ACTIVE SOUND LOCALISATION THROUGH BAYESIAN INFERENCE

Glen Adam McLachlan

Active sound localisation through Bayesian inference

by

Glen Adam McLachlan

Thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the Faculty of Science at the University of Antwerp

Supervisors:

Prof. Dr. Herbert Peremans and Prof. Dr. Bart Partoens

November, 2024

Preface

The problem of perception is not just a scientific one, it is deeply philosophical. It raises profound questions about the nature of reality, consciousness, and the relationship between our minds and the external world. These are obviously not questions that I will be able to answer in my doctoral thesis—the questions that I *did* set out to investigate are demanding enough—yet they provide a captivating context in which the following scientific findings can be viewed.

In the *Critique of Pure Reason*, Immanuel Kant introduces the metaphysical theory of transcendental idealism. He poses that there lies a deep chasm, an impassable one, dividing 'truth' and the representation of that 'truth' in our minds. Kant describes two worlds. There is the world of things in themselves, the 'noumenal', that exist beyond the limitations and capabilities of our senses. Then there is the 'phenomenal', the world of appearances, a crude map of the real world that our mind and senses produce so that we can attempt to navigate it.

In other words, we do not have direct access to the world as it is. Rather, we experience the world that we infer from our eyes, ears and mental faculties. This implies that there exists a limit to our knowledge and comprehension of reality; an understanding line. In the same way that a cat will never understand a reality TV show, we may never be able to understand the fabric of existence, simply because it exceeds the hardware that we have available to detect and process it. We have, after all, not evolved to measure the fabric of existence; we are designed to recognise which fruits are poisonous and which aren't.

To me this is a humbling thought. It emphasises the importance of human experience, the subjective, over objective truth, as objective truth is by definition out of our reach. That is, of course, not to say that the objective should be dismissed and thrown out of the window. As researchers, we develop models of reality; simplifications and approximations, laden with assumptions. Many of such models have been paramount to human progress, but I believe that Kant's point of view helps us remain wary to not fall into that trap of arrogance, of thinking that we fully understand something. It leaves the door ajar for new discoveries and interpretations.

In this thesis, there will be many parallels with this dichotomy; the 'truth' versus the 'perceived'. Bayesian inference, a key concept explored in this work, takes the subjective experience into

consideration. It explains why the exact same sensory information can elicit varying responses among individuals when the prior beliefs and expectations of those individuals are not the same. In the end, what I have developed is a model for the problem of dynamic sound localisation. It is only a small component of the larger phenomenon of multisensory perception, but it is this connection to the broader human experience that has truly inspired me.

With all of that said, I would finally like to introduce this thesis with a personal favourite quote by Nietzsche that is not only relevant to the subject of perception, but to the scientific method in its entirety: “There are no facts, only interpretations”.

Antwerp, Belgium, 18/06/2024 **Glen McLachlan**

Acknowledgements

I doubt that any individual holding a PhD degree would claim their path was smooth sailing. However, speaking with many fellow PhD candidates has underscored the privilege of my own circumstances. Sleepless nights and small mental breakdowns aside, each moment of this unforgettable journey has been deeply fulfilling, and if I could, I would do it all over again.

First and foremost, I wish to express my sincere thanks to my supervisors, Prof. Dr. Herbert Peremans and Prof. Dr. Bart Partoens, who not only guided me through each step along the way, but also treated me like a valued colleague as much as their student.

I am also grateful for the many insightful (also the less-insightful but equally entertaining) conversations with Dr. Jonas Reijniers and Dr. Alejandro Osses Vecchi. They have shown me that it is possible to deliver work of impeccable quality without taking yourself too seriously, something I think all should aspire to.

My sincere thanks to Dr. Piotr Majdak and Michael Mihocic from the Acoustics Research Institute in Vienna, as well as to Prof. Dr. Ville Pulkki and Pedro Lladó Gonzalez from the Acoustics Lab at Aalto University, for their invaluable support and warm hospitality during my research visits.

A great deal of thanks is owed to the board of the Young Acousticians Network, with a special mention to Diogo Pereira, for the opportunities and friendships that emerged from our work together. I believe we have done an excellent job fostering a supportive community for young acousticians, and I look forward to seeing you all at the next conference.

I thank the funding agencies that made this research possible: the Research Foundation Flanders (grant no. G023619N), the Agency for Innovation and Entrepreneurship, and the European Union (project SONICOM, grant no. 101017743, RIA action of Horizon 2020).

On a more personal note, my deepest appreciation goes to my parents, Mirjam and Scott, my brother, Kyle, and my close friends who have always shown their unwavering support during this academic journey: Dr. Nick Bleisch, Yamini Chitale, Stefan Cornelissen, Jurjen Faber,

Michiel Geluykens, Jeroen Korste, Niels Lejeune, Annemarije Makkinga, Claudia Marciano, Ulrike Schrijvers, Jente Snoeck, Karla Steinecke and the lovely people from the Antwerp Running Crew.

Finally, I would like to conclude my acknowledgments with Zofia Iwaszkiewicz, the talented artist behind this thesis cover and my dearest friend. As I write this, she is courageously battling Hodgkin's lymphoma. Zofia, you are the kindest and strongest person I know, and I am incredibly proud of you.

Abstract

While the human auditory system is proficient on its own at discerning the direction of incoming sounds, it operates in concert with other sensory modalities to reach accurate spatial awareness. Many studies have investigated the integration of auditory and visual information, but much less attention has been given to the importance of proprioceptive and vestibular information in the localisation process. The vestibular and proprioceptive systems aid in discerning self-motion from source motion and, through this, can stabilise perception and provide additional cues for sound localisation.

The aim of this PhD thesis was to better understand how head movement and position estimation affects sound localisation. To this end, an ideal-observer model, based on Bayesian principles, was developed as a tool to predict dynamic sound localisation in humans and to test how performance is affected by the available information.

Behavioural experiments were conducted in conjunction with model simulations to determine the acoustic cues and head motions that are relevant to dynamic sound localisation. The results from the psychoacoustic experiments were found to be in general agreement with the model output, though some quantitative differences indicated that dynamic sound localisation may involve processes that can be considered non-ideal.

These studies offer valuable insights for the field of psychoacoustics and for auditory engineering applications in modern technologies such as hearing aids and virtual or augmented reality.

Keywords: *Dynamic Sound Localisation*, *Spatial Hearing*, *Head Movement*, *Bayesian Inference*, *Behavioural Experiments*, *Statistical Models*.

List of Publications

A McLachlan, G., Majdak, P., Reijniers, J., and Peremans, H. (2021). Towards modelling active sound localisation based on Bayesian inference in a static environment. Acta Acustica, 5, 45. doi: 10.1051/aacus/2021039.

B McLachlan, G., Majdak, P., Reijniers, J., Mihocic, M., and Peremans, H. (2023). Dynamic spectral cues do not affect human sound localization during small head movements. Frontiers in Neuroscience, 17, 1027827. doi: 10.3389/fnins.2023.1027827.

C McLachlan, G. and Peremans, H. (2023). Modelling dynamic sound localisation through Bayesian inference: a sensitivity analysis. In Forum Acusticum, A11-01: Spatial Hearing: Modeling and Applications - Part I (pp. 267-272). Italy: Turin. doi: 10.61782/fa.2023.0115.

D McLachlan, G., Majdak, P., Reijniers, J., Mihocic, M., and Peremans, H. (2024). Bayesian active sound localisation: to what extent do humans perform like an ideal-observer?

- manuscript submitted to PLOS Computational Biology on 25/04/2024. - manuscript revised on 25/09/2024.

E McLachlan, G., Lladó P., and Peremans, H. (2024). Head rotations follow those of a truncated Fick gimbal during an auditory guided visual search task.

- manuscript submitted and accepted in Journal of Neurophysiology.

Contribution Statement

All papers included in the thesis have been coauthored with other researchers. Below I list the contributions of each individual author.

A The first author, Glen McLachlan, conducted the literature review, developed the MATLAB code, produced the figures and wrote the paper. The remaining authors provided feedback and reviewed the paper. The final author, Herbert Peremans, supervised the project and proposed the theoretical framework.

B The first author, Glen McLachlan, conceptualised the behavioural experiments, conducted the experiments, ran the data analysis, produced the figures and wrote the paper. The second author, Piotr Majdak, supervised the experiments as principal investigator. The fourth author, Michael Mihocic, developed the software for the experiments and assisted in conducting the experiments. The final author, Herbert Peremans, supervised the project. All authors reviewed the paper.

C The first author, Glen Mclachlan, developed the MATLAB code, produced the figures and wrote the paper. The second author, Herbert Peremans, supervised the project and reviewed the paper.

D The first author, Glen Mclachlan, developed the MATLAB code, conceptualised the behavioural experiments, conducted the experiments, ran the data analysis, produced the figures and wrote the paper. The second author, Piotr Majdak, supervised the experiments as principal investigator. The fourth author, Michael Mihocic, developed the software for the experiments and assisted in conducting the experiments. The final author, Herbert Peremans, supervised the project. All authors reviewed the paper.

E The first author, Glen Mclachlan, conceptualised the paper, conducted the experiments, analysed the data, produced the figures and wrote the paper. The second author, Pedro Lladó, conducted the experiments and reviewed the paper. The third author, Herbert Peremans, supervised the project and reviewed the paper.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AMT Auditory Modeling Toolbox 61, 76, 99

dILD dynamic interaural level difference 44, 45, 57–59

dITD dynamic interaural time difference . . . . . xi, 44–46, 49, 53, 56–60

dMSS dynamic monaural spectral shape 44–46, 53, 57–60

FBC front-back confusion . . . . . xvi, 3, 51, 54–57, 59, 86–88, 99

HMD head-mounted display 46, 47, 50, 51, 84

HRTF head-related transfer function ix, xi, 6, 7, 12, 18, 20, 21, 23, 40, 47–50, 55, 56, 64, 65, 76, 77, 82, 84, 88

ILD interaural level difference 6, 14, 17–20, 22, 44–46, 49, 58–60, 82, 99

ITD interaural time difference x, xi, 4, 6, 10, 13, 17–19, 22–24, 26, 30–32, 34, 40–42, 44–46, 49, 56–58, 60, 63, 67, 69, 76–78, 81–83, 87, 91, 92, 97, 99

JND just-noticeable difference . . . . . xii, 67, 77, 79, 83, 93

MAP maximum a posteriori 9, 30, 39, 81, 82

MSS monaural spectral shape . . . . . xi, 44–46, 49, 53, 56–60

PDF probability density function . . . . . xi, 8–10, 30–33, 35–39, 41, 65, 66, 77, 80, 81

QE quadrant error . . . . . xii, xiii, xvi, 86, 87, 91–99

RMSE root mean square error . . . . . xii, xiii, xvi, 86, 87, 92–95, 97–99

UDC up-down confusion . . . . . xvi, 86–88

VBAP virtual-base amplitude panning 47–49, 85

General Introduction

It may appear a trivial task to determine the location of a sound; in daily life we constantly make estimates of objects in our perceived environment without consciously paying much attention to them. However, the ability of humans to accurately infer the three-dimensional location of objects based solely on auditory cues is nothing short of astonishing.

The information transmitted from our sensory organs to our brain is typically noisy and ambiguous, which leads to variability and inaccuracies in perceptual responses. This noisiness depends on both external factors (e.g., environmental noise and reverberation) and internal factors (e.g., neuronal noise). To deal with this, the brain requires an internal model capable of representing sensory uncertainty.

Moreover, the auditory system does not allow for the intrinsically spatial representation of sensory information that the visual and somatosensory systems benefit from. Instead of a topographical mapping at the sensory level, spatial information is encoded in acoustic cues as a function of frequency and time, which the auditory system then needs to somehow decode.

Motivation

Studies and models of sound localisation typically only consider the acoustic cues available to humans under static conditions, where neither the source, nor the head undergoes any movement. However, real-life environments are dynamic, where both the listener and the objects around them are capable of movement. The assumption of a stationary setting and listener therefore lacks ecological validity and fails to account for numerous crucial factors that influence sound localisation.

First, in addition to acoustic information, a model for dynamic sound localisation needs to process vestibular and proprioceptive information. Despite our own movement, we perceive the external world as stable. For the auditory system to account for this, there must be a coordinate

transformation from the head-centred auditory cues into a stable, world-centred frame of reference, which requires positional information about the head (Vliegen et al., 2004). Vestibulo-proprioceptive information is also necessary to disambiguate the motion of a source from apparent motion that results from head movements, which would in theory produce similar acoustic cues (Wightman and Kistler, 1999).

Second, the dynamic acoustic cues obtained from self-motion have been proven beneficial for sound localisation and externalisation (Wallach, 1940; Perrett and Noble, 1997a; Brimijoin et al., 2013; Thurlow and Runge, 1967; Jiang et al., 2019). Their major contribution is in eliminating front-back confusions (FBCs), although they can improve elevation estimation under certain conditions too. Head movements are particularly useful when the listener localises under poor acoustic conditions, e.g. in reverberant (Giguère and Abel, 1993) or virtual environments (Perrett and Noble, 1997a).

Third, head motion control requires its own movement model, as the head is limited to certain anatomical and behavioural constraints. Humans show trends in e.g. the magnitude and acceleration profile of head rotation (Kim et al., 2013a) and the common axes around which the head is rotated (Thurlow et al., 1967). A framework that implements head movements can also help investigate 'closed-loop' listening strategies during sound localisation, in which the type of head movement depends on the acoustic input received. For example, specific movement strategies may be applied to 'triangulate' a source, to decrease interference from reverberation, or to attend to a single (moving) source in a complex listening environment.

The development of binaural hearing aids and virtual and augmented reality technologies is generating an increasing need to better understand the complex dynamic processes that underlie auditory perception (McAnally and Martin, 2014; Carlile and Leung, 2016). The principal aim of the current work is therefore to expand on existing sound localisation research by examining and modelling the role of head movement. This not only leads us to a broader understanding of sound localisation, but also of the interconnected systems that are responsible for human multisensory perception.

Scope

Sound localisation can be subdivided into several categories (see Fig. 1). First there is the distinction between static and dynamic localisation, which exclude and include any form of movement, respectively. Within dynamic localisation, there are the separate problems of source motion and self-motion. Within this manuscript, localisation using self-motion is used interchangeably with the term 'active' sound localisation. Finally, self-motion can be divided into

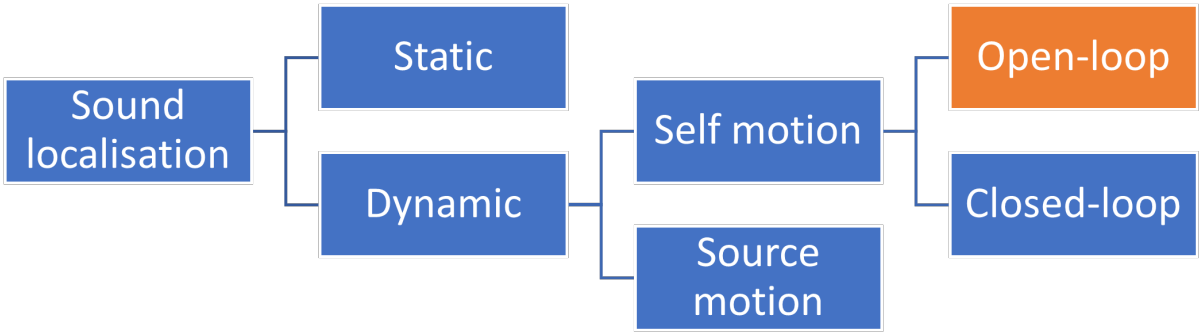

Fig. 1 Subcategories of sound localisation. The scope of the project is aimed at small open-loop head rotations.

open-loop and closed-loop listening. These are terms borrowed from control theory. Motion is considered closed-loop when the listener adjusts the head movements in response to the perceived sound in a feedback loop, e.g., rotating the head to better understand a speaker. Open-loop motion is movement that is made regardless of the incoming source signal, e.g., movements made while walking. The scope of this project was limited to open-loop self-motion, as this is the first significant step to integrating acoustic and proprioceptive information. More specifically, the types of movement investigated were small rotations ($< 10°$), as they are most relevant from an ecological perspective (i.e., most abundant) and from an application perspective (i.e., large rotations may involve more complex processes). Furthermore, localisation is restricted to *directional* localisation, leaving distance perception out of consideration.

Acoustic cues for sound localisation

The human anatomy allows for several acoustic cues to be retrieved from an incoming acoustic signal. These cues are complementary in the sense that they are most informative under different conditions, and their reliability varies depending on the characteristics of the sound and the acoustic environment. The acoustic cues can be divided into binaural and monaural cues, which help localisation along the lateral (left/right) and polar dimensions (up/down and front/back), respectively. See Fig. 2 for a visualisation of the lateral-polar coordinate system.

Binaural cues, which result from the spatial separation of the two ears and the acoustic shadow created by the head, play a crucial role for the localisation of the lateral position of a sound source. Directional information is obtained through the interaural time difference (ITD) and

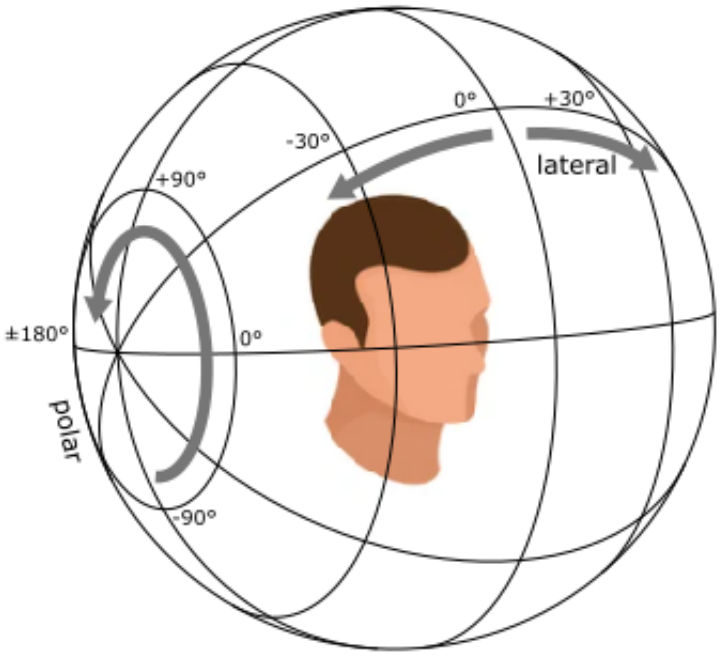

Fig. 2 Lateral-polar coordinate system, as used in the rest of this thesis. Lateral angles range between $\pm 90°$, polar angles range between $\pm 180°$. The lateral angle of $0°$ describes sources located on the median plane. The polar angle of $0°$ describes sources located on the horizontal plane at eye level. Figure obtained from Paper A.

the interaural level difference (ILD). Fig. 3 shows that, at 10° azimuth, the ITD is smaller than 0.1 ms. It is quite remarkable that humans can perceive such small differences in time and infer directional information from them. However, the binaural cues do not provide enough information for accurate localisation beyond the horizontal plane.

Monaural cues, which result from the filtering properties of the outer ear, head and torso, carry additional information on the polar position of the source. The direction-dependent character of this filtering is often expressed as a set of head-related transfer functions (HRTFs) in the frequency domain or as head-related impulse responses in the time domain. An HRTF dataset is derived from acoustic measurements of filter characteristics of the listener's ears, typically obtained with recordings from two microphones placed at the entrance of the ear canals while a broadband probe sound is presented from systematically varying locations. HRTFs are subject-dependent and, for that reason, high-resolution audio reproduction for virtual environments requires the use of HRTFs recorded from a listener's own ears, although the use of simulated or non-individualised HRTFs is also actively investigated. A further description of the acoustic cues for sound localisation is given in Chapter A.

Bayesian observer models

The 'inverse problem' or 'under-determination problem' describes the long standing challenge of understanding how the sensory system achieves a unique and stable percept of the true state of the world (the distal stimulus) from ambiguous sensory signals (the proximal stimulus) (Pizlo, 2001). In the context of sound localisation, this translates to the observer using the available acoustic cues and vestibulo-proprioceptive information to infer the true location of a sound source.

As a possible solution, Von Helmholtz (1867) posed the idea of perception as 'unconscious inference', by which he proposed that the brain resolves sensory ambiguity through built-in knowledge of the scene and the sensory organs, and uses this knowledge to automatically and unconsciously infer the properties of objects.

Bayesian theory formalises this process of inference into an elegant framework, where prior knowledge about the world is combined with new sensory observations, while taking into account the individual uncertainty of each contributing factor. This is done in an optimal manner, i.e., by minimising a predefined loss function. An additional advantage of this approach is that a Bayesian model does not just predict a response, but also predicts the uncertainty on this response.

