theta) + \\ C_{22}(l,k) \Big) |S(l,k)|^{2}. \tag{E.17}$$

Inserting $\hat{\alpha}_{\mathrm{left}}$ into $\mathcal{L}$ gives us:

$$\mathcal{L}_{\mathrm{ff}}(\underline{\underline{\mathbf{R}}}(l); \boldsymbol{\Upsilon}_{\mathrm{ff}}(\theta), D_{\mathrm{left}}(\theta)) = \frac{f^{2}_{\mathrm{ff}}(\boldsymbol{\Upsilon}_{\mathrm{ff}}(\theta), D_{\mathrm{left}}(\theta))}{g_{\mathrm{ff}}(\boldsymbol{\Upsilon}_{\mathrm{ff}}(\theta))}. \tag{E.18}$$

From Eq. (E.16) it can be seen that for a given θ, $f_{\mathrm{ff}}(\boldsymbol{\Upsilon}_{\mathrm{ff}}(\theta), D_{\mathrm{left}}(\theta))$ is an IDFT, which can be evaluated efficiently, with respect to $D_{\mathrm{left}}(\theta)$, while $g_{\mathrm{ff}}(\boldsymbol{\Upsilon}_{\mathrm{ff}}(\theta))$ is a simple summation. Therefore, computing $\mathcal{L}_{\mathrm{ff}}$ for a given θ results in a discrete-time sequence corresponding to different values of $D_{\mathrm{left}}(\theta)$. Since θ is unknown, we consider a discrete set Θ of different θs, and compute $\mathcal{L}$ for each $\theta \in \Theta$ using an IDFT. Evaluating $\mathcal{L}$ for all $\theta \in \Theta$ results in a 2-dimensional discrete grid as a function of different values of θ and D_{left}. The MLEs of θ and D_{left} are then found from the global maximum:

$$\left[\hat{\theta}_{\mathrm{ff}}, \hat{D}_{\mathrm{left}}\right] = \underset{\theta \in \Theta, D_{\mathrm{left}}}{\arg\max}\, \mathcal{L}_{\mathrm{ff}}(\underline{\underline{\mathbf{R}}}(l); \boldsymbol{\Upsilon}_{\mathrm{ff}}(\theta), D_{\mathrm{left}}(\theta)). \tag{E.19}$$

5.2 The spherical-head model DoA estimator

The derivation of the DoA estimator based on the spherical-head model is analogous to the free-field-far-field DoA estimator. We assume, as in the free-field-far-field model, that the delay and the attenuation of acoustic channels are frequency-independent, and we replace $D_{\mathrm{right}}(\theta)$ and $\alpha_{\mathrm{right}}(\theta)$ with functions of $D_{\mathrm{left}}(\theta)$ and $\alpha_{\mathrm{left}}(\theta)$, respectively, using Eqs. (E.9) and (E.10):

$$\begin{aligned} D_{\mathrm{right}}(\theta) &= \Delta D_{\mathrm{sp}}(\theta) + D_{\mathrm{left}}(\theta) \\ &= -\frac{a}{2c}\left(\sin(\theta) + \theta\right) + D_{\mathrm{left}}(\theta), \qquad \text{(E.20)} \\ \alpha_{\mathrm{right}}(\theta) &= \Gamma_{\mathrm{sp}}(\theta) \alpha_{\mathrm{left}}(\theta) \\ &= 10^{\frac{6.5 \sin(\theta)}{20}} \alpha_{\mathrm{left}}(\theta). \qquad \text{(E.21)} \end{aligned}$$

Inserting Eqs. (E.20) and (E.21) into Eq. (E.14) makes $\mathcal{L}$ independent of $D_{\text{right}}(\theta)$ and $\alpha_{\text{right}}(\theta)$, i.e. we have $\mathcal{L}(\underline{\underline{\mathbf{R}}}(l); \mathbf{\Upsilon}_{\text{sp}}(\theta), \alpha_{\text{left}}(\theta), D_{\text{left}}(\theta))$. As for the free-field-far-field model, to find the MLE of $\alpha_{\text{left}}(\theta)$ as a function of the other parameters, we solve $\frac{\partial \mathcal{L}}{\partial \alpha_{\text{left}}(\theta)} = 0$. The resulting MLE of $\alpha_{\text{left}}(\theta)$ can be expressed as

$$\hat{\alpha}_{\text{left}}(\theta) = \frac{f_{\text{sp}}(\mathbf{\Upsilon}_{\text{sp}}(\theta), D_{\text{left}}(\theta))}{g_{\text{sp}}(\mathbf{\Upsilon}_{\text{sp}}(\theta))}, \tag{E.22}$$

where

$$\begin{aligned} f_{\text{sp}}(\mathbf{\Upsilon}_{\text{sp}}(\theta), D_{\text{left}}(\theta)) = \sum_{k=1}^{N} \Big(& C_{11}(l,k) R_{\text{left}}(l,k) + \\ & C_{12}(l,k) R_{\text{right}}(l,k) + \big(C_{21}(l,k) R_{\text{left}}(l,k) + \\ & C_{22}(l,k) R_{\text{right}}(l,k)\big) \Upsilon^*_{\text{sp}}(k,\theta) \Big) \times \\ & S^*(l,k) e^{\frac{j2\pi k D_{\text{left}}(\theta)}{N}}, \end{aligned} \tag{E.23}$$

and

$$\begin{aligned} g_{\text{sp}}(\mathbf{\Upsilon}_{\text{sp}}(\theta)) \quad = \quad & \sum_{k=1}^{N} \Big(C_{11}(l,k) + 2C_{21}(l,k) \Upsilon^*_{\text{sp}}(k,\theta) + \\ & \qquad \Gamma^2_{\text{sp}}(\theta) C_{22}(l,k) \Big) |S(l,k)|^2. \end{aligned} \tag{E.24}$$

Inserting Eq. (E.22) into $\mathcal{L}(\underline{\underline{\mathbf{R}}}(l); \mathbf{\Upsilon}_{\text{sp}}(\theta), \alpha_{\text{left}}(\theta), D_{\text{left}}(\theta))$ gives us:

$$\mathcal{L}_{\text{sp}}(\underline{\underline{\mathbf{R}}}(l); \mathbf{\Upsilon}_{\text{sp}}(\theta), D_{\text{left}}(\theta)) = \frac{f^2_{\text{sp}}(\mathbf{\Upsilon}_{\text{sp}}(\theta), D_{\text{left}}(\theta))}{g_{\text{sp}}(\mathbf{\Upsilon}_{\text{sp}}(\theta))}. \tag{E.25}$$

Again, it can be seen that $f_{\text{sp}}(\mathbf{\Upsilon}_{\text{sp}}(\theta), D_{\text{left}}(\theta))$ in Eq. (E.23) is an IDFT with respect to $D_{\text{left}}(\theta)$, and $g_{\text{sp}}(\mathbf{\Upsilon}_{\text{sp}}(\theta))$ is a simple summation for a given θ. As before, for a given θ, evaluating $\mathcal{L}_{\text{sp}}$ results in a discrete-time sequence corresponding to different discrete values of $D_{\text{left}}(\theta)$. Since θ is unknown, we consider a discrete set Θ of different θs, and compute $\mathcal{L}$ for each $\theta \in \Theta$ using an IDFT. The MLEs of θ and D_{left} are then found from the global maximum:

$$\left[\hat{\theta}_{\text{sp}}, \hat{D}_{\text{left}}\right] = \underset{\theta \in \Theta, D_{\text{left}}}{\arg\max}\, \mathcal{L}_{\text{sp}}(\underline{\underline{\mathbf{R}}}(l); \mathbf{\Upsilon}_{\text{sp}}(\theta), D_{\text{left}}(\theta)). \tag{E.26}$$

5.3 The measured-RTF model DoA estimator

In the measured-RTF model, we assume that a database Θ_{ms} of measured frequency-dependent RTFs, labeled by their corresponding directions, for the specific user, is available. The DoA estimator using this model is based on evaluating $\mathcal{L}$ for the different RTFs in Θ_{ms}. The DoA label of the RTF, which gives the highest likelihood is the MLE of the target DoA.

To evaluate $\mathcal{L}$ for each $\mathbf{\Psi}_{\text{ms}}(\theta) \in \Theta_{\text{ms}}$, we assume the parameters of the acoustic transfer function related to the "sunny" microphone is frequency independent. The "sunny" microphone is the microphone which is not in the "shadow" of the head, if we assume the sound is coming from the direction θ. To be more precise, when we evaluate $\mathcal{L}$ for $\mathbf{\Psi}_{\text{ms}}(\theta)$ corresponding to the directions on the left side of the head ($\theta \in [-90°, 0°]$), the acoustic transfer function parameters related to the left microphone, i.e. $\alpha_{\text{left}}(\theta)$ and $D_{\text{left}}(\theta)$, are assumed to be frequency independent. Similarly, when we evaluate $\mathcal{L}$ for $\mathbf{\Psi}_{\text{ms}}(\theta)$ corresponding to the directions on the right side of the head ($\theta \in (0°, +90°]$), the acoustic transfer function parameters related to the right microphone, i.e. $\alpha_{\text{right}}(\theta)$ and $D_{\text{right}}(\theta)$, are assumed to be frequency independent. Note that this evaluation strategy can be carried out in practice; it requires no prior knowledge about the true DoA.

This assumption about the "sunny" microphone is reasonable, because if the sound is really coming from direction θ, the signal received by the "sunny" microphone is almost unaltered by the head and torso of the user, i.e. this resembles a free-field situation. As shown below, this assumption allows us to use an IDFT for evaluation of $\mathcal{L}$. Note that this frequency-independency assumption is only related to the acoustic channel parameters from the target to one of the microphones. The RTFs between microphones are allowed to be frequency-dependent.

To evaluate $\mathcal{L}$ for $\mathbf{\Psi}_{\text{ms}}(\theta)$ where $\theta \in [-90°, 0°]$, let us replace $\alpha_{\text{right}}(k, \theta)$ and $D_{\text{right}}(k, \theta)$ in $\mathcal{L}$ with functions of $D_{\text{left}}(\theta)$ and $\alpha_{\text{left}}(\theta)$, respectively:

$$\begin{aligned}
\alpha_{\text{right}}(k,\theta) &= \Gamma_{\text{ms}}(k,\theta)\alpha_{\text{left}}(\theta), && \text{(E.27)}\\
D_{\text{right}}(k,\theta) &= \Delta D_{\text{ms}}(k,\theta) + D_{\text{left}}(\theta) \\
&= \frac{-N}{2\pi k}\left(\Phi_{\text{ms}}(k,\theta) + 2\pi\rho\right) + D_{\text{left}}(\theta), && \text{(E.28)}
\end{aligned}$$

where ρ is a phase unwrapping factor. This makes $\mathcal{L}$ independent of H_{right} parameters. Afterwards, as before, to make $\mathcal{L}$ independent of $\alpha_{\text{left}}(\theta)$, we find the MLE of $\alpha_{\text{left}}(\theta)$ as functions of other parameters in $\mathcal{L}$ by solving $\frac{\partial \mathcal{L}}{\partial \alpha_{\text{left}}(\theta)} = 0$. The obtained MLE of $\alpha_{\text{left}}(\theta)$ is:

$$\hat{\alpha}_{\text{left}}(\theta) = \frac{f_{\text{ms,left}}(\mathbf{\Psi}_{\text{ms}}(\theta), D_{\text{left}}(\theta))}{g_{\text{ms,left}}(\mathbf{\Psi}_{\text{ms}}(\theta))}, \tag{E.29}$$

where

$$\begin{aligned} f_{\text{ms,left}}(\mathbf{\Psi}_{\text{ms}}(\theta), D_{\text{left}}(\theta)) = \sum_{k=1}^{N} \Big(& C_{11}(l,k)R_{\text{left}}(l,k) + \\ & C_{12}(l,k)R_{\text{right}}(l,k) + \big(C_{21}(l,k)R_{\text{left}}(l,k) + \\ & C_{22}(l,k)R_{\text{right}}(l,k)\big)\Psi^{*}_{\text{ms}}(k,\theta)\Big) \times \\ & S^{*}(l,k)e^{\frac{j2\pi k D_{\text{left}}(\theta)}{N}}, \end{aligned} \tag{E.30}$$

and

$$\begin{aligned} g_{\text{ms,left}}(\mathbf{\Psi}_{\text{ms}}(\theta)) \quad = \quad \sum_{k=1}^{N} \Big(& C_{11}(l,k) + 2C_{21}(l,k)\Psi^{*}_{\text{ms}}(k,\theta) + \\ & \Gamma^{2}_{\text{ms}}(\theta)C_{22}(l,k)\Big)|S(l,k)|^{2}. \end{aligned} \tag{E.31}$$

Substituting $\hat{\alpha}_{\text{left}}(\theta)$ in $\mathcal{L}$ leads to

$$\mathcal{L}_{\text{ms,left}}(\underline{\underline{\mathbf{R}}}(l); \mathbf{\Psi}_{\text{ms}}(\theta), D_{\text{left}}(\theta)) = \frac{f^{2}_{\text{ms,left}}(\mathbf{\Psi}_{\text{ms}}(\theta), D_{\text{left}}(\theta))}{g_{\text{ms,left}}(\mathbf{\Psi}_{\text{ms}}(\theta))}.$$

Analogously, to evaluate $\mathcal{L}$ for $\mathbf{\Psi}_{\text{ms}}(\theta)$ where $\theta \in (0^{\circ}, +90^{\circ}]$, if we replace $\alpha_{\text{left}}(k,\theta)$ and $D_{\text{left}}(k,\theta)$ in $\mathcal{L}$ with functions of $\alpha_{\text{right}}(\theta)$ and $D_{\text{right}}(\theta)$, respectively, and go through the similar process, we end up with

$$\mathcal{L}_{\text{ms,right}}(\underline{\underline{\mathbf{R}}}(l); \mathbf{\Psi}_{\text{ms}}(\theta), D_{\text{right}}(\theta)) = \frac{f^{2}_{\text{ms,right}}\Big(\mathbf{\Psi}_{\text{ms}}(\theta), D_{\text{right}}(\theta)\Big)}{g_{\text{ms,right}}(\mathbf{\Psi}_{\text{ms}}(\theta))},$$

where

$$\begin{aligned} f_{\text{ms,right}}(\mathbf{\Psi}_{\text{ms}}(\theta), D_{\text{right}}(\theta)) = \sum_{k=1}^{N} \Big(& C_{21}(l,k)R_{\text{left}}(l,k) + \\ & C_{22}(l,k)R_{\text{right}}(l,k) + \big(C_{11}(l,k)R_{\text{left}}(l,k) + \\ & C_{12}(l,k)R_{\text{right}}(l,k)\big)(\Psi^{*}_{\text{ms}})^{-1}(k,\theta)\Big) \times \\ & S^{*}(l,k)e^{\frac{j2\pi k D_{\text{right}}(\theta)}{N}}, \end{aligned} \tag{E.32}$$

and

$$\begin{aligned} g_{\text{ms,right}}(\mathbf{\Psi}_{\text{ms}}(\theta)) = \sum_{k=1}^{N} \Big(& C_{22}(l,k) + 2C_{12}(l,k)(\Psi^{*}_{\text{ms}}(k,\theta))^{-1} + \\ & \Gamma^{-2}_{\text{ms}}(\theta)C_{11}(l,k)\Big)|S(l,k)|^{2}. \end{aligned} \tag{E.33}$$

Regarding Eqs. (E.30) and (E.32), $f_{\text{ms,left}}(\mathbf{\Upsilon}_{\text{ms}}(\theta), D_{\text{left}}(\theta))$ and $f_{\text{ms,right}}(\mathbf{\Upsilon}_{\text{ms}}(\theta)$ $, D_{\text{right}}(\theta))$ can be seen to be IDFTs with respect to $D_{\text{left}}(\theta)$ and $D_{\text{right}}(\theta)$, respectively. Therefore, for a given θ, evaluating $\mathcal{L}_{\text{ms,left}}$ or $\mathcal{L}_{\text{ms,right}}$ results in a discrete-time sequence corresponding to different discrete values of $D_{\text{left}}(\theta)$ or $D_{\text{right}}(\theta)$. Therefore, evaluating $\mathcal{L}$ for all $\mathbf{\Upsilon}_{\text{ms}}(\theta) \in \Theta_{\text{ms}}$ results in a 2-dimensional discrete grid. The MLEs of θ and D_{left} or D_{right} are then found from the global maximum:

$$\left[\hat{\theta}_{\text{ms}}, \hat{D}\right] = \underset{\mathbf{\Upsilon}_{\text{ms}}(\theta) \in \Theta_{\text{ms}}, D}{\arg\max}\ \mathcal{L}_{\text{ms}}(\underline{\underline{\mathbf{R}}}(l); \mathbf{\Upsilon}_{\text{ms}}(\theta), D(\theta)), \tag{E.34}$$

where

$$\mathcal{L}_{\text{ms}}(\underline{\underline{\mathbf{R}}}(l); \mathbf{\Upsilon}_{\text{ms}}(\theta), D(\theta)) =$$
$$\begin{cases} \mathcal{L}_{\text{ms,left}}(\underline{\underline{\mathbf{R}}}(l); \mathbf{\Upsilon}_{\text{ms}}(\theta), D_{\text{left}}(\theta)) & , \theta \in [-90^\circ, 0^\circ] \\ \mathcal{L}_{\text{ms,right}}(\underline{\underline{\mathbf{R}}}(l); \mathbf{\Upsilon}_{\text{ms}}(\theta), D_{\text{right}}(\theta)) & , \theta \in (0^\circ, +90^\circ] \end{cases}.$$

6 Simulation Results

In this section, we evaluate the performance of the estimators in simulation experiments. Specifically, we study the effects of the target sound DoA θ, the signal-to-noise ratio (SNR), the frame length, the noise type and the reverberation.

6.1 Implementation

The simulation parameters are generally as follows: the sampling frequency is 16 kHz, the DFT order $N = 512$, $w(n)$ is a Hamming window, the length of the window $w(n)$ is the same as the DFT order N, $A = \frac{N}{2}$, and the microphone distance $a = 16.4\,\text{cm}$. Moreover, to evaluate the likelihood functions, the noise CPSD matrix $\mathbf{C}_v(l,k)$ must be known. In the following, the procedure for estimating $\mathbf{C}_v(l,k)$ is outlined.

Estimating the noise CPSD matrix

to estimate $\mathbf{C}_v(l,k)$ in practice, we use $S(l,k)$, which is available at the HAS, as a voice activity detector. Specifically, access to $S(l,k)$ allows us to determine the time-frequency regions in $\boldsymbol{R}(l,k)$, where the target speech is essentially absent, and to adaptively estimate $\mathbf{C}_v(l,k)$ via recursive averaging [17, 30].

Alg. 1 shows the procedure for estimating $\mathbf{C}_v(l,k)$. If the difference between the maximum energy $S_{max}(k)$ in frequency bin k of the target signal observed so far and the energy of $S(l,k)$ in dB is larger than a certain threshold δ_{th}, we assume the target signal to be absent in frame l and frequency bin

Algorithm 1: Estimation of $\mathbf{C}_v(l,k)$

```
Input  : $\boldsymbol{R}(l,k)$, $S(l,k)$
Output: $\mathbf{C}_v(l,k)$
if $S_{max}(k) - 20\log_{10}|S(l,k)| > \delta_{\text{th}}$ then
    /* Target signal is almost absent                              */
    $\mathbf{C}_v(l,k) = \eta\boldsymbol{R}(l,k) * \boldsymbol{R}(l,k)^{\text{H}} + (1-\eta)\mathbf{C}_v(l-1,k)$;
else
    $\mathbf{C}_v(l,k) = \mathbf{C}_v(l-1,k)$;
end
if $S_{max}(k) < 20\log_{10}|S(l,k)|$ then
    $S_{max}(k) = 20\log_{10}|S(l,k)|$
else
    $S_{max}(k) = S_{max}(k) + 10\log_{10}(\beta)$
end
```

k. Hence, $\boldsymbol{R}(l,k)$ is noise dominated in this time-frequency region. Therefore, the estimate of $\mathbf{C}_v(l,k)$ is updated via exponential smoothing with a smoothing factor $0 < \eta < 1$. On the other hand, if the difference is smaller than the threshold δ_{th}, the target signal is assumed to be present in $\boldsymbol{R}(l,k)$. Therefore, the estimate of $\mathbf{C}_v$ is not updated, i.e. $\mathbf{C}_v(l,k) = \mathbf{C}_v(l-1,k)$. Finally, we update $S_{max}(k)$ if needed, or use a forgetting factor $0 < \beta < 1$ to adapt $S_{max}(k)$ with the possible changes in the target signal over time, e.g. if the target talker has changed, or if the target talker stops speaking. We use $\delta_{\text{th}} = 25\,\text{dB}$, $\eta = 0.9$ and $\beta = 0.95$ in the implementation.

6.2 Acoustic setup

To simulate real world scenarios, we use the database of head related impulse responses (HRIRs) and binaural room impulse responses, provided by [31]. We use a subset of the database for the frontal-horizontal plane $\theta \in \Theta = \{-85°, -80°, \cdots, +85°\}$ measured with behind-the-ear (BTE) hearing aids mounted behind the ears of a head-and-torso simulator (HATS). We consider only the frontal-horizontal plane because in practice, the target talker is usually located at the front of the user. Moreover, because of the head symmetry and the microphone locations, the estimators suffer from front-back confusions, as humans do [32]. Therefore, considering only the frontal plane allows to avoid the influence of the front-back confusions on the estimators performance. To simulate a signal from a particular position, we convolve the signal with the corresponding impulse response.

As a target signal, we consider a four-minute speech signal composed of two male and two female voices from the TSP database [33]. To evaluate the performance of the estimators in different noisy situations, we consider four different noise types: car-interior noise, speech-shaped noise, large-crowd noise, and bottling-factory-hall noise. These noise types cover noise signals with low-frequency content (the car-interior noise), high-frequency content (the bottling-factory-hall noise), stationary noises (the speech-shaped noise) and non-stationary noises (the large-crowd noise). The long-term power spectrum of the target signal emitted at the target position and the noise signals received at the left microphone are depicted in Fig. E.2. To simulate a large-crowd noise field, we play back simultaneously 72 different speech signals from 72 different positions, which are uniformly distributed on a circle in the horizontal plane centered at the HATS. Similarly, for the speech-shaped noise and the bottling-factory-hall noise, we play back different realizations of the considered noise signal from all 72 considered positions simultaneously. The car-interior noise field, however, is a binaural recording measured by BTE hearing aids mounted behind the ears of a HATS placed on the passenger seat of a car driving in a city. The wide-band SNR, to be reported for each simulation experiment, is expressed relative to the left-ear microphone signals.

6.3 Performance metric

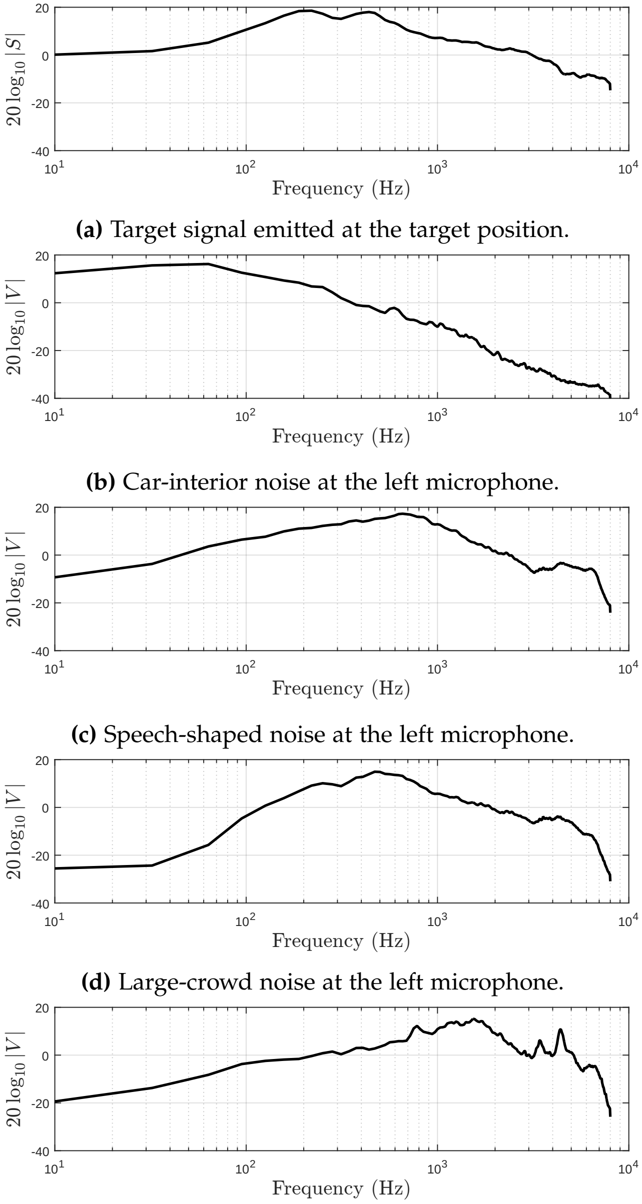
As a performance metric, we use the mean absolute error (MAE) of the DoA estimation, given by:

$$\mathrm{MAE} = \frac{1}{\mathrm{L}} \sum_{j=1}^{\mathrm{L}} |\theta - \hat{\theta}_j|, \tag{E.35}$$

where $\hat{\theta}_j$ is the estimated DoA for the j^{th} frame of the signal, and L is the number of target-active frames (the target-inactive frames are disregarded).