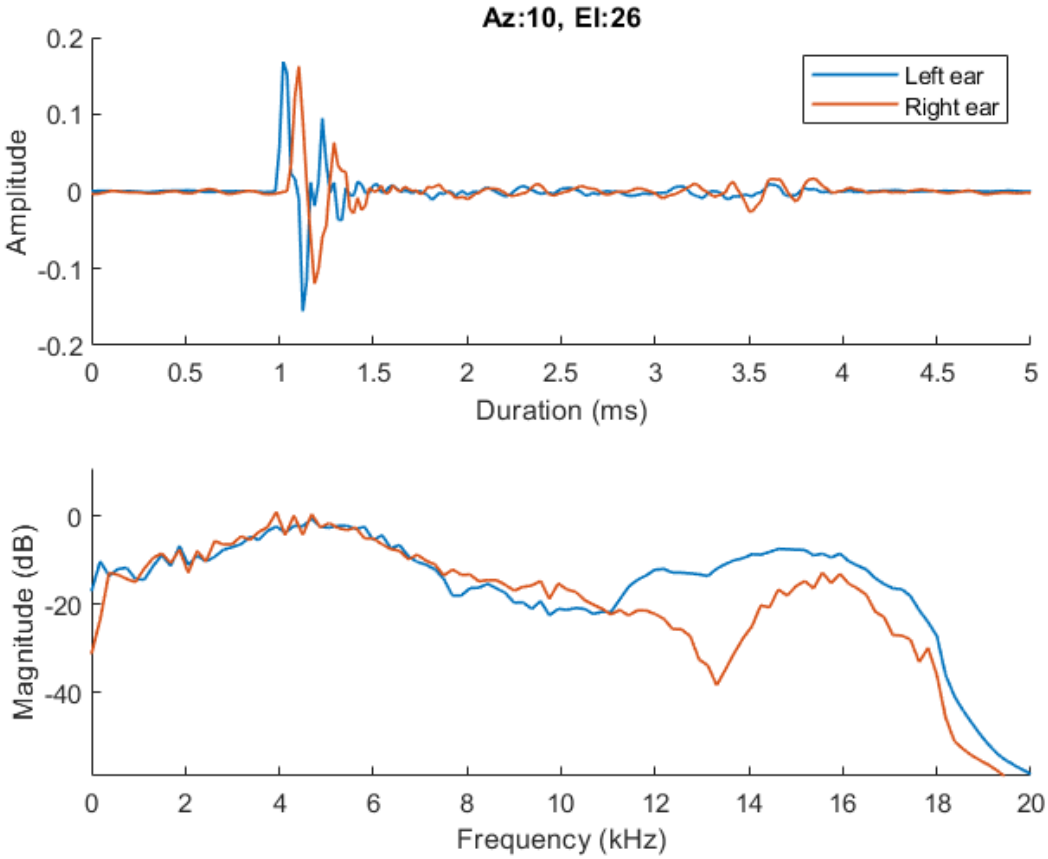

Fig. 3 Impulse responses in the time-domain (above) and magnitude spectra in the frequency-domain (below) of the measured left and right ear HRTFs for source direction $(az, el) = (10°, 26°)$.

Numerous studies have reinforced the notion that humans operate akin to Bayesian observers, offering invaluable insights into multisensory integration (Angelaki et al., 2009; Ernst and Bülthoff, 2004; Ernst, 2007; Kayser and Shams, 2015), causal inference (Beierholm et al., 2007; Ma and Rahmati, 2013; Körding et al., 2007) and motor control (Körding and Wolpert, 2004). As a 'holy grail', we might eventually describe the overarching concept of auditory scene analysis (Bregman, 1994), by which the auditory system organises sound into perceptually meaningful elements, as a Bayesian process (Hambrook et al., 2017).

Bayes' rule

Fundamentally, the Bayesian observer is defined by two key components. First, the prior distribution, $p(\psi)$, which is the probability of a hypothesis ψ to be true, *prior* to any observations made. Priors are constrained by natural or learned stimulus statistics (Wei and Stocker, 2015). As this is a probability distribution, each value of ψ has a corresponding probability $p(\psi)$. This represents the observer's belief about how frequently a certain stimulus value will occur. Second is the likelihood function, $p(y|\psi)$, which is the probability of making observation y, given the hypothesis ψ. This captures the encoding accuracy in the sensory representation of the observer.

For sound localisation, the prior $p(\psi)$ represents the probability of encountering a particular source direction, independent of the current sensory information. Which translates to the belief that the possible location of an object is not necessarily uniformly distributed over space. The likelihood $p(y|\psi)$ represents the probability of perceiving particular acoustic input y, given the source direction ψ.

Through Bayes' theorem, the prior and likelihood functions are combined to compute the posterior distribution $p(\psi|y)$, which quantifies how much the observer should believe ψ after considering the observed acoustic information y:

$$p(\psi|y) = \frac{p(\psi)\, p(y|\psi)}{p(y)} \tag{1}$$

where $p(y)$ is a normalisation constant so that $\int p(\psi|y)\, d\psi = 1$. This equation assigns a probability to every direction on the sphere, which makes the posterior a probability density function (PDF).

Besides including prior knowledge, Bayesian processes can help understand how different cues within one modality, or cues between two different modalities, combine into a single estimate (Seilheimer et al., 2014; Ernst, 2006).

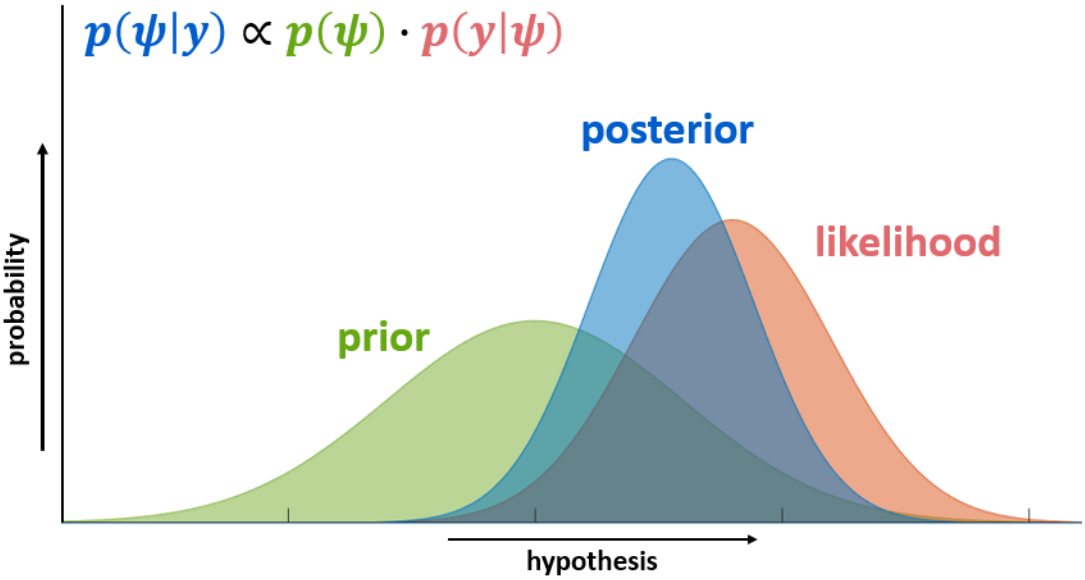

Fig. 4 Bayes' rule. The green curve represents a Gaussian prior and the red curve represents the likelihood, obtained from sensory information confounded by Gaussian noise. The resulting posterior, represented by the blue curve is then also Gaussian.

To combine cues the observer needs to weigh one cue against the other. Calculating this optimally in a Bayesian way means that the weighting will depend on the relative uncertainties in the cues. In order to do this, the observer makes (a priori) assumptions about the distribution of the errors on the sensory signals. In the most simple case of two independent cues A and B, confounded with Gaussian noise with variances σ_A^2 and σ_B^2, the optimal Bayesian observer will multiply cue A with weight w_A and cue B with weight $w_B = 1 - w_A$:

$$w_A = \frac{1/\sigma_A^2}{1/\sigma_A^2 + 1/\sigma_B^2}. \tag{2}$$

Decision rules

To obtain a point estimate from the posterior PDF, the Bayesian observer requires a loss function. The loss function $\mathcal{L}\left(\psi, \hat{\psi}\right)$ outputs a single value which quantifies the *loss* the observer will sustain by choosing the target estimate $\hat{\psi}$ instead of the targets's true value ψ, i.e., it represents the cost associated with the prediction errors. An optimal decision rule is one where the observer minimises the loss on average over a set of responses. Note that the definition of the loss function will influence the observer's decision.

A common and very straightforward decision strategy is the maximum a-posteriori (MAP) rule (Bassett and Deride, 2019). This is an 'all or nothing' approach, which leads the observer

to always select the maximum of the posterior PDF, in the form:

$$\mathcal{L}\left(\psi, \hat{\psi}\right) = \begin{cases} 0, & \text{if } \hat{\psi} = \psi, \\ 1, & \text{otherwise,} \end{cases} \tag{3}$$

Ideal-observer model for static sound localisation

This research builds upon the work of Reijniers et al. (2014), who proposed a Bayesian ideal-observer model for two-dimensional static sound localisation, i.e., without source or head movement.

By determining the posterior using Bayes' rule and taking into account possible probabilistic dependencies between the cues, the ideal-observer model functions as a 'ceiling' to human sound localisation performance and provides a benchmark to which behavioural data can be compared. Hence, any degradation in performance is solely due to the absence of the relevant spatial information in the received acoustic signals and in the a priori information available to the listener.

The model predictions were shown to be qualitatively in good agreement with the actual human localisation performance in a static setting, as assessed in a meta-analysis of many localisation experiments. Note, however, that accordance with an ideal observer's performance does not mean that the brain functions as a Bayesian observer. Bayesian modelling in psychophysical research is aimed at predicting and systematising performance, not on the mechanism that gives rise to the performance (Colombo and Seriès, 2012).

In the work by Reijniers et al. (2014), all spatial information was considered to be encoded by the ITD and by the two monaural spectra received at each ear. This may not necessarily correspond with the true information processing inside the auditory system, but the model framework specifically enabled exploration of the spatial information that might be lost by omitting or reshaping particular cues, along with their respective noise models, rather than the original acoustic signals received at both ears.

In this work, the static localisation model was enhanced on two important fronts. First, the input was no longer a single, static binaural measurement, but the model was extended such that it can deal with a sequence of binaural measurements. Second, the model was extended to process information on the orientation of the head and the uncertainty on this orientation.

Thesis Outline

In paper A, the state of the art in modelling dynamic sound localisation was reviewed, with a particular focus on Bayesian inference. Then, a theoretical Bayesian framework for dynamic and active sound localisation in a static environment was introduced. As a proof of concept, the results from two simplified numerical simulations were presented. This paper served as the foundation on which the subsequent work was built.

The goal of paper B was to investigate which acoustic cues are utilised by human listeners to localise sounds using small head movement. To this end, a dynamic sound localisation experiment was conducted, in which seven subjects performed a localisation task under four different stimulus conditions and three different movement conditions. The main finding in this study was that human listeners do not process dynamic changes in the monaural spectral cues to localise sounds when utilising small head movements, as conflicting spectral cues during movement did not affect performance. Furthermore, it was shown that (small) pitch rotations do not affect localisation performance, as was predicted by the dynamic localisation model in paper A.

Paper C presented a sensitivity analysis of the Bayesian model for dynamic sound localisation, in order to better understand the direction-dependent effect of each individual noise parameter. The insights gained from this study were used to identify possible problems or limitations in the presented model. These insights assisted in the analyses conducted in the following paper.

Paper D brought together all previous work in a spatial comparison between the dynamic localisation model and extensive data obtained from a second localisation experiment. Most importantly, the model parameters were set a priori, based on results from various psychoacoustic and proprioceptive studies, i.e., without any post-hoc parameter fitting to behavioral results. The model output matched the behavioural data on several localisation metrics, which showed that the model parameters corresponded with internal processes and could be derived and estimated experimentally. Further, specific effects of the sampling rate, the spatial prior and sizes of various model uncertainties were tested. This revealed a number of interesting effects, providing new insights on modelling the human integration of acoustic and proprioceptive information during sound localisation.

Finally, paper E addressed the topic of head movement behaviour during localisation tasks. The experiment conducted in this work investigated the differences and similarities in head movements made by different subjects, with the purpose of finding and proposing a general head rotation model. It was found that subjects generally adhered to the same laws of movement, with the roll component of the rotation vector a linear function of the combined yaw and pitch

components. Substantial differences were found between subjects in the slope of these linear relationships, but a statistical test revealed that including these individual differences in a movement model did not significantly improve it.

Thesis Contributions

This thesis provides several contributions to dynamic sound localisation research. We present these contributions here.

Bayesian model for dynamic sound localisation. The cornerstone of this thesis is the model for dynamic localisation, which is the first sound localisation model to include head position information and uncertainty. As it integrates information over time, it can be used in the future to model the temporal aspects of sound localisation as well. The structure of the model allows the user to easily test different settings, such as different noise parameters, spatial priors or decision rules. To this end, the code is well documented for further use. The MATLAB code for the model is publicly available in the Auditory Modeling Toolbox as `mclachlan2024` (Majdak et al., 2022). The AMT is available at www.amtoolbox.org.

LocaDyn test platform. We developed an application under the name LocaDyn within the custom software framework ExpSuite, which aids researchers to set up dynamic localisation experiments for various types of stimulation (multi-channel via loudspeakers or binaural via headphones). It allows for control of visual interfaces via virtual reality and for head movement to be tracked. This application was used for two of our localisation experiments and is publicly available on SourceForge (Majdak and Mihocic, 2022).

Experimental data. The data from papers B, D and E has been formatted and documented for replicability and further research. This data includes measured HRTFs, dynamic localisation data for several different stimulus and movement conditions, and head tracker data. The data from D contains a high number of repetitions per subject, per source direction, which allows for elaborate spatial statistics. The study from E provides additional head tracker and torso tracker data.

Spatial analysis and plotting tool. For paper D, the localisation data was visualised and analysed spatially. This has been done in the past for static data, though with a lower number of repetitions per direction (Carlile et al., 1997). The source code for the plotting tool was obtained from the Spak library (Leong and Carlile, 1998), and was modernised and integrated into the AMT (Majdak et al., 2022) as `plot_mclachlan2024`. The plotting tool allows for visualisation of the biases, distributions and reversal errors of responses for source directions over the full sphere around a listener. This tool was then used to evaluate direction-dependent

localisation performance.

Sensitivity analysis and model validation. In our work, the output of the dynamic localisation model was tested in two ways. First, in paper C, a sensitivity analysis provided insights on the relationships between the model's various noise parameters on its localisation performance. Second, in paper D, the parameter values were selected a priori, derived from psychouacoustic and kinaesthetic experiments, and the results were compared to the localisation data that we obtained through localisation experiments. These results showed good agreement between the model output and the behavioural data, thereby validating the model.

Future Work

While the findings in this Ph.D. thesis offer significant new insights into dynamic sound localisation, they also highlight several interesting avenues for further investigation. Below are suggestions for future research that can build upon the work presented here.

Acoustic feature space. While it is well-established that humans use ITD and spectral cues to localise sound, it is not fully understood to what extent they are used under different conditions. Thus, the feature space assumed in the present model is by no means definitive. For example, it has been suggested that humans solely use binaural cues for lateralisation and spectral cues for polar estimation through separate neural pathways (Hofman and Van Opstal, 1998; Ege et al., 2018). The information retrieved from the spectral content of the source can also be different from the assumed cues in this work. For example, previous work found that some neurons in the dorsal inferior colliculus of cats are sensitive to the positive spectral gradient of incoming sound, rather than center frequencies of peaks and notches (Reiss and Young, 2005). This was tested in a model for localisation on the sagittal plane and led to plausible model predictions (Baumgartner et al., 2014).

Closed-loop localisation. As described in the scope, the majority of the research in this thesis was restricted to small open-loop head movements. This is an appropriate first step to investigating active sound localisation. However, it does not include the possibility for different task-dependent localisation strategies. For example, it was found that the movements made by hearing impaired listeners differ substantially from those made by normal hearing listeners Brimijoin et al. (2010). It is also not self-evident that localisation performance remains the same for larger head movements or for different time scales. In other words, we cannot confidently assume that the outcome of a large head rotation equals the sum of a series of smaller rotations. It is well established that auditory information is integrated across several timescales, but the windows of each timescale and the conditions they apply to are not fully

understood (De Boer, 1985). Several studies suggest a sliding temporal window of integration of 150–200 ms (Yabe et al., 1998; Shinozaki et al., 2003). On the other hand, for interaural differences, a much higher temporal resolution is required to detect minute changes. For example, previous findings suggest that ILD sensitivity depends on binaural integration within a much smaller 3 ms temporal window (Brown and Tollin, 2016). Integrating over longer periods of time for such rapidly changing cues may not be beneficial to human listeners.

Complex acoustic environments. The problem of dynamic sound localisation is twofold: there is the effect of head movements and the effect of source movement Carlile and Leung (2016). The present work assumed single, stationary sound sources. Natural environments will often contain several sources that cannot always be assumed to be motionless. Extending this work to multiple sources and source movement increases the dimensionality of the problem significantly. The model will not only have to estimate the source's location, but also its movement direction, velocity and acceleration. These new variables will also have their own sets of prior distributions (such as the slow motion prior Senna et al. (2015)), which will require more behavioural research to reliably quantify.

Pointing error. A significant part of the errors measured in localisation experiments may be due to a response or 'pointing' error (Majdak et al., 2010). Pointing errors are likely not uniformly distributed over space, but it is difficult to differentiate between errors resulting from the auditory system and from the response method. Previous work found a progressive increase in horizontal errors with azimuth, a negative horizontal bias for rear locations, and a negative vertical bias for sources located at high elevations (Bahu et al., 2016). The prevalence of these trends depended on the pointing method. Pointing errors have been addressed to some extent in previous models for sound localisation (Baumgartner et al., 2014; Barumerli et al., 2023), though a uniform distribution was assumed.

Visual information. The present research only considers acoustic and proprioceptive information. Of course, information from the visual system is a very important factor in the context of multisensory perception, both for source localisation (Van Wanrooij et al., 2010) and for causal inference (Shams and Beierholm, 2010; Mendonça et al., 2016). For example, there are differences in spatial biases between the visual and auditory system. In the horizontal plane, for visual stimuli, biases were found towards the centre of the visual field. For auditory stimuli, biases were directed towards the periphery instead (Odegaard et al., 2015). Furthermore, higher weights have been found in general for visual information, when compared to auditory information (Battaglia et al., 2003a).

Non-ideal observers. Because human perception is limited by both the nature of internal

computations and its physiological hardware, significant departures from optimality may be expected. Nevertheless, Bayesian ideal observer models are an excellent benchmark against which human performance in any well defined perceptual task can be compared, and provide a first approximation to human performance that has been surprisingly effective (Kersten et al., 2004; Ernst and Banks, 2002; Helbig and Ernst, 2007). Moreover, the instances where modelled and behavioural results differ can be used to identify our limits and find the situations where humans specifically function sub-optimally (Stengård and Van den Berg, 2019).

PAPER A

Towards modelling active sound localisation based on Bayesian inference in a static environment

Glen McLachlan[1], Piotr Majdak[2], Jonas Reijniers[1], Herbert Peremans[1]

[1] Department of Engineering Management, University of Antwerp, Belgium
[2] Acoustics Research Institute, Austrian Academy of Sciences, Austria

Abstract - Over the decades, Bayesian statistical inference has become a staple technique for modelling human multisensory perception. Many studies have successfully shown how sensory and prior information can be combined to optimally interpret our environment. Because of the multiple sound localisation cues available in the binaural signal, sound localisation models based on Bayesian inference are a promising way of explaining behavioural human data. An interesting aspect is the consideration of dynamic localisation cues obtained through self-motion. Here we provide a review of the recent developments in modelling dynamic sound localisation with a particular focus on Bayesian inference. Further, we describe a theoretical Bayesian framework capable to model dynamic and active listening situations in humans in a static auditory environment. In order to demonstrate its potential in future implementations, we provide results from two examples of simplified versions of that framework.

A.1 Introduction

Sound localisation is a primary function of the human auditory system. Besides its well established evolutionary advantages (Avan et al., 2015), it is a crucial process for attention control and self-orientation. Proper understanding and implementation of the cues responsible for localisation is relevant for a range of modern audio applications, such as binaural hearing aids and three-dimensional audio displays for augmented or virtual reality (Blauert and Braasch, 2020).

The binaural nature of the auditory system is of high importance for localisation of the lateral position (Fig. A.1) of a sound source (Blauert, 1997). Humans obtain information about the source through the interaural differences in time of arrival (interaural time difference, ITD) and level (interaural level difference, ILD). However, the ITD and ILD cues do not provide enough information for accurate localisation beyond the horizontal plane, as several source locations will give rise to nearly the exact same binaural cues in the so called 'cones of confusion' (Tobias, 2012). Monaural spectral cues, which result from the filtering properties of the outer ear, head and torso, carry additional information on the polar position of the source (Fig. A.1). This spectral information aids in resolving the ambiguity in the binaural cues (Wightman and Kistler, 1997).

The aforementioned ITD, ILD and spectral cues can be considered 'static', as they are usually obtained in a situation where neither the source, nor the head undergoes any movement. However, in addition to these static cues, the auditory system can also utilise 'dynamic' cues, which are obtained by either sound source or head movement. Thus, dynamic cues can be defined as the changes in static cues during motion. Dynamic cues are beneficial during sound localisation, especially for resolving front-back confusions (e.g. Perrett and Noble, 1997a). They aid localisation in some way, but their importance relative to static cues is still an active point of research. Furthermore, dynamic cues obtained from self-motion, e.g., head movements, bring the additional challenge of processing vestibulo-proprioceptive information. As a result, the majority of state of the art models for sound localisation do not include the use of head movements (Kim et al., 2013a).

There are many available models, which each focus on a specific aspect of binaural localisation, such as processing of binaural cues (Macpherson, 1991; Willert et al., 2006), spectral cues (Baumgartner et al., 2014; Reijniers et al., 2014), or reverberant environments (Braasch, 2002). Over the past decade, machine learning techniques have also been applied to the modelling problem (e.g. May et al., 2010; Ma et al., 2017). Despite promising results, these techniques require substantial amounts of training data and can be difficult to understand due to their black-box nature (Kothig et al., 2019).

Bayesian inference is a method to optimally combine information about a multivariate system, when relying on noisy observations only. Bayesian inference has often been shown to take place not only in human multisensory perception (Alais and Burr, 2004; Battaglia et al., 2003b; Ernst and Banks, 2002; Knill and Pouget, 2004; Shams et al., 2005) but also human perception based on multiple cues within a single modality (Jacobs, 1999; Bülthoff and Mallot, 1990; Landy et al., 1995). Because of the multiple sound localisation cues available in the binaural signal, sound localisation models based on Bayesian inference seem to be a promising way of explaining behavioural human data. Temporal integration and learning can be modelled through recursive Bayesian estimation, where probabilities and estimates are updated recursively over time with incoming measurements (Cox and Fischer, 2015; Zonooz et al., 2018).

This article has two main purposes. First, we review the relevance of dynamic cues and their role in existing models of sound localisation, with a particular focus on the implementation of Bayesian inference. Second, we describe a recursive theoretical framework for dynamic listening through Bayesian inference. This framework aims at modelling dynamic listening situations which involve stationary sound sources in combination with head movements.

A.2 Static listening

A.2.1 Acoustic features and perceptual cues

Sound source localisation consists of determining the position of a source in three dimensions comprising two angles and the distance. In the interaural-polar coordinate system, the two angles are defined as the lateral and polar angles relative to a single pole passing through the two ears, i.e., the interaural axis. The interaural-polar system as used in (Majdak et al., 2010; Baumgartner et al., 2014; Barumerli et al., 2020) with a fixed distance between the listener and the source is illustrated in Fig. A.1. The static physical acoustic cues for localisation are captured by the (binaural) head-related transfer functions (HRTFs), which describes the filtering of the sound for a given direction by the listener's anatomy as recorded at the two ear drums.

For sound-source localisation along the lateral angle, the two main cues are the ITDs and the ILDs which are caused by the wave propagation time difference and the shadowing effects of the head, respectively (Blauert, 1997). ITD as a function of lateral angle roughly follows a sine shape, with zeroes on the median plane and maxima on the interaural axis (Shaw, 1974). This means that small displacements around the median plane produce larger changes in ITD than the same displacements at the lateral sides of the head. The ILD calculated for a spherical head model does not show maxima on the interaural axis, but for locations 45° on either side of that axis (Shaw, 1974). For narrowband sounds, the ILD cues are dominant at the middle

to high-frequency range of human hearing, and the ITD cues are particularly important for low frequencies (Rayleigh, 1907). This is known as the duplex theory of sound localisation. For broadband sounds, which encompasses most natural signals, both cues have substantial weight, but ITD dominates for most listeners (Macpherson and Middlebrooks, 2002).

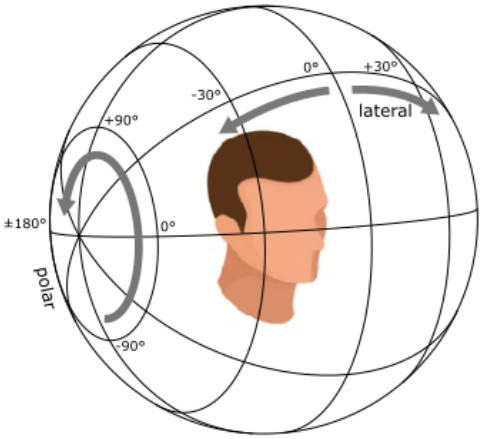

Fig. A.1 Interaural-polar coordinate system as used in Majdak et al. (2010); Baumgartner et al. (2014); Barumerli et al. (2020). The lateral angle of $0°$ describes sources located on the median plane. The polar angle of $0°$ describes sources located on the horizontal plane at eye level. Note that in this system, the lateral angle increases to the left, providing the advantage that for sources located at eye level, the lateral angle coincides with the azimuth angle of the widely used spherical coordinate system.

ITDs and ILDs produced by a sound at one location are ambiguous cues as they can also be produced by a sound at any location on the surface of a cone centred on the interaural axis, an effect known as the 'cone of confusion' (Blauert, 1997). Thus, in addition to these interaural broadband cues, the asymmetric and convoluted shape of the outer ears functions as a direction-dependent filter by causing frequency-dependent interference before sound waves reach the ear drums. The spectral cues introduced at each ear provide spatial information along the sagittal planes that helps to disambiguate the cones of confusion (Wightman and Kistler, 1997) which results in smaller elevation errors and a reduced so-called quadrant error rate, i.e., rate of confusing the spatial quadrant of the source direction, including the confusions between front and back and top and bottom (Middlebrooks, 1999; Majdak et al., 2010). Thus, the interaural-polar coordinate system provides a simple but complete representation of all sound directions from the perceptual perspective (Morimoto and Aokata, 1984), in which the lateral angle depends mostly on interaural cues and polar angle depends mostly on monaural spectral cues (see Fig.A.1).

Despite the varying contributions from different spectral regions, incoming sound must comprise sufficient energy in the relevant frequency region to make use of the spectral cues by the

auditory system (King and Oldfield, 1997; Zonooz et al., 2019). The human pinna's most prominent spectral notch related to the sound's polar angle falls within the 6 to 9 kHz band, which varies systematically and monotonically with polar angle (Zonooz et al., 2019). Acoustic features above 9 kHz still dependent on the polar angle, but they vary in a much more complex way. As a general upper limit, frequencies up to l6 kHz are evaluated by the auditory system in order to localise the direction of a sound (Hebrank and Wright, 1974). As for the lower frequency limit, sounds below 4 kHz have wavelengths that are too large to be affected by the dimensions of the pinnae and, thus, the resonances are direction independent (Jiang et al., 2019). Additionally, the effectiveness of monaural cues is highly listener specific, due to individual head and ear morphology (Wenzel et al., 1993). This is a prominent issue in 3D auditory displays for sound presentation over headphones (McAnally and Martin, 2014) as such systems require listener-specific HRTFs to reproduce the spectral cues with full accuracy.

Note that in this article we focus on the direction, and put less attention on the third dimension, the distance. On the one hand, distance perception is closely linked to sound reverberation (Zahorik et al., 2005). On the other hand, our proposed framework is expandable to consider additional variables and more complex problems. For example, in the near field (distances below 1 m), ILDs become a significant cue for the disambiguation of source location (Shinn-Cunningham et al., 2000) and this information could be used to extend our considerations to more complex localisation scenarios. Also, auditory motion parallax can be exploited to assess the relative distances of two sound sources (Genzel et al., 2018).

A.2.2 Ill-posed problem and prior information

Even for sound sources that meet the requirements above, polar angle estimation is still argued to be a mathematically ill-posed problem (Ege et al., 2018), as the spectrum of the signal at the eardrum results from a time-domain convolution of two unknowns: the actual source spectrum and the particular direction-dependent HRTF. This means that a priori knowledge of the source spectrum and/or direction helps to differentiate between spectral cues resulting from the source properties and from the filtering by the pinnae. Thus, a listener with an a priori knowledge is better able to estimate the pinna filtering characteristics from an incoming sound and associate those characteristics with the appropriate source position.

Despite the ill-posed problem for sound's polar-angle estimation, human localisation performance can be accurate and precise for most sound directions. Thus, to estimate the most likely direction, the auditory system seems to complement the acoustic cues with non-acoustic information about sounds and the environment. For example, the auditory system considers certain parts of the sensory information to be more reliable than others, such as different weightings on

different frequency bands (Zonooz et al., 2019). With respect to sound localisation, priors emphasising the central directions helped to describe the systematic underestimation of peripheral source directions in owls (Fischer and Peña, 2011). There also appears to be a clear mapping between frequency and elevation estimation, where high pitch is consistently mapped to high positions and vice versa (Parise et al., 2014). A priori assumptions such as the HRTF being unique for each sound elevation and natural source spectra not resembling HRTFs helped in modelling the process of sound localisation (Ege et al., 2019).

Interestingly, sound-localisation mechanisms seem to be independent along the horizontal and vertical dimensions, providing evidence that they may be embedded as distinct strategies to deal with spatial uncertainty in the acoustic environment (Ege et al., 2018). In the same study, azimuth estimation did not require a prior. Conversely, elevation estimation did require a prior in the form of a Gaussian spatial distribution centred around the horizontal plane. This is in line with the current understanding of multisensory perception where priors are independently encoded (Beierholm et al., 2009). The elegant inclusion of such prior information is a major advantage of a Bayesian framework.

In vision, priors have been discovered, e.g., observers tend to underestimate the speed of an object as they initially assume them to move stationary or move slowly (Weiss et al., 2002). These priors may also apply to audition. In fact, an analogous prior for low velocity of auditory sources has already been suggested (Senna et al., 2015; Freeman et al., 2017).

A.3 Dynamic listening

Our acoustic environment is in constant motion, due to both animate objects and listener movements. This makes the dynamic listening problem twofold (Carlile and Leung, 2016): 1) How is motion perceived and encoded in the auditory system? 2) How can a listener disambiguate moving sources from the apparent motion caused by head rotation? Both questions will be addressed in this section.

It is important to distinguish here between ‘passive’ dynamic listening and ‘active’ dynamic listening, particularly because several definitions exist in different fields related to audition (Barnett-Cowan and Harris, 2011; Cooke et al., 2007; van der Heijden et al., 2018; Portello et al., 2014). According to our definition, passive dynamic listening involves a dynamic acoustic environment without the employment of head movements, i.e., dynamic cues are solely produced by moving sound sources. In contrast, in the active dynamic listening situation listeners can rotate their heads and obtain dynamic cues even from static sound sources.

There are three degrees of freedom considered in most research related to dynamic listening: head rotation around the z-axis (yaw), head turn around the y-axis (pitch) and head tilt around the x-axis (roll), see Fig. A.2. Note that in this article, we focus on the directional localisation process, thus we do not consider head translations, which are usually related to distance estimation (Lu and Cooke, 2011).

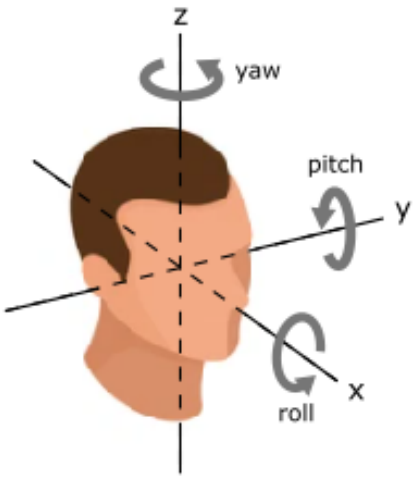

Fig. A.2 Three degrees of freedom in head orientation: yaw (head rotation), roll (head pivot), and pitch (head tip).

A.3.1 Acoustic features and perceptual cues

Wallach first suggested that dynamic ITDs and ILDs associated with head rotation are used to refine localisation accuracy, especially along the cones of confusion (Wallach, 1940). He argued that head yaw rotations would eliminate front-back ambiguity due to the contrasting change in the interaural cues provided by a stationary source (Macpherson, 2013). The contribution of dynamic cues to resolve front from back has since been empirically shown multiple times (Wightman and Kistler, 1999; Thurlow and Runge, 1967; Begault et al., 2001), especially in conditions in which spectral cues are not fully accessible to the auditory system (Ashby et al., 2014; Morikawa et al., 2013; McAnally and Martin, 2014). Dynamic interaural cues contribute more to front-back resolution than dynamic monaural spectral cues (Burger, 1958) and dynamic ITD is a more salient cue than dynamic ILD (Macpherson, 2013).

Not only yaw rotations produce a strong dynamic cue, head rolls provide supplementary information to resolve up-down confusions (Jiang et al., 2019). The contribution of yaw and roll to the process of sound localisation based on ITD can be investigated with the so-called ITD angular rate, i.e., $dITD/d\alpha$ with α being the source angle along a given rotation axis. ITD angular rate describes the change in ITD caused by the change in the source direction (Pavão et al., 2020). Fig. A.3 shows ITD angular rates for the three rotation axes and sources placed over the

full sphere. The ITDs were calculated from the HRTFs of a mannequin (KU 100, Neumann, Germany) available from the THK SOFA database (Bernschütz, 2013). Figure A.3 shows that yaw and roll induce large ITD rates providing a strong cue to resolve the cone of confusion. The head pitch, on the other hand, does not seem to evoke significant ITD rates, i.e., it does not provide dynamic cues to sufficiently resolve the cone of confusion.

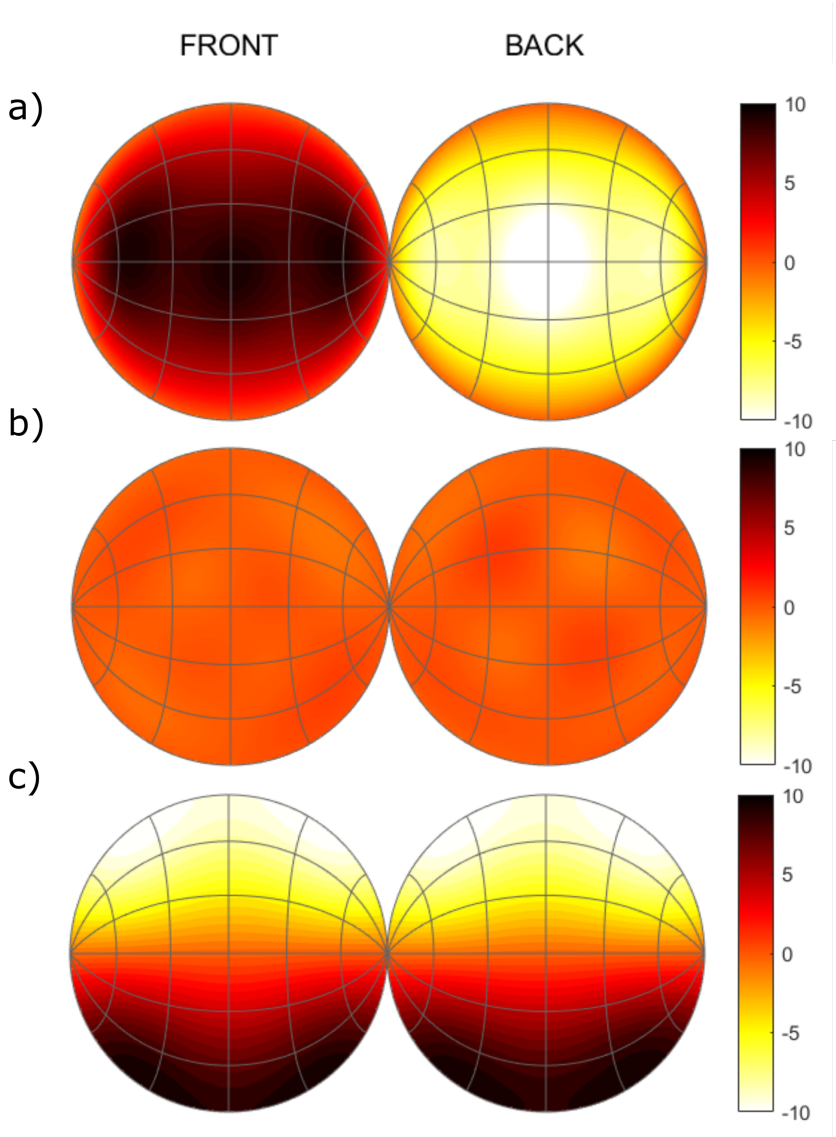

Fig. A.3 ITD angular rates ($dITD/d\alpha$) in µs per degree. Here $d\alpha$ corresponds with a positive rotation around the **a)** z-axis (yaw), **b)** y-axis (pitch), and **c)** x-axis (roll). Left and right panels represent source locations in the front and back of the head, respectively.

Dynamic cues also help in estimating the elevation of a sound source. As it can be deduced from Fig. A.3a, the ITD angular rates caused by head yaw depend on the source elevation angle (Wallach, 1940). They are large for sources placed on the horizontal plane and nearly zero for

those placed directly above or below a listener. This relation is shown in more detail in Fig. A.4, which shows the ITD angular rate for head yaw ($dITD/d\alpha$ with α being the lateral angle) as a function of the polar angle for sources located on the median plane. The auditory system is able to evaluate these ITD rate differences and associate them with the sound elevation (Wallach, 1940; Perrett and Noble, 1997a; McAnally and Martin, 2014; Kim et al., 2013a; Ashby et al., 2014). The sensitivity of this feature varies with the elevation. In reference to Fig. A.4, the ITD yaw rate is largest for sources on the horizontal plane and the smallest for sources above and below the listener. The opposite is true for the slope of the ITD yaw rate, i.e., the slope is steeper for higher polar angles. The steepness of the slope may explain why elevation estimation based on dynamic ITD only improves for elevations greater than $30°$ above or below the horizontal plane (Perrett and Noble, 1997a). Generally, the relation between the dynamic ITD rate and elevation perception seems to be quite complex, as it seems to further depend on the stimulus bandwidth (Ashby et al., 2014) and might even be supported by dynamic spectral cues (McAnally and Martin, 2014), but not in a monaural listening situation (Hirahara et al., 2021).

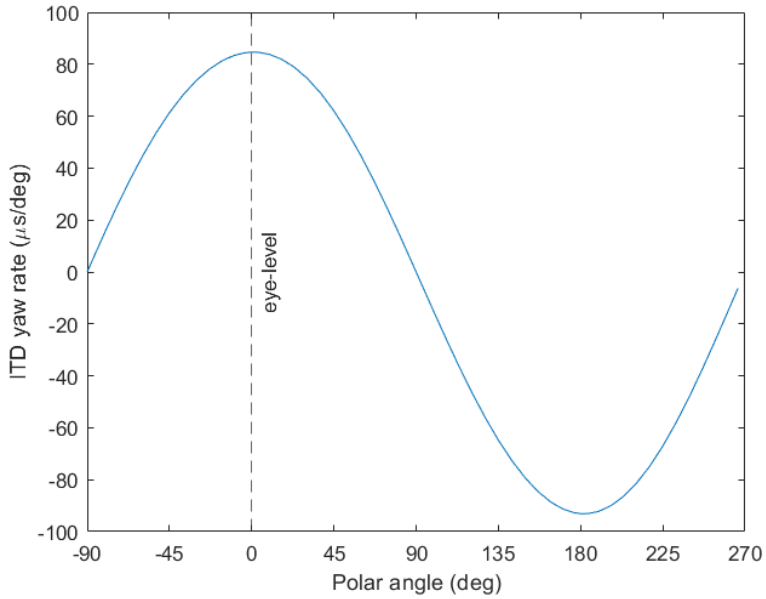

Fig. A.4 ITD yaw rates (in µs per degree) calculated for the median plane as a function of the sound-source polar angle. Note the sinusoidal shape indicating that the ITD yaw rates do not change linearly with polar angle, showing the largest values at eye level (polar angle of zero), but the largest changes above or below the listener (polar angle of ± 90).

Head movements do not always improve localisation. Brief sounds played during an ongoing head movement may even degrade localisation accuracy (Leung et al., 2008; Honda et al., 2016). During rapid head turns, the auditory space can be perceived as distorted or 'smeared'

(Cooper et al., 2008), which indicates that rotation speed may be a relevant parameter in the process of sound localisation. Interestingly, those distortions only occurred when the sounds were presented near the end of the head turn indicating a complex interaction between the head rotation and perceived auditory space. On top of that, all the spatial cues can vary temporally and our brains need to integrate the information somehow in order to obtain a stable image of the environment. Unfortunately, it is not completely clear yet how the brain accomplishes this task (Gerken et al., 1990).

So far, there is no evidence for peripheral neurons sensitive to auditory motion (Carlile and Best, 2002; Carlile et al., 2014; Freeman et al., 2014), in contrast to those found in the visual system. Still, humans are able to faithfully track a sound's unpredictable movements in the horizontal plane with smooth-pursuit responses of the head, which in turn supports the existence of a pursuit system for auditory head-tracking (Calvo et al., 2021). This is supported by neurons in the midbrain (inferior colliculus and medial geniculate nucleus) sensitive to dynamic cue changes. This suggests the existence of a higher-level neural network estimating sound motion, similar to that of third-order (acceleration) motion detectors found in vision for cats (Al'tman et al., 1985), bats (Pollak, 2012), guinae pigs (Ingham et al., 2001), and barn owls (Wagner and Takahashi, 1992; McAlpine et al., 2000). These networks are heavily modulated by attention (Boucher et al., 2004) and have been measured by means of electroencephalography (EEG) (Kreitewolf et al., 2011), providing further evidence for higher cognitive processes involved in decoding sound velocity in humans. Taken together, sound motion is most probably tracked by sampling the estimated source position and integrating that information by higher stages of the auditory system (Middlebrooks, 2015; Carlile and Leung, 2016), rather than by a continuous measurement of sound velocity in the peripheral stages.

It is generally accepted that the auditory system depends on a type of 'temporal integration' (Loveless et al., 1996). It is important here to distinguish between the operation of mathematical integration (as the term 'temporal integration' seems to imply) and the actual process in the analysis of time-variant information. In cognitive sciences, temporal integration considers a variety of models working on various time scales (Teng et al., 2016). For example, in the 'multiple looks' model, samples or 'looks' are taken from the acoustic features, stored in memory, and can be selectively accessed and processed (Viemeister and Wakefield, 1991). When applied to the process of sound localisation, the auditory system seems to integrate acoustic information over a duration of approximately 5 ms to form a single look, which are then combined through a leaky integrator, with a stable composite estimation requiring a stimulus duration of approximately 80 ms (Hofman and Van Opstal, 1998).