6.4 Competing methods

We compare the proposed estimators with the methods proposed in [18] and [17]. As outlined in Section 1, the method proposed in [18], which we refer to as the cross-correlation-based method, is simple because it does not take the ambient noise characteristics and the head shadowing effect into account. However, to model the curved path between the microphones, the distance between the microphones is assumed to be 25.2 cm, which is larger than the actual microphones distance. This particular distance is used because it leads to the best performance [18]. On the other hand, the method proposed in [17], called MLSSL, is a complex method. It takes the ambient noise characteristics into account by a maximum likelihood approach, and it exploits the details of the head shadowing effect via a database of HRTFs. In the MLSSL implementation, we use the same measured HRTF database, which is used to build the measured-RTF model.



(a) Target signal emitted at the target position.

(b) Car-interior noise at the left microphone.

(c) Speech-shaped noise at the left microphone.

(d) Large-crowd noise at the left microphone.

(e) Bottling-factory-hall noise at the left microphone.

Fig. E.2: Long-term power spectrum of the signals.

6.5 Results and discussions

Influence of the target DoA

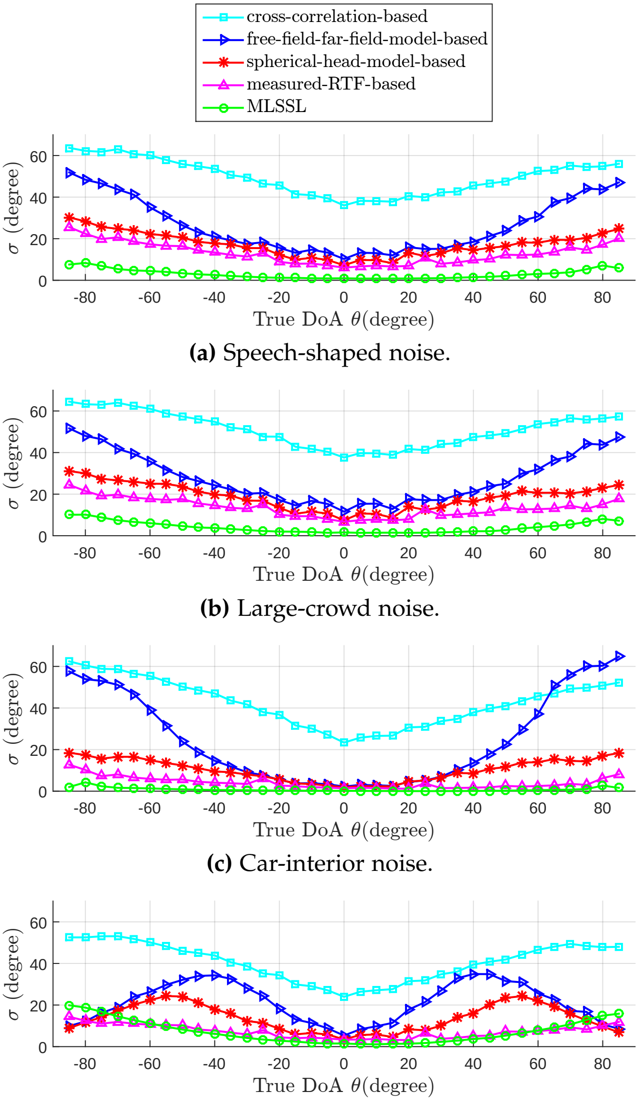
Fig. E.3 compares the performance of the DoA estimators as a function of θ in an anechoic situation at SNR of 0 dB in different noise fields. As can be seen, the performance of all the estimators proposed in this paper are markedly more accurate than the performance of the cross-correlation-based method proposed in [18].

The poor performance of the cross-correlation-based method can be partly explained by the fact that the conventional cross-correlation technique is a maximum-likelihood optimal TDoA estimator for the situation, where the noise is white and Gaussian [34]. However, the frequency characteristics of the considered noise fields, shown in Fig. E.2, are different from a white noise. This difference degrades considerably the performance of the cross-correlation-based method.

Among the estimators proposed in this paper, the estimator based on the free-field-far-field model has the worst performance because it does not consider the shadowing effect of the user's head. In contrast, the spherical-head-model-based estimator models the head shadowing effect and improves the performance of the DoA estimation significantly, especially when the target is located at the sides of the HATS ($\theta \approx \pm 85^\circ$), because this is where the shadowing effect of the head has the highest impact. When the user-specific, measured RTFs are available, even better performance can be achieved, because the influence of the head and torso is modeled more accurately.

Finally, as can be seen in Fig. E.3, the performance of MLSSL is better than the performance of the measured-RTF-based estimator. This is because the exact HRTFs corresponding to the target locations are in the database searched by MLSSL, i.e. a highly idealized situation. Frequency-dependent HRTFs, as used in MLSSL, represent the acoustic transfer functions more accurately than the signal model used in the measured-RTF-based method, where the parameters of the acoustic channel between the target source and the microphone which is not in the head "shadow" are assumed to be frequency independent.

Another point to be made from Fig. E.3 is that, similar to the sound source localization performance of humans [32], the general performance of the estimators when the target is at the sides (i.e. $\theta \approx \pm 90$) is worse than when the target is at the front ($\theta \approx 0^\circ$). This is because the HRTFs (RTFs) corresponding to the front vary stronger within a certain angular range than the HRTFs (RTFs) corresponding to the sides [35]. In other words, when $\theta \in [-90^\circ, -75^\circ]$ or $\theta \in [75^\circ, 90^\circ]$, it is more probable to confuse the true HRTF (RTF) with the nearby HRTFs (RTFs).



(a) Speech-shaped noise.

(b) Large-crowd noise.

(c) Car-interior noise.

(d) Bottling-factory-hall noise.

Fig. E.3: Performance as a function of θ in an anechoic situation at SNR of 0 dB for different noise fields. The distance between the user and the target source is 300 cm. The HRTF database used for generation of the target signal is identical to the HRTF database used by MLSSL and the HRTFs used to build the measured-RTF model.

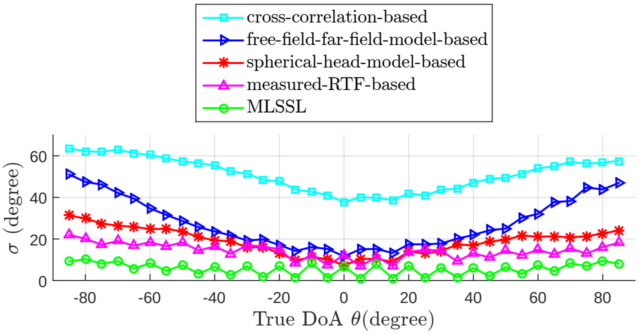


Fig. E.4: Performance as a function of θ in an anechoic situation at SNR 0 dB in the large-crowd noise field. The HRTF database used by MLSSL and the measured-RTF database do not have any entries for every other considered θs for simulation.

Influence of the resolution of the databases

In practice, none of the entries in the HRTF database used by MLSSL or none of the entries in the RTF database used by the measured-RTF-based method can be expected to represent the actual DoA or distance of the target. Here, we investigate the performance of the estimators in these situations.

First, let us consider situations where the exact θ are not represented in the databases. To assess the performance of MLSSL and the measured-RTF-based estimator in these situations, we constructed reduced databases by eliminating every other entry from the MLSSL HRTF database and from the measured-RTF-model database. In other words, there is no entry in the databases for half of the considered target θs. Fig. E.4 shows the performance of the estimators in this case. Comparing Fig. E.4 with Fig. E.3b shows that when the exact θ is not in the databases, the performance of MLSSL and the measured-RTF-based estimator degrade, as expected. However, most often, they succeed in finding the database entry closest to the target θ.

Next, we consider situations where the HRTFs corresponding to the actual distance between the target and the user are not in the database searched by MLSSL or in the HRTF database used to build the measured-RTF model. Fig. E.5 shows the performance in such a situation, where the actual distance between the user and the target is 300 cm, but the employed HRTF database is for the case where the target is 80 cm away from the user (the database contains HRTFs for all the considered directions). It can be seen that the performance of MLSSL degrades dramatically in this situation: MLSSL is extremely sensitive to these HRTF mismatches. However, when the same HRTF database is used to build the measured-RTF model, the performance of the measured-RTF-based method degrades only slightly compared with Fig. E.3. This robustness to the distance mismatches is because the measured

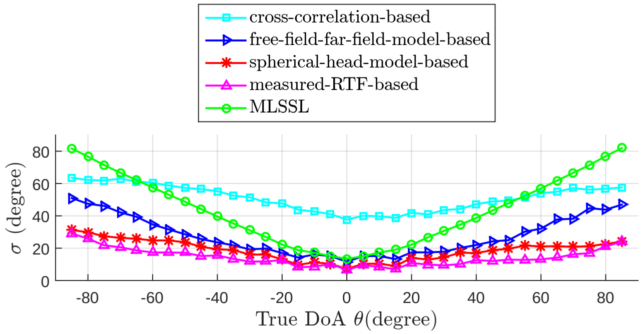


Fig. E.5: Performance as a function of θ in an anechoic situation at SNR 0 dB in the large-crowd noise field. The distance between the user and the target source is 300 cm. The HRTF database used by MLSSL and the HRTF database used to build the measured-RTF model are for the case where the target is 80 cm away from the user.

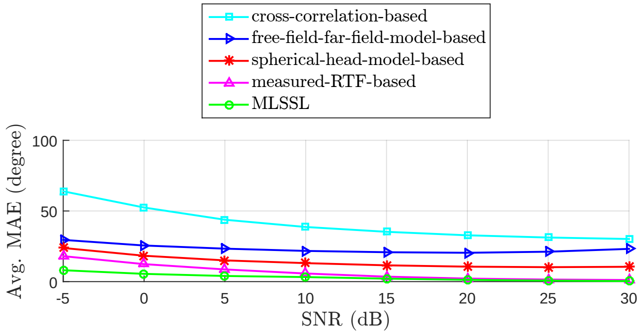


Fig. E.6: Performance as a function of SNR in the same situation as in Fig. E.3. The MAE is averaged over all considered θs.

RTFs are relatively distance independent. Therefore, the database used by the measured-RTF-based method can be just a function of the DoA, leading to a significant reduction of both memory and search complexity over the MLSSL method.

Influence of SNR

The SNR is another factor which generally influences the estimation performance. Fig. E.6 shows the performance for different SNRs in terms of the MAE averaged over all considered θs in an anechoic situation in a large-crowd noise field. As expected, the higher the SNR, the better the performance. Moreover, as can be seen, the general performance order of Fig. E.3 remains at different SNRs; however, the performance of the proposed measured-RTF-based method is almost the same as the performance of the MLSSL at

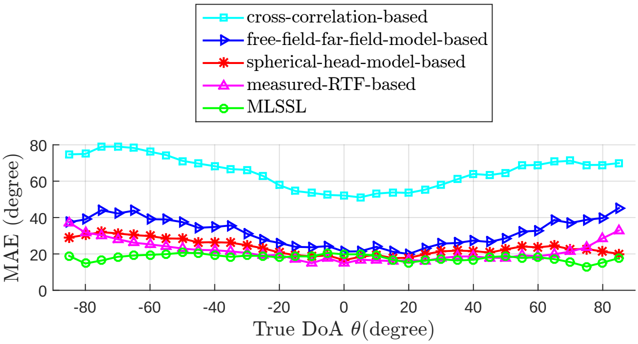


Fig. E.7: Performance as a function of θ in a reverberant office with a reverberation time T_{60} of around 500 ms at SNR of 0 dB. The target is one meter away from the user. The HRTF database used by MLSSL, and the HRTFs used to build the measured-RTF model are "dry" and "clean" HRTFs for the case where the target is 80 cm away.

high SNRs.

Influence of reverberation

Many speech communication situations occur indoor, where reverberation exists. Therefore, it is important to study the impact of reverberation on the performance of the estimators. Fig. E.7 shows the performance of the DoA estimators as a function of θ in a reverberant office ($T_{60} \approx 500$ ms) at SNR of 0 dB in a large-crowd noise field. In contrast to Fig. E.3, performance of all the estimators is reduced because none of them directly considers and models the reverberation. Even though, on average, the general performance order of Fig. E.3 remains, the performance of the spherical-head-model-based method, the measured-RTF method and the MLSSL method approach each other. This is partly because the available "clean" HRTF database used by MLSSL and used to build the measured-RTF model are for the case where the target is 80 cm away while the actual distance of the target is 100 cm in the simulations.

Influence of the window length

Another factor which influences the performance of the estimators is the window (frame) length. Generally, at the cost of higher computational overhead and longer algorithmic delay, longer window lengths must lead to better performance because: 1) greater window lengths provide more observations, which reduces the variance of the estimates in a noisy situation, 2) the MTF approximation (Eq. E.3) depends on the window length: the greater the window length, the better the approximation [24], and 3) greater window lengths

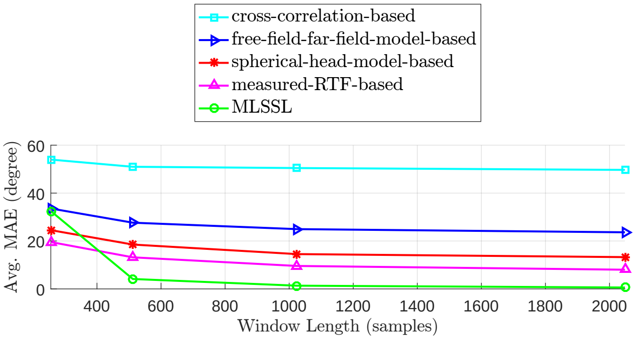


Fig. E.8: Performance as a function of N in the same condition as in Fig. E.3. The MAE is averaged over all considered θs.

strengthen the assumption that DFT coefficients are independent across frequencies (this assumption was used to write the simplified likelihood function in Eq. (E.5)). On the other hand, increasing the window length may violate the assumption implicitly made in Eq. (E.5) that signals are stationarity within a window duration.

Fig. E.8 shows the performance of the DoA estimators as a function of window length. The results are consistent with the expectations: greater window lengths lead to better performance. Interestingly, even though MLSSL has better performance at longer window lengths, its performance is apparently very sensitive to smaller window lengths and deteriorates dramatically compared with the proposed estimators performance.

Influence of non-individualized HRTF databases

MLSSL and the measured-RTF-based method rely on HRTF databases measured for a specific user, and so far, we have presented their performance when user-specific databases are available. In some situations, measuring HRTFs for each user is impractical; however, it is possible to measure the HRTFs for a HATS beforehand. Therefore, in this part, we would like to compare the performance of the estimators in two different cases: 1) individualized: user-specific HRTF databases are available. 2) non-individualized: user-specific HRTF databases are not available; however, the corresponding databases measured for a HATS is available.

For the simulation, we use the HRTFs measured for binaural BTE hearing aids for five different persons (three males and two females) and a HATS. The HRTFs are measured in an anechoic situation for the frontal-horizontal plane.

Fig. E.9 shows the performance of the estimators for the considered cases at an SNR of 0 dB in the large-crowd noise field. As can be seen, MLSSL is

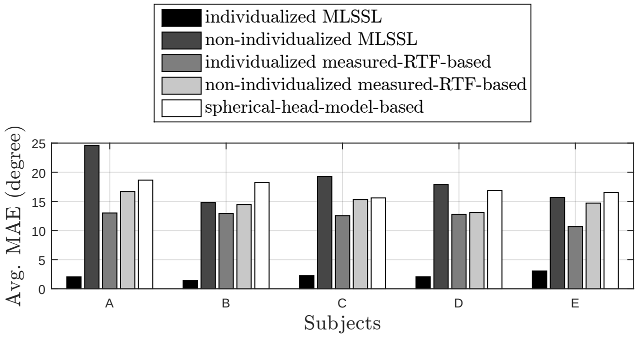


Fig. E.9: Influence of non-individualized HRTF databases on the DoA estimators. The SNR is 0 dB in the large-crowd noise field. The MAE is averaged over all considered θs.

very sensitive to the mismatches in user-specific HRTF database. It has the best performance for all the users (subjects) when the user-specific HRTFs are available (the individualized case), but its performance degrades significantly when the HATS database is used for the DoA estimation (the non-individualized case). On the other hand, the measured-RTF-based method is much less sensitive. Overall, the measured-RTF-based method performs markedly better than MLSSL in the non-individualized case (when only the HATS database is available for the DoA estimation). The performance of the measured-RTF-based method in the non-individualized case is also better than the spherical-head-model-based method, which does not depend on any user-specific databases.

centage of a column being occupied by floats.

Informed estimator vs. uninformed estimator

To demonstrate the benefits of access to the noise-free target signal, here we compare the performance of the proposed "informed" DoA estimators with the performance of a recently developed "uninformed" DoA estimator [22], which we refer to as Braun's method. As mentioned in Section 1, Braun's method is a narrow-band estimator based on the measured-RTF model for the "uninformed" DoA estimation problem, i.e. where the clean target signal is not available. Regarding Eq. (E.3), it has been shown in [22] that the minimum mean square error (MMSE) estimator of the RTF between the two microphones at a particular frequency bin is given by:

$$\hat{\Psi}_{i,j}(k,\theta) = \frac{\phi_{R_{i,j}} - \phi_{V_{i,j}}}{\phi_{R_{j,j}} - \phi_{V_{j,j}}}, \tag{E.36}$$

where i and j are microphone indexes, $\phi_{R_{i,j}} = \mathrm{E}\{R_i(l,k)R_j^*(l,k)\}$ and $\phi_{V_{i,j}} = \mathrm{E}\{V_i(l,k)V_j^*(l,k)\}$. To make the estimate more robust, Braun's method aver-

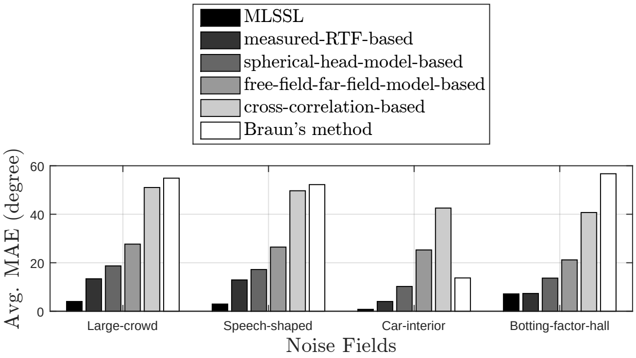


Fig. E.10: Comparison of the "informed" DoA estimators with an "uninformed" DoA estimator proposed in [22], in different noise fields. The simulation was done in the same conditions as in Fig. E.3. The MAE is averaged over all considered θs.

ages the RTF estimate over the microphone index permutations, i.e.

$$\bar{\Psi}_{i,j}(k,\theta) = \frac{1}{2}\left\{\hat{\Psi}_{i,j}(k,\theta) + \hat{\Psi}_{j,i}^{-1}(k,\theta)\right\}. \tag{E.37}$$

Regarding the measured-RTF model Θ_{ms}, Braun's method estimates the DoA θ of the target signal at a particular frequency bin by

$$\hat{\theta}_{\text{Braun}} = \underset{\Psi_{\text{ms}}(k,\theta)\in\Theta_{\text{ms}}}{\arg\min} \sum_{i,j\in\mathcal{M}} W_{i,j}|\bar{\Psi}_{i,j}(k,\theta) - \Psi_{\text{ms}}(k,\theta)|, \tag{E.38}$$

where the set $\mathcal{M}$ contains all microphone pair combinations, and $W_{i,j}$ is a weighting factor for the $\{i,j\}$-th pair. In our setup, because we only have one microphone pair, we drop $W_{i,j}$ and consider $i =$ right and $j =$ left. Moreover, because the target in our problem is at the same position in all frequency bins, we modify the cost function as follows, to integrate the information of all frequency bins:

$$\hat{\theta}_{\text{Braun}} = \underset{\boldsymbol{\Psi}_{\text{ms}}(\theta)\in\Theta_{\text{ms}}}{\arg\min} \sum_{k=0}^{N-1} |\bar{\Psi}_{i,j}(k,\theta) - \Psi_{\text{ms}}(k,\theta)|. \tag{E.39}$$

To implement Braun's method, we used the same measured-RTF model as used by the proposed "informed" measured-RTF-based estimator. Moreover, as proposed in [22], to estimate $\phi_{R_{i,j}}$, a recursive averaging technique with a time constant of 50 ms was used. Finally, to estimate $\phi_{V_{i,j}}$ used in Braun's method, we use the estimation of $\mathbf{C}_v$ outlined in Section 6.1.

Fig. E.10 shows the performance of the proposed "informed" DoA estimators vs. Braun's method. Clearly, the proposed DoA estimators, which have access to the noise-free target signal, perform markedly better than Braun's method, which does not have access to the noise-free signal. Moreover, in large-crowd noise, speech-shaped noise and bottling-factory-hall noise fields, the cross-correlation-based estimator, which is an "informed" estimator with low computational complexity, performs slightly better than Braun's method, which has relatively higher computational overhead. However, the estimation error of Braun's method significantly decreases in the car-interior noise, which is relatively stationary low frequency noise (c.f. Fig. E.2b). At the cost of higher computational complexity, the performance of Braun's method could be improved to some extent by measuring the positive definiteness of $\mathbf{Q}(l,k) = \mathrm{E}\left\{\boldsymbol{R}(l,k)\boldsymbol{R}^{\mathrm{H}}(l,k)\right\} - \mathbf{C}_v(l,k)$, before subtracting the correlations in Eq. (E.36). In cases where $\mathbf{Q}(l,k)$ is not positive definite, the nearest positive definite matrix [36] of $\mathbf{Q}(l,k)$ could be used to modify the estimate of $\mathbf{C}_v(l,k)$ used in Eq. (E.36).

7 Conclusion and Future Work

In this paper, we proposed three maximum-likelihood-based DoA estimators for a hearing aid system (HAS) which has access to the noise-free target signal via a wireless microphone. The proposed DoA estimators are based on three different models of the direction-dependent relative transfer functions (RTFs) between the HAS' microphones. These RTF models, which we call i) the free-field-far-field model, ii) the spherical-head model, and iii) the measured-RTF model, represent, with increasing accuracy and complexity, the head shadowing effect of the user's head on impinging signals. We showed that the considered signal model and the RTF models allowed the likelihood function to be calculated efficiently via inverse discrete Fourier transform techniques. In simulation experiments, we analyzed the influences of the true DoA, SNR, window length and reverberation on the performance of the proposed estimators. Moreover, we compared the performance of the estimators with the methods proposed in [18] and [17], which we refer to as the cross-correlation-based method and MLSSL, respectively. The cross-correlation-based method does not take ambient noise characteristics and head shadowing effects into account while MLSSL does take noise characteristics and detailed head shadowing effects into account via a user-specific HRTF database. Simulation results showed that all the DoA estimators proposed in this paper markedly outperform the cross-correlation-based method, while MLSSL outperform the proposed DoA estimators, when the user-specific HRTFs corresponding to the actual location of the target is in the HRTF database used by MLSSL; this is obviously a highly ideal case. We showed that MLSSL is very sensitive to mismatches between the HRTF database and the actual target source distance and the particular user. These mismatches deteriorate the MLSSL performance dramatically while the proposed estimators generally perform well.

Among the DoA estimators proposed in this paper, the measured-RTF-based method provides the lowest DoA estimation error robustly across different noise fields, DoAs, SNRs, and window lengths. In situations where the user-specific measured RTFs or the measured RTFs for a head-and-torso simulator (HATS) are not available, the spherical-head-model-based estimator provides a good performance and is robust against changing physical characteristics and, hence, HRTFs of users.

The proposed estimators rely on spatiospectral signal characteristics, which are assumed fixed across a short (in the range of milliseconds) duration. It is a topic of future research to extend the estimators to take temporal characteristics of the acoustic scene into accounts, e.g. by modeling the relative movement of the user's head and the target source.

References

[1] A. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT press, 1994.

[2] A. Bayat, M. Farhadi, A. Pourbakht, H. Sadjedi, H. Emamdjomeh, M. Kamali, and G. Mirmomeni, "A comparison of auditory perception in hearing-impaired and normal-hearing listeners: an auditory scene analysis study," *Iranian Red Crescent Medical Journal*, vol. 15, no. 11, 2013.

[3] J. M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 2, 2003, pp. 1228–1233.

[4] J. A. MacDonald, "A localization algorithm based on head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4290–4296, 2008.

[5] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[6] J. Kotus, K. Lopatka, and A. Czyzewski, "Detection and localization of selected acoustic events in acoustic field for smart surveillance applications," *Multimedia Tools and Applications*, vol. 68, no. 1, pp. 5–21, 2014.

[7] S. Goetze, T. Rohdenburg, V. Hohmann, B. Kollmeier, and K.-D. Kammeyer, "Direction of arrival estimation based on the dual delay line approach for binaural hearing aid microphone arrays," in *International*

Symposium on Intelligent Signal Processing and Communication Systems, Nov 2007, pp. 84–87.

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

[9] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction(src) on a large-aperture microphone array," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, April 2007, pp. I–121–I–124.

[10] R. Schmidt, "A signal subspace approach to multiple emitter location and spectral estimation," Ph.D. dissertation, Stanford University, 1981.

[11] R. Badeau, G. Richard, and B. David, "Fast adaptive esprit algorithm," in *IEEE/SP Workshop on Statistical Signal Processing*, July 2005, pp. 289–294.

[12] J. C. Murray, H. Erwin, and S. Wermter, "Robotics sound-source localization and tracking using interaural time difference and cross-correlation," in *Proceedings of NeuroBotics Workshop*, 2004, pp. 89–97.

[13] Y. Huang, J. Benesty, and J. Chen, *Time Delay Estimation and Source Localization*. Springer Berlin Heidelberg, 2008, pp. 1043–1063.

[14] F. keyrouz, "Advanced binaural sound localization in 3-D for humanoid robots," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 9, pp. 2098–2107, Sept 2014.

[15] C. Vina, S. Argentieri, and M. Rébillat, "A spherical cross-channel algorithm for binaural sound localization," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 2921–2926.

[16] M. Zohourian and R. Martin, "Binaural speaker localization and separation based on a joint ITD/ILD model and head movement tracking," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2016, pp. 430–434.

[17] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Maximum likelihood approach to "informed" sound source localization for hearing aid applications," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2015, pp. 16–20.

[18] G. Courtois, P. Marmaroli, M. Lindberg, Y. Oesch, and W. Balande, "Implementation of a binaural localization algorithm in hearing aids: specifications and achievable solutions," in *Audio Engineering Society Convention 136*, April 2014, paper 9034.

[19] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Informed TDoA-based direction of arrival estimation for hearing aid applications," in *Proceedings of IEEE Global Conference on Signal and Information Processing*, 2015, pp. 953–957.

[20] ——, "Informed direction of arrival estimation using a spherical-head model for hearing aid applications," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2016, pp. 360–364.

[21] J. Jensen, M. S. Pedersen, M. Farmani, and P. Minnaar, "Hearing system," U.S. Patent 20 160 112 811, April 21, 2016.

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

[23] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *The Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048–3058, 1998.

[24] Y. Avargel, "Linear system identification in the short-time Fourier transform domain," Ph.D. dissertation, Israel Institute of Technology, 2008.

[25] 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.

[26] D. R. Brillinger, *Time Series: Data Analysis and Theory*. Society for Industrial and Applied Mathematics (SIAM), 2001.

[27] R. Woodworth, *Experimental Psychology*. Holt, New York, 1938.

[28] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68–77, Jan 2010.

[29] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space," in *Audio Engineering Society Convention 107*, 1999, paper 5026.

[30] R. L. Bouquin-Jeannes, A. A. Azirani, and G. Faucon, "Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 484–487, Sep 1997.

[31] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, “Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses,” *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 298605, pp. 1–10, 2009.

[32] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, 1997.

[33] P. Kabal, “TSP speech database,” Department of Electrical and Computer Engineering, McGill University, Tech. Rep., 2002.

[34] D. Avitzour, “Time delay estimation at high signal-to-noise ratio,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 27, no. 2, pp. 234–237, Mar 1991.

[35] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, “On the influence of microphone array geometry on HRTF-based sound source localization,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2015, pp. 439–443.

[36] N. J. Higham, “Computing the nearest correlation matrix—a problem from finance,” *IMA Journal of Numerical Analysis*, vol. 22, no. 3, pp. 329–343, 2002.

Paper F

Bias-compensated informed sound source localization using relative transfer functions

Mojtaba Farmani, Michael Syskind Pedersen, Zheng-Hua Tan,
and Jesper Jensen

The paper has been submitted to
Journal of IEEE/ACM Transactions on Audio, Speech, and Language Processing

The layout has been revised.

Abstract

In this paper, we consider the problem of estimating the target sound direction of arrival (DoA) for a hearing aid system (HAS), which can connect to a wireless microphone worn by the talker of interest. The wireless microphone "informs" the HAS about the noise-free target speech. To estimate the DoA, we consider a maximum likelihood (ML) approach, and we assume that a database of DoA-dependent relative transfer functions (RTFs) has been measured in advance and is available. The proposed DoA estimator is able to take the available noise-free target speech, ambient noise characteristics and the shadowing effect of the user's head on the received signals into account, and it supports both monaural and binaural microphone array configurations. Moreover, we analytically analyze the bias in the proposed estimator and introduce a modified estimator, which has been compensated for the bias. We demonstrate that the proposed method has lower computational complexity and better performance than recent RTF-based estimators. Further, to decrease the number of parameters required to be wirelessly exchanged between the hearing aids (HAs) in binaural configurations, we propose an information fusion strategy, which avoids transmitting microphone signals between the HAs. An important benefit of the proposed IF strategy is that the number of parameters to be exchanged between the HAs is independent of the number of HA microphones. Finally, we investigate the performance of variants of the proposed estimator extensively in different noisy and reverberant situations.

1 Introduction

The auditory scene analysis (ASA) ability in humans allows us to focus intentionally on a sound source, while suppressing other unrelated sound sources, which are usually present simultaneously in realistic acoustic scenes [1]. Sensorineural hearing-impaired listeners lose this ability to some extent and face difficulties in interacting with the environment [2]. In an attempt to retrieve the normal interactions of the hearing impaired users with the environment, hearing aid systems (HASs) may carry out some of the ASA tasks, which are carried out by a healthy auditory system.

This paper studies sound source localization (SSL)—one of the main tasks in ASA—in a hearing aid context. SSL using microphone arrays has been investigated extensively in various applications, such as robotics [4–7], video conferencing [8–10], surveillance [11, 12], and hearing aids [13–15]. In most of these applications, the noise-free target sound is not accessible, e.g. [4–15]. However, modern HASs can connect to a wireless microphone worn by the target talker to access an essentially noise-free version of the target signal emitted at the target talker's position [3, 16–21]. This feature introduces the "informed" SSL problem considered in this paper.

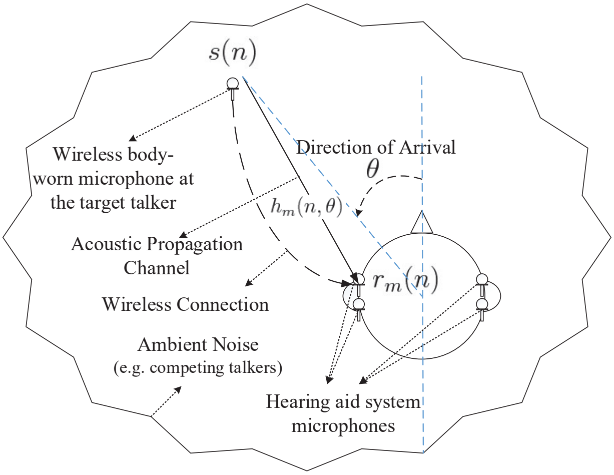


Fig. F.1: An "informed" SSL scenario for a HAS using a wireless microphone. $r_m(n)$ is the noisy received signal at microphone m, $s(n)$ is the noise-free target signal emitted at the target location, and $h_m(n,\theta)$ is the acoustic channel impulse response between the target talker and microphone m. $s(n)$ is available at the HAS via the wireless connection, and the HAs are also connected to each other wirelessly. The goal is to estimate the direction of arrival θ [3].

Fig. F.1 depicts an exemplar situation considered in this paper. The HAS includes two wirelessly connected hearing aids (HAs) and a wireless microphone. The HAs are mounted behind each ear of the user, and the wireless microphone is worn by the target talker. The target signal $s(n)$ is generated at the target location, propagates through the acoustic channel $h_m(n,\theta)$, and reaches microphone m of the binaural HAS. Due to additive ambient noise, the signal $r_m(n)$ captured by microphone m is a noisy signal. Further, the signal $s(n)$ is transmitted wirelessly to the HAS. In this setup, we aim to estimate the target signal direction of arrival (DoA) θ. Estimation of the target sound DoA would allow the HAS to enhance the spatial rendering of the acoustic scene, e.g. by imposing the corresponding binaural cues on the wirelessly received target sound [17, 18].

The "informed" SSL problem for hearing aid applications was first studied in [16]. The method proposed in [16] is based on estimation of time difference of arrivals (TDoAs) of microphone signals. More precisely, this method employs a cross-correlation technique to estimate the time difference of arrival (TDoA), then considers a sine law to map the estimated TDoA to a DoA estimate. This approach has relatively low computational load; however, it does not take the shadowing effect of the user's head and potential ambient noise characteristics into account. This degrades the DoA estimation performance markedly [3, 19].

To consider the head shadowing effect and ambient noise characteristics for the “informed” SSL, a maximum likelihood (ML) approach has been proposed in [19] using a database of measured head related transfer functions (HRTFs[1]) labeled by their corresponding DoA. To estimate the DoA, this approach, called MLSSL (maximum likelihood sound source localization), looks for the HRTF entry in the database, which maximizes the likelihood of the observed microphone signals. MLSSL has relatively high computational load, but it performs effectively under severely noisy conditions, when detailed personal HRTFs for different directions and different distances are available [3, 19]. On the other hand, when the personal HRTFs are not available, or when the HRTFs corresponding to the actual distance of the target are not covered in the database, the estimation performance of MLSSL degrades [3].

In [3], a new ML approach, which also considers head shadowing effects and ambient noise characteristics, has been proposed for “informed” SSL using a database of measured relative transfer functions (RTFs). An RTF is the ratio between two HRTFs [23] and can easily be obtained from the measured HRTFs [3]. Compared with MLSSL, this new approach has lower computational load, and provides more robust performance, when an individualized database is not available [3]. RTFs, in comparison with HRTFs, are almost independent of the distance between the target talker and the user, especially in far-field situations [3, 24]. The distance independency of RTFs reduces the required memory and the computational load of the estimator proposed in [3] compared with MLSSL. This is because, to estimate the DoA, the estimator in [3] must search in an RTF database, which is only a function of DoA, while MLSSL must search in an HRTF database which is a function of both DoA and distance.

In this paper, we propose an ML approach that uses a database of measured RTFs to estimate the DoA. Unlike the estimator proposed in [3], which considers a binaural configuration using exactly two microphones (one microphone in each HA), the proposed method works for any $M \geq 2$ microphones in both monaural and binaural configurations. Further, compared with [3], the proposed method decreases the computational complexity and the number of parameters required to be wirelessly transmitted between the HAs, while maintaining—and in some situations, even improving—the estimation accuracy. To decrease the computational load, we relax some of the constraints used in [3] for modeling the acoustic transfer function between the target and a reference microphone. This relaxation makes the signal model more realistic, and we show that it also allows us to formulate the problem in a way that decreases the computational complexity. To decrease

[1] An HRTF is formally defined as “a specific individuals left or right ear far-field frequency response, as measured from a specific point in the free field to a specific point in the ear canal” [22]. Here, an HRTF refers to the frequency response from a target source to the microphone of a hearing aid system [3, 19].

the number of parameters required to be wirelessly exchanged between the HAs in binaural configurations, we propose an information fusion strategy, which transmits posterior DoA probabilities between the HAs instead of entire signal frames. Finally, we analytically derive the bias in the estimator, and propose a closed-form bias-compensation strategy, resulting in an unbiased estimator.

The structure of this paper is as follows. In Secs. 2 and 3, the signal model and the ML framework are presented, respectively. Afterwards, in Sec. 4, the proposed "informed" DoA estimators using the ML framework is derived. In Sec. 5, we analytically derive the bias of the DoA estimator and propose a bias-compensation strategy. In Sec. 6, we propose an information fusion strategy to decrease the wireless communication between the HAs in binaural configurations. In Sec. 7, the performance of variants of the proposed estimator is studied extensively and compared with existing algorithms using experimental simulations. Lastly, we conclude the paper in Sec. 8.

2 Signal Model

In Fig. F.1, the noisy signal r_m received at microphone m of the HAS is given by:

$$r_m(n) = s(n) * h_m(n,\theta) + v_m(n), \qquad m = 1, 2, \cdots M, \tag{F.1}$$

where $s(n)$ is the noise-free target signal emitted at the target talker's position, $h_m(n,\theta)$ is the acoustic channel impulse response between the target talker and microphone m, and $v_m(n)$ is an additive noise component. Further, n is the discrete time index, and $*$ indicates the convolution operator.

In the short time Fourier transform (STFT) domain, Eq. (F.1) can be approximated as [25]:

$$R_m(l,k) = S(l,k)H_m(k,\theta) + V_m(l,k), \tag{F.2}$$

where

$$R_m(l,k) = \sum_n r_m(n)w(n-lA)e^{-\frac{j2\pi k}{N}(n-lA)},$$

denotes the STFT of $r_m(n)$, where l and k are frame and frequency bin indexes, respectively, N is the discrete Fourier transform (DFT) order, A is the decimation factor, $w(n)$ is the windowing function, and $j = \sqrt{-1}$ is the imaginary unit. Similarly, $S(l,k)$ and $V_m(l,k)$ denote the STFT of $s(n)$ and $v_m(n)$, respectively, and are defined analogously to $R_m(l,k)$. Moreover,

$$\begin{aligned} H_m(k,\theta) &= \sum_n h_m(n,\theta)e^{-\frac{j2\pi kn}{N}} \\ &= \alpha_m(k,\theta)e^{-\frac{j2\pi k}{N}D_m(k,\theta)}, \end{aligned} \tag{F.3}$$

denotes the discrete Fourier transform (DFT) of $h_m(n,\theta)$, where $\alpha_m(k,\theta)$ is a positive real number which denotes the frequency-dependent attenuation factor due to propagation effects, and $D_m(k,\theta)$ is the frequency-dependent propagation time, in samples, from the target talker to microphone m.

As mentioned, Eq. (F.2) is an approximation of Eq. (F.1) in the STFT domain. This approximation is known as the multiplicative transfer function (MTF) approximation [25], and its accuracy depends on the length and smoothness of $w(n)$: the longer and the smoother the analysis window $w(n)$, the more accurate the approximation [25].

To rewrite Eq. (F.2) into vector form, let $\boldsymbol{d}(k,\theta) = [d_1(k,\theta), d_2(k,\theta), \cdots, d_M(k,\theta)]^{\mathrm{T}}$ denote a vector of RTFs defined with respect to a reference microphone, as

$$d_m(k,\theta) = \frac{H_m(k,\theta)}{H_u(k,\theta)}, \qquad m = 1, \cdots, M,$$

where u is the index of the reference microphone. Moreover, let $\boldsymbol{R}(l,k) = [R_1(l,k), R_2(l,k), \cdots, R_M(l,k)]^{\mathrm{T}}$, and $\boldsymbol{V}(l,k) = [V_1(l,k), V_2(l,k), \cdots, V_M(l,k)]^{\mathrm{T}}$. Now, Eq. (F.2) can written as:

$$\boldsymbol{R}(l,k) = S(l,k)H_u(k,\theta)\boldsymbol{d}(k,\theta) + \boldsymbol{V}(l,k). \tag{F.4}$$

3 Maximum Likelihood Framework

To define the likelihood function, we assume that the additive noise vector $\boldsymbol{V}(l,k)$ follows a zero-mean circularly-symmetric complex Gaussian distribution, i.e. $\boldsymbol{V}(l,k) \sim \mathcal{N}(0, \mathbf{C}_v(l,k))$, where $\mathbf{C}_v(l,k) = \mathrm{E}\{\boldsymbol{V}(l,k)\boldsymbol{V}^{\mathrm{H}}(l,k)\}$, and where $\mathrm{E}\{.\}$ and the superscript H represent the expectation and Hermitian transpose operators, respectively. Since we assume that the target signal is picked up without any noise by the wireless microphone, we consider $S(l,k)$ as a deterministic and known variable at the HAS. Moreover, $H_u(k,\theta)$ and $\boldsymbol{d}(k,\theta)$ are also considered deterministic, but unknown. Further, $\mathbf{C}_v(l,k)$ is assumed to be known (in Sec. 7.1, we briefly explain a simple and robust method for estimating $\mathbf{C}_v(l,k)$). Hence, from Eq. (F.4), it follows that:

$$\boldsymbol{R}(l,k) \sim \mathcal{N}\left(S(l,k)H_u(k,\theta)\boldsymbol{d}(k,\theta), \mathbf{C}_v(l,k)\right).$$

Furthermore, let us assume that the noisy observations are independent across frequencies (to be precise, this assumption is valid, when the correlation time of the signal is short compared with the frame length [26, 27]). Accordingly, the likelihood function for frame l is given by:

$$p(\underline{\underline{\mathbf{R}}}(l); \boldsymbol{H}_u(\theta), \underline{\underline{\mathbf{d}}}(\theta)) =$$
$$\prod_{k=0}^{N-1} \frac{1}{\pi^M \det\left[\mathbf{C}_v(l,k)\right]} e^{\{-(\boldsymbol{Z}(l,k))^{\mathrm{H}}\mathbf{C}_v^{-1}(l,k)(\boldsymbol{Z}(l,k))\}},$$

where det[.] denotes the matrix determinant, and

$$\begin{aligned}\underline{\underline{\mathbf{R}}}(l) &= [\ \boldsymbol{R}(l,0),\ \boldsymbol{R}(l,1),\ \cdots,\ \boldsymbol{R}(l,N-1)\ ],\\ \boldsymbol{H}_u(\theta) &= [\ H_u(0,\theta),\ H_u(1,\theta),\ \cdots,\ H_u(N-1,\theta)\ ],\\ \underline{\underline{\mathbf{d}}}(\theta) &= [\ \boldsymbol{d}(0,\theta),\ \boldsymbol{d}(1,\theta),\ \cdots,\ \boldsymbol{d}(N-1,\theta)\ ],\\ \boldsymbol{Z}(l,k) &= \boldsymbol{R}(l,k) - S(l,k)H_u(k,\theta)\boldsymbol{d}(k,\theta).\end{aligned}$$

To simplify the expressions, we consider the log-likelihood function and drop terms independent of θ. Therefore, the reduced log-likelihood function is given by:

$$\begin{aligned}&\mathcal{L}(\underline{\underline{\mathbf{R}}}(l);\boldsymbol{H}_u(\theta),\underline{\underline{\mathbf{d}}}(\theta))) =\\ &\qquad\sum_{k=0}^{N-1}\{-(\boldsymbol{Z}(l,k))^{\mathrm{H}}\mathbf{C}_v^{-1}(l,k)(\boldsymbol{Z}(l,k))\}.\end{aligned} \tag{F.5}$$

The ML estimate of θ at frame l is found by maximizing $\mathcal{L}$ with respect to θ. In the following, we derive the proposed DoA estimator.

4 The Proposed DoA Estimator

To derive the proposed estimator, we assume a database Θ of pre-measured $\underline{\underline{\mathbf{d}}}$s labeled by their corresponding θ_i is available. To be more precise, $\Theta = \left\{\underline{\underline{\mathbf{d}}}(\theta_1),\underline{\underline{\mathbf{d}}}(\theta_2),\cdots,\underline{\underline{\mathbf{d}}}(\theta_I)\right\}$, where I is the number of entries in Θ, is assumed to be available for the DoA estimation. To find the ML estimate of θ, the proposed DoA estimator evaluates $\mathcal{L}$ for each $\underline{\underline{\mathbf{d}}}(\theta_i) \in \Theta$. The MLE of θ is the DoA label of the $\underline{\underline{\mathbf{d}}}$, which leads to the largest log-likelihood. In other words,

$$\hat{\theta} = \arg\max_{\underline{\underline{\mathbf{d}}}(\theta_i)\in\Theta} \mathcal{L}(\underline{\underline{\mathbf{R}}}(l);\boldsymbol{H}_u(\theta),\underline{\underline{\mathbf{d}}}(\theta_i)).$$

To be able to evaluate $\mathcal{L}$ efficiently for different $\underline{\underline{\mathbf{d}}}(\theta_i)$, we first make $\mathcal{L}$ independent of $\boldsymbol{H}_u$. To do so, we find the maximum likelihood estimate (MLE) of $\boldsymbol{H}_u$, as a function of the other variables, and replace the MLE back into $\mathcal{L}$. Solving $\frac{\partial\mathcal{L}}{\partial H_u(k,\theta)} = 0$ for $H_u(k,\theta)$ leads to

$$\hat{H}_u(k,\theta) = \frac{\boldsymbol{d}^H(k,\theta)\mathbf{C}_v^{-1}(l,k)\boldsymbol{R}(l,k)}{S(l,k)\boldsymbol{d}^H(k,\theta)\mathbf{C}_v^{-1}(l,k)\boldsymbol{d}(k,\theta)}.$$

Inserting $\hat{H}_u(k,\theta)$ into $\mathcal{L}$ gives

$$\mathcal{L}(\underline{\underline{\mathbf{R}}}(l);\underline{\underline{\mathbf{d}}}(\theta)) \propto \sum_{k=1}^{N-1}\frac{|\boldsymbol{R}^H(l,k)\mathbf{C}_v^{-1}(l,k)\boldsymbol{d}(k,\theta)|^2}{\boldsymbol{d}^H(k,\theta)\mathbf{C}_v^{-1}(l,k)\boldsymbol{d}(k,\theta)}, \tag{F.6}$$

where $|.|$ denotes the magnitude of a complex number. Note that terms independent of θ have been omitted.

Regarding Eq. (F.6), $\mathcal{L}$ is now independent of $\boldsymbol{H}_u$, but surprisingly, it is also independent of the clean target signal $S(l,k)$, which is available in the set up considered in this paper. In other words, the accessible information of $S(l,k)$ does not play any direct role in the estimator (Eq. (F.6))—the indirect role of $S(l,k)$ is in estimating $\mathbf{C}_v(l,k)$ as explained in Sec. 7.1. To investigate the reason, let us consider the signal model presented in Eq. (F.4). The factor $S(l,k)H_u(k,\theta)$ represents the clean signal received at the reference microphone. Even though $S(l,k)$ is known, the product $S(l,k)H_u(k,\theta)$ is entirely unknown, because $H_u(k,\theta)$ is an unknown variable. In other words, replacing $S(l,k)H_u(k,\theta)$ with a deterministic but unknown dummy variable $X(l,k)$, which represents the unknown clean signal received at the reference microphone, leads to an "uninformed" signal model. Finding the MLE of $X(l,k)$, and replacing the result in the corresponding log-likelihood function, would lead to an equation similar to Eq. (F.6). Hence, the current signal model of H_u completely spoils the knowledge of $S(l,k)$, when forming the product $S(l,k)H_u(k,\theta)$. In the following, we explain how this problem can be confronted by imposing certain constraints on $\boldsymbol{H}_u$.

To solve the problem mentioned above and to exploit the accessible $S(l,k)$ in the DoA estimator, we assume that $\boldsymbol{H}_u$ is related to a "sunny" microphone [3]. In other words, when the method evaluates $\mathcal{L}$ for $\underline{\underline{\mathbf{d}}}$s corresponding to directions to the left side of the head, $\boldsymbol{H}_u$ is related to a microphone in the left HA, and when the method evaluates $\mathcal{L}$ for $\underline{\underline{\mathbf{d}}}$s corresponding to directions to the right side of the head, $\boldsymbol{H}_u$ is related to a microphone in the right HA (note that this evaluation strategy is practically operational and requires no prior knowledge about the true DoA). Further, in contrast to the method proposed in [3], which assumes that both the attenuation α_u and the delay D_u of the "sunny" HRTF are frequency independent, we remove the frequency-independency constraint on the delay D_u. Removing this constraint makes the signal model more realistic, because the head presence generally introduces a frequency-dependent delay on the received signals, which is more in-line with human head acoustics [24]. When evaluating $\mathcal{L}$, we will show that this "sunny" HRTF model allows us to simply sum over all frequency bins instead of computing an inverse discrete Fourier transform (IDFT) as proposed in [3]. This decreases the estimator's computational complexity because an IDFT has an order of complexity of $N \log N$ [28], while summing over all frequencies has an order of complexity of N.

Regarding Eq. (F.3), H_u can be written as a function of its parameters, i.e. $\alpha_u(k,\theta)$ and $D_u(k,\theta)$. Let us assume that the attenuation $\alpha_u(k,\theta)$ is frequency independent, i.e. $\alpha_u(k,\theta) = \alpha_u(\theta)$, and collect the $D_u(k,\theta)$ values in a vector $\boldsymbol{D}_u(\theta) = [D_u(0,\theta), D_u(1,\theta), ..., D_u(N-1,\theta)]^{\mathrm{T}}$. This allows us to write the log-likelihood function as $\mathcal{L}(\underline{\underline{\mathbf{R}}}(l); \alpha_u(\theta), \boldsymbol{D}_u(\theta), \underline{\underline{\mathbf{d}}}(\theta))$.