More specifically, static elevation estimation seems to require 40 to 80 ms of broadband in-

put(Hofman and Van Opstal, 1998). For static lateralisation, stable performance can be achieved with stimuli as short as 3 ms (Vliegen et al., 2004). During dynamic sound-source localisation, the sound localisation system seems to require a minimum 100 ms of input to yield an improved estimate (likely due to the process of vestibular-auditory integration), with the stimulus duration above 100 ms further improving the localisation performance (Macpherson, 2013; Perrett and Noble, 1997a).

Doppler shift, i.e., the frequency shift caused by the motion of the sound source and/or the listener, is an additional dynamic cue that must be noted. Interestingly, within a single frequency band, the binaural Doppler equation results in mathematically equivalent results as the ITD angular rate (Baumann et al., 2015). Despite its implementation in robotic systems (Kumon and Uozumi, 2011), there seems to be no evidence that human use the Doppler effect to localise sound sources. When considering moving sources, however, humans are indeed able to utilise the Doppler shift as a cue for velocity discrimination (Lutfi and Wang, 1999).

A.3.2 Integration of proprioceptive information

Listeners are capable of dissociating self-motion from source motion, with the largest apparent difference being the additional sensory feedback from the vestibular and proprioceptive systems in the case of self-motion (Carlile and Leung, 2016). Consequently, the contribution of self-motion implies the consideration of proprioceptive information in modelling the localisation process.

In the human auditory system, acoustic cues are encoded in an egocentric representation, i.e., head-centred reference frame (Vliegen et al., 2004). In the process of spatial inference, the frame of reference needs to be transformed from egocentric to allocentric, i.e., world-centred information about the environment (Schechtman et al., 2012). The auditory system is able to compensate for head rotations during the perception of sound-source motion, though this compensation seems to be incomplete (Freeman et al., 2017). Complementary information from other senses, integrated with the acoustic input can help to better estimate the allocentric spatial properties of the environment. In fact, mechanisms responsible for building an allocentric frame of reference are based on multisensory processing (Lewald and Karnath, 2000; Viaud-Delmon and Warusfel, 2014; Yost et al., 2015).

Thus, it is not surprising that in the process of sound localisation, information needs to be integrated from many systems such as the vestibular system, proprioception or from efference copies of motor commands (Goossens and Van Opstal, 1999). For example, performance in a dynamic spatial auditory task improved when dynamic cues were generated by self-induced

head motion rather than by the source itself (Brimijoin and Akeroyd, 2014). However, front-back confusions that initially did not resolve with source movement were in fact resolved when source movements were controlled by the listener (Wightman and Kistler, 1999), suggesting that head movements may not be required to produce dynamic cues to resolve front–back ambiguity and, instead, that the listener's priors, e.g., additional information on the direction of the source, contribute strongly.

The extent to which the various information channels contribute to the process of spatial calibration remains an open question. There is a strong indication that vestibular-auditory integration takes place in the sound localisation process as indicated by the requirement of a long stimulus duration (in the range of 100 ms) for an effective use of dynamic auditory cues (Macpherson, 2013). However, in that study, listeners rotated their head at a constant velocity and the stimulus was played when the head orientation entered a selected spatial window. Because of this, movement initiation and acceleration, in which vestibulo-proprioceptive mechanisms may play a prominent role, were not tested. Experiments that made subjects orient themselves "straight ahead" found that proprioceptive input from the neck region does significantly impact the subjective body orientation in humans, even though the effect was smaller than that found for vestibular stimulation (Karnath et al., 1994). In order to clarify whether the vestibulo-proprioceptive information integrated with acoustic cues is derived from vestibular or proprioceptive systems, several head and and body movement conditions were tested in a sound localisation experiment (Kim et al., 2013b). The proprioceptive information did not improve localisation indicating that the vestibular inputs are sufficient to inform the auditory system about head movement. In line with these findings, work by Genzel et al. (2016) shows that auditory updating is dominated by vestibular signals, though they did find significant contributions from proprioception/efference copy. Even eye position (Lewald and Ehrenstein, 1996) and audiovestibular interaction (Van Barneveld and John Van Opstal, 2010) seem to affect the spatial localisation, which further complicates the understanding of the contribution of vestibulo-proprioceptive information to the process of sound localisation. It is apparent, however, that the vestibular system is dominant in many of the tested scenarios.

A.3.3 Active-listening strategies

Borrowing from control theory terminology, active listening can be subdivided into open-loop and closed-loop listening. In open-loop listening head movements do not depend on the sound source. For closed-loop listening, the listener adjusts the head movements in response to the perceived sound in a feedback loop, adapting the movement for the duration of the sound signal. This makes closed-loop listening a task-dependent problem. Closed-loop listening can be beneficial to 'triangulate' a source, to decrease interference from reverberation, and to attend to

a single (moving) source in a complex listening environment (Cooke et al., 2007). Naturally, closed-loop listening strategies are only possible if there is enough time to react. In a dynamic listening task, localisation accuracy could be improved with signals as short as 50 ms, but only in the cases where the listener responds with a head movement within the duration of the stimulus (Perrett and Noble, 1997a). Note that, besides head movement, eyes can also be moved as a reaction to a sound (Goossens and Van Opstal, 1997), but because they do not change the auditory signal, we do not consider them in this article.

In an unconstrained listening situation, i.e., any head movements allowed, listeners utilise yaw more often than pitch or roll (Thurlow et al., 1967; Morikawa et al., 2013). This is in line with the observation that yaw produces the most informative dynamic cues, compare Fig. A.3. In a closed-loop listening situation, listeners can also orient their head towards the source. Indeed, in another listening task, a majority of the subjects rotated their head toward the direction of the source (Muir and Field, 1979; Fuller, 1992). This behaviour may be beneficial for several reasons such as the spatial centring the acoustic image of a sound and the alignment of the visual system with the source of the sound (Brimijoin et al., 2010). The horizontal localisation is best in the area around the frontal half of the median plane (Mills, 1958; Oldfield and Parker, 1984) and thus may be considered as a neuro-computational auditory fovea, which somewhat resembles the visual ocular pursuit system (Calvo et al., 2021). Following this, a listener's intention may be to orient their head such that the source direction is within the field of highest spatial resolution (Blauert, 1997). Furthermore, listeners also tend to make reversals in head movements, i.e., rotating their head back and forth (Muir and Field, 1979; Fuller, 1992). By doing so, a continuum of dynamic cues is produced, which when integrated, may improve the estimation of the position of a stationary sound source.

It may be unnecessary to consider all physically attainable orientations of the head, because a confined area around the initial position covers the majority of head positions in natural listening situations. Even though humans can rotate their heads on the yaw axis as far as $\pm 70^\circ$, the listeners do not seem to rotate their heads to this extent (Kim et al., 2013a). Small head rotations (up to $\pm 16^\circ$) already significantly reduce the rate of front-back confusions, though larger movements are required to also significantly reduce elevation errors (McAnally and Martin, 2014). Head movements are smaller for broadband noise than for narrowband noise (Morikawa et al., 2013), indicating an inverse relationship between spectral content and the required rotation angle.

It is important to note that all the aforementioned studies report large individual differences in the head movements. The optimal manner of obtaining dynamic information may be subject-dependent because of differences in morphology and hearing capabilities. It is, however, rather

likely that untrained and fully unconstrained listeners do not inherently know how to utilise dynamic cues. In a speech perception experiment, untrained listeners did not make optimal use of the dynamic cues (Grange and Culling, 2016). In fact, some listeners did not move at all, some rotated directly to near-optimum orientations, while others moved gradually and erratically. After being instructed on the head movements, listeners' behaviour became more coherent and performance improved indicating that listeners are capable of quickly learning new strategies in order to optimise their head movements. However, there seem to be only little advantage from 'free' (i.e., no instructions) over 'forced' (i.e., an instructed direction and speed) rotations (Perrett and Noble, 1997a; Thurlow and Runge, 1967). In summary, inclusion of individual listener strategies in an active listening model would require the consideration of a task-dependent variable driving the head orientation, freely chosen at each moment of time.

A.4 Bayesian models

There is a general consensus that in order to estimate a sound-source direction, the human auditory system performs a comparison between incoming acoustic features and their learned representation (Middlebrooks, 1992, 1999; Reijniers et al., 2014; Zonooz et al., 2019). In other words, the models assume a template-matching process, in which the auditory system maintains a stored library of templates of the acoustic information associated with each sound-source direction. When a stimulus is perceived, the listener then compares it to the templates. Given some prior assumptions, the localisation estimate then corresponds to the direction for which the template fits most closely.

This procedure can be well represented in the Bayesian framework, in which the probability of an occurring event may be affected by prior knowledge about the event, e.g., how frequently a stimulus previously occurred at a given position. The inclusion of a prior probability is what distinguishes this method from other interpretations of probability. A multitude of studies on multimodal perception suggests that the brain uses a Bayesian approach to combine stimuli during estimation of spatial localisation (Shams et al., 2005; Alais and Burr, 2004; Körding et al., 2007; Gu et al., 2008; Ursino et al., 2017) and to learn and adapt to changes in the environment (Körding and Wolpert, 2004; Stocker and Simoncelli, 2006; Hudson et al., 2007; Carlile et al., 2014; Ege et al., 2019).

A.4.1 Bayesian estimation

In Bayes' theorem (in terms of probability density functions),

$$p(\psi|\mathbf{y}) = \frac{p(\mathbf{y}|\psi)p(\psi)}{p(\mathbf{y})},$$

the posterior probability density function (PDF) $p(\psi|\mathbf{y})$ of direction ψ of a source given acoustic information $\mathbf{y}$ depends on three factors: 1) The likelihood $p(\mathbf{y}|\psi)$, representing the PDF of acoustic information $\mathbf{y}$ being observed for a source at direction ψ; 2) The prior PDF $p(\psi)$, representing assumptions on the result, derived from the past experience on the parameter to be estimated; and 3) The denominator $p(\mathbf{y})$, representing the PDF of acoustic information $\mathbf{y}$ being observed and assumed to be a normalisation constant inferred from $\int p(\psi|\mathbf{y})d\psi = 1$, so that the area under the posterior PDF integrates to 1.

When formulating the sound source localisation problem as a Bayesian decision problem, the listener first determines the posterior PDF given both prior and sensory information. Next, a loss function is defined on the set of source directions and by using the posterior PDF to minimise the expected loss the 'best' estimate of the source direction is determined.

If the loss function specifies the minimisation of the probability of error, the optimal Bayesian decision rule selects the maximum of the posterior PDF, a strategy known as the maximum-a-posteriori (MAP) strategy (Bahl et al., 1974). Note that in the special case of the prior being a uniform PDF, the MAP strategy obtains the same result as maximum likelihood estimation, which returns the parameter value ψ that maximises the likelihood. Interestingly, the auditory system does not always rely on a point estimate like the MAP rule and a random sampling strategy from the posterior PDF seems to better explain the localisation process in some conditions (Ege et al., 2018).

Bayesian inference is a widely used approach in investigating various auditory effects. For example, Bayesian inference can help to reconstruct localisation cues from listener responses to random cue spectra (Hofman and Van Opstal, 2002), or to investigate how fluctuations of binaural cues in realistic noisy listening conditions affect localisation performance (Nix and Hohmann, 2006), or to investigate the trading between accuracy and precision in the sound-localisation process (Ege et al., 2018).

Bayesian inference and the template-matching procedure have been combined to model sound localisation based on ITDs and spectral acoustic features of a stationary source (Reijniers et al., 2014). That Bayesian ideal-observer model was able to predict empirical sound localisation errors, reproducing patterns observed in human localisation experiments. It can be seen as a first step in modelling active dynamic localisation and is, in fact, a simplified example for our model later (see Sec. A.5.5).

A.4.2 Recursive Bayesian estimation

There is a variety of methods to introduce time dependency to Bayesian models (Barber et al., 2011), especially when it comes to derive a decision in complex dynamic systems (Mark et al., 2018). Recursive Bayesian inference is one of these techniques and can be applied to fit a statistical model to data in a series of steps (Särkkä, 2013). The methodology of recursive Bayesian inference can be defined as a two-step process that recursively cycles through a prediction step (which "predicts" a prior PDF of current state based on the old estimate) and an update step (which "updates" the state PDF to form a posterior estimate by taking the newly available measurement into account).

A popular way to approximate the recursive Bayesian estimation in discrete state-space is through linear-Gaussian models, i.e., Kalman filters, which assume that the true state of system model $\mathbf{X}(t_i)$ at time t_i linearly evolves from the state at time t_{i-1}. In order to process non-linear systems, adaptations of the Kalman filter can be used, such as extended Kalman filters or unscented Kalman filters (Wan et al., 2001). The Kalman filter and its variants update the process mean, i.e., the state, and its variance at each iteration, making it the optimum (minimum error) estimator when the noises are Gaussian. However, multimodal noises may bring the Kalman filter to instability, which prevents it from converging to the mean. In order to handle such multimodal ambiguities as well as non-linear models, the particle filter method (or sequential Monte Carlo) has been developed (Li, 2014).

The switch between linear and non-linear systems can even be required when solving a single problem. For example, in an investigation of auditory-based prey capture by the barn owl (Cox and Fischer, 2015), both linear and non-linear recursive Bayesian estimations were used to predict a source's future direction, given a sequence of sensory observations and a prior PDF for direction and angular velocity. A linear relationship between prey direction and ITD was assumed for prey in the frontal hemisphere allowing for the use of a Kalman filter. For more lateral sound directions, however, the linear approximation did not apply and a particle filter was required to compute the Bayesian prediction. Note that while that model is a dynamic model, it utilises dynamic cues from source motion only, not from head movements.

Recursive Bayesian inference was also used to model the external nucleus of the inferior colliculus (Willert et al., 2006), which is thought to be responsible for the transformation from a frequency-specific code for spatial cues into a topographic code for space (Cohen and Knudsen, 1999). A similar mechanism was able to explain how an input with a limited spectrum can improve elevation estimation after training (Zonooz et al., 2018).

In the field of auditory scene analysis (Bregman, 1994), recursive Bayesian estimation has been

used to improve the process of resolving individual sources in a complex acoustic environment (Hambrook et al., 2017). That model is based on dynamic ITDs and uses a simple recursive approach, in which the prior PDF of sound positions is stored in a spectrospatial map and the incoming short-time maps constitute new evidence. As a simplification, the dynamic cues of the head movements are translated to equivalent inverse changes in the source position. The model also assumes perfect control of the head orientation. Thus, it does not reflect a realistic dynamic listening situation, in which the actual head orientation needs to be a random variable with an uncertainty, because of the inherent error in the sensorimotor system. Still, it can be considered as a case of 'idealised' dynamic listening.

Much can also be learnt from research in robotics, as there are many existing models for multisensory integration and motion strategies (Luo and Chang, 2011; Schymura et al., 2014, 2015; Ma et al., 2017; May et al., 2015). However, many of these studies have focused on developing artificial auditory systems that are less applicable to research on human binaural listening, such as the use of microphone arrays (Aarabi, 2003; Valin et al., 2007). Nonetheless, the techniques applied can prove useful for more biologically plausible models, especially for dynamic listening. For example, performance of an extended Kalman filter has been improved by introducing additional a priori information verifying the consistency of location estimation at each time step (Kumon and Uozumi, 2011).

A.5 Modelling active listening in a static environment

In this section, we propose a state-space model to describe the problem of active sound localisation, explain the generative model used in the Bayesian framework, and derive the posterior PDFs of the head orientation and source direction. We also demonstrate the feasibility of our model in two simplified examples.

In our model, we limit the source to be stationary with respect to the head movements within the considered temporal interval in the model. This is justified by assuming that our framework is a part of a larger framework of causal inference, in which the listener tests various hypotheses on the auditory environment and at the moment of probing the environment, the most probable assumption is an environment consisting of non-moving auditory objects. This assumption is further upheld by the empirical data supporting the 'slow-motion prior' in listening (Senna et al., 2015; Freeman et al., 2017).

We allow the listener to actively control the head movements. While head rotations can be described in a general way by using quaternions, in our article, we use a simplified description by limiting head movements to yaw rotations only. Finally, we assume a 'multiple looks' model

whereby the listener collects information at discrete time steps during the motion and updates the evidence step by step.

A.5.1 State-space model

We describe the listener as a dynamic state-space process, where in all instances of listening, the listener needs to determine the posterior PDF of the source direction and that PDF needs to be updated recursively as more evidence or information becomes available. This recursive estimation process is important in dynamic models including temporal integration and relies on the Markov assumption: the future is independent of the past given the present (Fosler-Lussier, 1998). The state-space representation utilises state variables, of which the values evolve over time in a way that depends on their current value (i.e. state) and on the input variables. Using this definition and the Markov assumption, we denote the true state of the system as $\mathbf{X}$. This true state is hidden to the listener, who can only observe the state via a noisy measurement that we denote as $\mathbf{y}$,

$$\begin{aligned} \mathbf{X}(t_i) &= g(\mathbf{X}(t_{i-1}), u(t_{i-1}), \delta_u(t_i)), \\ \mathbf{y}(t_i) &= h(\mathbf{X}(t_i), \delta_x(t_i)), \end{aligned} \tag{A.1}$$

where g and h are called the system model and the measurement model, respectively, u is the system control function, and δ_u and δ_x are the system and measurement noise, respectively.

In the context of a sound localisation process with the source positioned in the far-field, the state information required by the listener to localise a source consists of the head orientation $\theta_H(t)$ and the source direction ψ a 2D vector defined by the lateral and polar angles of the source

$$\mathbf{X}(t) = (\theta_H(t), \psi)^T,$$

with both the head orientation and the source direction measured relative to the torso that we assume as the link between the egocentric and allocentric spatial systems.

The control signal $u(t)$ is represented by the speed of rotating the head, i.e., $u(t) = \omega_z(t)$. Thus, the discrete-time dynamic state-space model g is formulated as

$$\mathbf{X}(t_{i+1}) - \mathbf{X}(t_i) = (\omega_z(t_i)\Delta t + \delta_u, \psi(t_{i+1}) - \psi(t_i))^T, \tag{A.2}$$

with Δt the time step of the 'multiple looks' model and δ_u the noise on the self-motion representing the difference between the intended and the actually executed head movement. Note that, in a stationary auditory environment, the difference between the previous and current sound

direction, $\psi(t_{i+1})$ and $\psi(t_i)$, respectively, is zero.

The measurement equation of the proposed state-space model consists of two components

$$\mathbf{y}(t_i) = (y_A(t_i), y_H(t_i))^T.$$

The first component, $y_A(t_i)$, describes the acoustic features used by the auditory system in the process of sound localisation, as described in Sec. A.3. For example, when resolving front-back confusions, the system may rely on dynamic ITDs only, resulting in $y_A(t_i) = ITD(\theta_H(t_i), \psi) + \delta_A$. In that example, the noise in the acoustic features is modelled as unbiased Gaussian noise $\delta_A \sim \mathcal{N}(0, \sigma_{y_{ITD}})$.

The second component, $y_H(t_i)$, describes the measurement of the head orientation as $y_H(t_i) = \theta_H(t_i) + \delta_H$ and assumes that the listener knows the head orientation relative to the torso up to some additive (unbiased) Gaussian noise $\delta_H \sim \mathcal{N}(0, \sigma_{y_H})$.

While the aforementioned noise sources are described as additive, control of movement (Todorov, 2005) and stimulus perception (Stern and Johnson, 2010) are generally assumed to not only be endowed with additive but also with multiplicative noise, in which the standard deviation of the noise is linearly related to the amplitude of the signal. A possible solution in the case of our framework can be transforming the signals to a space of constant variance.

Note that we make no assumptions about the linearity of the acoustic component of the measurement equation. Note also that we do assume that the head orientation and acoustic measurement processes are independent, in particular that the acoustic features are not used by the listener to estimate the head orientation. This is directly exploited in the following sections: in order to translate the acoustic measurements into source direction, the listener needs information about the head orientation and we simplify the active sound localisation problem considerably by first estimating the head orientation (Sec. A.5.3), and subsequently making use of that information in the process of estimating the source position (Sec. A.5.4). In order to retain the assumption of independent noise in the underlying measurement processes, we assume the delays involved in the process to be broadband.

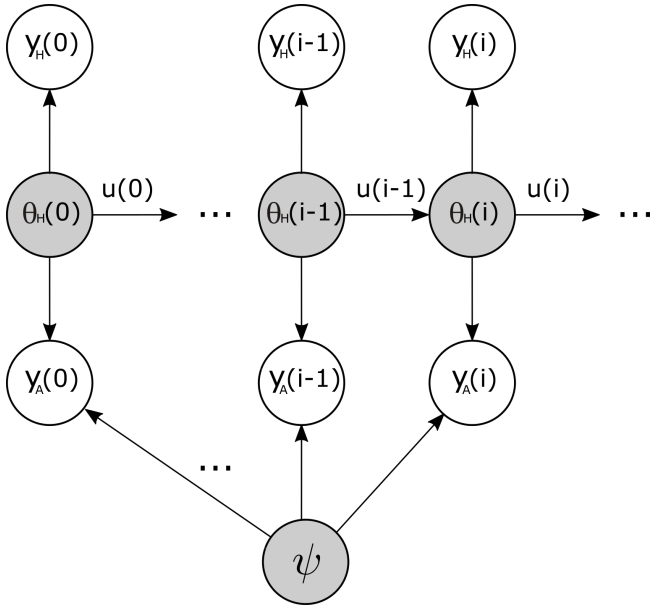

Fig. A.5 Bayesian network describing the dynamic listening situation. The white and grey circles represent observed and hidden variables, respectively. The arrows denote conditional dependencies. ψ denotes the stationary sound source direction.