As before, to evaluate $\mathcal{L}$ efficiently for different $\underline{\underline{\mathbf{d}}}(\theta_i)$, we eliminate $\alpha_u(\theta)$ and $\boldsymbol{D}_u(\theta)$ by substituting their MLEs into $\mathcal{L}$. To find the MLE of α_u, we solve $\frac{\partial \mathcal{L}}{\partial \alpha_u(\theta)} = 0$, which leads to

$$\hat{\alpha}_u(\theta) = \frac{\sum_{k=1}^{N-1}(A(l,k,\theta) + A^*(l,k,\theta))}{\sum_{k=1}^{N-1} B(l,k,\theta)}, \tag{F.7}$$

where

$$A(l,k,\theta) = S^*(l,k)\boldsymbol{d}^H(k,\theta)\mathbf{C}_v^{-1}(l,k)\boldsymbol{R}(l,k)e^{\frac{j2\pi k}{N}D_1(k,\theta)},$$
$$\text{and } B(l,k,\theta) = |S(l,k)|^2\boldsymbol{d}^H(k,\theta)\mathbf{C}_v^{-1}(l,k)\boldsymbol{d}(k,\theta).$$

Replacing Eq. (F.7) into $\mathcal{L}$ gives

$$\mathcal{L}(\underline{\underline{\mathbf{R}}}(l);\boldsymbol{D}_u(\theta),\underline{\underline{\mathbf{d}}}(\theta)) = \frac{\left(\sum_{k=1}^{N-1}(A(l,k,\theta) + A^*(l,k,\theta))\right)^2}{4\sum_{k=1}^{N-1} B(l,k,\theta)}. \tag{F.8}$$

Now, let us find the MLE of $D_u(k,\theta)$ by solving $\frac{\partial \mathcal{L}}{\partial D_u(k,\theta)} = 0$ and replace the result into $\mathcal{L}$. The MLE of $D_u(k,\theta)$ is given by

$$\hat{D}_u(k,\theta) = \frac{N}{j4\pi k}\log\left(\frac{C^*(k,\theta)}{C(k,\theta)}\right), \tag{F.9}$$

where

$$C(k,\theta) = S^*(l,k)\boldsymbol{d}^H(k,\theta)\mathbf{C}_v^{-1}(l,k)\boldsymbol{R}(l,k).$$

Substituting Eq. (F.9) into Eq. (F.8) leads to the final result

$$\tilde{\mathcal{L}}(\underline{\underline{\mathbf{R}}}(l);\underline{\underline{\mathbf{d}}}(\theta)) = \frac{\left(\sum_{k=1}^{N-1}|S^*(l,k)\boldsymbol{d}^H(k,\theta)\mathbf{C}_v^{-1}(l,k)\boldsymbol{R}(l,k)|\right)^2}{\sum_{k=1}^{N-1}|S(l,k)|^2\boldsymbol{d}^H(k,\theta)\mathbf{C}_v^{-1}(l,k)\boldsymbol{d}(k,\theta)}, \tag{F.10}$$

which only depends on the unknown $\underline{\underline{\mathbf{d}}}(\theta)$. Note that in contrast to Eq. (F.6), the available clean target signal $S(l,k)$ also contributes in the derived log-likelihood function. Further, the evaluation of Eq. (F.10) only requires summation across frequencies in contrast to computing IDFTs, which are required in the method proposed in [3]. Now, we can find the MLE of θ by

$$\hat{\theta} = \underset{\underline{\underline{\mathbf{d}}}(\theta_i)\in\Theta}{\arg\max}\,\tilde{\mathcal{L}}(\underline{\underline{\mathbf{R}}}(l);\underline{\underline{\mathbf{d}}}(\theta_i)).$$

5 Bias Investigation

Generally, it is desirable that the proposed log-likelihood function does not have any intrinsic bias towards any specific direction. In this section, we analytically show that the proposed log-likelihood function is indeed biased, and we propose a closed-form bias-compensated log-likelihood function. To do so, we consider situations, where the target signal is almost absent, or the signal to noise ratio (SNR) is approaching $-\infty$. In these situations, the likelihood values corresponding to the different directions in Θ should ideally be equal and independent of θ. To be more precise, when the target signal is almost absent or when the SNR $\rightarrow -\infty$, we have $\boldsymbol{R}(l,k) \approx \boldsymbol{V}(l,k)$, and

$$\lim_{SNR \rightarrow -\infty \text{ or } S \rightarrow 0} \tilde{\mathcal{L}}(\underline{\underline{\mathbf{R}}}(l); \underline{\underline{\mathbf{d}}}(\theta)) = \frac{\left(\sum_{k=1}^{N-1} |S^*(l,k)\boldsymbol{d}^H(k,\theta)\mathbf{C}_v^{-1}(l,k)\boldsymbol{V}(l,k)| \right)^2}{\sum_{k=1}^{N-1} |S(l,k)|^2 \boldsymbol{d}^H(k,\theta)\mathbf{C}_v^{-1}(l,k)\boldsymbol{d}(k,\theta)}. \tag{F.11}$$

In these situations, the expected value of $\tilde{\mathcal{L}}$ with respect to $\boldsymbol{V}$ should ideally be constant with respect to θ, or equivalently, it should be independent of $\underline{\underline{\mathbf{d}}}(\theta)$.

In the following, we derive a closed-form expression for the expected value of Eq. (F.11) with respect to the only random variable in this expression, i.e. $\boldsymbol{V}$. For notational convenience, we omit the frequency and the frame indexes. From the assumption that $\boldsymbol{V}$ follows a zero-mean circularly-symmetric complex Gaussian distribution, i.e. $\boldsymbol{V} \sim \mathcal{N}(0, \mathbf{C}_v)$ (Sec. 3), we have

$$S^* \boldsymbol{d}^H(\theta)\mathbf{C}_v^{-1}\boldsymbol{V} \quad \sim \quad \mathcal{N}(0, |S|^2 \boldsymbol{d}^H(\theta)\mathbf{C}_v^{-1}\boldsymbol{d}(\theta)).$$

Since $S^* \boldsymbol{d}^H(\theta)\mathbf{C}_v^{-1}\boldsymbol{V}$ follows a circularly-symmetric complex Gaussian distribution, its real and imaginary parts, i.e. $X = \text{Re}\{S^* \boldsymbol{d}^H(\theta)\mathbf{C}_v^{-1}\boldsymbol{V}\}$ and $Y = \text{Im}\{S^* \boldsymbol{d}^H(\theta)\mathbf{C}_v^{-1}\boldsymbol{V}\}$, are independent and also follow Gaussian distributions [29],

$$X \sim \mathcal{N}(0, \frac{1}{2}|S|^2 \boldsymbol{d}^H(\theta)\mathbf{C}_v^{-1}\boldsymbol{d}(\theta)),$$

and

$$Y \sim \mathcal{N}(0, \frac{1}{2}|S|^2 \boldsymbol{d}^H(\theta)\mathbf{C}_v^{-1}\boldsymbol{d}(\theta)).$$

Therefore, $U = \sqrt{X^2 + Y^2} = |S^* \boldsymbol{d}^H(\theta)\mathbf{C}_v^{-1}\boldsymbol{V}|$ follows a Rayleigh distribution [29], where

$$\text{E}\{U\} = \sqrt{\frac{\pi}{4}|S|^2 \boldsymbol{d}^H(\theta)\mathbf{C}_v^{-1}\boldsymbol{d}(\theta)},$$

and

$$\text{Var}\{U\} = \frac{4-\pi}{4}|S|^2 \boldsymbol{d}^H(\theta)\mathbf{C}_v^{-1}\boldsymbol{d}(\theta),$$

where Var{.} denotes the variance operator. Reinvoking the assumption that observations are independent across frequencies (Sec. 3), we have

$$\mathrm{E}\left\{\sum_{k=1}^{N-1} U\right\} = \sum_{k=1}^{N-1} \sqrt{\frac{\pi}{4}|S|^2 \boldsymbol{d}^H(\theta)\mathbf{C}_v^{-1}\boldsymbol{d}(\theta)}, \tag{F.12}$$

and

$$\mathrm{Var}\left\{\sum_{k=1}^{N-1} U\right\} = \sum_{k=1}^{N-1} \frac{4-\pi}{4}|S|^2 \boldsymbol{d}^H(\theta)\mathbf{C}_v^{-1}\boldsymbol{d}(\theta). \tag{F.13}$$

From Eqs. (F.12) and (F.13), the expected value of the numerator of Eq. (F.11) with respect to $\boldsymbol{V}$ is given by F

$$\mathrm{E}\left\{\left(\sum_{k=1}^{N-1} U\right)^2\right\} = \left(\mathrm{E}\left\{\sum_{k=1}^{N-1} U\right\}\right)^2 + \mathrm{Var}\left\{\sum_{k=1}^{N-1} U\right\}.$$

Hence, the expected value of $\tilde{\mathcal{L}}$ with respect to $\boldsymbol{V}$, when the target signal is almost absent, or the SNR is approaching $-\infty$, is given by

$$\mathrm{E}\left\{\tilde{\mathcal{L}}(\underline{\underline{\mathbf{R}}}(l);\underline{\underline{\mathbf{d}}}(\theta))\right\}\Bigg|_{SNR\to-\infty \text{ or } S\to 0} = \frac{\left(\sum_{k=1}^{N-1}\sqrt{\frac{\pi}{4}|S|^2\boldsymbol{d}^H(\theta)\mathbf{C}_v^{-1}\boldsymbol{d}(\theta)}\right)^2}{\sum_{k=1}^{N-1}|S|^2\boldsymbol{d}^H(\theta)\mathbf{C}_v^{-1}\boldsymbol{d}(\theta)} + \frac{4-\pi}{4}. \tag{F.14}$$

Unfortunately, Eq. (F.14) shows that the expected value of $\tilde{\mathcal{L}}$ with respect to $\boldsymbol{V}$ in the considered situations is not independent of $\underline{\underline{\mathbf{d}}}(\theta)$, and hence, $\tilde{\mathcal{L}}$ is biased. However, a bias-compensated log-likelihood function can be defined simply by subtracting this expectation (Eq. (F.14)) from the log-likelihood (Eq. (F.10)), i.e.

$$\bar{\mathcal{L}}(\underline{\underline{\mathbf{R}}}(l);\underline{\underline{\mathbf{d}}}(\theta)) = \frac{\left(\sum_{k=1}^{N-1}|S^*\boldsymbol{d}^H(\theta)\mathbf{C}_v^{-1}\boldsymbol{R}|\right)^2}{\sum_{k=1}^{N-1}|S|^2\boldsymbol{d}^H(\theta)\mathbf{C}_v^{-1}\boldsymbol{d}(\theta)} - \frac{\left(\sum_{k=1}^{N-1}\sqrt{\frac{\pi}{4}|S|^2\boldsymbol{d}^H(\theta)\mathbf{C}_v^{-1}\boldsymbol{d}(\theta)}\right)^2}{\sum_{k=1}^{N-1}|S|^2\boldsymbol{d}^H(\theta)\mathbf{C}_v^{-1}\boldsymbol{d}(\theta)} - \frac{4-\pi}{4}, \tag{F.15}$$

and the bias-compensated MLE of θ is given by

$$\bar{\theta} = \arg\max_{\underline{\underline{\mathbf{d}}}(\theta_i)\in\Theta} \bar{\mathcal{L}}(\underline{\underline{\mathbf{R}}}(l);\underline{\underline{\mathbf{d}}}(\theta_i)).$$

6 Reducing the Wireless Communication in Binaural Configurations

In binaural configurations, up to this point, we assumed that the signals received by all the microphones of both HAs are available at the "master" HA (the HA which performs the DoA estimation). This means that one of the HAs should transmit the signals received by its microphones to the other HA (the "master" HA). In this section, we aim to decrease the number of parameters required to be exchanged between the HAs in binaural configurations, where the number of microphones in each HA is at least two.

The trivial way to completely eliminate the wireless communication between the HAs is that each HA estimates the DoA independently using the signals received by its own microphones. This way is expected to degrade the estimation performance notably, because the number of signal frames considered for estimating the DoA has been decreased, compared with a system utilizing all microphone signals. Moreover, this can cause the HAs to have different and inconsistent estimates of the DoA.

Instead, in this section, we present an information fusion (IF) strategy, which needs some wireless communication between the HAs, but does not need to transmit all the microphone signals between the HAs. Moreover, using the proposed IF strategy, the number of parameters required to be exchanged between the HAs is independent of the number of microphones in a HA, i.e. increasing the number of microphones does not change the number of parameters to be exchanged. To do so, we assume that each HA evaluates $\bar{\mathcal{L}}$ locally for each $\underline{\underline{\mathbf{d}}}(\theta_i) \in \Theta$, using the signals picked up by its own microphones. In other words, for each $\underline{\underline{\mathbf{d}}}(\theta_i) \in \Theta$, two evaluations of $\bar{\mathcal{L}}$ are performed, one in the left HA and one in the right HA (let us denote them as $\bar{\mathcal{L}}_{\text{left}}$ and $\bar{\mathcal{L}}_{\text{right}}$, respectively). Afterwards, one of the HAs, e.g. the right HA, transmits the evaluation values of $\bar{\mathcal{L}}_{\text{right}}$ for all $\underline{\underline{\mathbf{d}}}(\theta_i) \in \Theta$ to the "master" HA, i.e. the left HA. To estimate the DoA, the "master" HA uses an IF technique, which will be defined later in this section, to combine $\bar{\mathcal{L}}_{\text{left}}$ and $\bar{\mathcal{L}}_{\text{right}}$ values. This strategy decreases the number of parameters to be transmitted between the HAs, because instead of transmitting all microphone signals, only I different evaluations of $\bar{\mathcal{L}}$ corresponding to different $\underline{\underline{\mathbf{d}}}(\theta_i) \in \Theta$, must be transmitted, at each time frame (typically, I is much smaller than the signal frame length.).

The main idea in fusing $\bar{\mathcal{L}}_{\text{left}}$ and $\bar{\mathcal{L}}_{\text{right}}$ is to approximate the joint likelihood $p\left(\underline{\underline{\mathbf{R}}}_{\text{left}}(l), \underline{\underline{\mathbf{R}}}_{\text{right}}(l); \underline{\underline{\mathbf{d}}}(\theta_i)\right)$, where $\underline{\underline{\mathbf{R}}}_{\text{left}}(l)$ and $\underline{\underline{\mathbf{R}}}_{\text{right}}(l)$ respectively represent the signals received by the microphones of the left HA and the right

HA. To do so, we use the following conditional probabilities:

$$p\left(\underline{\underline{\mathbf{R}}}_{\text{left}}(l);\underline{\underline{\mathbf{d}}}(\theta_i)\right) \propto \exp\left(\bar{\mathcal{L}}_{\text{left}}\left(\underline{\underline{\mathbf{R}}}_{\text{left}}(l);\underline{\underline{\mathbf{d}}}(\theta_i)\right)\right),$$

$$p\left(\underline{\underline{\mathbf{R}}}_{\text{right}}(l);\underline{\underline{\mathbf{d}}}(\theta_i)\right) \propto \exp\left(\bar{\mathcal{L}}_{\text{right}}\left(\underline{\underline{\mathbf{R}}}_{\text{right}}(l);\underline{\underline{\mathbf{d}}}(\theta_i)\right)\right).$$

It should be noted that normalization is necessary to ensure that the sum of all posterior probabilities equals unity in each HA.

In general, to calculate $p(\underline{\underline{\mathbf{R}}}_{\text{left}}(l), \underline{\underline{\mathbf{R}}}_{\text{right}}(l); \underline{\underline{\mathbf{d}}}(\theta_i))$, the covariance between $\underline{\underline{\mathbf{R}}}_{\text{left}}(l)$ and $\underline{\underline{\mathbf{R}}}_{\text{right}}(l)$ must be known; and to estimate this covariance matrix, the microphones' signals must be transmitted between the HAs. However, if we assume $\underline{\underline{\mathbf{R}}}_{\text{right}}(l)$ and $\underline{\underline{\mathbf{R}}}_{\text{left}}(l)$ are conditionally independent of each other given $\underline{\underline{\mathbf{d}}}(\theta_i)$, there is no need to transfer the signals between the HAs, because the joint probability is the product of the two marginals,

$$\begin{aligned} p\left(\underline{\underline{\mathbf{R}}}_{\text{left}}(l), \underline{\underline{\mathbf{R}}}_{\text{right}}(l);\underline{\underline{\mathbf{d}}}(\theta_i)\right) = \\ p\left(\underline{\underline{\mathbf{R}}}_{\text{left}}(l);\underline{\underline{\mathbf{d}}}(\theta_i)\right) \times p\left(\underline{\underline{\mathbf{R}}}_{\text{right}}(l);\underline{\underline{\mathbf{d}}}(\theta_i)\right). \end{aligned} \tag{F.16}$$

Strictly speaking, $\underline{\underline{\mathbf{R}}}_{\text{right}}(l)$ and $\underline{\underline{\mathbf{R}}}_{\text{left}}(l)$ are conditionally independent of each other given $\underline{\underline{\mathbf{d}}}(\theta_i)$, when $\boldsymbol{V}_{\text{left}}$ and $\boldsymbol{V}_{\text{right}}$ are independent of each other ($\boldsymbol{V}_{\text{left}}$ and $\boldsymbol{V}_{\text{right}}$ represent noise signals received by microphones of the left HA and the right HA, respectively). Clearly, in cases where $\boldsymbol{V}_{\text{left}}$ and $\boldsymbol{V}_{\text{right}}$ contain only internal microphone noise, i.e. when ambient noise is essentially absent, $\boldsymbol{V}_{\text{left}}$ and $\boldsymbol{V}_{\text{right}}$ can be assumed independent of each other. Moreover, when the noise field can be approximated as isotropic, an approximation which is often used to model commonly encountered reverberation [30], $\boldsymbol{V}_{\text{left}}$ and $\boldsymbol{V}_{\text{right}}$ can be considered independent, especially for frequencies higher than approximately 600 Hz, e.g. [31].

Based on Eq. (F.16), the estimate of θ is given by

$$\check{\theta} = \arg\max_{\underline{\underline{\mathbf{d}}}(\theta_i)\in\Theta} p(\underline{\underline{\mathbf{R}}}_{\text{right}}, \underline{\underline{\mathbf{R}}}_{\text{left}}; \underline{\underline{\mathbf{d}}}(\theta_i)).$$

7 Simulation Results

In this section, we assess and compare the performance of the variants of the proposed estimator with existing methods in simulation experiments. Particularly, we investigate the effects of microphone array configuration, signal-to-noise ratio (SNR), noise type and reverberation on performance.

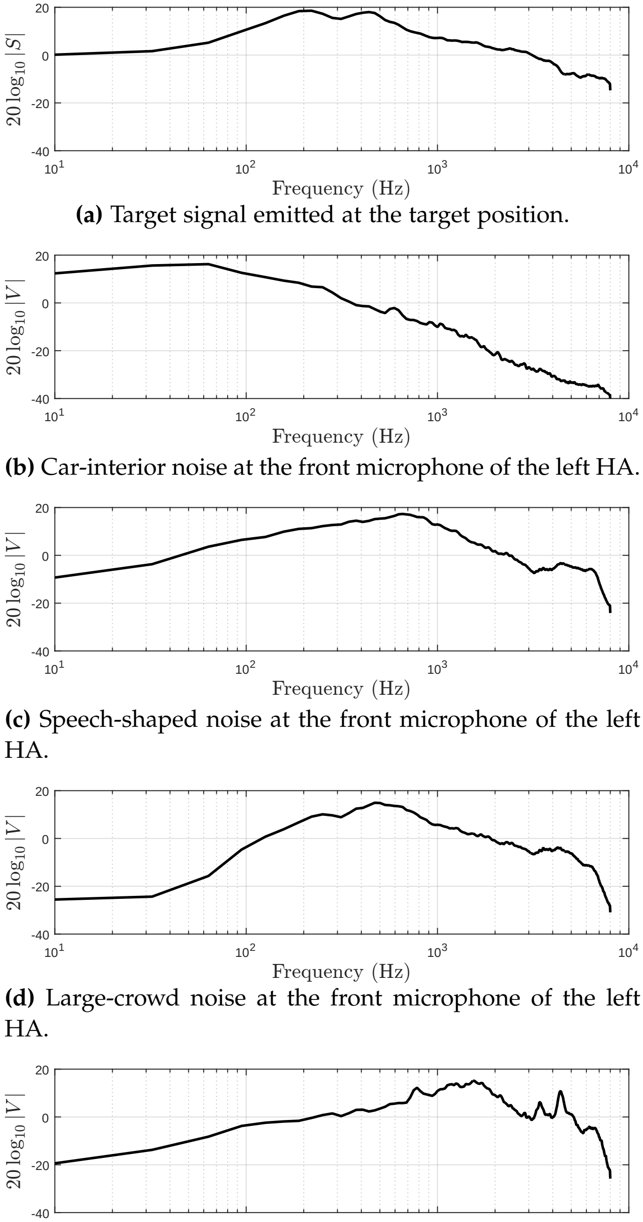
7.1 Implementation

The parameters used in all simulation experiments are as follows: the sampling frequency is 16 kHz, the DFT order $N = 512$, $w(n)$ is a Hamming window with a length of 512 samples, the DFT order N is the same as the window length (i.e. $N = 512$) and the decimation factor $A = \frac{N}{2}$. Moreover, to evaluate the likelihood functions, the noise CPSD matrix $\mathbf{C}_v(l,k)$ must be known. To estimate $\mathbf{C}_v(l,k)$, we use the same procedure as in [3]. Briefly, this method uses $S(l,k)$, which is available at the HAS, as a voice activity detector (VAD) to determine the time-frequency regions in $\boldsymbol{R}(l,k)$ where the target speech is essentially absent. Based on these noise-dominant regions, we adaptively estimate $\mathbf{C}_v(l,k)$ via recursive averaging [3].

7.2 Acoustic setup

To simulate realistic acoustic scenarios, we use the database of head related impulse responses (HRIRs) and binaural room impulse responses (BRIRs), made available by [32]. We only use the part of the database, which corresponds to the horizontal plane, $\Theta = \{-175^\circ, -170^\circ, -165^\circ, \cdots, +180^\circ\}$, and which is measured with behind-the-ear (BTE) hearing aids mounted behind the ears of a head-and-torso simulator (HATS). To generate a signal from a desired position, we convolve the signal with the corresponding impulse response.

Similar to [3], as a target signal, a four-minute speech signal is used consisting of two male and two female voices from the TSP database [33]. To assess the performance of the estimator in different noisy situations, we consider four different noise fields: car-interior noise, speech-shaped noise, large-crowd noise, and bottling-factory-hall noise. These noise fields cover noisy situations with different characteristics [3], e.g. with low-frequency content (the car-interior noise), with high-frequency content (the bottling-factory-hall noise), statistically stationary (the speech-shaped noise) and statistically non-stationary (the large-crowd noise). To generate the large-crowd noise field, the speech-shaped noise field and the bottling-factory-hall noise field, different realizations of the considered noise signals are played back simultaneously from spatial positions, which are uniformly distributed on a circle in the horizontal plane centered at the HATS. Further, the car-interior noise field was measured binaurally by BTE hearing aids mounted behind the ears of a HATS placed on the passenger seat of a car driving in a city. Fig. F.2 shows the long-term power spectrum of the target signal measured at the target position and the noise signals received at the front microphone of the left hearing aid. The wide-band SNR reported for each simulation experiment is defined in terms of the signals of the front microphone of the left HA.



(a) Target signal emitted at the target position.

(b) Car-interior noise at the front microphone of the left HA.

(c) Speech-shaped noise at the front microphone of the left HA.

(d) Large-crowd noise at the front microphone of the left HA.

(e) Bottling-factory-hall noise at the front microphone of the left HA.

Fig. F.2: Long-term power spectrum of the signals [3].

7.3 Performance metric

To measure performance of estimators, we use the absolute error (AE) metric, given by:

$$\mathrm{AE}_l = f\{\theta_l - \hat{\theta}_j\}, \tag{F.17}$$

where θ_l and $\hat{\theta}_l$ are the true DoA and the estimated DoA at time frame l, respectively. The function $f\{.\}$ is a circular wrapping function that gives the absolute error and guarantees that the error is smaller than the maximum possible error, i.e. $180°$.

To report the results, we generally use box plots [34], where the bottom and top of a box are the first and third quartiles, and the band inside a box is the median of the results. Moreover, we represent the mean absolute error (MAE) of the results with a circle on the box plots. It should be noted that we only report the results of the target-active frames. This reflects performance that is achievable in practical systems, because an accurate VAD is available in informed situations.

7.4 Competing methods

To compare the performance of the proposed estimator with existing methods, we consider the estimators proposed in [3] and [14], which we refer to them as the *measured-RTF-based method* and *Braun's method*, respectively.

The *measured-RTF-based method* is an "informed" estimator. We consider this particular estimator amongst "informed" estimators, because the *measured-RTF-based method* is the most recent one, which, as reported in [3], performs more accurately and more robustly than other "informed" estimators proposed in [16, 19–21]. Further, the proposed method and *measured-RTF-based method* both are based on a database of measured RTFs; however, the proposed method is bias-compensated, employs a more realistic model of the acoustic transfer function between the target and the "sunny" microphone, and is computationally cheaper than the *measured-RTF-based method*.

Braun's method is a narrow-band "uninformed" DoA estimator, which also uses a database of measured RTFs, for hearing applications. Comparing the performance of the proposed method with *Braun's method* shows the advantage of having access to the noise-free target signal. To make the implementation choices clear, in the following, we briefly explain *Braun's method*.

Braun's method

With respect to the signal model presented in Eq. (F.2), it has been shown that the minimum mean squared error (MMSE) estimator of the RTF between the two microphones at a particular frequency bin is given by [14]:

$$\hat{d}_{i,j}(k) = \frac{\phi_{R_{i,j}} - \phi_{V_{i,j}}}{\phi_{R_{j,j}} - \phi_{V_{j,j}}},$$

where i and j are microphone indexes, $\phi_{R_{i,j}} = \mathrm{E}\{R_i(l,k)R_j^*(l,k)\}$ and $\phi_{V_{i,j}} = \mathrm{E}\{V_i(l,k)V_j^*(l,k)\}$. Averaging the RTF estimate over the microphone index permutation, i.e.

$$\bar{d}_{i,j}(k) = \frac{1}{2}\left\{\hat{d}_{i,j}(k) + \hat{d}_{j,i}^{-1}(k)\right\},$$

makes the estimate more robust [14]. *Braun's method* estimates the DoA θ of the target signal at a particular frequency bin using a database Θ of the measured-RTFs labeled by their corresponding DoA by

$$\hat{\theta}_{\text{Braun}} = \underset{\underline{\mathbf{d}}(\theta_u)\in\Theta}{\arg\min} \sum_{i,j\in\mathcal{M}} W_{i,j}(k)|\bar{d}_{i,j}(k) - d_{i,j}(k,\theta_u)|,$$

where the set $\mathcal{M}$ contains all microphone pair combinations, and $W_{i,j}$ is a weighting factor for the $\{i,j\}$-th pair. In [14], the performance of the DoA estimator, which uses $W_{i,j}(k) = 1$, was compared with the performance of the DoA estimator, which uses a weighting function based on coherent-to-diffuse ratios. The simulation results in [14] show that both estimators perform similarly. Therefore, we consider $W_{i,j}(k) = 1$ in the implementation of *Braun's method*. Moreover, because the target source is located at the same position for all frequency bins, we change the cost function as follows, to combine the information of all frequency bins:

$$\hat{\theta}_{\text{Braun}} = \underset{\underline{\mathbf{d}}(\theta_u)\in\Theta}{\arg\min} \sum_{k=0}^{N-1} |\bar{d}_{i,j}(k) - d(k,\theta_u)|.$$

To estimate $\phi_{R_{i,j}}$ used in *Braun's method*, a recursive averaging technique with a time constant of 50 ms was used, as proposed in [14]. Moreover, to estimate $\phi_{V_{i,j}}$, we use the recursive estimate of $\mathbf{C}_v$ as described in Sec. 7.1.

7.5 Results and discussion

Biased vs bias-compensated log-likelihoods

In this part, we numerically compare the biased log-likelihood function $\tilde{\mathcal{L}}$ proposed in Eq. (F.10) with the bias-compensated log-likelihood function $\bar{\mathcal{L}}$ proposed in Eq. (F.15). To do so, we consider a situation where the true DoA is 0°, and the SNR is -100 dB in a large-crowd noise field.

As mentioned in Sec. 5, at very low SNRs, ideally, we would expect all DoAs in Θ to be equally likely. Fig. F.3 shows the normalized log-likelihood

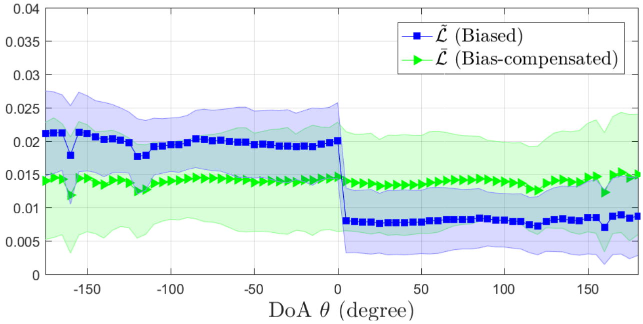


Fig. F.3: Biased vs. bias-compensated log-likelihood at SNR of −100 dB in a large-crowd noise field and in an anechoic situation. The log-likelihoods are evaluated using the signals of four microphones (two microphones in each HA) in a binaural configuration. The log-likelihood values are normalized by the sum of the log-likelihood values of all entries in Θ, and are averaged over all signal frames. The shaded areas represent the standard deviations. The true DoA is 0°.

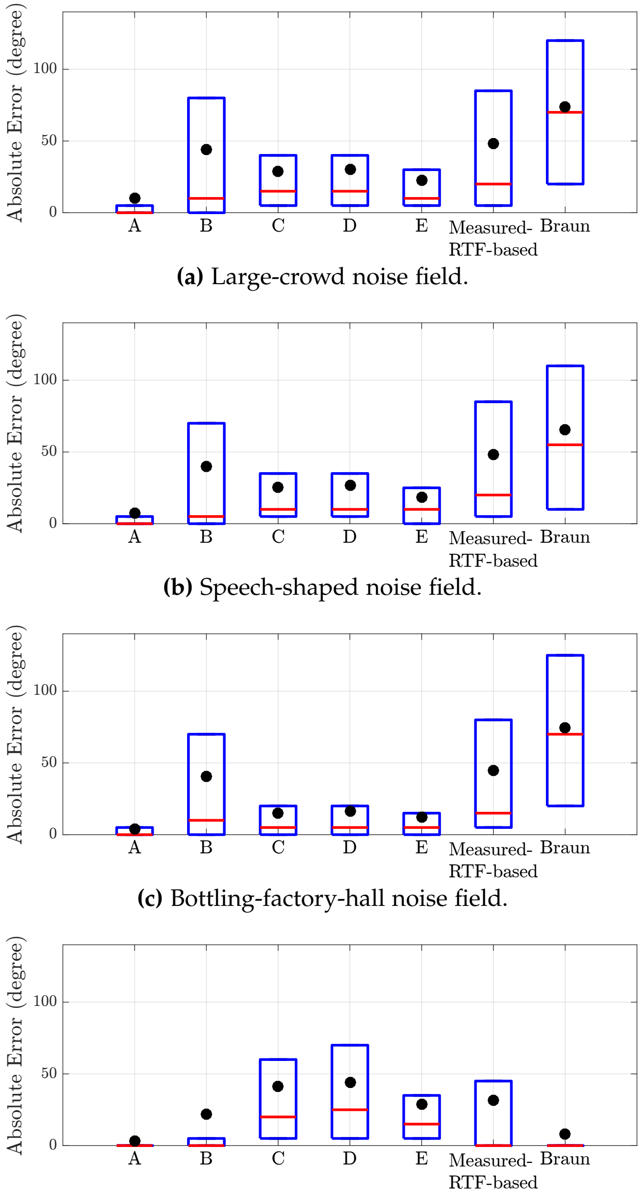
values as a function of $\underline{\mathbf{d}}(\theta_i) \in \Theta$. As Fig. F.3 shows, the uncompensated log-likelihood function $\tilde{\bar{\mathcal{L}}}$ (Eq. (F.10)) is biased towards directions to the left side of the user, i.e. $\theta_i \in [-175°, 0°]$. In contrast, the bias-compensated log-likelihood function $\bar{\mathcal{L}}$ (Eq. (F.15)) is essentially uniformly distributed across DoAs. Therefore, for the remaining simulation results, we only consider the bias-compensated version of the proposed method.

Influence of the microphone array geometry

In general, the microphone array configuration influences the performance of the DoA estimators [35]. In this part, we investigate the influence of different microphone array configurations on the performance of the proposed method (cf. Table F.1). These configurations require different degrees of wireless communication between the HAs. More precisely, for each time frame, the monaural configurations, indicated by ‘C’ and ‘D’ in Table F.1, do not need any wireless information exchange between the HAs. The binaural configurations indicated by ‘A’, ‘B’ and ‘E’ need $2N$ signal samples, N signal samples and I log-likelihood values, respectively, to be transmitted between HAs (N is the window length, and I is the number of the entries in the RTF database ($N \gg I$)).

Fig. F.4 shows the performance of the proposed method based on the different configurations mentioned in Table F.1. For comparison, the performance of the *measured-RTF-based method* [3] and *Braun's method* [14] are also shown. The *measured-RTF-based method* considers a binaural configuration similar to configuration ‘B’, while *Braun's method* considers a binaural configuration similar to configuration ‘A’.

To determine whether there are any statistically significant differences



(a) Large-crowd noise field.

(b) Speech-shaped noise field.

(c) Bottling-factory-hall noise field.

(d) Car-interior noise field.

Fig. F.4: The box plot of the performance of the proposed method based on the different microphone array configurations in Table F.1. The circles represent the MAEs. The simulation experiment was done in an anechoic situation at SNR of 0 dB in different noise fields.

Table F.1: Different microphone array geometries.

Config.	Details
A	a binaural configuration using the signals of four microphones—two microphones in each HA.
B	a binaural configuration using the signals of two microphones—one microphone in each HA.
C	a monaural configuration using the signals of two microphones in the left HA.
D	a monaural configuration using the signals of two microphones in the right HA.
E	a binaural configuration using the signals of four microphones and exploiting the IF technique proposed in Sec. 6 to decrease the wireless communication.

between the MAEs of the methods, we performed a one-way analysis of variance (ANOVA) test [36, 37]. Before the test, we transformed the AEs of the methods as

$$\mathrm{AE}_j' = \log(\mathrm{AE}_j + 1), \tag{F.18}$$

to equalize the variances of the AEs of the different methods [36, 38]. As a result, the ANOVA test rejected the hypothesis that all MAEs are identical ($p < 10^{-10}$), in all the noise fields. Moreover, as a post hoc test, multiple pairwise comparison test (Tukey HSD [36]) was applied to identify the statistical differences between the MAEs of any two methods. The results of the test revealed that the MAEs of all the methods are statistically different from each other ($p < 10^{-5}$), except for the MAEs of the configurations 'C' and 'D' in the speech-shape noise field ($p = 0.23$). Therefore, we can conclude from the simulation results shown in Fig. F.4 that:

i) The performance of configuration 'A' is the best, at the cost of higher computational load and wireless exchange of two microphones' signals.

ii) While configurations 'B', 'C' and 'D' use signals of two microphones, the performance of the monaural configurations, i.e. configurations 'C' and 'D', is generally better than the performance of the binaural configuration 'B' (except for the interior-car noise field). The reason for the better performance of the monaural configurations will be explained later in this section.

iii) Configuration E, which uses the information fusion approach and needs less parameters than configurations 'A' and 'B' to be wirelessly exchanged between the HAs, performs generally better than the monaural configurations and the binaural configuration 'B'.

iv) The performance of the proposed method based on configuration 'B' is slightly better than the performance of the *measured-RTF-based method*, while the computational complexity of the *measured-RTF-based method* is higher than the proposed method.

v) The performance of the proposed method based on configuration 'A' is, as expected, significantly better than the performance of *Braun's method*. The access to the clean target signal gives the proposed method a large advantage compared with *Braun's method.*

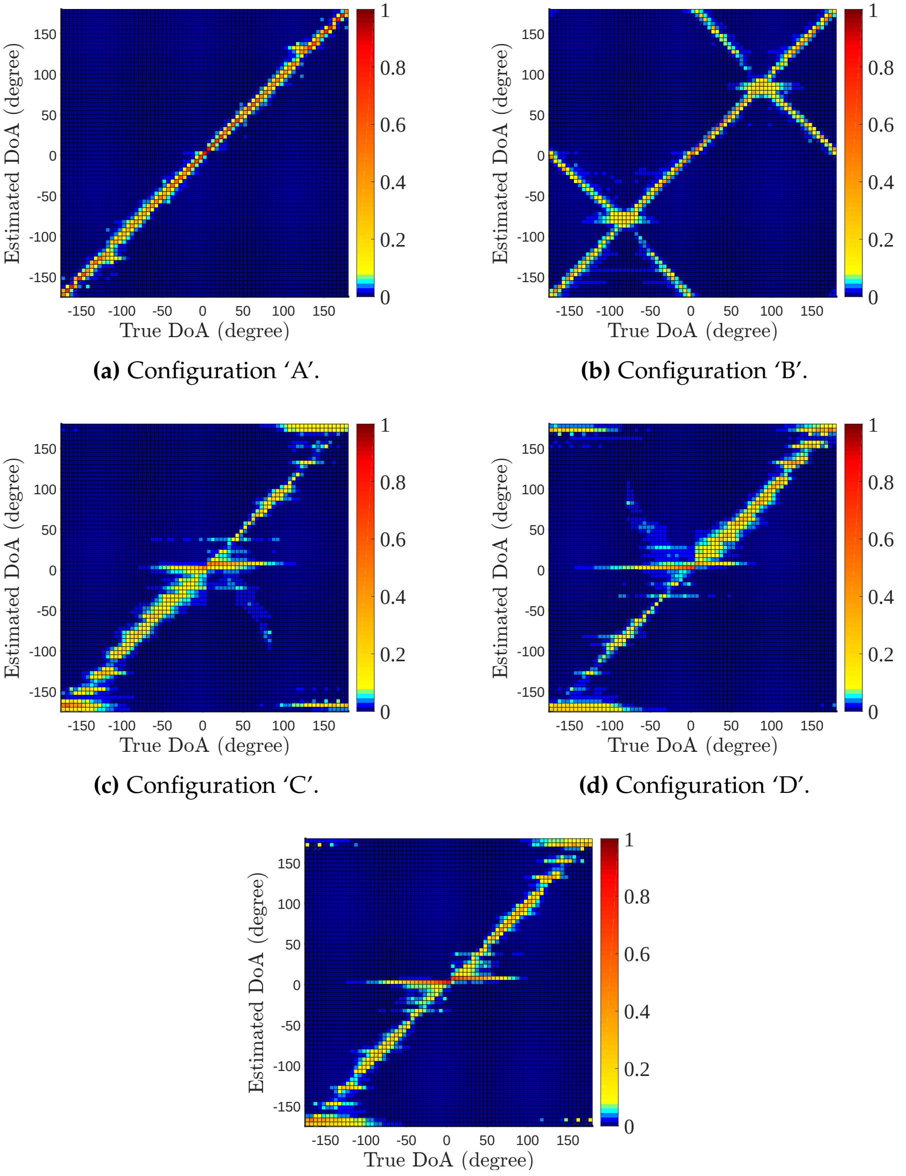
To study the performance of the proposed method in more detail, we plot in Fig. F.5 the confusion matrices of the proposed method based on different configurations. Each column of the matrices relates to a particular true DoA, and represents the normalized histogram (probability) of the estimated DoAs for that particular true DoA.

The confusion matrices depicted in Fig. F.5 demonstrate the following points:

i) In Fig. F.5a, the clear diagonal of the confusion matrix shows that the configuration 'A' is very effective in finding the true DoAs. If the proposed method based on this configuration cannot find the true DoA, the estimated DoA is generally close to the true DoA.

ii) In Fig. F.5b, the two parallel anti-diagonal lines in the confusion matrix show that the proposed method based on configuration 'B' suffers from front-back confusions, as humans do [39]. In other words, when the proposed method based on this configuration cannot find the true DoA, then the most probable choice is located on the other side of the head with respect to the axis between the considered microphones. The front-back confusions are because of the symmetry of the head with respect to the microphone array placement. The front-back confusions cause large estimation errors, especially for the DoAs located in the front or back of the HATS.

iii) Figs. F.5c and F.5d show the confusion matrices of the proposed method based on the monaural configurations. These matrices demonstrate that when the true DoA is from the same side of the head as where the HA is positioned (i.e. $\theta \in [-175°, 0°]$ in Fig. F.5c and $\theta \in [0°, 180°]$ in Fig. F.5d), the proposed method based on monaural configurations performs decently. However, when the true DoA is from the other side of the head (i.e. $\theta \in [0°, 180°]$ in Fig. F.5c and $\theta \in [-175°, 0°]$ in Fig. F.5d), the proposed method suffers somewhat from left-right confusions, especially when the true DoA is at the sides. This is partly because of head-shadowing effect.



(a) Configuration 'A'.

(b) Configuration 'B'.

(c) Configuration 'C'.

(d) Configuration 'D'.

(e) Configuration 'E'.

Fig. F.5: Confusion matrices of the proposed method based on different configurations mentioned in Table F.1. The simulation experiment conditions are the same as in Fig. F.4a.

Table F.2: Different cases considered to study the influence of the resolution of the RTF database.

Case I	for all the considered locations of the target talker, there is a representative entry in the RTF database searched by the proposed method.
Case II	there is no entry in the RTF database for every other considered DoAs, i.e. there is no entry in the database for half of the considered DoAs.
Case III	the actual distance of the target from the user is different from the target distance used in constructing the RTF database: the actual distance of the target from the user is 300 cm, while the RTF database is measured for a target distance of 80 cm. For all the DoAs, there is a representative RTF in the database.

iv) Fig. F.5e shows that when we combine the information of the monaural configurations using the IF technique proposed in Sec. 6, the left-right confusions, which occurs in monaural configurations, can largely be resolved.

v) As reported in Fig. F.4, on average, the proposed method based on configuration 'B' performs worse than the proposed method based on configurations 'C' and 'D'. The reason can be explained by comparing Fig. F.5b with Figs. F.5c and F.5d. A large part of the estimation errors is because of front-back confusions for configuration 'B' and left-right confusions for configurations 'C' and 'D'. Comparing Fig. F.5b with Figs. F.5c and F.5d demonstrates that the possibility of the front-back confusions for configuration 'B' is higher than the possibility of left-right confusions for configuration 'C' or 'D'. This is because front-back confusions for configuration 'B' may occur in the estimations of all the considered DoAs, while left-right confusions for configurations 'C' and 'D' often occur for a smaller subset of DoAs—the DoAs located in the shadow of the head.

Influence of the resolution of the RTF database

In practice, most of the time, none of the entries in the RTF database, searched by the proposed DoA estimator, represent exactly the actual DoA or distance of the target. Here, we study the performance of the proposed estimator in these situations. To do so, we compare the performance of the proposed method for three different cases outlined in Table F.2.

Fig. F.6 shows the performance of the proposed method for these cases in different configurations. In all configurations, the one-way ANOVA test on the transformed AEs (Eq. (F.18)) shows that the MAEs of different cases

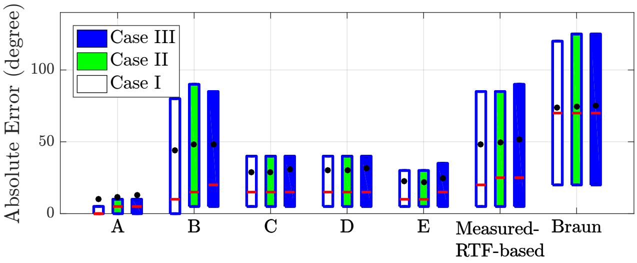


Fig. F.6: Performance of the proposed method based on different microphone array configurations (Table F.1) for different RTF-database resolutions mentioned in Table F.2. The simulation experiment was done in an anechoic situation at an SNR of 0 dB in a large-crowd noise field.

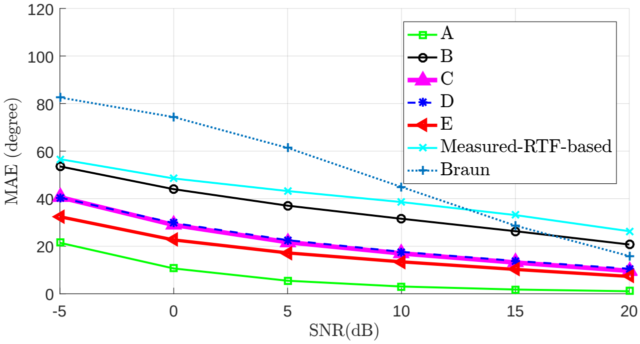


Fig. F.7: Performance of the proposed method based on the different microphone array configurations mentioned in Table F.1 for different SNRs. The simulation experiment were done in an anechoic situation in a large-crowd noise field.

are statistically different ($p < 10^{-10}$). Pairwise Tukey HSD tests for each configuration show that the MAEs of any two cases are different ($p < 10^{-9}$), except the MAEs of Case I and Case II in *Braun's method* ($p = 0.99$). Even though the MAEs of different cases are statistically different, the differences are small. This suggests that the performance of the proposed method is robust against the mismatches between the actual location of the target and the RTF database.

Influence of SNR

Another factor which generally affects the estimation performance is SNR. Fig. F.7 shows the estimation performance for different SNRs in terms of the MAE averaged over all considered DoAs in an anechoic situation in a large-crowd noise field. As expected, the estimation performance of all methods is improving by increasing the SNR. Moreover, the general performance order of Fig. F.4a remains at different SNRs; however, at high SNRs, the perfor-

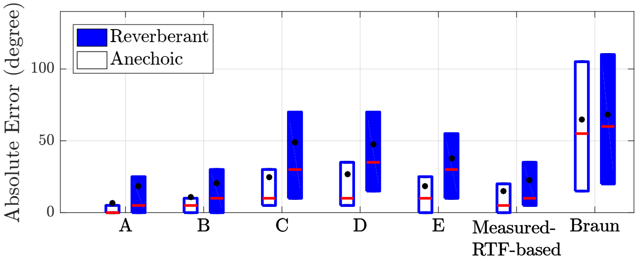


Fig. F.8: Performance of the proposed method based on the different microphone array configurations mentioned in Table F.1 in both anechoic and reverberant situations. The simulation was done in a large-crowd noise field at an SNR of 0 dB.

mance of *Braun's method*, which uses a configuration similar to configuration 'A', is improving more than the performance of the other methods. In other words, at a sufficiently high SNR, the "uninformed" *Braun's method* using four noisy microphone signals reveals more about the target location than the "informed" estimator using two microphone signals.

Influence of reverberation

In practice, HASs must operate in reverberant situations. Therefore, we investigate the impact of reverberation on the performance of the proposed estimator. To simulate a reverberant environment, we use the BRIRs measured in an office (T60 $\approx$ 500 ms) [32]. Because the BRIRs are only available for the front-horizontal half-plane, we confine the proposed method to search in the database for the DoAs related to this half-plane.

Fig. F.8 shows the performance of the different configurations of the proposed method in both anechoic and reverberant situations. For each configuration, to analyze the statistical difference of the MAEs of the anechoic and reverberant situations, a two-sample t-tests on the transformed AEs (Eq. (F.18)) has been applied. The test results show that the MAEs of the anechoic and reverberant situations are statistically different ($p < 10^{-10}$) in all configurations. We can conclude that performance of all the methods is degraded in the reverberant situation, compared with the performance in the anechoic situation. This is because none of the methods directly consider and model the reverberation. One way to explicitly take the reverberation into account is to model it as a highly time-varying isotropic noise field, e.g. [30, 40].

Another point, which is visible by comparing Fig. F.8 with Fig. F.4a is that the performance of the proposed method based on configuration 'B' for the anechoic situation has been improved significantly. This is because the RTF database has been confined to the DoAs related to the front-horizontal half-plane. This restriction prevents front-back confusions and improves markedly the performance of the proposed method based on configuration 'B'. Obviously, in many practical situations, the target signal may not be safely assumed to be originating from the frontal-horizontal half-plane, in which case performance similar to that of Fig. F.4a must be expected.

Influence of non-individualized RTFs

So far, we have presented performance, when the RTF database measured for the specific user is available. In practice, measuring the RTFs for each user might be difficult; however, it is often possible to measure the RTFs of a HATS beforehand. Therefore, in this part, we aim to compare the performance of the estimators for two different cases:

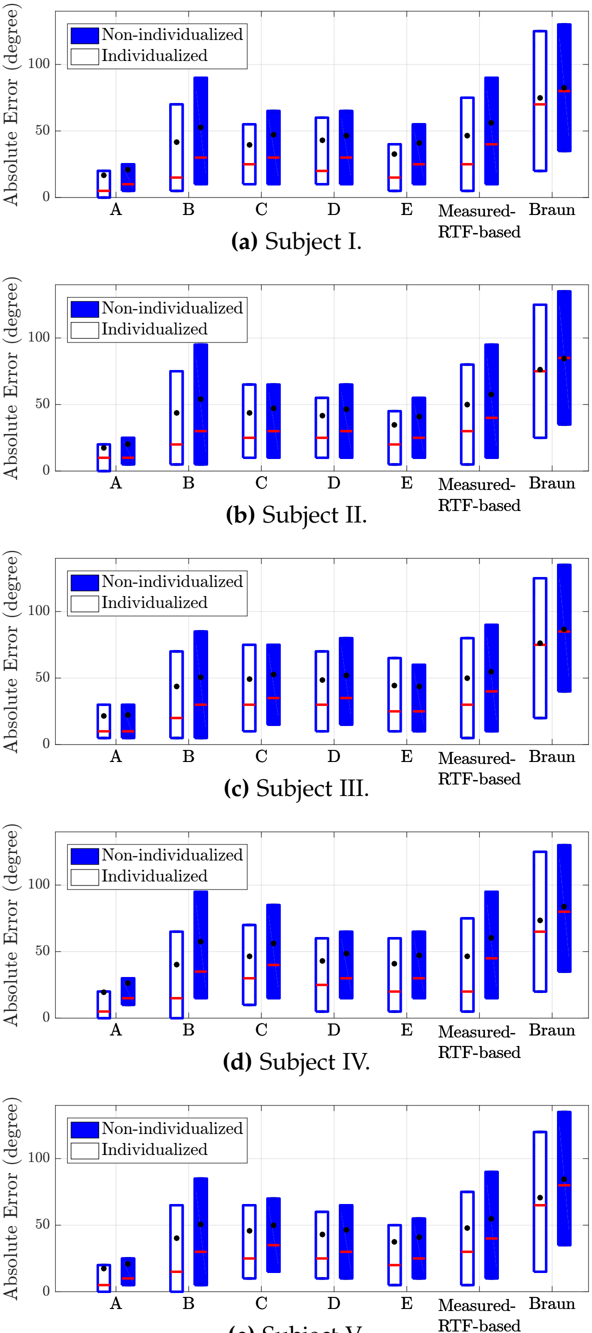
1. Individualized: user-specific databases are available.
2. Non-individualized: user-specific databases are not available; however, an RTF database measured for a HATS is available.