A.5.2 Generative model

In the framework of Bayesian inference, we assume that the listener wants to determine the source direction based on all prior information about the environment and on all sensory information collected during the head movement. To this end, the listener first determines the posterior PDF of the source direction. The desired posterior PDF, taking into account all information available at time t_i will be denoted as,

$$p_{t_i} = p(\psi \mid u(0:i-1), \mathbf{y}(0:i)), \tag{A.3}$$

with $u(0:i-1)$ denoting the sequence of control signals, i.e., rotation speed, applied at times t_0 until t_{i-1} with the time step Δt, and $\mathbf{y}(0:i)$ the sequence of sensor readings, i.e., acoustic features and head orientation, collected at times t_0 until t_i. Figure A.5 illustrates the generative model describing this posterior PDF. Note that, while we assume the source to be stationary, we make explicit the time varying nature of our knowledge about the source direction by taking into account all relevant information available at time t_i. Thus, we refer to this distribution by the shorthand p_{t_i}.

In Bayesian decision theory, the posterior PDF is usually used to minimise a loss function, that describes the optimal point estimate of the source direction. In closed-loop listening, the posterior PDF, computed at each update of the recursive process, can be further used as the input for a head-movement strategy. An example is the smooth posterior mean strategy (Schymura

et al., 2015), which makes the listener steer the head on a smooth trajectory towards the posterior mean of the source position during each iteration. While the definition of a relevant loss function and its minimisation is both of theoretical interest and required for practical implementations (see our numerical examples in Sec. A.5.5), it is beyond the scope of this article having its main focus on the derivation of the posterior PDF.

A.5.3 Estimation of head orientation

We model the dynamic process of the head rotation by

$$\begin{aligned} \theta_H(t_{i+1}) &= \theta_H(t_i) + \omega_z(t_i)\Delta t + \delta_u, \\ y_H(t_{i+1}) &= \theta_H(t_{i+1}) + \delta_H, \end{aligned}$$

with the initial head orientation given by a normally distributed variable $\theta_H(t_0) \sim \mathcal{N}(\theta_0, \sigma_0)$ representing the listener's uncertainty about the initial head orientation. The additive noise on both the movement equation and the sensor equation is assumed to be zero-mean white Gaussian noise $\delta_u \sim \mathcal{N}(0, \sigma_u)$ and $\delta_H \sim \mathcal{N}(0, \sigma_{y_H})$.

Making use of Bayes' rule and taking into account all head rotations executed as well as all sensor readings collected so far, the PDF of the head orientation at time t_{i+1} during the head movement can be shown to be Gaussian and given by

$$p(\theta_H(t_{i+1}) \mid y_H(0:i+1), u(0:i)) = \mathcal{N}(\hat{\theta}_H(t_{i+1}), \sigma_{\theta_H(t_{i+1})}), \quad \text{(A.4)}$$

with mean and variance

$$\begin{aligned} \hat{\theta}_H(t_{i+1}) &= (1-K)\cdot(\hat{\theta}_H(t_i) + \omega_z(t_i)\Delta t) + K \cdot y_H(t_{i+1}), \\ \sigma^2_{\theta_H(t_{i+1})} &= (1-K)\cdot(\sigma^2_{\theta_H(t_i)} + \sigma^2_u), \end{aligned}$$

and

$$K = \frac{\sigma^2_{\theta_H(t_i)} + \sigma^2_u}{\sigma^2_{\theta_H(t_i)} + \sigma^2_u + \sigma^2_{y_H}}.$$

The prior required to initiate the recursive process is based on two components: the prior knowledge available to the listener about the sound source direction $p(\psi)$ and the prior knowledge $p(\theta_H(0) \mid y_H(0))$ available to the listener about the initial head orientation. We assume here that $\theta_H(0)$ is a Gaussian distribution centred on an initial head orientation $\hat{\theta}_H(0)$ as described in Eq. A.6, but that may depend on the actual experiment being modelled.

Note the recursive nature of these equations as well as their correspondence with a Kalman filter implementation of the head orientation estimation process.

A.5.4 Estimation of sound-source direction

Here we derive a recursive expression for Eq. A.3 describing the posterior PDF p_{t_i} at time $t_i = t_{i-1} + \Delta t$ in terms of the prior PDF $p_{t_{i-1}}$ derived at time t_{i-1} combined with the extra information from the most recent 'look' in the sequence of 'multiple looks' collected during the head movement. We assume the sensor readings and the control signals to be available to the estimation process as

$$\mathbf{y}(0:i) = ((y_A(t_0), y_H(t_0))^T, (y_A(t_1), y_H(t_1))^T, \cdots \\ (y_A(t_i), y_H(t_i))^T)$$

and

$$u(0:i-1) = [\omega_z(t_0), \omega_z(t_1) \cdots \omega_z(t_{i-1})],$$

i.e., we assume a varying speed of head rotation around the yaw-axis (which remains constant during each time-step Δt).

In the first step, as the source direction is part of the full state $\mathbf{X} = (\theta_H, \psi)^T$, we derive the desired PDF described by Eq. A.3 from the joint full-state PDF by marginalisation over all possible head orientations

$$p_{t_i} = p(\psi \mid \mathbf{y}(0:i), u(0:i-1)) = \\ \int_{\theta_H} p_X(\psi, \theta_H(t_i) \mid \mathbf{y}(0:i), u(0:i-1)) d\theta_H$$

This operation allows us to correctly take into account the effect of the remaining head orientation uncertainty on the source estimation. The joint PDF p_X can be expanded as

$$p_X(\psi, \theta_H(t_i) \mid (y_A(t_i), y_H(t_i))^T, \mathbf{y}(0:i-1), u(0:i-1)) = \\ p(\psi \mid \theta_H(t_i), y_A(t_i), \mathbf{y}(0:i-1), u(0:i-1)) \times \\ p(\theta_H(t_i) \mid y_H(t_i), \mathbf{y}(0:i-1), u(0:i-1)).$$

Our knowledge of the head orientation taking into account all relevant data up until time t_i can be described by a reformulation of Eq. A.4

$$p(\theta_H(t_i) \mid y_H(0:i), u(0:i-1)) = \mathcal{N}(\hat{\theta}_H(t_i), \sigma_{\theta_H(t_i)}).$$

In the second step, by using Bayes' rule, we rewrite the first term of p_X

$$p(\psi \mid \theta_H(t_i), y_A(t_i), \mathbf{y}(0:i-1), u(0:i-1)) = \frac{p(y_A(t_i) \mid \theta_H(t_i), \psi) \times p(\psi \mid \mathbf{y}(0:i-1), u(0:i-1))}{p(y_A(t_i) \mid \theta_H(t_i), \mathbf{y}(0:i-1), u(0:i-1))},$$

and simplify it with

$$p(\psi \mid \mathbf{y}(0:i-1), u(0:i-1)) = p_{t_{i-1}},$$

i.e., the posterior distribution on the source direction we determined at time t_{i-1}.

By combining these results we obtain the recursive expression for the posterior PDF

$$\begin{aligned} p(\psi \mid \mathbf{y}(0:i), u(0:i-1)) = p_{t_i} = C \cdot p_{t_{i-1}} \times \\ \int_{\theta_H} p(\theta_H(t_i) \mid y_H(0:i), u(0:i-1)) \times \\ p(y_A(t_i) \mid \theta_H(t_i), \psi) d\theta_H \end{aligned} \tag{A.5}$$

with the normalisation constant C derived from the posterior PDF

$$\int_{\psi} p(\psi \mid \mathbf{y}(0:i), u(0:i-1)) d\psi = 1.$$

The recursive process in Eq. A.5 is comparable to Hambrook et al. (2017) and expresses our new state of knowledge about the source direction p_{t_i} based on the previous state of knowledge $p_{t_{i-1}}$ with the extra knowledge obtained in the most recent 'look'. This extra knowledge takes into account not only the acoustic measurement but also the current head orientation estimate.

The recursive processes is initiated with the source direction PDF from Eq. A.3 derived from the initial acoustic measurement performed at time t_0

$$p_{t_0} = p_{\psi}(\psi \mid (y_A(t_0), y_H(t_0))^T).$$

This PDF, following a similar derivation as the one described for Eq. A.5, is given by

$$\begin{aligned} p_{t_0} = C \cdot p(\psi) \times \int_{\theta_H} p(\theta_H(t_0) \mid y_H(t_0)) \times \\ p(y_A(t_0) \mid \theta_H(t_0), \psi) d\theta_H, \end{aligned} \tag{A.6}$$

with $p(\theta_H(t_0) \mid y_H(t_0)) = \mathcal{N}(\hat{\theta}_H(t_0), \sigma_{\theta_H(t_0)})$ representing the initial (uncertain) head orientation and $p(\psi)$ being the prior PDF on the source direction. Note that at this very first moment

t_0 the head motion is not made use of, as at that moment only the current head orientation is known (up to some uncertainty). As with Eq. A.5, the constant C can be derived from the normalisation of this posterior PDF.

Depending on the behavioural task and listener's priors on the environment, this prior information $p(\psi)$ (in Eq. (A.6)) may substantially modify the model predictions. For example, it may constrain the possible source directions to a sub-region of the full sphere around the listener's head, e.g., the frontal hemisphere or the horizontal plane. This can be modelled by choosing $p(\psi)$ accordingly.

A.5.5 Numerical examples

The presented concept is mathematically consistent, yet its complete numerical evaluation is not trivial: it depends on the considered acoustic features and needs to consider many listening situations. Thus, it deserves separate discussions in future articles. In this article, we illustrate the explanatory power of the concept by numerically applying it to two examples in a simplified setting, using an MAP estimator to convert the posterior PDF into a point estimate. Note that the MAP estimator is just one of the possibilities to obtain a point estimate from the PDF.

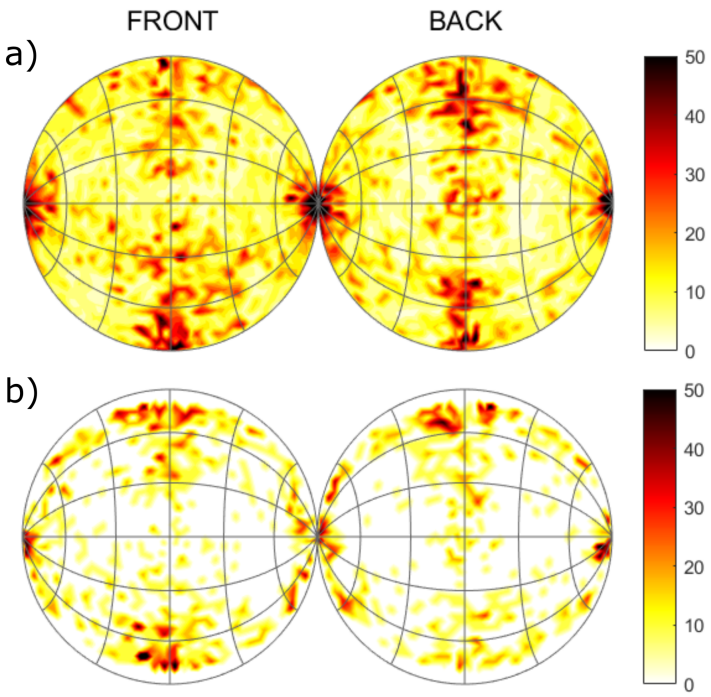

Fig. A.6 Example 1: Predictions obtained from 10 simulated trials of the simplified concept without head movements, head orientation straight ahead. **a)** Polar angle errors (in degrees). **b)** Front-back confusion rates (in %); rates for target directions near the frontal plane are not shown for clarity .

In the first example, we simplify our concept to the sound localisation with head oriented straight ahead and without any head movements (not even noise), i.e., $u = 0$, and $y_H = 0$, respectively. This simplifies Eq. A.2 to $\mathbf{X}(t_{i+1}) = \mathbf{X}(t_i)$ and Eq. A.3 to $p = p_\psi(\psi \mid y_A)$. The resulting model corresponds to the ideal-observer model by Reijniers et al. (2014). Results from that model were replicated by using our implementation of the simplified concept. To replicate the results from the original study, we used the same acoustic features y_A: ITDs, the summed binaural spectral information, and the interaural spectral information. We also used the same noise parameters tuned to the ITD thresholds and absolute hearing thresholds. For the input signal, we used Gaussian white noise filtered with HRTFs from Sec. A.3.1, which were also used as templates in the model. 2354 target directions on the spherical grid were considered. Per target sound direction, 10 trials were simulated and averaged to obtain polar errors and front-back confusion rates. Figure A.6 shows the polar errors and front-back confusion rates obtained with that simplified model. The polar errors are largest above and below the listener, which is qualitatively in line with the observations obtained from actual localisation experiments (Best et al., 2011). The large polar errors near the interaural axis can be attributed to the disproportional changes in polar angles even for small changes of the source direction (Majdak et al., 2010).

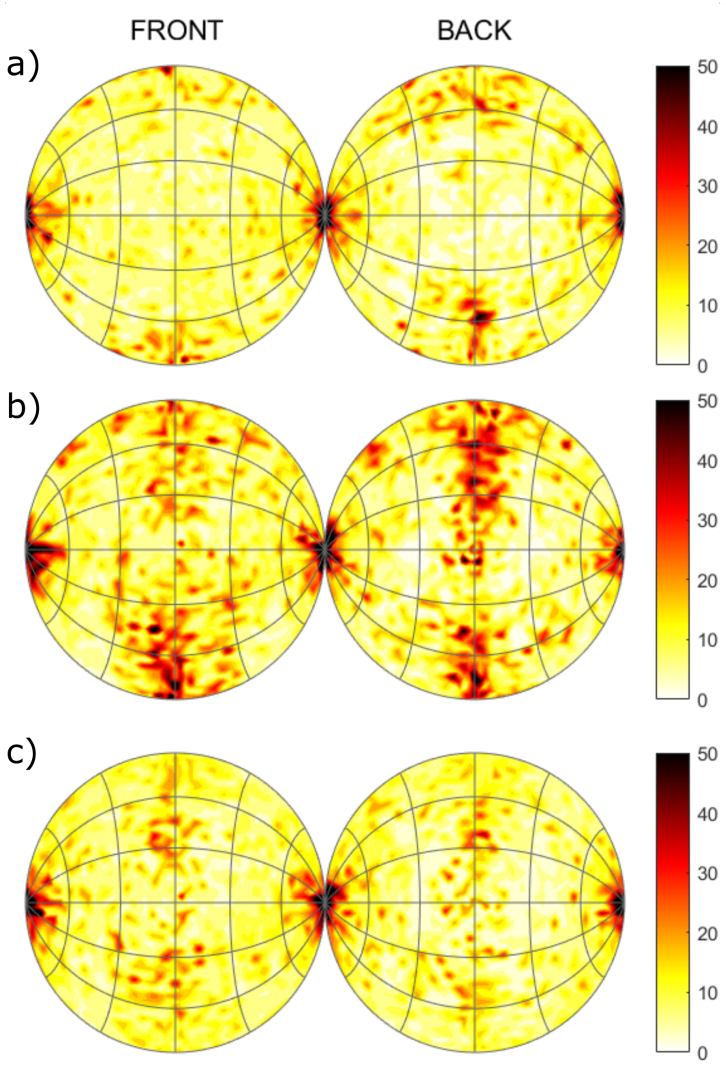

Fig. A.7 Example 2: Polar angle errors (in degrees) obtained from 10 simulated trials of the simplified concept with a 10° turn and perfectly determinable head orientation. **a)** Head yaw rotations. **b)** Head pitch turns. **c)** Head roll tilts. Left and right panels represent source locations in the front and back of the head, respectively.

The second example demonstrates the recursive nature of the estimation process in our concept. We reduce the concept to head rotations that are small (i.e., up to 10°), open-loop, and at a constant speed (i.e., $u = \omega_z$), so we can assume a linear relationship between ITD and head-rotation angle (Cox and Fischer, 2015). Head position is assumed to be exactly known, i.e., a deterministic y_H not confounded by any noise. Further, we used the broadband ITD as acoustic feature, i.e., $y_A(t_i) = ITD(\theta_H(t_i), \psi) + \delta_A$ and used a uniform distribution for the prior PDF. Taking into account these simplifications, Eq. A.3 reduces to $p_{t_i} = p(\psi \mid \mathbf{y}(0:i))$ with $\mathbf{y}(t_i) = (y_A(t_i), \theta_H(t_i))$. The predictions were calculated for yaw, roll, and pitch separately, with identical simulation parameters from the first example. Figures A.7 and A.8 show

the polar errors and front-back confusion rates, respectively. The results show that head yaw significantly reduces front-back confusions, though roll also shows a slight improvement. Head yaw and roll also reduced polar errors, with most of the reduction for sources located at eye level. All these findings are in line with empirical data (Perrett and Noble, 1997a; McAnally and Martin, 2014). Head pitch, on the other hand, did not show much improvement, neither in reducing the localisation errors nor the front-back confusions, both being in line with the observation of small ITD rates when tilting the head (see A.3c). These two examples demonstrate the feasibility of the concept.

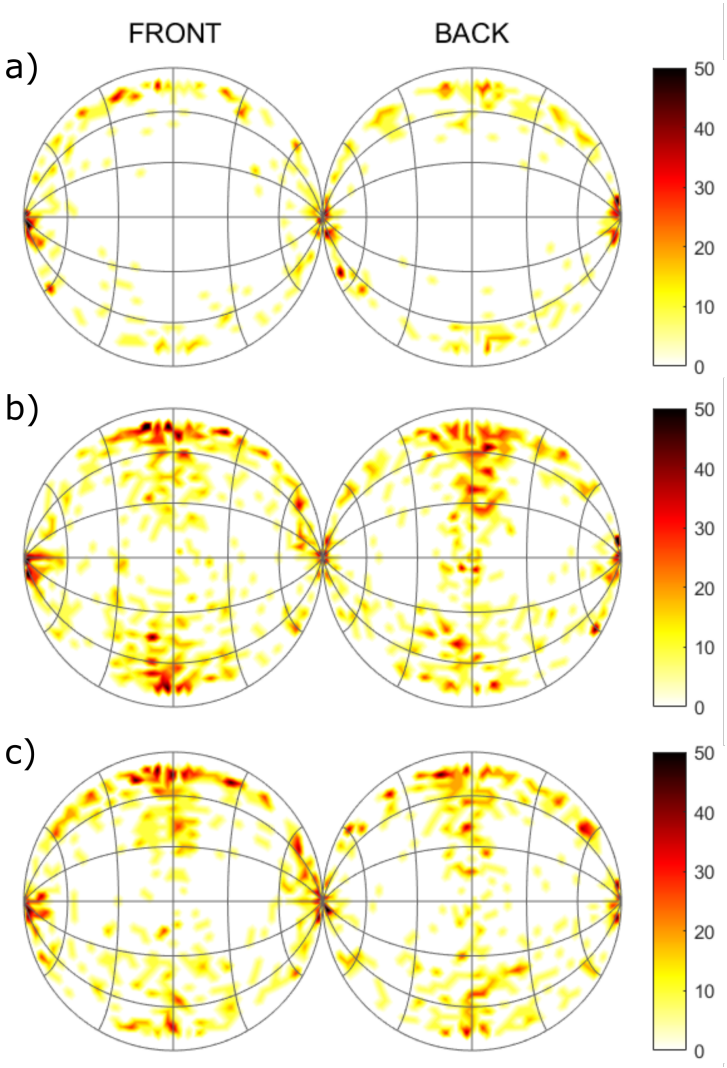

Fig. A.8 Example 2: Front-back confusion rates (in %) obtained from the simplified concept with a 10° turn and perfectly determinable head orientation. **a)** Head yaw rotations. **b)** Head pitch turns. **c)** Head roll tilts. For clarity rates for target directions near the frontal plane are not shown. Left and right panels represent source locations in the front and back of the head, respectively.

A.6 Conclusions

This article briefly reviews the recent literature on modelling active dynamic sound localisation, which complements the well-documented sound localisation based on static acoustic features with cues from self-motion or source motion. The review focuses on Bayesian inference because of its prominent role in recent multimodal cognitive models and high potential in modelling dynamic cognitive processes. We have defined the term of active dynamic sound localisation, describing the localisation process in which the listener actively updates the head orientation to facilitate the localisation process.

Further, we described a theoretical Bayesian modelling framework based on the independent estimation of acoustic features and head rotations. In order to show the feasibility of the concept, we provide two short examples of simplified versions of the concept, for which numeric implementations are available. While these two examples do not fully validate the concept in all its aspects, they demonstrate the potential of the proposed concept towards a general dynamic sound localisation model.

Future work will involve model extensions along different directions. First, model parameters will need to be fine-tuned through sensitivity analyses and comparisons to empirical data in order to quantitatively fit the predictions to the sound localisation performance of humans. Second, an implementation of a closed-loop version of the model will be required to completely test the concept. Here questions related to listening strategies will become relevant. Additionally, our current concept considers stationary sources only. It can be extended to dynamic auditory environments by integrating e.g. a multiscale network (Ferreira and Lee, 2007), expanding our concept to a general framework of active sound localisation in dynamic auditory environments.