For the simulation experiment, we use the HRTFs measured by binaural BTE hearing aids for five different persons (three males and two females) and a HATS.

Fig. F.9 shows the performance of the methods for different subjects at an SNR of 0 dB in the large-crowd noise field. For each subject and for each configuration, to determine the statistical differences between the MAEs of the individualized and non-individualized cases, we applied a two-sample t-tests on the transformed AEs (Eq. (F.18)). The results show that the MAEs of the individualized and non-individualized cases for all subjects and in all configurations are statistically different ($p < 10^{-10}$). However, as can be seen in Fig. F.9, the performance of the proposed method degrades relatively slightly in the absence of the individualized databases. This means the performance of the proposed method appears to be robust to inaccuracies in the RTF database.

8 Conclusion

In this paper, we proposed a target source DoA estimator for a hearing aid system (HAS) which has access to the noise-free target signal via a wireless microphone. The proposed method is based on a pre-measured database of relative transfer functions (RTFs). Each measured RTF entry in the database has been labeled by its corresponding DoA, and the proposed method uses a maximum likelihood approach to find the RTF entry, which maximizes the likelihood of the received signals. The label of the RTF entry is considered as the estimate of the DoA. Moreover, we analytically investigated the bias of the estimator, and proposed an estimator which has been compensated



(a) Subject I.

(b) Subject II.

(c) Subject III.

(d) Subject IV.

(e) Subject V.

Fig. F.9: Influence of non-individualized databases on the DoA estimators for five different subjects. The SNR is 0 dB in the large-crowd noise field. The MAE is averaged over all considered DoAs.

for the bias. We showed that the proposed estimator is computationally cheaper and performs better than other recent RTF-based DoA estimators. The proposed method supports any microphone array configurations, with $M \geq 2$ microphones, both monaural and binaural. Our simulation experiments for hearing aid applications suggest that the binaural configuration using four microphones (two microphones in each hearing aid (HA)) provides the best performance at the cost of higher computational complexity and wireless communication, while the monaural configurations using two microphones suffer from left-right confusions, and the binaural configuration using two microphones (one microphone in each HA) suffers from front-back confusions. To decrease the number of parameters required to be wirelessly exchanged between the HAs in binaural configurations, where the number of microphones in each HA is at least two, we proposed an information fusion technique, which avoids transmitting microphones' signals between the HAs. An important benefit of the proposed IF strategy is that the number of parameters required to be exchanged between the HAs is independent of the number of microphones in the HA.

As a topic of future research, we aim to extend the proposed estimator to take temporal characteristics of the acoustic scene into accounts, e.g. by modeling and tracking the relative movement of the user's head and the target source.

References

[1] A. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT press, 1994.

[2] A. Bayat, M. Farhadi, A. Pourbakht, H. Sadjedi, H. Emamdjomeh, M. Kamali, and G. Mirmomeni, "A comparison of auditory perception in hearing-impaired and normal-hearing listeners: an auditory scene analysis study," *Iranian Red Crescent Medical Journal*, vol. 15, no. 11, 2013.

[3] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Informed sound source localization using relative transfer functions for hearing aid applications," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 611–623, March 2017.

[4] J. C. Murray, H. Erwin, and S. Wermter, "Robotics sound-source localization and tracking using interaural time difference and cross-correlation," in *Proceedings of NeuroBotics Workshop*, 2004, pp. 89–97.

[5] J. M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in

Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, vol. 2, 2003, pp. 1228–1233.

[6] J. A. MacDonald, "A localization algorithm based on head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4290–4296, 2008.

[7] F. keyrouz, "Advanced binaural sound localization in 3-D for humanoid robots," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 9, pp. 2098–2107, Sept 2014.

[8] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1997, pp. 187–190.

[9] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[10] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[11] J. Stachurski, L. Netsch, and R. Cole, "Sound source localization for video surveillance camera," in *IEEE International Conference on Advanced Video and Signal Based Surveillance*, Aug 2013, pp. 93–98.

[12] J. Kotus, K. Lopatka, and A. Czyzewski, "Detection and localization of selected acoustic events in acoustic field for smart surveillance applications," *Multimedia Tools and Applications*, vol. 68, no. 1, pp. 5–21, 2014.

[13] M. Zohourian and R. Martin, "Binaural speaker localization and separation based on a joint ITD/ILD model and head movement tracking," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2016, pp. 430–434.

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

[15] S. Goetze, T. Rohdenburg, V. Hohmann, B. Kollmeier, and K.-D. Kammeyer, "Direction of arrival estimation based on the dual delay line approach for binaural hearing aid microphone arrays," in *International*

Symposium on Intelligent Signal Processing and Communication Systems, Nov 2007, pp. 84–87.

[16] G. Courtois, P. Marmaroli, M. Lindberg, Y. Oesch, and W. Balande, "Implementation of a binaural localization algorithm in hearing aids: specifications and achievable solutions," in *Audio Engineering Society Convention 136*, April 2014, paper 9034.

[17] J. M. Kates, K. H. Arehart, and R. K. Muralimanohar, "Improving externalization in remote microphone systems," in *Poster presented at International Hearing Aid Research Conference*, Aug 2016.

[18] J. Jensen, M. S. Pedersen, M. Farmani, and P. Minnaar, "Hearing system," U.S. Patent 20 160 112 811, April 21, 2016.

[19] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Maximum likelihood approach to "informed" sound source localization for hearing aid applications," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2015, pp. 16–20.

[20] ——, "Informed TDoA-based direction of arrival estimation for hearing aid applications," in *Proceedings of IEEE Global Conference on Signal and Information Processing*, 2015, pp. 953–957.

[21] ——, "Informed direction of arrival estimation using a spherical-head model for hearing aid applications," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2016, pp. 360–364.

[22] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space," in *Audio Engineering Society Convention 107*, 1999, paper 5026.

[23] 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.

[24] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *The Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048–3058, 1998.

[25] Y. Avargel, "Linear system identification in the short-time Fourier transform domain," Ph.D. dissertation, Israel Institute of Technology, 2008.

[26] 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.

[27] D. R. Brillinger, *Time Series: Data Analysis and Theory*. Society for Industrial and Applied Mathematics (SIAM), 2001.

[28] P. Duhamel and M. Vetterli, "Fast fourier transforms: a tutorial review and a state of the art," *Signal processing*, vol. 19, no. 4, pp. 259–299, 1990.

[29] A. Papoulis and S. Pillai, *Probability, Random Variables, and Stochastic Processes*, ser. McGraw-Hill series in electrical engineering: Communications and signal processing. New York: McGraw-Hill, 2002.

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

[31] M. Jeub and P. Vary, "Binaural dereverberation based on a dual-channel wiener filter with optimized noise field coherence," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2010, pp. 4710–4713.

[32] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 298605, pp. 1–10, 2009.

[33] P. Kabal, "TSP speech database," Department of Electrical and Computer Engineering, McGill University, Tech. Rep., 2002.

[34] D. F. Williamson, R. A. Parker, and J. S. Kendrick, "The box plot: a simple visual method to interpret data," *Annals of internal medicine*, vol. 110, no. 11, pp. 916–921, 1989.

[35] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "On the influence of microphone array geometry on HRTF-based sound source localization," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2015, pp. 439–443.

[36] G. E. P. Box, J. S. Hunter, and W. G. Hunter, *Statistics for experimenters: design, innovation, and discovery*, ser. Wiley series in probability and statistics. Wiley-Interscience, 2005.

[37] M. W. Fagerland, "t-tests, non-parametric tests, and large studies—a paradox of statistical practice?" *BMC Medical Research Methodology*, vol. 12, no. 78, pp. 1–7, 2012.

[38] L. Onyiah, *Design and Analysis of Experiments: Classical and Regression Approaches with SAS*, ser. Statistics: A Series of Textbooks and Monographs. CRC Press, 2008.

[39] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, 1997.

[40] S. Braun and E. A. 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. 34, pp. 1–14, 2015.

Paper G

TDoA-based self-calibration of dual-microphone arrays

Mojtaba Farmani, Richard Heuasdents, Michael Syskind Pedersen, and Jesper Jensen

The paper has been published in the
Proceedings of European Signal Processing Conference, pp. 617–621, 2016.

The layout has been revised.

Abstract

We consider the problem of determining the relative position of dual-microphone subarrays. The proposed solution is mainly developed for binaural hearing aid systems (HASs), where each hearing aid (HA) in the HAS has two microphones at a known distance from each other. However, the proposed algorithm can effortlessly be applied to acoustic sensor network applications. In contrast to most state-of-the-art calibration algorithms, which model the calibration problem as a non-linear problem resulting in high computational complexity, we model the calibration problem as a simple linear system of equations by utilizing a far-field assumption. The proposed model is based on target signals time-difference-of-arrivals (TDoAs) between the HAS microphones. Working with TDoAs avoids clock synchronization between sound sources and microphones, and target signals need not be known beforehand. To solve the calibration problem, we propose a least squares estimator which is simple and does not need any probabilistic assumptions about the observed signals.

1 Introduction

Performance of many signal processing algorithms using microphone arrays depends on the knowledge of the microphone array geometry. For example, in [1, 2], the microphone array geometry is needed to estimate the direction of arrival (DoA) of the target sound for a binaural hearing aid system (HAS). A binaural HAS consists of two hearing aids (HAs) mounted on the ears of a user. Different heads radii and varying shapes of pinnae of users cause uncertainties about the geometry of the microphone array, e.g. the distance between the HAs, which degrade performance of the DoA estimation algorithms.

The microphone array calibration problem is the problem of determining the relative locations of the microphones in a microphone array. This problem has been studied using different types of measurements such as received signal strength (RSS) [3], time-of-arrival (ToA) [4–6], and time-difference-of-arrival (TDoA) [7]. Among these, TDoA is a suitable choice for HAS applications because it is less vulnerable to reverberation [4], does not require clock-synchronization between sources and microphones, and does not require the time of emission of the target signals.

Different techniques have been proposed to solve the calibration problem. Multi-dimensional scaling (MDS) [8] is one of the earliest methods that implicitly needs each node (HA) to be a compound of a microphone and a sound source, a requirement which in general is not satisfied in HA applications. Another approach has been proposed in [9] based on singular value decomposition (SVD) that finds the coordinates of the microphones up to an invertible matrix by assuming that sources are in the far-field. Finding

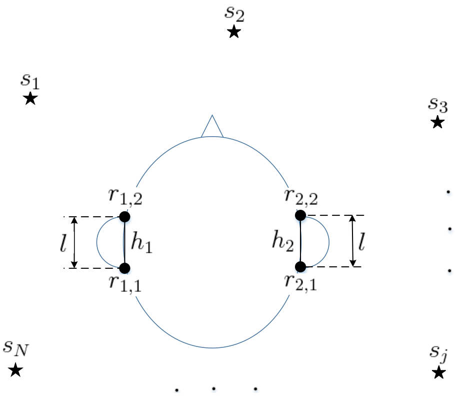


Fig. G.1: A typical scenario of microphone array calibration problem for a binaural HAS. We aim to find the relative locations of h_1 and h_2 using signals received from sound sources $s_1, s_2, ..., s_N$ which are distributed randomly around the user.

the appropriate invertible matrix is a non-linear optimization problem [9], which might be trapped in local minima. An SVD-based approach has also been proposed in [10], which avoids the far-field assumption but requires co-location of one of the sources and one of the microphones for a closed-form solution. Recently, an alternative approach was proposed [11] that solves the localization problem for a minimal case, where minimal number of microphones and sound sources are required to solve the problem, without imposing any co-location constraint. However, for overdetermined cases, where more sound sources or microphones than the minimal case are available, an additional non-linear optimization is still required. In [12] a closed-form solution has been proposed for an overdetermined case based on ToA measurements, for which synchronization of sources and microphones is needed. Lately, a new approach has been proposed [6] where pairs of microphones are set on a rigid rack, similar to the problem considered in this paper. However, the approach in [6] is based on ToA measurements which are not suitable for HAS applications.

Fig. G.1 shows an exemplary scenario of the problem considered in this paper. There are two HAs h_k, $k = 1, 2$, each with two microphones $r_{k,1}$ and $r_{k,2}$. The distance l between $r_{k,1}$ and $r_{k,2}$ is known, but the relative locations of h_1 and h_2 are unknown (we define the location of h_k as the center of its microphones axis). We aim to find the relative locations of h_1 and h_2 using the signals received by the HAs microphones from N sound sources $s_1, s_2, ..., s_N$. We assume that N is known and, at each time frame, exactly one sound source is active. This assumption is reasonable in HA applications, because when the HAS user moves his/her head, the relative location of a sound source with respect to the microphone array will change, which can be interpreted as a new sound source originating from a different relative location. Therefore, the user's head movements ensure sound signals from several different relative locations as needed.

The main contribution of this paper is in modeling the microphone array calibration problem as a linear system by utilizing a special far-field assumption. The proposed model is based on target signals TDoAs, which do not need clock synchronization between sound sources and microphones, and knowledge of target signals is not necessary. The latter point means that special calibration signals are unnecessary, and we can use signals which are naturally present, e.g. speech signals, for the calibration. To solve the modeled calibration problem, we use a least squares (LS) estimator, which additionally provides estimates of the sound sources locations. The proposed method effectively exploits the extra information about the microphones distance in a HA and needs only two sources when considering the horizontal plane, i.e. two dimensions. For simplicity, we will discuss our estimator in 2D. However, the generalization to three dimensions is straightforward.

2 Problem Formulation

Let $t_{k,i,j}$ denote the ToA of the target signal generated by source s_j received at receiver $r_{k,i}$ (microphone $i \in \{1,2\}$ of hearing aid $h_k \in \{h_1, h_2\}$), which is given by

$$t_{k,i,j} = \frac{\|r_{k,i} - s_j\|_2}{c} + t_j + \delta_{k,i}, \tag{G.1}$$

where $\|.\|_2$ denotes the Euclidean norm, c is the sound speed, t_j is the emission time at source j, and δ_i is the internal delay of microphone $r_{k,i}$. If we assume that the internal delays of the HAS microphones are equal, i.e. $\delta_{k,i} = \delta$ for all i and k, the TDoA of the target signal generated by source j received at $r_{k,i}$ and $r_{u,w}$ (microphone $w \in \{1,2\}$ of hearing aid $h_u \in \{h_1, h_2\}$) is

$$\Delta_{k,i,u,w,j} = t_{k,i,j} - t_{u,w,j} = \frac{\|r_{k,i} - s_j\|_2}{c} - \frac{\|r_{u,w} - s_j\|_2}{c}.$$

Hence, the TDoA depends only on the locations of the sources and the receivers, and it is independent of the δ and t_js. In the following, we will estimate the relative locations of the HAs using TDoAs and a special far-field assumption.

2.1 Far-field assumption

Let $d_{k,j}$ denote the distance between s_j and h_k. In HAS applications, the $d_{k,j}$s are usually much larger than the microphones distance within a HA, i.e. $d_{k,j} \gg l$. Therefore, we can assume that the DoAs of the target sounds

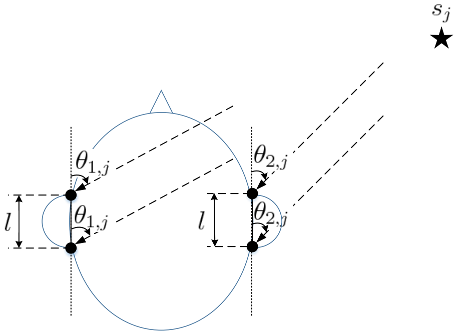


Fig. G.2: The special far-field assumption considered in this paper.

for the microphones of a HA are almost equal (see Fig. G.2). However, we assume the target distances are not much larger than the diameters of the user's head, which means $\theta_{1,j}$ and $\theta_{2,j}$ are not necessarily equal.

The far-field assumption and the given estimated TDoAs allow us to estimate $\theta_{k,j}$, $k = 1, 2$ (see Fig. G.2), up to a sign as follows:

$$\begin{aligned}\hat{\Delta}_{k,2,k,1,j} &= \frac{l}{c}\cos\left(\hat{\theta}_{k,j}\right) \\ &\Rightarrow \tilde{\theta}_{k,j} = \pm\hat{\theta}_{k,j} = \pm\arccos\left(\frac{c}{l}\hat{\Delta}_{k,2,k,1,j}\right),\end{aligned} \tag{G.2}$$

where $\hat{\Delta}_{k,2,k,1,j}$ is the estimated TDoA between $r_{k,2}$ and $r_{k,1}$ for the target signal from s_j. Note that the DoAs are expressed clockwise with respect to the microphones axis. Moreover, we define the TDoA of the target signal from s_j between midpoint of h_1 and h_2 as $\Delta_j = \frac{\hat{\Delta}_{2,1,1,1,j}+\hat{\Delta}_{2,2,1,2,j}}{2}$ to estimate $\Delta d_j = d_{2,j} - d_{1,j}$ as

$$\Delta d_j \approx \Delta_j c. \tag{G.3}$$

Therefore, there are three known parameters for each source s_j: $\tilde{\theta}_{1,j}$, $\tilde{\theta}_{2,j}$ and Δd_j, which leads to $3N$ known parameters in total. On the other hand, the locations of the sound sources, h_1 and h_2 are unknown. Without loss of generality, we will assume $h_1 = [0, 0]^{\mathrm{T}}$, and we estimate locations of h_2 and $\{s_1, ..., s_N\}$ with respect to h_1. As a consequence, we have $2N + 2$ unknown in a two-dimensional scenario, and the calibration problem is solvable when $3N \geq 2N + 2$, i.e. $N \geq 2$.

3 Localization Algorithm

In this section, we propose an algorithm to estimate the relative locations of h_1 and h_2 using the known parameters. The relation between s_j and h_k, $k = 1, 2$, can be written as

$$s_j = h_k + d_{k,j} \left[\sin(\theta_{k,j}) \quad \cos(\theta_{k,j}) \right]^{\mathrm{T}}, \tag{G.4}$$

which allows us to formulate the relative location of h_2 as

$$h_2 = \begin{bmatrix} X \\ Y \end{bmatrix} = h_1 + d_{1,j} \begin{bmatrix} \sin(\theta_{1,j}) \\ \cos(\theta_{1,j}) \end{bmatrix} - d_{2,j} \begin{bmatrix} \sin(\theta_{2,j}) \\ \cos(\theta_{2,j}) \end{bmatrix}. \tag{G.5}$$

From Eq. (G.3), we have $d_{2,j} = d_{1,j} + \Delta d_j$. Therefore,

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} d_{1,j} \sin(\theta_{1,j}) - (d_{1,j} + \Delta d_j) \sin(\theta_{2,j}) \\ d_{1,j} \cos(\theta_{1,j}) - (d_{1,j} + \Delta d_j) \cos(\theta_{2,j}) \end{bmatrix}. \tag{G.6}$$

Considering the second row of Eq. (G.6), we can express $d_{1,j}$ as a function of Y and Δd_j:

$$d_{1,j} = \frac{Y + \Delta d_j \cos(\theta_{2,j})}{\cos(\theta_{1,j}) - \cos(\theta_{2,j})}. \tag{G.7}$$

Substitution of Eq. (G.7) into the first row of Eq. (G.6) leads to:

$$\begin{bmatrix} \cos(\theta_{1,j}) - \cos(\theta_{2,j}) \\ -\sin(\theta_{1,j}) + \sin(\theta_{2,j}) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} X \\ Y \end{bmatrix} = \Delta d_j \sin(\theta_{1,j} - \theta_{2,j}), \tag{G.8}$$

and considering N sound sources together leads to a linear system of equations

$$\mathbf{A} h_2 = \boldsymbol{b}, \tag{G.9}$$

where $A \in \mathcal{R}^{N \times 2}$ and $\boldsymbol{b} \in \mathcal{R}^N$. The first and second columns of row j of $\mathbf{A}$ are $A_{j1} = \cos(\theta_{1,j}) - \cos(\theta_{2,j})$, $A_{j2} = -\sin(\theta_{1,j}) + \sin(\theta_{2,j})$ respectively, and row j of $\boldsymbol{b}$ is $b_j = \Delta d_j \sin(\theta_{1,j} - \theta_{2,j})$.

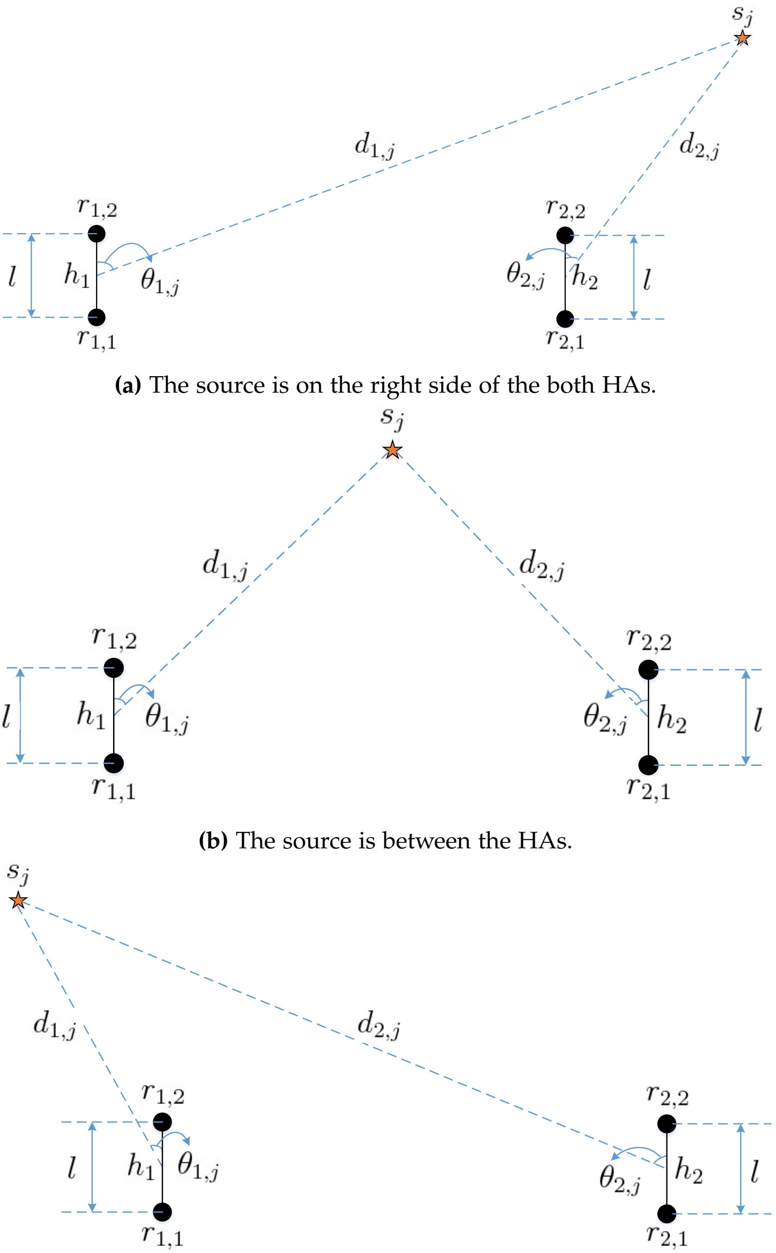
Because in practice observations are always noisy, to obtain the location of h_2 based on Eq. (G.9), we will compute a LS estimate of h_2 which is given by

$$\hat{h}_2 = \mathbf{A}^+ \boldsymbol{b}, \tag{G.10}$$

where $\mathbf{A}^+$ denotes the pseudo-inverse of $\mathbf{A}$. and straightforwardly, the LS estimators of $s_j \in \{s_1, s_2, ..., s_N\}$ can be obtained by replacing $\hat{h}_2$ in Eqs. (G.7) and (G.4), respectively.

One remaining issue is that, as showed in Sec. 2.1, we can estimate $\theta_{k,j}$ only up to a sign (see Eq. (G.2)). Therefore, for each s_j, three different cases are conceivable (see Fig. G.3):

- Case 1: s_j is on the right sides of h_1 and h_2 (Fig. G.3a), i.e. $\theta_{1,j} = +\hat{\theta}_{1,j}$ and $\theta_{2,j} = +\hat{\theta}_{2,j}$.



(a) The source is on the right side of the both HAs.

(b) The source is between the HAs.

(c) The source is on the left side of the both HAs.

Fig. G.3: Different relative locations of a sound source with respect to a binaural HAS.

- Case 2: s_j is between h_1 and h_2 (Fig. G.3b), i.e. $\theta_{1,j} = +\hat{\theta}_{1,j}$ and $\theta_{2,j} = -\hat{\theta}_{2,j}$.
- Case 3: s_j is on the left sides of h_1 and h_2 (Fig. G.3c), i.e. $\theta_{1,j} = -\hat{\theta}_{1,j}$ and $\theta_{2,j} = -\hat{\theta}_{2,j}$.

We can distinguish Case 1 and Case 3 by Δ_j:

- If $\Delta_j > 0$, the target signal reached h_1 before h_2, i.e. case 3 cannot be the case.
- If $\Delta_j < 0$, the target signal reached h_2 before h_1, i.e. case 1 cannot be the case.

However, cases 1 and 2, and cases 2 and 3 are not distinguishable from each other based on Δd_j. In other words:

$$\left[\theta_{1,j}, \theta_{2,j}\right] = \begin{cases} [\pm\hat{\theta}_{1,j}, -\hat{\theta}_{2,j}], & \text{if } \Delta d_j > 0 \\ [+\hat{\theta}_{1,j}, \pm\hat{\theta}_{2,j}], & \text{otherwise} \end{cases}. \tag{G.11}$$

Therefore, for each source, we have two different cases which cannot be distinguished based on Δd_j. To resolve this ambiguity, we solve the calibration problem for all possible combinations of different cases of the $\theta_{k,j}$s, and the combination of the cases that can justify all the estimated parameters best is the solution to the problem. Two different cases for each source result in 2^N different combinations of cases considering all sources. Therefore, the problem must be solved for 2^N different combinations of the cases, and the best combination b^* is given by:

$$b^* = \arg\min_{b \in \mathcal{B}} \sum_{j=1}^{N} \|\Delta d_j - \hat{\Delta} d_{j,b}\|_2, \tag{G.12}$$

where $\mathcal{B}$ is the set of all possible combinations of the cases, and $\hat{\Delta} d_{j,b} = \hat{d}_{2,j,b} - \hat{d}_{1,j,b}$, where $\hat{d}_{1,j,b}$ is obtained by Eq. G.7 for combination b and $\hat{d}_{2,j,b} = \|\hat{h}_{2,b} - \hat{s}_{j,b}\|_2$, ($\hat{h}_{2,b}$ and $\hat{s}_{j,b}$ denote the estimated locations of h_2 and s_j for combination b, respectively). The outputs of the localization algorithm are $\hat{h}_{2,b^*}$ and $\{\hat{s}_{1,b^*}, ..., \hat{s}_{N,b^*}\}$.

3.1 TDoA estimation

The last issue is how to estimate the TDoAs upon which the above algorithm relies. The most well-known approach for time delay estimation (TDE) is based on the Generalized Cross Correlation (GCC) method [13]: the GCC of two correlated signals has a maximum at a lag τ corresponding to the delay.

Let $r_{k,i,j}(n)$ and $r_{u,w,j}(n)$ denote the signals received from source j by microphone i of hearing aid k, and microphone w of hearing aid u, respectively. Furthermore, let $R_{k,i,j}(f)$ and $R_{u,w,j}(f)$ denote their discrete Fourier transforms (DFTs), respectively. The GCC is then given by [13]:

$$\mathcal{R}_{k,i,u,w,j}(\tau) = \sum_{f=1}^{M} \psi(f) R^*_{k,i,j}(f) R_{u,w,j}(f) e^{j2\pi f \tau}, \tag{G.13}$$

where M is the DFT order, $*$ represents complex conjugation and $\psi(.)$ is a weighting function. Then, the estimated $\Delta_{k,i,u,w,j}$ is given by:

$$\hat{\Delta}_{k,i,u,w,j} = \arg\max_{\tau} \;\; \mathcal{R}_{k,i,u,w,j}(\tau). \tag{G.14}$$

Because microphone array calibrations are usually performed in high SNR situations, we simply use the conventional cross-correlation method for TDoA estimation, i.e. $\psi(f) = 1$ for all f in Eq. (G.13). However, to improve the TDE performance in noisy situations, there are more complex weighting functions which take into account the noise characteristics [13].

Because TDoAs are estimated based on sampled signals, the estimation accuracy is limited by the sampling interval. Moreover, the small distance between the microphones of a HA limits the possible discrete TDoA values. Therefore, subsample TDE is necessary, and we need interpolation methods to tackle this problem [14, 15]. In this paper, we use the cubic spline method [16] to interpolate the microphone signals before computing the GGC.

4 Simulation Results

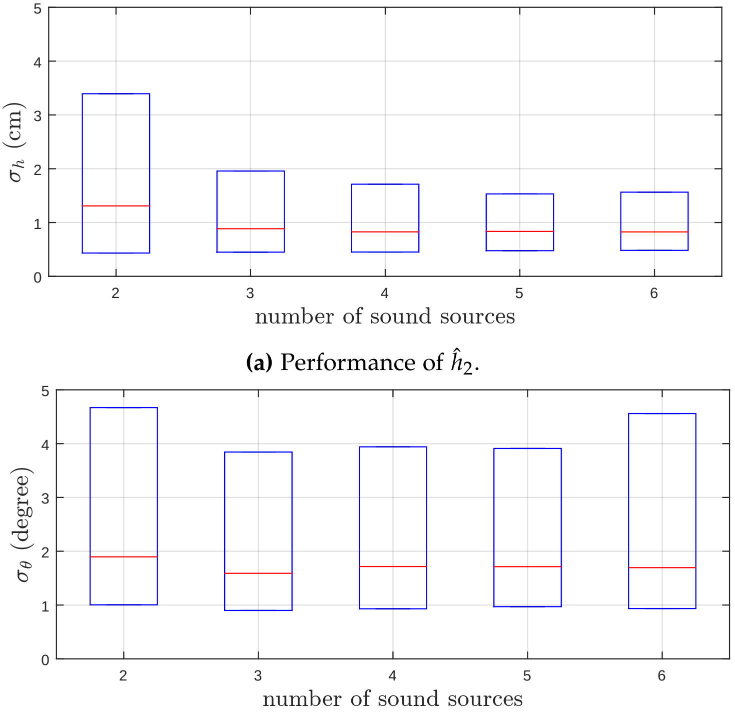
4.1 Setup

To evaluate the performance of the proposed algorithm, we consider a free-field situation, i.e. head presence is ignored in the simulations. Moreover, we set $l = 1\,\text{cm}$ and consider the head diameter, or more precisely, the distance between h_1 and h_2 to be 16 cm. We distribute the sound sources randomly according to a uniform distribution on a disc or a circle (depending on the experiment) around the user. We use the TSP database [17] for generating speech sound sources. The sampling frequency is 48 kHz, the estimation window length is 1024 samples, and we run the simulations for 200 different realizations. The number of query points for interpolation between each two consecutive sample points of the signal is 100.

4.2 Performance measures

To evaluate the estimated location of h_2, we use

$$\sigma_h = \|h_2 - \hat{h}_2\|_2, \tag{G.15}$$



(a) Performance of $\hat{h}_2$.

(b) Performance of the estimated DoAs.

Fig. G.4: The box plot of the performance of the proposed algorithm as a function of number of sound sources. The bottom and top of the boxes are the first and third quartiles, and the bands inside the boxes are the median.

where $\|.\|$ denotes the 2-norm. As another performance metric, we use the mean absolute error of the obtained DoAs:

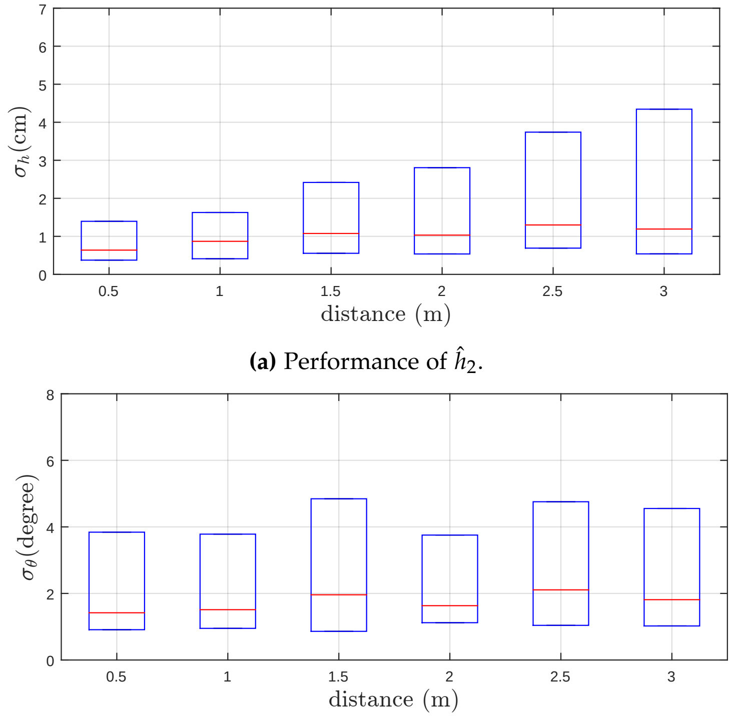
$$\sigma_\theta = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{|\theta_{1,j}-\tilde{\theta}_{1,j}|+|\theta_{2,j}-\tilde{\theta}_{2,j}|}{2}\right), \tag{G.16}$$

where $\tilde{\theta}_{1,j}$ and $\tilde{\theta}_{2,j}$ obtained from $h_1 = [0,0]^{\mathrm{T}}$, $\hat{h}_{2,b^*}$ and $\hat{s}_{j,b^*}$, and $\theta_{1,j}$ and $\theta_{2,j}$ are the true DoAs of the target signal from s_j to h_1 and h_2, respectively.

To demonstrate the results, we use box plots (Figs. G.4 and G.5), where the bottom and top of the box are the first and third quartiles, and the band inside the box is the median.

4.3 Results and discussion

The effect of the number of sound sources on the proposed algorithm has been shown in a box plot in Fig. G.4. As can be seen, increasing the number of



(a) Performance of $\hat{h}_2$.

(b) Performance of the estimated DoAs.

Fig. G.5: The box plot of the performance of the proposed algorithm as a function of the distance of the sound sources from the user.

sound sources from two to three would improve the estimation performance. However, increasing the number of the sound sources to more than three does not offer any advantages because the fundamental subsample error of the TDoA estimation cannot be overcome by increasing the number of the sound sources. Overall, the estimated medians of σ_h and σ_θ are around 1 cm and 2 degree, respectively. It should be mentioned that $d_j \in [0.5, 1.5]$ in these simulations.

Fig. G.5 shows the box plot of the proposed algorithm as a function of d_j. We distribute three sound sources randomly on a circle centered at the user's head for different distances. Generally, increasing the distance degrades the performance because the distance increment would put the sound sources in a far-field situation regarding both HAs—we modeled the problem in a way that the sound sources are in far-field with respect to each HA individually, not both HAs. Overall, as before, the estimated medians of σ_h and σ_θ are around 1 cm and 2 degree, respectively.

5 Conclusion and Future Work

In this paper, we studied the microphone array calibration problem for binaural hearing aid systems. The proposed algorithm is based on the estimated TDoAs of the target signals received by hearing aid microphones. We used a far-field assumption to model the problem as a linear system, and we proposed a least squares estimator to estimate the locations. As future work, we plan to study the proposed algorithm under more realistic situations by considering presence of the head, microphone noise and reverberation.

References

[1] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Informed TDoA-based direction of arrival estimation for hearing aid applications," in *Proceedings of IEEE Global Conference on Signal and Information Processing*, 2015, pp. 953–957.

[2] ——, "Informed direction of arrival estimation using a spherical-head model for hearing aid applications," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2016, pp. 360–364.

[3] M. Chen and others., "Energy-based position estimation of microphones and speakers for ad hoc microphone arrays," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 22–25.

[4] R. Heusdens and N. D. Gaubitch, "Time-delay estimation for TOA-based localization of multiple sensors," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2014, pp. 609–613.

[5] N. D. Gaubitch *et al.*, "Auto-localization in ad-hoc microphone arrays," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2013, pp. 106–110.

[6] S. Zhayida, S. Burgess, Y. Kuang, and K. Åström, "TOA-based self-calibration of dual-microphone array," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 791–801, Aug 2015.

[7] R. Kaune, "Accuracy studies for TDOA and TOA localization," in *Proceedings of International Conference on Information Fusion*, July 2012, pp. 408–415.

[8] S. T. Birchfield and A. Subramanya, "Microphone array position calibration by basis-point classical multidimensional scaling," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1025–1034, 2005.

[9] S. Thrun, "Affine structure from sound," in *Advances in Neural Information Processing Systems*, 2005, pp. 1353–1360.

[10] M. Crocco, A. D. Bue, and V. Murino, "A bilinear approach to the position self-calibration of multiple sensors," *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 660–673, Feb 2012.

[11] Y. Kuang, S. Burgess, A. Torstensson, and K. Åström, "A complete characterization and solution to the microphone position self-calibration problem," in *Proceedings of of IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 3875–3879.

[12] M. Pollefeys and D. Nister, "Direct computation of sound and microphone locations from time-difference-of-arrival data," in *Proceedings of of IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2008, pp. 2445–2448.

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

[14] F. Viola and W. F. Walker, "A spline-based algorithm for continuous time-delay estimation using sampled data," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 52, no. 1, pp. 80–93, 2005.

[15] X. Lai and H. Torp, "Interpolation methods for time-delay estimation using cross-correlation method for blood velocity measurement," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 46, no. 2, pp. 277–290, March 1999.

[16] C. B. Moler, *Numerical Computing with MATLAB: Revised Reprint*, ser. SIAM e-books. Society for Industrial and Applied Mathematics, 2008.

[17] P. Kabal, "TSP speech database," Department of Electrical and Computer Engineering, McGill University, Tech. Rep., 2002.

Paper H

Concurrent localization of sound sources and dual-microphone sub-arrays using ToFs

Mojtaba Farmani, Richard Heuasdents,
Michael Syskind Pedersen, Zheng-Hua Tan and Jesper Jensen

The paper has been published in the
Proceedings of International Conference on Information Fusion, pp. 1931–1936,
2016.

Abstract

In this paper, we present a localization algorithm which simultaneously estimates the locations of the target sound sources and dual-microphone sub-arrays using time-of-flight (ToF) measurements of the target signals. The proposed solution is mainly developed for a binaural hearing aid system (HAS) which consists of two hearing aids (HAs) mounted on the ears of a user. Each HA has two microphones at a known distance from each other, but the relative locations of the HAs are unknown. In this paper, we aim to find the relative locations of the HAs and the locations of the target sources. The main contribution of this paper is in modeling the localization problem as a simple linear system of equations, which markedly decreases the computation overhead in contrast to most state-of-the-art localization algorithms, which model the problem as a non-linear problem resulting in higher computational complexity. The proposed algorithm requires at least two sound sources with different locations to solve the localization problem in 2D.

1 Introduction

In many microphone array applications, knowledge of the relative positions of the microphones is required. For example, in [1, 2], estimation of the direction of arrival (DoA) of target sounds for a binaural hearing aid system (HAS) relies on the knowledge of the microphone array geometry. A binaural HAS consists of two hearing aids (HAs) mounted on the ears of a user. Different heads radii and varying shapes of pinnae of users [3] result in uncertainties about the geometry of the microphone array, which deteriorate performance of the DoA estimation algorithms.

The microphone array localization problem, also called the microphone array calibration problem, deals with determining the relative locations of the microphones in a microphone array using the signals received from some sound sources, whose positions are unknown. The state-of-the-art solutions are based on different types of measurements such as received signal strength (RSS) [4], time-of-flight (ToF, sometimes called time-of-arrival (ToA)) [5–9], time-difference-of-arrival (TDoA) [10, 11], and angle-of-arrival [12]. Among these, we use ToF measurements because they can be relatively easily estimated in the setup considered in this paper. The considered setup is an advanced Hearing Aid System (HAS) which can connect to a wireless microphone worn by a target talker (cf. Fig. H.1). The wireless microphone provides an almost noise-free content of the target signals and their time of emissions, which are necessary for the ToFs estimations.

Given the inter-node distances (obtained by multiplying the ToFs by the sound speed), different techniques have been proposed to solve the localization problem. Multi-dimensional scaling (MDS) [13] is one of the earliest

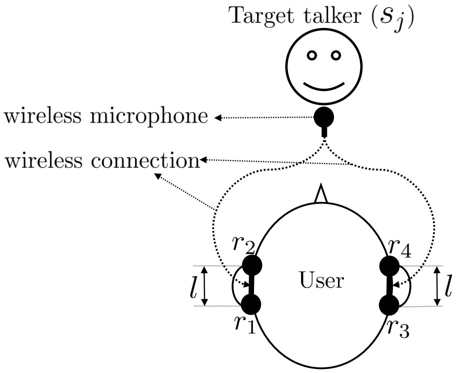


Fig. H.1: The binaural hearing aid system (HAS) considered in this paper. It consists of two hearing aids (HAs) and four microphones: r_1, r_2, r_3 and r_4. Moreover, the HAS can connect to a wireless microphone worn by a target talker. The distance l between the microphones within a HA is known, but the relative locations of the microphones are unknown. We aim to find the relative locations of $r_i, i = 1, \cdots, 4$, and the locations of the target talker over time using the signals received by the HAS microphones from the target talker.

methods that implicitly needs each node to be a compound of a microphone and a sound source, a requirement which in general is not satisfied in HA applications. Alternatively, one could consider a likelihood maximization approach [14] which leads to a non-convex optimization problem and might be trapped in local minima. A singular value decomposition (SVD)-based approach has been proposed in [15] that finds the coordinates of the microphones up to an invertible matrix by assuming that sources are in the far-field. Finding the appropriate invertible matrix is a non-linear optimization problem [15]. Recently, a closed-form solution has been proposed in [9], which avoids the far-field assumption but requires co-location of one of the sources and one of the microphones and also needs both microphones and sources to be spread on the 3D space, i.e. they do not lie on a plane; however, microphones' locations in HAS scenarios usually span a 2D plane. Recently, an alternative approach was proposed [16] that solves the localization problem for a minimal case, where minimal number of microphones and sound sources are required to solve the problem, without imposing any co-location constraint. However, for overdetermined cases, where more sound sources or microphones than the minimal case are available, an additional non-linear optimization is still required. In [8], a matrix-factorization-based solution has been proposed which needs at least five microphones and 10 sound sources in 3D and uses a non-linear least-squares approach to refine the estimations. Lately, a new approach has been proposed [7] where pairs of microphones are set on a rigid rack, similar to the problem considered in this paper. However, the approach in [7] do not tackle the problem when both the sound sources and the microphones lie on the same plane. In [10], a far-field assumption has been employed to propose a TDoA-based localization algorithm for dual-microphone sub-arrays.

The main contribution of this paper is in modeling the localization problem of the dual-microphone sub-arrays as a linear system of equations and avoiding any far-field assumptions. The proposed model is based on target signals' ToFs, which can be estimated relatively easily in the setup considered in this paper. The proposed method effectively exploits the extra information about the microphones distance in a HA and needs only two sound sources with different locations (or more precisely, two different relative positions of the target talker with respect to the microphone array) in 2D. The relative positions of the sound sources are also estimated simultaneously by the proposed solution. We will discuss our estimator in 2D because HAS scenarios usually span a 2D plane. However, the generalization to three dimensions is relatively straightforward.

2 Problem Formulation

Fig. H.1 shows the binaural HAS considered in this paper. The HAS consists of two HAs and has four microphones denoted by r_1, r_2, r_3 and r_4. Moreover, the HAS can connect to a wireless microphone worn by the target talker. The distance l between the microphones within a HA is known, but the relative locations of the microphones are unknown. We aim to find the relative locations of the HAS microphones (r_1, r_2, r_3 and r_4) using the signals received by the HAS microphones from the target talker over N different time frames: $\{s_j : j = 1, \cdots, N\}$, where s_j denotes the relative position of the target talker with respect to the microphone array at time frame j. In HAS applications, because of possible movements of the target talker and the movements of the user's head, the relative location of the target talker with respect to the microphone array will change over time. This change can be interpreted as a new sound source originating from a different relative location. We assume that at each time frame the relative location of the target talker with respect to the microphone array will change, and in this paper, we treat the different relative locations of the target talker over N time frames as N different sound sources.

The ToF of a signal is the time that it takes to travel the distance between the source and the receiver. In other words, the ToF of a signal generated by source j received at receiver i is

$$t_{i,j} = \frac{\|r_i - s_j\|_2}{c}, \tag{H.1}$$

where $\|.\|_2$ denotes the Euclidean norm, and c is the sound speed. Assuming all the clocks in the system are synchronized, $t_{i,j}$ can be estimated by

$$\hat{t}_{i,j} = \hat{q}_{i,j} - \tau_j, \tag{H.2}$$

where $\hat{q}_{i,j}$ denotes the estimated reception time (time of arrival) at receiver i of the signal generated by source j, and τ_j denotes the emission time at source j. In the HAS setup considered in this paper, $\hat{q}_{i,j}$ can be estimated by the cross-correlation between the noise-free target signal transferred via the wireless microphone and the signal received by microphone i of the HAS [17], and τ_j is provided by the wireless microphone worn by the target talker. Now, we can estimate the distances between r_i and s_j via

$$\hat{d}_{i,j} = \hat{t}_{i,j}\, c. \tag{H.3}$$

In the following, we will estimate the relative locations of the microphones and sound sources using the obtained distances.

In the localization problem considered in this paper, locations of the sound sources (the relative locations of the target talker over N time frames) and the HAS microphones are unknown. Without loss of generality, we assume $r_1 = [0, 0]^{\mathrm{T}}$ and $r_2 = [0, l]^{\mathrm{T}}$, and we estimate locations of r_3, r_4 and $\{s_j, j = 1, \cdots, N\}$ with respect to r_1. Therefore, we have $2N + 4$ unknown parameters in a two-dimensional scenario. On the other hand, for each s_j, we have four different distances between the microphones and s_j. Moreover, the distance between r_3 and r_4 is also known. Therefore, we have $4N + 1$ known parameters. Subsequently, the localization problem is solvable when $2N + 4 \leq 4N + 1$, or more precisely $N \geq 2$.

3 Localization Algorithm

In this section, we propose a self-localization algorithm to estimate the relative locations of the microphones and sound sources using $\{\hat{d}_{i,j} : i = 1, \cdots, 4; j = 1, \cdots, N\}$. Fig. H.2a shows an exemplary relative positions of the microphones and source j. As can be seen, r_1, r_2 and s_j forms a triangle which its three sides ($d_{1,j}, d_{2,j}$ and l) are almost known. Similarly, r_3, r_4 and s_j also forms another triangle which its three sides ($d_{3,j}, d_{4,j}$ and l) are almost known. Therefore, using the law of cosines [18], we can estimate $\theta_{1,j}, \theta_{3,j}$ and $\theta_{4,j}$ up to a sign by:

$$\tilde{\theta}_{1,j} = \pm\hat{\theta}_{1,j} = \pm \arccos\left(\frac{l^2 + \hat{d}^2_{1,j} - \hat{d}^2_{2,j}}{2\, l\, \hat{d}_{1,j}}\right), \tag{H.4}$$

$$\tilde{\theta}_{3,j} = \pm\hat{\theta}_{3,j} = \pm \arccos\left(\frac{l^2 + \hat{d}^2_{3,j} - \hat{d}^2_{4,j}}{2\, l\, \hat{d}_{3,j}}\right), \tag{H.5}$$

$$\tilde{\theta}_{4,j} = \pm\hat{\theta}_{4,j} = \pm\pi \mp \arccos\left(\frac{l^2 + \hat{d}^2_{4,j} - \hat{d}^2_{3,j}}{2\, l\, \hat{d}_{4,j}}\right). \tag{H.6}$$

It should be mentioned that the input argument of arccos(.) should be in the interval of $[-1, 1]$, and if the argument lies outside this interval because of the estimation errors, we will truncate the argument. Moreover, $\theta_{i,j}$s are expressed clockwise with respect to the microphones axis.