PAPER B

Dynamic spectral cues do not affect human sound localisation during small head movements

Glen McLachlan[1], Piotr Majdak[2], Jonas Reijniers[1], Michael Mihocic[2], Herbert Peremans[1]

[1] Department of Engineering Management, University of Antwerp, Belgium
[2] Acoustics Research Institute, Austrian Academy of Sciences, Austria

Abstract - Natural listening involves a constant deployment of small head movement. Spatial listening is facilitated by head movements, especially when resolving front-back confusions, an otherwise common issue during sound localisation under head-still conditions. The present study investigated which acoustic cues are utilised by human listeners to localise sounds using small head movements (below $\pm 10^{\circ}$ around the center). The stimulus conditions included the interaural time difference (ITD), interaural level difference (ILD), monaural spectral shape (MSS) and their dynamic counterparts: dynamic ITD (dITD), dynamic ILD (dILD) and dynamic MSS (dMSS). They were presented under three movement conditions: no movement, head rotations over the yaw axis and over the pitch axis. The results show that even small yaw rotations provide a remarkable decrease in front-back confusion rate, whereas pitch rotations did not show much of an effect. Furthermore, MSS cues improved localisation performance even in the presence of dITD cues. However, performance was similar between stimuli with and without dMSS cues. This indicates that human listeners utilise MSS cues before the head moves, but do not rely on dMSS cues to localise sounds when utilising small head movements.

B.1 Introduction

Human sound localisation experiments and models have historically predominantly focused on passive localisation, i.e., where the head is held still. However, head movements have repeatedly been shown to improve sound localisation (Wallach, 1940; Wightman and Kistler, 1999), especially for sources with little or distorted spectral content (Morikawa et al., 2013; Perrett and Noble, 1997b; Kato et al., 2003). This has led to increasing interest in the influence of head movements on the localisation performance.

Unfortunately, modelling active sound localisation is significantly more complex than its passive counterpart. The three major acoustic cues that humans use during sound localisation are the interaural time difference (ITD), interaural level difference (ILD), and monaural spectral shape (MSS) cues (Blauert, 1997). In spatially static listening conditions, i.e., static sound source and without head movements, ITDs and ILDs contribute to the sound localisation in the lateral dimension and listener-specific MSS cues are required to achieve sound localisation performance in sagittal planes (Majdak et al., 2020). In a dynamic environment, i.e., with moving sources and listeners, the acoustic cues that the auditory system processes change over time, and thus all have a dynamic counterpart: dynamic ITD (dITD), dynamic ILD (dILD) and dynamic MSS (dMSS) cues. In theory this gives the auditory system many acoustic cues from which the source's location can be inferred, but whether this happens in practice is currently a point of debate. Moreover, active listening requires a separate movement model to complement the acoustic model (McLachlan et al., 2021). This demands insights on the vestibular and motor systems and the way that humans coordinate these systems in conjunction with the auditory system.

Due to these added complexities, it is of great interest to simplify auditory modelling wherever possible. For example, we may consider which cues are essential to localisation performance and which could be omitted. Modelling dITD is relatively easy due to its near-linear link with the lateral position of the sound (Cox and Fischer, 2015), whereas the link between the dMSS and source direction seems to be more complex and would result in a considerably more complicated localisation model. The movement model can also be simplified, e.g., by restricting it to only small head rotations along a single axis. Such small rotations can be approximated as constant velocity, modelling the movements as a simple first-order derivative for the dynamic acoustic cues without having to consider translation or acceleration. Further, small head rotations can be seen as the first step in understanding and modelling complex movement behaviour, because any natural head rotation can be decomposed into a sequence of smaller rotations. Outside the context of modelling, small head movements are important to investigate explicitly because of their constant presence in natural listening (Carlile and Leung, 2016), and from

an evolutionary standpoint it may be expected that the brain exploits these movements if they provide additional information.

For horizontal movement of sound sources, the minimum audible movement angle is about 2° at velocities up to $15^\circ/s$ and about 8° at $115^\circ/s$. When motion is restricted to the vertical, minimum audible movement angles are substantially larger at all velocities, ranging from 6° to 12° (Saberi and Perrott, 1990). This data suggests that listeners are more sensitive to (dynamic) binaural cues (ITD and ILD) than to (dynamic) spectral cues (MSS), though all three can be perceivable within a head rotation of 10°, depending on velocity. These results quantify movement discrimination, but it is still unclear how this affects localisation performance.

The purpose of this study was to investigate the effects of small head movements (up to $\pm 10^\circ$) on sound localisation for normal hearing subjects, and determine which dynamic cues are responsible for these effects. Based on current evidence, the hypothesis was that dITD is the dominant dynamic acoustic cue. Further, this cue was expected to have an even more profound positive effect in conditions in which spectral cues are not available. Finally, it was also expected that dMSS cues do not affect localisation performance. This hypothesis is supported by the findings that dMSS cues during monaural listening (Hirahara et al., 2021) and pitch rotations (i.e., rotations along the interaural axis) (Thurlow and Runge, 1967; Kato et al., 2003) have no effect on sound localisation.

B.2 Methods

B.2.1 Subjects

Seven normal-hearing subjects (6 female, 1 male) participated in the experiment. Their absolute hearing thresholds were within the average (± 1 standard deviation) of the age-relevant norms (Corso, 1959; Stelmachowicz et al., 1989) within the frequency range from 0.125 to 16 kHz. The age range of the subjects was between 22 and 30 years. None of the subjects were familiar with this type of experiment.

B.2.2 Apparatus

The experiment was conducted in the loudspeaker array studio of the Acoustics Research Institute of the Austrian Academy of Sciences in Vienna. The studio is a semi-anechoic room consisting of 91 speakers (E301, KEF Inc.) distributed over the sphere within the elevation angles from -47° to 90°. The room was equipped with a head-mounted display (HMD, Oculus Rift, CV1, Meta Inc.) for the visual presentation of virtual reality and three infrared cameras for the tracking of the listener within the six degrees of freedom. The HMD was worn in all

conditions. The setup also provided equipment for the measurement of listener-specific HRTFs by inserting microphones (KE 4-211-2, Sennheiser Inc.) into the ear canals and measuring the responses from the individual loudspeakers. The HRTF measurement procedure corresponded to that from Majdak et al. (2010), without the HMD in place.

The experiment was controlled by a computer running a 64-bit Windows 10, equipped with a 8-core, 3.6-GHz CPU (i7-11700KF, Intel Inc.), 16 GB of RAM, and a graphic card with dedicated 8 GB of RAM (GeForce RTX 3070, NVIDIA Inc.). The computer was controlled by custom software framework ExpSuite version 1.1 which provides modules for various types of stimulation (multi-channel via loudspeakers, binaural via headphones), for controlling visual interfaces via virtual reality, for tracking systems, and many other functionalities (Majdak and Mihocic, 2022). The HRTF measurements were run by the ExpSuite application AMT@ARI version 7.0.31. The experiment was run by the ExpSuite application LocaDyn version 0.8. All ExpSuite applications are freely available (Majdak and Mihocic, 2022).

Free-field signals were presented via the 91 loudspeakers, each driven by an amplifier (Sonible d:24, sonible GmbH) connected to a computer via sound interface (RME MADIface USB, RME Audio AG). In order to create virtual sound sources appearing from an arbitrary direction, vector-base amplitude panning (VBAP, Pulkki, 1997) was applied. For a requested direction of the virtual sound, three loudspeaker positions surrounding that direction were found and the amplitude gains for those loudspeakers were calculated. VBAP was implemented within the ExpSuite module YAMI100 version 1.3 running within the puredata version 0.49 (Puckette et al., 1996) environment. Note that only static virtual sound sources were used in this experiment.

The binaural signals were presented via open headphones (HD-650, Sennheiser Inc.). The spatialisation of the binaural virtual sources was done in real-time in the ExpSuite module SOFAlizer for Unity version 1.6 (Jenny et al., 2022) running within the Unity version 2020.3.34f1 environment. The binaural signals were updated in real time by capturing the subject's position and orientation with the tracking system of the head-mounted display. The tracking system provides an accuracy of 0.76 cm (in a sitting position, inside a room, Borrego et al., 2018), and a latency below 6 ms (Becher et al., 2018).The subject's position and orientation were recorded for later analyses.

B.2.3 Stimuli

The acoustic stimulus used in this experiment was a wideband (20 to 20000 Hz) white noise burst, gated with a 5-ms cosine ramp. The duration was 2000 ms for static listening and was gated off after 10° of head rotation for dynamic listening.

41 directions were selected to test the effects of localisation over different planes (see Fig. B.1). These directions were based on the results of a model for dynamic localisation (McLachlan et al., 2021) and the results from (Perrett and Noble, 1997b), which show the largest improvement of errors through head movement for sources around the median plane, especially for sources at high elevations. A higher source density was selected around this area of interest. The sources were further distributed over several sagittal planes, following the so-called 'cones of confusion'. Because of the left-right symmetry, positions on the midsagittal plane were tested twice, so that the set of tested positions consisted of 52 directions.

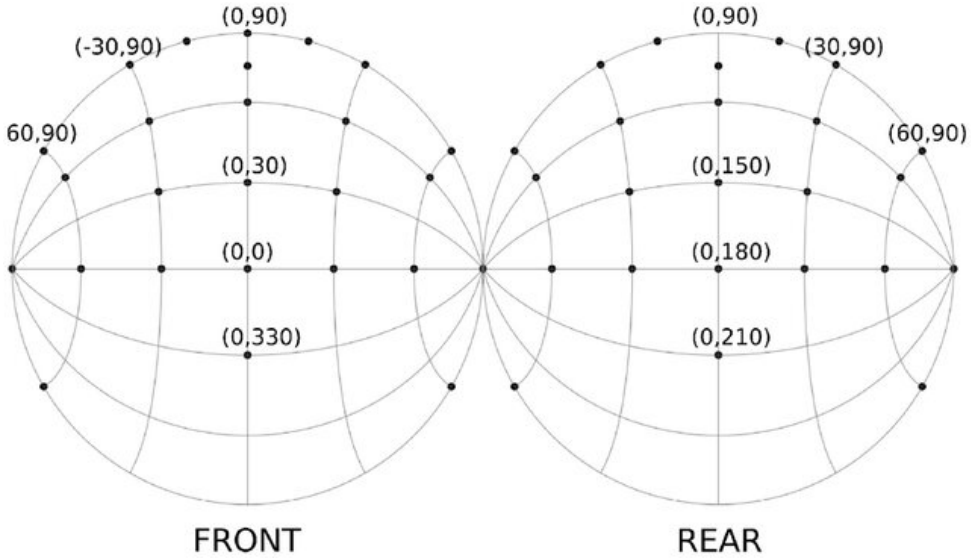

Fig. B.1 Spatial directions of the sound sources used in the experiment, plotted as two hemispheres cut through the frontal plane. In total 41 directions were used.

Playback over headphones used listener-specific HRTFs. The acoustically measured HRTFs were available for 91 spatial directions only. Thus, in order to create a smooth real-time dynamic listening environment, each listener-specific HRTF set was interpolated to a denser grid of 5762 positions. This dense grid was created by subdividing the faces formed by the sparse grid. The subdivision was iterated three times. An HRTF of the dense grid was calculated by removing the time of arrival by calculating its minimum-phase version, interpolating the spectrum by applying VBAP based on the sparse directions, and introducing a dense time of arrival modelled by the spherical-head time of arrival model (Ziegelwanger and Majdak, 2014). Figure B.2 shows the acoustically measured and the interpolated HRTFs in the top and center rows, respectively.

B.2.4 Tested Conditions

Four acoustic cue conditions were tested, named by their aimed spectral properties: full, flat, frozen and free-field. The first three conditions were run using a virtual acoustic display, i.e.,

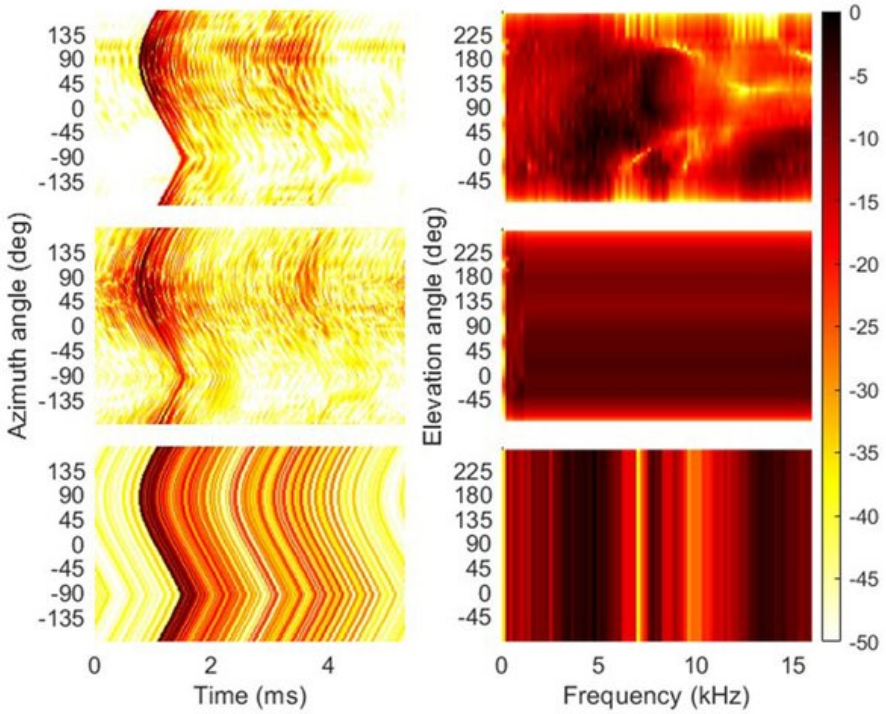

Fig. B.2 HRTFs of the subject NH257 as an example. Left column: Energy time curves of the impulse responses (in dB) calculated for the HRTFs of a left ear along the azimuth angle. Right column: Magnitude spectra of HRTFs (in dB) along the median plane as a function of the polar angle. Top row: HRTFs from the condition 'full'. Center row: HRTFs from the condition 'flat', here the MSS was significantly flattened, but the time of arrival, thus, the ITD and dITD cues remained unchanged. Bottom row: HRTFs from the condition 'frozen' to (0°,0°), i.e., for all spatial positions, the MSS was frozen to that of the spatial position in the front and at eye level, but the time of arrival, thus, the ITD and dITD cues, were identical to those from the actual spatial positions.

presented via headphones. The full condition used dense listener-specific HRTFs without any further processing. The flat condition used the same HRTFs as in the full conditions, but with flattened frequency-dependent contrasts between 1 and 16 kHz, leaving only (dynamic) ITD and ILD as acoustic cues. This processing was exactly as that in Baumgartner et al. (2017) for $C = 0$. The frozen condition used HRTFs with all cues available while the head was stationary, but considered only ITD changes when the head rotated. Thus, ILD and MSS cues remained static, i.e., frozen to the ILD and MSS for the head in the initial orientation. Figure B.2 shows the frozen HRTFs in the bottom row. The free-field condition refers to the loudspeaker-based stimulus presentation using VBAP to create a virtual sound source at the requested direction. This condition served as a sanity check for the headphone results, where, assuming an adequate headphone playback using listener-specific HRTFs, similar results were expected between the free-field and full conditions.

Three types of head rotations were tested: static (i.e., no head movement), yaw rotation, and pitch rotation. The two rotations were single-sided (i.e., from reference position $(0°, 0°)$ towards a specific point). The signal was gated off so that only $10°$ of the rotation contained acoustic information. The yaw rotation evokes large interaural changes whereas pitch rotation evokes minimal interaural changes but large monaural changes, enabling a clear comparison between the individual dynamic cues.

B.2.5 Procedure

B.2.5.1 Training

Each subject underwent acoustic training in the virtual reality environment in order to get familiar with the equipment and task, and to reach a baseline localisation performance (Middlebrooks, 1999). The training consisted of 300 trials with a 2000-ms white noise burst spatialised from a direction randomly selected from a uniform distribution over the available directions of the interpolated HRTF set.

At the start of each trial, the subjects oriented their head straight ahead by placing a crosshair over a visual target at $(0°, 0°)$ presented via the HMD. The stimulus was then played, during which the subjects kept their head still. At the end of the stimulus, the subjects pointed towards their perceived source direction with a hand-tracking device to provide their localisation estimate. Visual feedback was then provided of the actual source direction, and the stimulus repeatedly played until targeted a second time, so that the subjects could familiarise themselves with the available dynamic cues.

B.2.5.2 Static localisation

In the static localisation experiment, head movements were restricted by instructing the subjects to keep their head aligned with the reference position. Further, the binaural presentation did not consider subject's head movements. The experiment consisted of nine blocks (three repetitions of three stimulus conditions) of 52 trials, with each trial representing one tested sound direction. The stimulus was a 2000-ms white noise burst. Within each block, directions were selected from the grid in a random order, while ensuring that each direction was presented once. The order of blocks was ordered randomly across the listeners. No feedback was provided to the subject, neither on the subject's performance nor the actual sound-source positions.

B.2.5.3 Dynamic localisation

In the dynamic localisation experiment, subjects were instructed to make a specific one-sided rotation around either the yaw or the pitch axis, as soon as they heard the stimulus onset. At

the beginning of each trial, the subject was asked to rotate to the reference position (0 azimuth, 0 elevation) and an arrow was presented on the HMD to instruct the direction of head rotation, indicating the direction of either the pitch axis (up or down), or the yaw axis (left or right). The subject then confirmed to be ready. The head rotation speed was left unrestricted, but was monitored through the tracking system of the virtual reality headset and recorded for the analysis. Then the stimulus was played and the head orientation was recorded. After a head rotation of 10° was registered, the stimulus was again gated off over 5 ms, so that dynamic acoustic cues were only provided for 10° of rotation. Finally, the subject was asked to point to the perceived sound direction and confirm with a press of the button. This process was repeated for 12 blocks (3 repetitions for 4 stimulus conditions). Each block consisted of 208 trials (4 arrow directions for 52 source directions). All other details were as in the static-localisation task.

Note that this task of dynamic sound localisation differed from the head-sweep method used in dynamic sound localisation experiments (Macpherson, 2013). In this experiment, the movement was initiated after, not before, the stimulus onset. Therefore, this experiment did not test the isolated dynamic localisation performance, but instead tested the added benefit of head movements after a period of static localisation.

B.2.6 Data Analysis

Before investigating individual localisation performance metrics, head orientations recorded during the stimulus presentation (for dynamic and static free-field localisation) were checked for outliers. Responses linked with head rotations of less than 5° were excluded. Responses linked with rotations along the incorrect axis (i.e., axis orthogonal to the instructed axis) of more than 2° were also excluded.

Localisation performance was assessed by three metrics; lateral and polar precision errors (both in degrees), i.e., standard deviations (Middlebrooks, 1999; Majdak et al., 2010), and front-back confusions (FBC) rate, i.e., each response was labeled by 0 for the correct hemisphere and 1 for a reversal, then the rate (in %) of FBCs was calculated. In identification of front-back errors, target locations with an absolute lateral angle over 60° were ignored, and responses were allowed to cross the midline by 10°. For polar errors, the trials that resulted in an FBC were omitted before calculation to keep all metrics independent.

Each subject participated in all experiments, allowing an extensive within-subject analysis. The lateral and polar precision was analysed using a linear mixed-effects model. FBC data was analysed with a mixed-effects logistic regression. For both model types, the subject was treated

as a random variable. Stimulus type and rotation type, as well as the interaction between the two, were treated as fixed variables. As a follow-up analysis, the estimated (least-squares) marginal means were compared, using Tukey's adjustment for multiplicity. The statistical significance was considered below the levels of p of 0.05 as significant and p of 0.001 as highly significant. These statistical analyses were run in R version 4.2.1 (Müller, 2020).

B.3 Results

B.3.1 Head rotations

9984 responses were collected each for yaw and pitch rotation. They were checked for the head-movement related outliers. After the clean up, 8488 localisation responses were obtained for yaw rotation and 8364 estimates were obtained for pitch rotation. Subjects NH919 and NH1016 showed extremely slow head rotations compared to the other subjects, i.e., almost all trials employed a head rotation speed below $30°/s$. Still, there was no evidence that justified removing these data.

The maximum velocities of head rotations averaged over all subjects were low ($36.41 \pm 23.03°/s$ for yaw rotations and $28.54 \pm 17.50°/s$ for pitch rotations). Stimulus duration for dynamic listening was on average below one second with the averages of 645.53 ± 447.78 ms for yaw rotations and 802.12 ± 545.91 for pitch rotations.

Finally, reaction times of rotation initiation after the stimulus started were on average in the range between 100 and 200 ms. These reaction times are slightly lower than those found in previous studies, e.g., 200 to 300 ms in Perrett and Noble (1997a)). This can be explained by the button press before each trial, which may have primed the subject to be more responsive to the next stimulus.

B.3.2 Localisation performance

Figures B.3 and B.4 show the statistics of the localisation performance of all subjects. Both figures present the same data, but they are grouped differently for easier comparison across the types of head rotation (Fig. B.3) and the acoustic cues (Fig. B.4). Boxes show the lower and upper quartiles with the median represented by the line in the center, whiskers show the non-outlier minima and maxima, and the small circles show the outliers.