Let $\hat{r}_{3,j}$ and $\hat{r}_{4,j}$ denote the estimated relative location of r_3 and r_4 with respect to r_1 using the estimated distances from source j, respectively. Regarding Fig. H.2a, $\hat{r}_{3,j}$ and $\hat{r}_{4,j}$ are given by

$$\hat{r}_{3,j} = r_1 + \hat{d}_{1,j}\begin{bmatrix}\sin(\theta_{1,j})\\ \cos(\theta_{1,j})\end{bmatrix} - \hat{d}_{3,j}\begin{bmatrix}\sin(\theta_{3,j})\\ \cos(\theta_{3,j})\end{bmatrix}, \tag{H.7}$$

$$\hat{r}_{4,j} = r_1 + \hat{d}_{1,j}\begin{bmatrix}\sin(\theta_{1,j})\\ \cos(\theta_{1,j})\end{bmatrix} - \hat{d}_{4,j}\begin{bmatrix}\sin(\theta_{4,j})\\ \cos(\theta_{4,j})\end{bmatrix}, \tag{H.8}$$

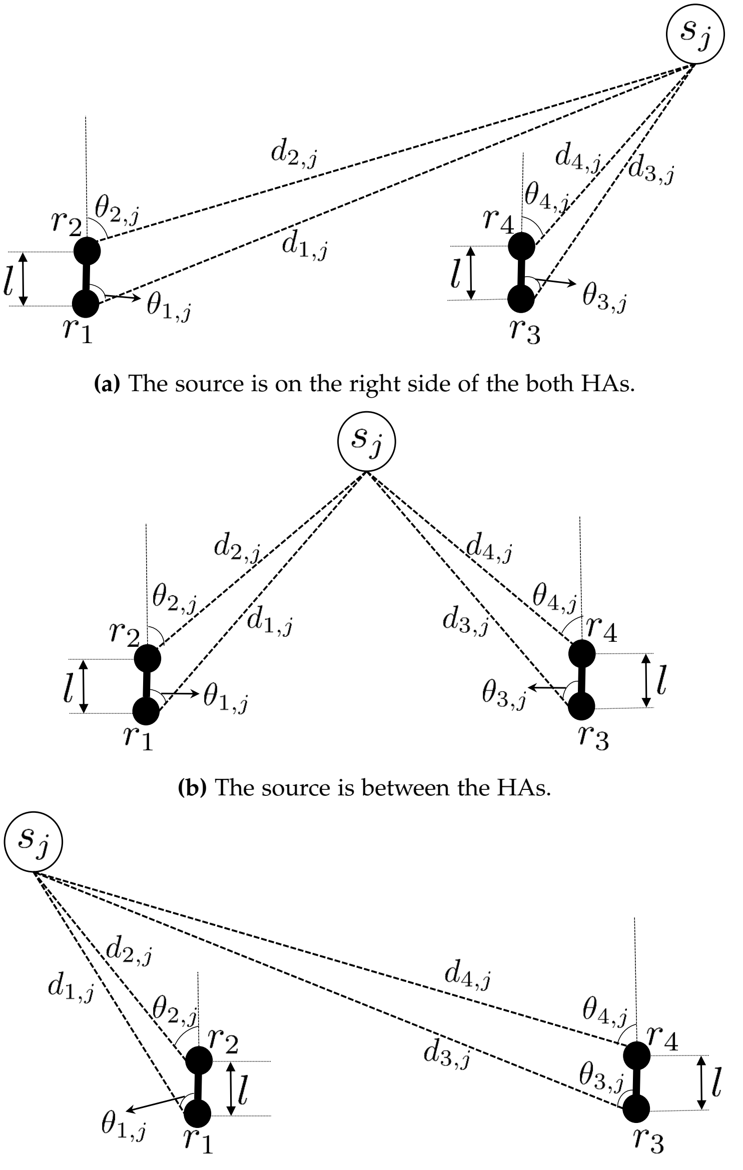
respectively. With perfect knowledge of the parameters, estimations of the relative locations using distances from different sources should be equal, i.e. $\hat{r}_{3,1} = \hat{r}_{3,2} = \cdots = \hat{r}_{3,N}$, and $\hat{r}_{4,1} = \hat{r}_{4,2} = \cdots = \hat{r}_{4,N}$. However, in practice, observations are always noisy; therefore, we consider

$$\hat{r}_3 = \frac{1}{N}\sum_{j=1}^{N}\hat{r}_{3,j}, \qquad \hat{r}_4 = \frac{1}{N}\sum_{j=1}^{N}\hat{r}_{4,j} \tag{H.9}$$

as the estimated relative locations of r_3 and r_4. Regrading the microphones' locations (the unknown variables), Eqs. (H.7), (H.8) and (H.9) form a linear system of equations, which is the core of the localization algorithm in this paper.

One remaining issue is that we can estimate $\theta_{i,j}$ only up to a sign $\big($cf. Eqs. (H.4), (H.5) and (H.6)$\big)$. Therefore, for each s_j, three different cases are conceivable (cf. Fig. H.2):

- Case 1: s_j is on the right sides of the HAS (cf. Fig. H.2a), i.e. $\theta_{1,j} = +\hat{\theta}_{1,j}$, $\theta_{3,j} = +\hat{\theta}_{3,j}$ and $\theta_{4,j} = +\hat{\theta}_{4,j}$.
- Case 2: s_j is between the two HAs (cf. Fig. H.2b), i.e. $\theta_{1,j} = +\hat{\theta}_{1,j}$, $\theta_{3,j} = -\hat{\theta}_{3,j}$ and $\theta_{4,j} = -\hat{\theta}_{4,j}$.



(a) The source is on the right side of the both HAs.

(b) The source is between the HAs.

(c) The source is on the left side of the both HAs.

Fig. H.2: Different relative locations of the target talker with respect to a binaural HAS.

- Case 3: s_j is on the left sides of the HAS (cf. Fig. H.2c), i.e. $\theta_{1,j} = -\hat{\theta}_{1,j}$, $\theta_{3,j} = -\hat{\theta}_{3,j}$ and $\theta_{4,j} = -\hat{\theta}_{4,j}$.

We can distinguish Case 1 and Case 3 by the differences between $\hat{d}_{i,j}$s:

- If $\hat{d}_{1,j} + \hat{d}_{2,j} > \hat{d}_{3,j} + \hat{d}_{4,j}$, case 3 cannot be the case.
- If $\hat{d}_{1,j} + \hat{d}_{2,j} < \hat{d}_{3,j} + \hat{d}_{4,j}$, case 1 cannot be the case.

However, cases 1 and 2, and cases 2 and 3 are not distinguishable from each other using the differences between $\hat{d}_{i,j}$s. In other words:

$$\left[\theta_{1,j}, \theta_{3,j}, \theta_{4,j}\right] = \begin{cases} [+\hat{\theta}_{1,j}, \pm\hat{\theta}_{3,j}, \pm\hat{\theta}_{4,j}], & \text{if } \Delta d_j > 0 \\ [\pm\hat{\theta}_{1,j}, -\hat{\theta}_{3,j}, -\hat{\theta}_{4,j}], & \text{otherwise} \end{cases}. \tag{H.10}$$

where $\Delta d_j = (\hat{d}_{1,j} + \hat{d}_{2,j}) - (\hat{d}_{3,j} + \hat{d}_{4,j})$. Therefore, for each source j, we have two different cases which cannot be distinguished based on Δd_j. Two different cases for each source result in 2^N different combinations of cases considering all sources.

Let $\mathcal{P}$ denote the set of all possible combinations of the indistinguishable cases considering all sources ($|\mathcal{P}| = 2^N$). We aim to find the true combination of the cases ($p^* \in \mathcal{P}$) which can justify all the measurements best.

For each $p \in \mathcal{P}$ and each source j, we can solve the localization problem using Eqs. (H.7) and (H.8). Therefore, we have N different estimators for r_3 and r_4 for each $p \in \mathcal{P}$:

$$\hat{R}_{3,p} = \{\hat{r}_{3,1,p}, \hat{r}_{3,2,p}, \cdots, \hat{r}_{3,N,p}\}, \tag{H.11}$$

$$\hat{R}_{4,p} = \{\hat{r}_{4,1,p}, \hat{r}_{4,2,p}, \cdots, \hat{r}_{4,N,p}\}, \tag{H.12}$$

where $\hat{r}_{3,j,p}$ and $\hat{r}_{4,j,p}$ denote the estimator of r_3 and r_4, respectively, using source j for $p \in \mathcal{P}$. The best combination is given when the differences between all the N estimators are minimum:

$$p^* = \arg\min_{p\in\mathcal{P}}\{\sigma^2_{\hat{R}_{3,p}} + \sigma^2_{\hat{R}_{4,p}}\} \tag{H.13}$$

where

$$\sigma^2_{\hat{R}_{3,p}} = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{r}_{3,j,p} - \hat{r}_{3,p}\right)^2, \tag{H.14}$$

$$\sigma^2_{\hat{R}_{4,p}} = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{r}_{4,j,p} - \hat{r}_{4,p}\right)^2, \tag{H.15}$$

where

$$\hat{r}_{3,p} = \frac{1}{N}\sum_{j=1}^{N}\hat{r}_{3,j,p}, \qquad \hat{r}_{4,p} = \frac{1}{N}\sum_{j=1}^{N}\hat{r}_{4,j,p}. \tag{H.16}$$

The outputs of the localization algorithm are $\hat{r}_{3,p^*}$ and $\hat{r}_{4,p^*}$. Moreover, the position of s_j, $j = 1, \cdots, N$, can be estimated using the estimated positions of each of the four microphones and the corresponding distances and angles, i.e. we have four estimators for each s_j. We consider the average of these four estimators as the estimator of s_j, i.e.

$$\hat{s}_j = \frac{1}{4}\sum_{i=1}^{4}\left(\hat{r}_{i,p^*} + \hat{d}_{i,j}\begin{bmatrix}\sin(\theta^*_{i,j})\\ \cos(\theta^*_{i,j})\end{bmatrix}\right), \tag{H.17}$$

where $\theta^*_{i,j}$ is the estimation of $\theta_{i,j}$ for the best combination p^*, and $\hat{r}_{1,p^*} = [0,0]^\mathrm{T}$ and $\hat{r}_{2,p^*} = [0,l]^\mathrm{T}$.

3.1 ToF estimation

In this section, we explain in more details how to estimate the ToFs upon which the above algorithm relies.

As mentioned in Sec. 2, the ToF between source j and microphone i, denoted by $t_{i,j}$, can be estimated by Eq. (H.2), where τ_j is given by the wireless microphone. Therefore, estimation of $t_{i,j}$ depends on $\hat{q}_{i,j}$. The more accurate the estimation of $\hat{q}_{i,j}$, the more accurate the estimation of $t_{i,j}$. In the following, we discuss how to estimate $\hat{q}_{i,j}$ in the considered setup using a Generalized Cross-Correlation (GCC) approach.

Let $x_j(n)$ denote the noise-free target signal emitted at s_j, and let $y_{i,j}(n)$ denote the noisy signal received from source j by microphone i of the HAS. It should be noted that $x_j(n)$ is available at the HAS via the wireless microphone. Furthermore, let $X_j(f)$ and $Y_{i,j}(f)$ denote the discrete Fourier transforms (DFTs) of $x_j(n)$ and $y_{i,j}(n)$, respectively. The GCC is then given by [19]:

$$\mathcal{R}_{i,j}(\tau) = \sum_{f=1}^{M} \psi(f) X^*_j(f) Y_{i,j}(f) e^{j2\pi \frac{f}{M}\tau}, \tag{H.18}$$

where M is the DFT order, $*$ represents complex conjugation and $\psi(.)$ is a weighting function. Then, $\hat{q}_{i,j}$ is given by:

$$\hat{q}_{i,j} = \arg\max_{\tau} \ |\mathcal{R}_{i,j}(\tau)|, \tag{H.19}$$

where $|.|$ gives the absolute value of its argument. In this paper, we simply use the conventional cross-correlation method for the ToF estimation, i.e. $\psi(f) = 1$ for all f in Eq. (H.18). However, to improve the estimation performance in noisy situations, there are more complex weighting functions which take into account the noise characteristics [19].

Because ToFs are estimated based on sampled signals, the estimation accuracy is limited by the sampling interval. Therefore, subsample estimation errors are unavoidable, and we use interpolation methods to mitigate this problem [20, 21]. In this paper, we use the cubic spline method [22] to interpolate the microphone signals before computing the GGC to decrease the subsample estimation errors.

4 Simulation Results

In this section, we evaluate the performance of the proposed localization algorithm in simulation experiments. Specifically, we study the effects of the SNR, the distance between the target talker and the user, and the number of the sound sources (N) on the proposed algorithm.

4.1 Setup

For evaluation, we consider a free-field situation, i.e. head presence is ignored in the simulations. Moreover, we set $l = 1.5\,\text{cm}$ and consider the head diameter, or more precisely, the distance between two hearing aids to be 16 cm. We distribute the sound sources randomly according to a uniform distribution on a disc or a circle (depending on the experiment) around the user. We use the TSP database [23] for generating speech sound sources. The sampling frequency is 48 kHz, the estimation window length is 4096 samples, and we run the simulations for 200 different realizations. The number of query points for interpolation between each two consecutive sample points of the signal is 100.

To simulate the noisy received signal $y_{i,j}(n)$, the following signal model

$$y_{i,j}(n) = x_j(n) * h_{i,j}(n) + v_{i,j}, \tag{H.20}$$

has been used, where $h_{i,j}(n)$ and $v_{i,j}(n)$ are the acoustic channel impulse response between source j and microphone i, and an additive noise component, respectively. Convolution operator is represented by $*$. We consider the additive noise component $v_{i,j}$ in Eq. (H.20) to be statistically independent of the target signal, and we generate $v_{i,j}$ as an independent and identically distributed zero-mean Gaussian random variable, i.e. $v_{i,j} \sim \mathcal{N}(0, \sigma_v^2)$.

4.2 Performance measures

To evaluate the estimated microphone locations, we use mean absolute error defined as

$$\sigma_e = \frac{1}{2} \sum_{i=3}^{4} \|r_i - \hat{r}_{i,p^*}\|_2. \tag{H.21}$$

Moreover, because in HAS applications DoAs of the target sounds are more important than the exact locations of the target sources, to evaluate the estimated positions of the sound sources, we use

$$\sigma_\theta = \frac{1}{4N} \sum_{j=1}^{N} \sum_{i=1}^{4} |\theta_{i,j} - \check{\theta}_{i,j}|, \tag{H.22}$$

where $\theta_{i,j}$ is the true DoA of the target signal from source j at microphone i, and $\check{\theta}_{i,j}$ is the estimated DoA given by

$$\check{\theta}_{i,j} = \arctan \left(\frac{\hat{u}_j - \hat{x}_i}{\hat{w}_j - \hat{y}_i} \right), \tag{H.23}$$

where $\hat{r}_{i,p^*} = [\hat{x}_i, \hat{y}_i]^{\mathrm{T}}$ and $\hat{s}_j = [\hat{u}_j, \hat{w}_j]^{\mathrm{T}}$.

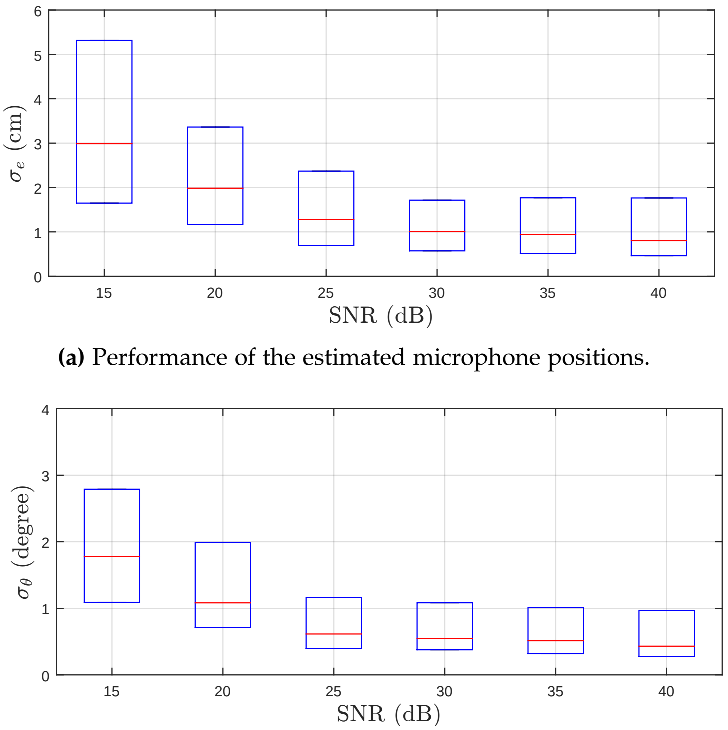
To demonstrate the results, we use box plots (Figs. H.3, H.4 and H.5), where the bottom and top of the box are the first and third quartiles, and the band inside the box is the median of the results obtained from different realizations.

4.3 Results and discussion

The effect of the SNR on the proposed algorithm has been shown in a box plot in Fig. H.3. As expected, the higher the SNR, the better performance of the localization algorithm. This is because, at higher SNRs, estimations of the ToFs and estimations of the distances between the sound sources and the microphones are more accurate. However, increasing the SNR to infinite would not lead to a zero error because the fundamental subsample error of the ToF estimation cannot be overcome by increasing the SNR. Nevertheless, if ToFs could in a way be estimated perfectly, then the estimation error of the proposed localization algorithm would be zero.

Fig. H.4 shows the box plot of the results of the proposed algorithm as a function of the distance between the target talker and the user. For these results, we consider $N = 3$ and for each realization, the three sound sources are distributed randomly on a circle centered at the user's head for different distances. As can be seen, generally, increasing the distance degrades the localization performance. Intuitively, it is because increasing the distance between the user and the target talker leads to a far-field situation, i.e. the distances between the microphones are negligible with respect to the distance between the user and the target, or in other words, the microphones' positions look almost the same from the position of the target talker. Therefore, determining the exact locations of the microphones is harder for the proposed localization algorithm. To be more precise, regarding Eqs. (H.7) and (H.8), the same estimation errors of $\theta_{i,j}$s result in higher localization errors at higher distances.

Performance of the proposed localization algorithm as a function of N has been shown in Fig. H.5. As expected, increasing N generally improves



(a) Performance of the estimated microphone positions.

(b) Performance of the estimated DoAs.

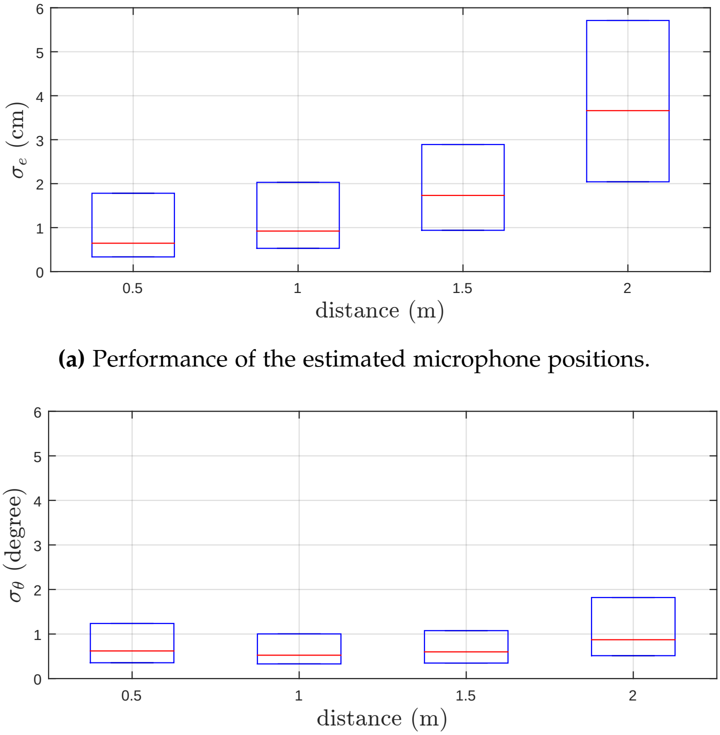
Fig. H.3: The box plot of the performance of the proposed algorithm as a function of SNR. The distance between the target talker and the user is in the range of 0.5 m to 1.5 m, and N is 3.

the localization performance because it provides more information. However, it costs higher computational overhead.

Overall, at SNRs around 30 dB and distances $d_{i,j} \approx 1$ m, the estimation error of the microphones locations is around 1 cm, and the estimation error of the target sounds DoAs is less than 1°.

5 Conclusion and Future Work

In this paper, we proposed a localization algorithm for dual-microphone subarrays considering hearing aid applications. The proposed localization algorithm is based on the estimated ToFs of the target signals received by the hearing aid microphones from sound sources whose locations are unknown. We modeled the problem as a linear system of equations and avoided any far-field assumption. We studied the impacts of different factors, such as SNR, distance of the sound sources from the microphones and number of



(a) Performance of the estimated microphone positions.

(b) Performance of the estimated DoAs.

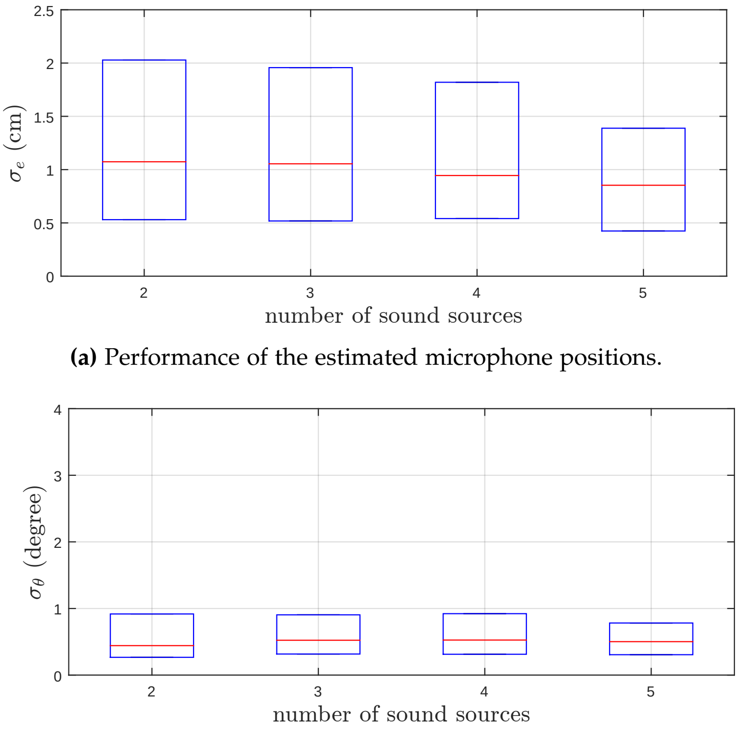
Fig. H.4: The box plot of the performance of the proposed algorithm as a function of distance between the target talker and the user. The SNR is 30 dB, and N is 3.

the sound sources, on the proposed algorithm. As future work, we plan to study the proposed algorithm under more realistic situations by considering presence of the head and reverberation.

References

[1] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Informed TDoA-based direction of arrival estimation for hearing aid applications," in *Proceedings of IEEE Global Conference on Signal and Information Processing*, 2015, pp. 953–957.

[2] ——, "Informed direction of arrival estimation using a spherical-head model for hearing aid applications," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2016, pp. 360–364.



(a) Performance of the estimated microphone positions.

(b) Performance of the estimated DoAs.

Fig. H.5: The box plot of the performance of the proposed algorithm as a function of N. The distance between the target talker and the user is in the range of 0.5 m to 1.5 m, and the SNR is 30 dB.

[3] J. H. Lee, S. Hwang, and C. L. Istook, "Analysis of human head shapes in the united states," *International Journal of Human Ecology*, vol. 7, no. 1, pp. 77–83, 2006.

[4] M. Chen and others., "Energy-based position estimation of microphones and speakers for ad hoc microphone arrays," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 22–25.

[5] R. Heusdens and N. D. Gaubitch, "Time-delay estimation for TOA-based localization of multiple sensors," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2014, pp. 609–613.

[6] N. D. Gaubitch *et al.*, "Auto-localization in ad-hoc microphone arrays," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2013, pp. 106–110.

[7] S. Zhayida, S. Burgess, Y. Kuang, and K. Åström, "TOA-based self-calibration of dual-microphone array," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 791–801, Aug 2015.

[8] M. Pollefeys and D. Nister, "Direct computation of sound and microphone locations from time-difference-of-arrival data," in *Proceedings of of IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2008, pp. 2445–2448.

[9] M. Crocco, A. D. Bue, and V. Murino, "A bilinear approach to the position self-calibration of multiple sensors," *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 660–673, Feb 2012.

[10] M. Farmani, R. Heusdens, M. S. Pedersen, and J. Jensen, "TDOA-based self-calibration of dual-microphone arrays," in *Proceedings of European Signal Processing Conference*, 2016, pp. 617–621.

[11] R. Kaune, "Accuracy studies for TDOA and TOA localization," in *Proceedings of International Conference on Information Fusion*, July 2012, pp. 408–415.

[12] F. Jacob, J. Schmalenstroeer, and R. Haeb-Umbach, "DOA-based microphone array postion self-calibration using circular statistics," in *Proceedings of of IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 116–120.

[13] S. T. Birchfield and A. Subramanya, "Microphone array position calibration by basis-point classical multidimensional scaling," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1025–1034, 2005.

[14] Y. T. Chan, H. Y. C. Hang, and P. C. Ching, "Exact and approximate maximum likelihood localization algorithms," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 1, pp. 10–16, Jan 2006.

[15] S. Thrun, "Affine structure from sound," in *Advances in Neural Information Processing Systems*, 2005, pp. 1353–1360.

[16] Y. Kuang, S. Burgess, A. Torstensson, and K. Åström, "A complete characterization and solution to the microphone position self-calibration problem," in *Proceedings of of IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 3875–3879.

[17] G. Jacovitti and G. Scarano, "Discrete time techniques for time delay estimation," *IEEE Transactions on Signal Processing*, vol. 41, no. 2, pp. 525–533, Feb 1993.

[18] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Dover Publications, 1964, vol. 55.

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

[20] F. Viola and W. F. Walker, “A spline-based algorithm for continuous time-delay estimation using sampled data,” *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 52, no. 1, pp. 80–93, 2005.

[21] X. Lai and H. Torp, “Interpolation methods for time-delay estimation using cross-correlation method for blood velocity measurement,” *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 46, no. 2, pp. 277–290, March 1999.

[22] C. B. Moler, *Numerical Computing with MATLAB: Revised Reprint*, ser. SIAM e-books. Society for Industrial and Applied Mathematics, 2008.

[23] P. Kabal, “TSP speech database,” Department of Electrical and Computer Engineering, McGill University, Tech. Rep., 2002.

ISSN (online): 2446-1628
ISBN (online): 978-87-7112-936-6

AALBORG UNIVERSITY PRESS