Fig. B.3 facilitates the comparison across the types of the head rotation by comparing across the rotation conditions, encoded by color, within each stimulus condition. Each column shows the performance metrics for no movement (blue), yaw rotation (red) and pitch rotation (yellow),

grouped by the available acoustic cues. Results of the statistical analysis of the differences across the types of head rotation (per acoustic cue) are presented in Tab. B.1.

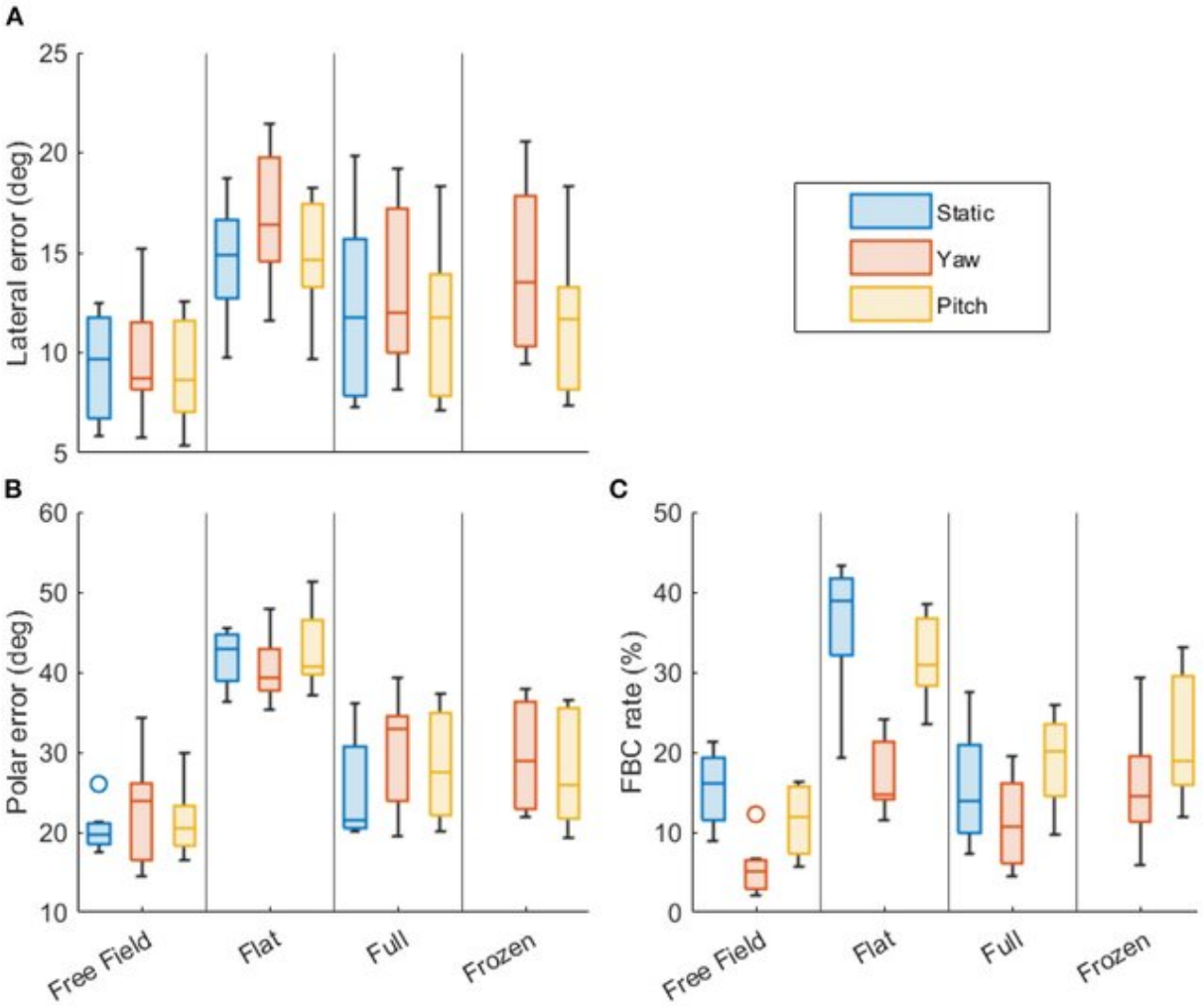

Fig. B.3 Localisation performance grouped by the available acoustic cues. (a): Lateral precision error (in degrees). (b): Polar precision error (in degrees). (c): Front-back confusion rate (in %). Lower values indicate better performance. Each boxplot shows the statistics of all subjects (median, first and third quartiles, minima and maxima, outliers).

Fig. B.4 facilitates the comparison across the available acoustic cues: Each column shows the performance metrics for the conditions ”flat” (dITD, blue), ”full” (dITD and dMSS, red), ”free field” (natural listening with all cues, yellow), and ”frozen” (dITD with MSS only, violet), grouped by the type of the head rotation. Results of the statistical analysis of the differences resulting from the available acoustics cues (per type of the head rotation) are presented in Tab. B.2.

Table B.1 Statistical significance of differences in three localisation performance metrics (lateral error, polar error, FBC rate) between the tested types of head rotation, grouped by the type of acoustic cues. Statistical significance is shown in bold.

Cue	Tested rotation	p lat err	p pol err	p FBC
Flat	static/yaw	0.6640	0.1172	**<.0001**
Flat	static/pitch	0.9999	1.0000	0.8634
Flat	yaw/pitch	**0.0007**	0.1979	**<.0001**
Full	static/yaw	0.9600	0.8780	0.2511
Full	static/pitch	0.9980	0.9988	0.6090
Full	yaw/pitch	**0.0073**	0.9999	**0.0010**
Free field	static/yaw	1.0000	0.6076	**<.0001**
Free field	static/pitch	1.0000	0.9604	0.4239
Free field	yaw/pitch	0.8441	1.0000	**0.0001**

Table B.2 Statistical significance of differences in three localisation performance metrics (lateral error, polar error, FBC rate) between the tested acoustic cues, grouped by the type of head rotation. Statistically significant results are printed in bold.

Rotation	Tested cue	p lat err	p pol err	p FBC
Static	full/flat	0.2540	**<.0001**	**<.0001**
Static	full/free field	0.3313	0.2498	1.0000
Yaw	full/flat	**0.0227**	**0.0089**	0.0868
Yaw	full/free field	0.0810	0.2736	**0.0001**
Yaw	full/frozen	0.9936	1.0000	**0.0098**
Yaw	frozen/flat	0.2236	**0.0093**	0.9997
Pitch	full/flat	0.0684	**0.0002**	**0.0014**
Pitch	full/free field	0.4723	0.3065	**0.0002**
Pitch	full/frozen	1.0000	1.0000	0.9129
Pitch	frozen/flat	0.1327	**0.0001**	**0.0085**

B.4 Discussion

B.4.1 General

As a first check, subject performance in conditions already tested in previous studies was analysed, with the goal to verify the general quality of this experiment's sound presentation. For the static listening situation in the free field, average localisation performance was 9.3°, 20.4°, and 15.4% (lateral error, polar error, and FBC rate, respectively). This is similar to the performance usually found in such experiments (e.g., 10.6°, 22.7°, 4.6% in Middlebrooks (1999)), with exception of the FBC rate. The larger FBC rate is surprising, however, it was not problematic for the experiment because it prevented floor effects by providing room for improvements when testing conditions including head rotations. The increased FBC appears to be related to the definition chosen to categorise FBCs. Indeed, when applying the FBC definition provided by

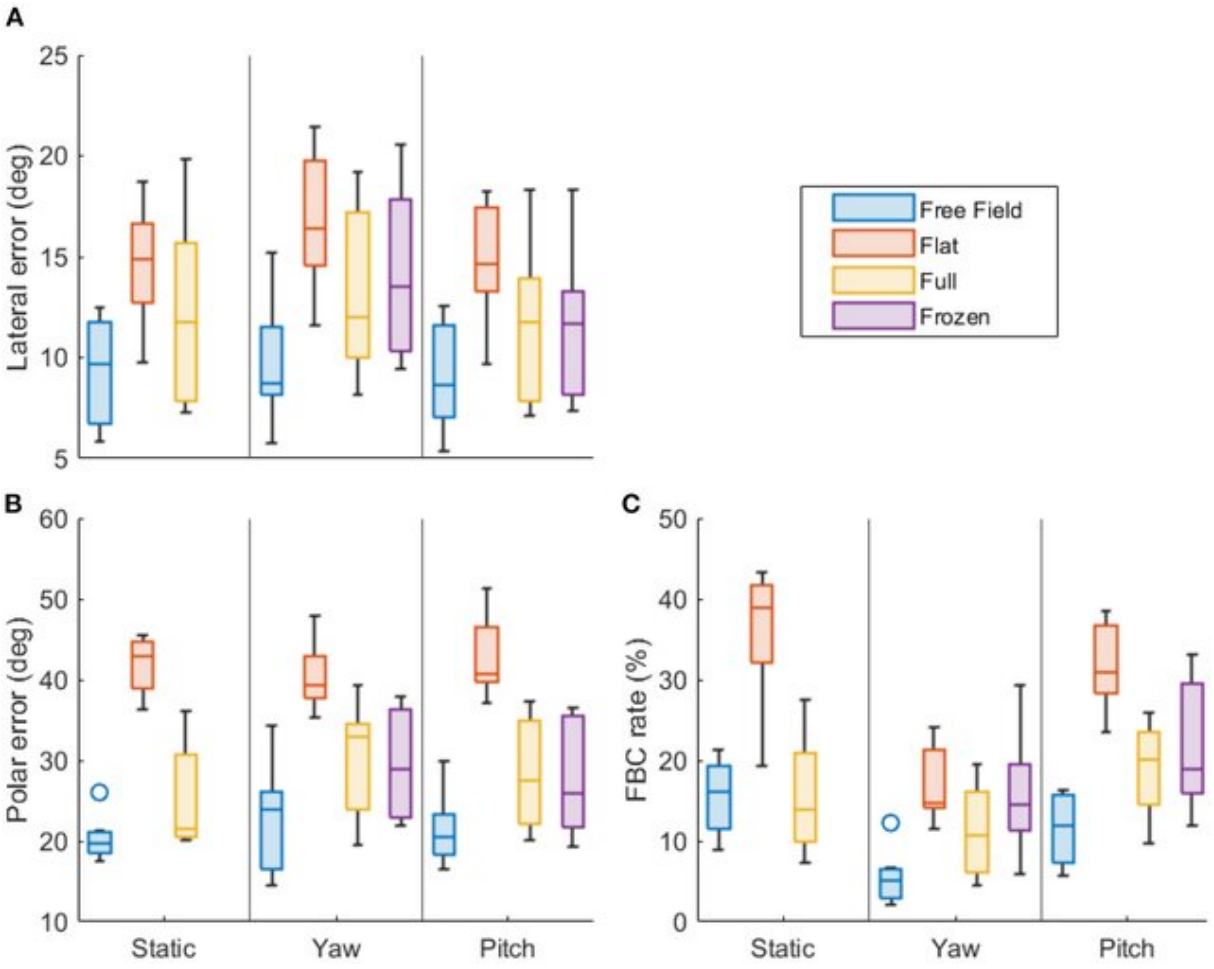

Fig. B.4 Localisation performance grouped by the type of head rotation. (a): Lateral precision error (in degrees). (b): Polar precision error (in degrees). (c): Front-back confusion rate (in %). Lower values indicate better performance. Each boxplot shows the statistics of all subjects (median, first and third quartiles, minima and maxima, outliers).

Middlebrooks (1999) the average FBC rate for static listening in the free field is 6.3%.

The comparison between the free-field and full conditions sheds light on the quality of the binaural rendering. In the static full condition, the performance was 12.0°, 25.7°, and 15.1% (lateral error, polar error, and FBC rate, respectively), which is a small but not significant difference to the static free-field condition (compare row #2 in Tab. B.2). This implies that the HRTF measurements, HRTF manipulations, and the static binaural spatialisation adequately simulated the loudspeaker-based reproduction. The localisation performance in the full condition was in the range (again excluding FBC rate) of that usually found in such experiments (e.g., 14.5°, 28.7°, and 7.7% in Middlebrooks (1999); 13.6°, 21.9°, and 12.8% in Majdak et al. (2011), calculated based on Majdak et al. (2010)). This indicates that the subjects here localised static sounds with the precision usually found in the literature.

In the dynamic conditions (yaw and pitch), no significant differences were found in the lateral and polar errors between the free-field and full conditions, indicating that the general presentation of dynamic cues worked as expected. However, there was a significant difference in FBC rates (compare rows #4 and #8 in Tab. B.2) and from the free-field condition to the full condition, FBC rates increased from 5% and 11% (yaw and pitch, respectively) to 10% and 20%. This implies that the dynamic spatialisation may have had some short-comings.

No differences were found between full and free-field FBC rates in the static condition, nor were found in the localisation precision. This suggests that the increased FBC rates in the dynamic conditions probably were not caused by problems in static MSS and ITD cues created by imperfections in the HRTF interpolation or differences between headphone and loudspeaker coloration. Artifacts in dITDs might have been the origin, though, for example because of a too large head-tracking latency, causing a lag between the actual subject's head position and not-yet correctly spatialised sound, manifesting itself in a slightly increased uncertainty about the plane of the sound direction, reflected in the increased FBC rates. While this problem may play a role in consumer spatialisation products, there is no evidence that it affects the interpretation of the experimental results. Most importantly, none of the subjects reported about dizziness, audio glitches, or loss of externalisation, which would indicate more serious problems with the dynamic sound reproduction.

A final important observation is the large variance across subjects in some of the conditions, indicating large inter-subject differences in the localisation performance. An analysis of the listener variability may provide valuable insights. It is, however, beyond the scope of this study.

B.4.2 Effect of the head rotations

In the free-field condition, there was no significant effect of head rotation on localisation precision, neither on lateral nor on polar errors. This condition best represents natural listening, suggesting that small head rotations in general do not affect the localisation precision. The lack of an effect on precision is consistent with McAnally and Martin (2014), who also showed that rotation does not affect lateral error, and that elevation error only decreases for rotations of 16° and larger. Previous studies that investigated larger pitch rotations than the ones produced here also found no reduction in elevation errors (Thurlow and Runge, 1967; Kato et al., 2003). In the flat and full conditions, a significant effect of head rotation was found, but only when comparing the lateral errors between yaw and pitch rotations (see rows #3 and #6 in Tab. B.1). These differences were approximately 2° (see Fig. B.3a), thus rather small. Interestingly, the lateral errors were larger in the yaw conditions indicating that the subjects had more difficulties to localise sound in the lateral dimension when rotating their head along that dimension. This

was only the case in the binaural reproduction. This small but significant effect might be related to some issues with the dynamic cue presentation in the study's binaural reproduction. As the lateral localisation performance is usually associated with ITD cues, this issue might indicate an imperfection in the reproduction of ITD and dITD cues. Note that polar errors did not show a significant difference between rotation conditions in any of the cue conditions. This indicates that imperfections with respect to spectral cues such as MSS and dMSS in the binaural reproduction were absent or imperceptible to the subjects. The absence of an effect on polar errors with head movements also suggest that small head movements do not induce the so-called Wallach cue, i.e., the rate of change in source azimuth angle relative to the change in head orientation, which has been hypothesised to provide information on elevation angle (Wallach, 1940; Perrett and Noble, 1997b). The general lack of an effect of head rotations on localisation precision shows that the information gained from dynamic cues due to small head rotations is either too small to improve the precision or, alternatively, is cancelled out by the increased uncertainty caused by head rotation. Both hypotheses align well with findings that lateral and polar errors decrease for head rotations larger than 45° only (McAnally and Martin, 2014).

For FBC rates, yaw rotations yielded reduced rates in all cue conditions, with the flat and free-field conditions showing a highly significant decrease of approximately 25% and 10%, respectively (see Fig. B.3c). This shows that small yaw rotations effectively aided in resolving FBCs, even when all acoustic cues were available. This is consistent with Wallach (1940) and Thurlow and Runge (1967). Pitch rotations, on the other hand, did not yield any significant changes in FBC rates in any stimulus conditions, indicating that pitch rotations do not help in resolving front from back.

Taken together, these results demonstrate that small head rotations do not improve localisation precision, however, yaw rotations, even as small as 10°, do help listeners to resolve front from back. This benefit is, however, most clearly visible in conditions without MSS cues (flat), but remains present even in the free field, when MSS cues are available. This means that dITD is still informational in the presence of MSS cues, and thus is not redundant. As hypothesised, pitch rotations do not affect the sound localisation performance at all. Pitch rotations follow the interaural axis, thus do not induce any dITD or dILD cue, leaving the dMSS as the only available dynamic cue. The lack of an effect caused by pitch rotations indicates that the subjects were unable to utilise dMSS in the present sound-localisation task. The effects of the dMSS and other cues are investigated in the following section.

B.4.3 Effects of the acoustic cues

For the lateral precision, errors in the flat condition were significantly larger than the full condition during yaw rotation (see row #3 in Tab. B.2). This suggests that, for some subjects, MSS or dMSS cues in the full condition may have provided additional information that improved lateral precision. There is, however, little evidence from literature that MSS cues contribute to the localisation of static sounds along the lateral angle, on contrary, MSS cues have been rather found as uncorrelated with the localisation along the lateral angle (Macpherson and Middlebrooks, 2002). Furthermore, it is also unlikely that dMSS caused this improvement, as no difference was found between the static and the yaw movement in the full-spectrum condition. However, the significant error increase might have originated from a potential conflict in the natural ITD-ILD combinations. The full condition provided natural ITD-ILD pairs, but the flat condition, because of the spectral flattening, might have reduced the ILDs, resulting in unnatural ITD-ILD pairs, yielding to a worse precision in localising sounds along the lateral dimension. In the static and pitch conditions there was no significant increase in lateral errors, indicating that if a potential conflict in the unnatural static ITD-ILD combinations affected the localisation then it was not much. The reason for the lateral error increase reaching the level of significance in the yaw condition might consist of two effects adding up: unnatural static ITD-ILD combinations (which alone increased the errors a little but did not cause any significance), and a potential conflict in the dynamic ITD-ILD pairs. This is supported by previous experiments on conflicting dynamic cues, such as the ”Wallach illusion” (Brimijoin and Akeroyd, 2012; Macpherson, 2011; Pöntynen et al., 2016) or the ”phantom walker illusion” (Martens et al., 2013), showing that conflicting dynamic interaural cues can be a cause for unstable location perception, potentially yielding increased localisation precision errors in the lateral dimension.

For the polar precision, significant differences were found between the flat and full conditions across all types of rotations (see rows #1, #3, and #7 in Tab. B.2). For the static condition, this is not surprising, as MSS cues are essential for localisation along the sagittal planes (Kistler and Wightman, 1992; Baumgartner et al., 2014). Similarly for pitch rotations, which did not provide dITD or dILD cues, this indicates that MSS cues help in localising sounds along the polar dimension. Note that this benefit is not necessarily evidence for the use of dMSS cues, as the improvement is similar to that in the condition without head movement. For the yaw rotation, the differences between flat and full conditions indicate that MSS cues provided additional information on the polar angle, despite the availability of dITD and dILD cues.

The analysis of the frozen condition can further help in disentangling the contribution of MSS and dMSS cues because the frozen condition provided the actual ITD and dITD cues, but the MSS cue was ’frozen’ to that of the initial head orientation, thus provided no dMSS cue. In the

frozen condition, the polar errors were at similar levels for both types of rotation (see Fig. B.4b) and there was no significant difference between the full and frozen conditions (see rows #5 and #9 in Tab. B.2). This indicates that the absence of the dMSS cue in the frozen condition was not relevant for the localisation precision in the polar dimension. When compared to the flat condition, however, the frozen condition yielded significantly improved localisation precision (see rows #6 and #10 in Tab. B.2), indicating that the static MSS cues were essential for the polar localisation precision even in the presence of head rotations.

For FBC rates, in the conditions without head rotations, there was a significant difference between the full and flat conditions, confirming the clear contribution of MSS cues when resolving front from back in a static sound-localisation task (Langendijk and Bronkhorst, 2002). In the conditions with pitch rotations, there was a significant difference between the conditions full and flat, but there was no significant difference between the conditions full and frozen (see rows #7 and #9 in Tab. B.2). This demonstrates that MSS cues, but not the dMSS, were essential to resolve front from back even when pitch rotations were involved. For the yaw rotations, the situation was different. No statistically significant differences were found between the flat and full conditions. This suggests that neither MSS nor dMSS cues provided a benefit to resolve front from back when the subject were allowed to rotate the head along the yaw axis. Thus, small head rotations seem to sufficiently have compensated for the absence of any spectral cues.

Next, a small but significant difference was found between the full and frozen conditions for the yaw rotation. The hypothesis behind this finding might be the effect of ILD cues, which remained constant over the course of the frozen stimulus, thus potentially creating a conflict between the dynamic interaural cues, i.e., dILD and dITD. These potential conflicts in the frozen condition may have slightly but significantly reduced the ability to resolve front from back. Note that such difference has not been found for pitch rotation, probably because the change of the interaural cues over this axis is minimal. This hypothesis is supported by findings of dILD as a perceivable cue, though it is mostly weaker than dITD cues (Pöntynen and Salminen, 2019). Experiments on conflicting dynamic cues, such as the ”Wallach illusion” (Brimijoin and Akeroyd, 2012; Macpherson, 2011; Pöntynen et al., 2016) or the ”phantom walker illusion” (Martens et al., 2013), have previously shown that conflicting dynamic interaural cues can be a cause for unstable location perception. Such illusions seem to be strongest when high-frequency spectral cues are absent, i.e., the spectral cues can correct for a conflicting dITD. Note that these studies did not address dITD versus dILD cues or the dynamic aspect of spectral cues explicitly. On the other hand, in another study involving an NoSπ task it was found that in the presence of conflicting dILD and dITD cues, dITD cues dominated for binaural signal detection (van der Heijden and Joris, 2010). However, it is not clear whether this result can be extrapolated to a dynamic localisation task as considered here.

The benefit of dynamic monaural spectral cues for sound localisation has been rarely discussed. A reduced elevation error has been observed for large pitch rotations during stimulus presentation (McAnally and Martin, 2014), suggesting an effectiveness of dynamic spectral cues with head rotation. Often, however, pitch rotation did not produce an improvement in localisation performance (Thurlow and Runge, 1967; Kato et al., 2003). In a monaural listening experiment, normal-hearing listeners with one ear plugged showed no benefit of head rotations when localising sounds located in the horizontal plane (Hirahara et al., 2021). However, single-sided deaf listeners do appear to utilise and even rely on changes in head position to induce changes in the monaural cues produced by the direction-dependent high-frequency attenuation resulting from acoustic head shadowing (Pastore et al., 2020).

Taken together, these results demonstrate the important contribution of static MSS cues to sound localisation performance, even during sound localisation involving head movements. The results also suggest that, for small head movements, dynamic changes of that cue are not evaluated by the auditory system. This has a direct implication for modeling active sound localisation in human listeners. Recently, a model of active sound localisation based on Bayesian inference was proposed (McLachlan et al., 2021). That model only implemented dITD as an additional dynamic cue, though the necessity for the consideration of other cues was at that point unclear. As it seems, further updates of MSS cues (dMSS cues) are not required when modelling human sound localisation, at least for small head movements.

B.5 Conclusions and future work

In this study, the influence of small head movements on localisation performance was investigated by means of three metrics. The results show no additional benefit of small head rotations (up to $\pm 10^{\circ}$) on lateral and polar localisation precision. Only yaw rotations significantly reduced the front-back confusions, whereas pitch rotations were of no help. This finding could be explained by the contribution of dynamic ITD cues (and, to a lesser extent, the dynamic ILD cues). These effects were most prominent for stimuli devoid of monaural spectral cues, but remained even when these cues were available. The analysis of the frozen spectrum condition, which provided the actual static and dynamic ITD cues, but 'froze' the monaural spectral cues to those from the initial head orientation, suggests that humans do utilise the static monaural spectral cues but are insensitive to their dynamic changes over the course of the stimulus. This is clearly supported by the results showing that monaural spectral cues, fixed to those from the initial head orientation and conflicting with the dynamic ITD cues, did not impair the localisation performance.

There are several directions that future work can move towards. First, due to the differences

between the binaural and loudspeaker performance in the dynamic listening conditions, it would be insightful to compare various loudspeaker-based conditions by using, for example, band-pass filtered instead of broadband signals. Second, a direction-dependent analysis of the localisation performance may reveal which spatial regions gain from the head movements most. Third, the results showed a large inter-subject variability. Thus, a listener-specific analysis of the contribution of static and dynamic cues may be interesting. A similar analysis has been done for the acoustic and non-acoustic factors contributing to the localisation performance in sagittal planes (Majdak et al., 2014). Finally, these findings will help to further develop the previously proposed model of active directional sound localisation (McLachlan et al., 2021).

Funding

This research was supported by the Research Foundation Flanders (FWO) under Grant no. G023619N, the Agency for Innovation and Entrepreneurship (VLAIO), and the European Union (project "SONICOM", grant number 101017743, RIA action of Horizon 2020).

Data Availability Statement

The data described in this article are publicly available as part of the Auditory Modeling Toolbox version 1.3 (AMT, Majdak et al., 2022) as the function `data_mclachlan2022`. In that AMT version, the function `exp_mclachlan2023` reproduces Figs. B.3 and B.4. The AMT is available at www.amtoolbox.org.

Acknowledgements

The authors are grateful to Stefan van Dongen and Dieter Heylen for their assistance in the design of the mixed-effects models for statistical analysis.

PAPER C

Modelling dynamic sound localisation through Bayesian inference: a sensitivity analysis

Glen McLachlan[1], Herbert Peremans[1]

[1] Department of Engineering Management, University of Antwerp, Belgium

Abstract - From a Bayesian perspective, sensory information in the brain is represented in the form of probability distributions. Inherent to these probability distributions is the representation of uncertainty due to sensory noise and ambiguity. Dynamic listening using head movements is a multisensory process which involves several sources of sensory uncertainty. In this study, we introduce the numerical implementation of a Bayesian dynamic sound localisation model and investigate how the model's sensory noise parameters affect its localisation performance over the 2D sphere assuming static sound-sources in the far field. We restrict ourselves to small, open-loop head rotations. Six noise parameters that describe both acoustic and proprioception measurements are proposed and investigated through a sensitivity analysis. The localisation performance is expressed in lateral error, polar error and front-back confusion rate. The results from this sensitivity analysis will be compared in the future to empirical data.

C.1 Introduction

Recently, we proposed a Bayesian framework to model sound localisation that includes self-motion, i.e., head movements (McLachlan et al., 2021). Fundamentally, it is an extension of the static ideal-observer model presented by Reijniers et al. (2014), which used the same Bayesian theory to determine the conditional probability distribution of the sound-source direction ψ, given the available acoustic input and prior information (Reijniers et al., 2014).

The present implementation expands the static model on two fronts. First, the model no longer relies on a single measurement, but instead makes an observation of the available cues at set intervals during stimulus presentation (Hofman and Van Opstal, 1998). This means that the the posterior distribution of the sound-source location can be recursively updated as more information becomes available. Second, the head position can be controlled over time. From this follows that not just acoustic information, but also proprioceptive information must be processed.

We previously provided a numerical implementation of the theoretical framework as a proof of concept, which qualitatively showed what information on source location could be gained from head movement (McLachlan et al., 2021). However, a more in-depth analysis of its output remained to be carried out. Moreover, this proof of concept relied on two significant simplifications: 1) it encoded interaural time differences (ITDs) as the only available dynamic acoustic cue, ignoring temporal changes in spectral cues and 2) it assumed the head positions to be fully deterministic, i.e., without uncertainty.

In this paper, we remove these two simplifications and use the model to study human dynamic localisation over the full 2D sphere in the far field (i.e., direction estimation) when presented with a broadband sound-source. This model is publicly available as `"mclachlan2023"` in the Auditory Modeling Toolbox (Majdak et al., 2022). Through a sensitivity analysis we will investigate the effect of the different model parameters on localisation performance.

C.2 Model extension

C.2.1 Acoustic and proprioceptive information

The present model assumes the same feature space of the acoustic input static ideal-observer model by Reijniers et al. (2014): $\mathbf{y}_A$, which consists of the noiseless "true" state of the acoustic information, $\mathbf{X}_A$, convolved with noise due to uncertainty caused by the auditory system or the environment:

$$\mathbf{y}_A = [X_{itd} + \delta_{itd}, \mathbf{X}_- + \delta_-, \mathbf{X}_+ + \delta_+], \tag{C.1}$$

$$\mathbf{X}_- = \mathbf{X}_L - \mathbf{X}_R, \tag{C.2a}$$

$$\mathbf{X}_+ = [\mathbf{X}_L + \mathbf{X}_R]/2, \tag{C.2b}$$

$$\mathbf{X}_{L/R} = \mathbf{S} + \mathbf{H}_{L/R}, \tag{C.2c}$$

where $\mathbf{X}_L$ and $\mathbf{X}_R$ are the frequencywise sum log-magnitudes of the sound source spectrum, $\mathbf{S}$, and the head-related transfer functions (HRTF), $\mathbf{H}_L$ and $\mathbf{H}_R$, for the left and right ear, respectively. $\mathbf{X}_-$ and $\mathbf{X}_+$ then correspond with the interaural spectral difference and an average of both monaural spectra, respectively. Note that this transformation is not strictly necessary, but aids to interpret and discuss the results. The noise sources are described as follows:

$$\delta_{itd} \sim \mathcal{N}(0, \sigma_{itd}^2) \tag{C.3a}$$

$$\delta_- \sim \mathcal{N}(0, \Sigma_-), \quad \Sigma_- = 2\sigma_I^2 \cdot \mathbf{I} \tag{C.3b}$$

$$\delta_+ \sim \mathcal{N}(0, \Sigma_+), \quad \Sigma_+ = (\sigma_I^2/2 + \sigma_S^2) \cdot \mathbf{I} + \sigma \tag{C.3c}$$

Variances σ_I^2, σ_S^2 and σ^2 model the noises on the spectral measurements, the subject's knowledge of the sound-source spectrum, and the cross-talk between adjacent frequency bands, respectively.

Finally, there is the proprioceptive component. Similar to $\mathbf{y}_A$, y_H denotes the noisy observation of the true state of the head orientation θ_H, which is applied to both azimuth and elevation. At each time step, θ_H is updated with a motor control signal u, which denotes a rotation of the head around the yaw or pitch axis. These variables are defined as:

$$y_H(t_i) = \theta_H(t_i) + \delta_H, \tag{C.4a}$$

$$\theta_H(t_{i+1}) = \theta_H(t_i) + u(t_i)\Delta t + \delta_u, \tag{C.4b}$$

δ_H and δ_u are the noises on the head orientation observation and the motor command, respectively. These noises are applied to both the azimuth and elevation angles and are defined as:

$$\delta_H \sim \mathcal{N}(0, \sigma_H^2), \tag{C.5a}$$

$$\delta_u \sim \mathcal{N}(0, \sigma_u^2), \tag{C.5b}$$

Note that, for easier notation, most equations above are not described as functions of time. In reality, new measurements are made at each time step (e.g., $\mathbf{y}_A(t_i)$) and the noise variance parameters can change over time (e.g., $\sigma_{itd}(t_i)$).

C.2.2 Recursive Bayesian estimation

We model temporal integration of acoustic and proprioceptive information through recursive Bayesian estimation, where probability density functions (PDFs) are updated recursively over time with incoming measurements.

Following Bayes' Theorem, this process can be simply written as:

$$p_{t_i} = C \cdot p_{t_{i-1}} \cdot M, \tag{C.6}$$

Turning to Bayesian terminology, p_{t_i} denotes the posterior PDF, $p_{t_{i-1}}$ denotes the prior PDF and M denotes the joint sensor model which computes the likelihood. C is a normalisation constant. Note that the prior at time step t_i equals the posterior from time step t_{i-1}. At the initiation of the recursive process, $p_{t_{i-1}} = p(\psi)$, which is the spatial prior, or the prior knowledge of the sound direction. A detailed description of this equation is given in McLachlan et al. (2021).

Fig. C.1 demonstrates the recursive process for 5 time steps with a time step size Δt of 5 ms. Here the left two columns accumulate into an increasingly narrow distribution. We see that, despite the large variance between each look, the cumulative distribution very quickly (after 25ms) decreases in spread.

This specific example shows the process that will result in a fairly successful localisation estimate, which is presented in Fig. C.2a. This does not necessarily need to happen. Fig. C.2 shows the results for three different iterations of the same model parameters. If system noise causes enough incorrect observations, then localisation can be affected either by an inability to narrow the distribution (Fig. C.2b), or by a narrowing of the distribution around an incorrect direction (Fig. C.2c). The first effect can be considered detrimental to precision, whereas the second effect is detrimental to accuracy.

C.3 Methods

A sensitivity analysis was conducted to determine how the different noise parameters affect the model's localisation performance. We considered three rotation conditions, 1527 target directions (distributed over the full sphere above an elevation of -30) and seven different individual HRTFs, obtained in an earlier study (McLachlan et al., 2023). Simulations were repeated 50 times per target direction per HRTF. For yaw and pitch rotation conditions, half of the simulations rotated the head in the positive direction and the other half rotated in the negative direction.

The input stimulus was a 100ms broadband white noise burst. For the movement conditions,

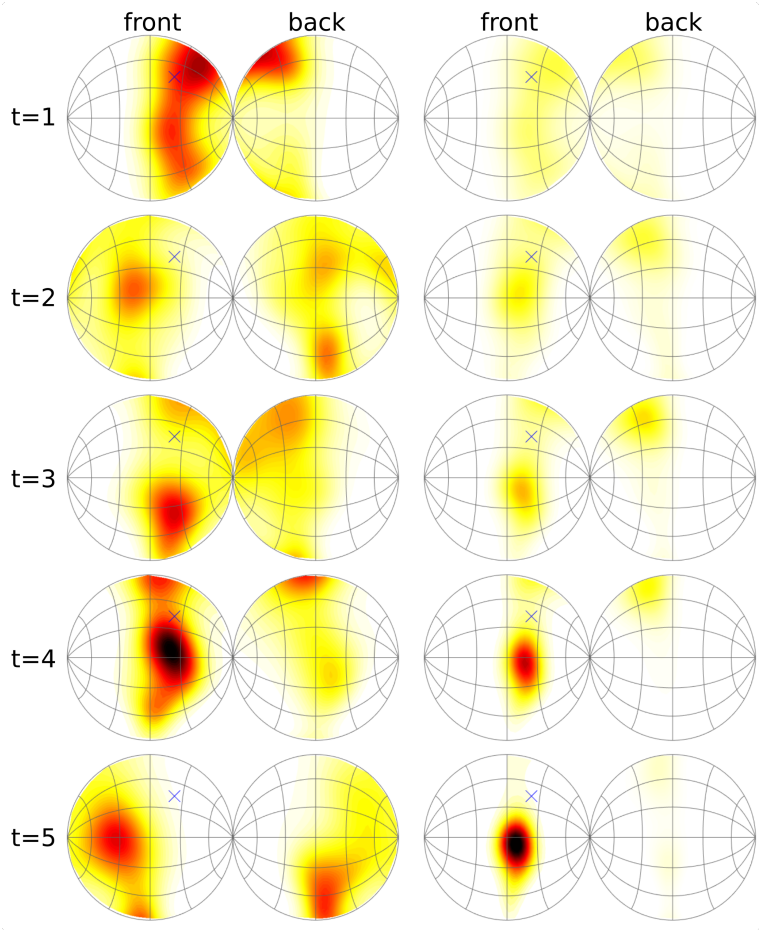

Fig. C.1 Probability density functions of the sound-source direction over the full sphere at time steps t=1-5, with the time between each step Δt set at 5ms. Left two columns: single look PDF at each time step. Right two columns: cumulative PDF, i.e., recursive posterior distribution at each time step. The blue 'x' marks the true source direction.

rotations of $10°$ were deployed at a constant speed of $100°/s$ along either the yaw or pitch axis. The initial head orientation was straight ahead, i.e., $0°$ azimuth and $0°$ elevation. The model worked with a time step size Δt of 5 ms, so the posterior was updated every 5 ms. The spatial prior was set to a uniform distribution to best visualise the effect of each tested parameter. To obtain a point estimate from the posterior PDF, we applied the maximum a-posteriori (MAP) estimate, which selects the mode of the distribution.

Localisation performance was evaluated based on three metrics: the lateral and polar root mean square error ϵ_L and ϵ_P (degrees) and the quadrant error ϵ_Q (% of trials), as defined in Middlebrooks (1999).

The tested parameters and their control values are presented in Tab. C.1. In preliminary simulations, it became apparent that the standard deviations of the noise models chosen in Reijniers

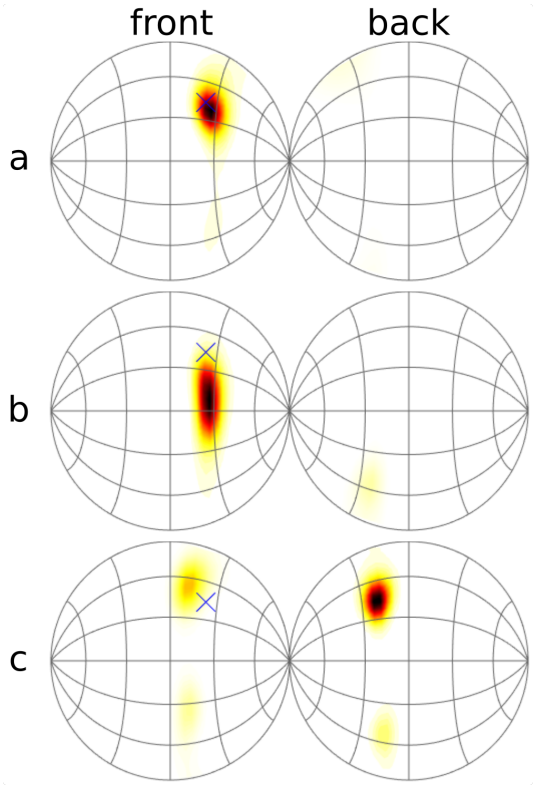

Fig. C.2 Three examples of probability density functions over the full sphere of the same sound-source direction at final time step t=21. a) accurate estimate, b) smeared estimate c) front-back confusion. The blue 'x' marks the true source direction.

et al. (2014) were too low to provide insightful results. For this reason, the control values for acoustic noise were set large enough to see the effects of each individual parameter. Every parameter was adjusted individually, because an investigation of the interaction effects would go beyond the scope of this work. For this same reason the proprioceptive control noise was set to zero: to prevent any interaction effects between proprioceptive and acoustic uncertainty during the analysis.

Table C.1 Noise parameters included in the sensitivity analysis, including descriptions of the signal they affect and their control values.

Noise on:	**Symbol**	**Value**
ITD look	σ_{itd}	3 JND
Spectrum look	σ_I	20 dB
Source knowledge	σ_S	20 dB
Head orientation look	σ_H	0°
Head motor signal	σ_u	0°

Table C.2 Lateral and polar root mean square error (ϵ_L, ϵ_P) and quadrant error rate (ϵ_Q) for five tested noise parameters and three rotation conditions, averaged over 7 virtual subjects, 1527 source directions and 50 repetitions. Values are rounded to one decimal place.

STATIC	ϵ_L	ϵ_P	ϵ_Q
control	4.1°	24.2°	8.1%
$2 \cdot \sigma_{itd}$	6.1°	25.4°	8.8%
$2 \cdot \sigma_I$	4.7°	30.1°	15.8%
$2 \cdot \sigma_S$	4.0°	25.9°	11.4%
$\sigma_H = 10$	4.7°	25.4°	9.3%
$\sigma_u = 2$	4.6°	21.8°	4.3%
YAW	ϵ_L	ϵ_P	ϵ_Q
control	4.6°	22.8°	5.1%
$2 \cdot \sigma_{itd}$	6.5°	24.7°	7.0%
$2 \cdot \sigma_I$	5.5°	28.8°	10.4%
$2 \cdot \sigma_S$	4.5°	24.4°	6.6%
$\sigma_H = 10$	5.3°	24.1°	6.1%
$\sigma_u = 2$	4.7°	20.7°	3.5%
PITCH	ϵ_L	ϵ_P	ϵ_Q
control	4.1°	24.3°	8.1%
$2 \cdot \sigma_{itd}$	6.1°	25.5°	8.8%
$2 \cdot \sigma_I$	4.7°	30.3°	15.9%
$2 \cdot \sigma_S$	4.1°	26.0°	11.5%
$\sigma_H = 10$	4.8°	25.4°	9.1%
$\sigma_u = 2$	4.6°	21.4°	4.0%

C.4 Results

Tab. C.2 presents the results of the sensitivity analysis, averaged over 7 virtual subjects, 1527 source directions and 50 repetitions. This serves as a starting point to determine the general effect of each noise parameter under different movement conditions. Fig. C.3 presents the same results for ϵ_L and ϵ_P, distributed over the full sphere for the static (no movement) condition. This figure provides an insight on the direction-dependent effects of each noise source. Because the spatial effects of ϵ_P and ϵ_Q were very similar, we decided to omit figures for ϵ_Q.

C.5 Discussion

Before beginning the discussion it must be noted that the purpose of this analysis is not to match the model output to empirical data, but to test the influence of each noise source to assist parametrisation of the model in a future stage. The noise sources were divided into

three categories: timing noise, spectral noise and proprioceptive noise. Each category will be discussed separately.

C.5.1 Control

Comparing between the three rotation conditions, it becomes apparent that yaw rotation provides a lot of information on the polar angle and the quadrant in which the sound-source is located. The decreases in ϵ_P and ϵ_Q can be explained by the hypothesis on the effect of head rotations posed in Wallach (1940). The sign of the change in ITD that accompanies a head rotation is an unambiguous indicator of the proper hemisphere, which reduces quadrant errors. Additionally, the rate of change in source azimuth angle relative to the change in head orientation can theoretically provide information on elevation angle. Pitch rotation does not affect localisation whatsoever, this agrees with the results from earlier studies (McLachlan et al., 2023).

C.5.2 Timing noise

In all rotation conditions, an increase in σ_{itd} causes a 50% increase in ϵ_L. This is no surprise: interaural cues are mostly informative about the lateral angle of a source. Perhaps more surprising are the small decreases in performance for the other two metrics. Looking at Fig. C.3a, we see that these small effects mostly take place behind the listener and away from the median plane. This suggests that there are small asymmetries in the ITD between the front and back hemispheres, which in the control condition sometimes (although rarely) provided information on the correct hemisphere. Note that the effect on ϵ_Q is larger during yaw rotation, because an increased σ_{itd} makes it more difficult to utilise the ITD rate of change, which, as mentioned before, is an indicator of the correct hemisphere.

C.5.3 Spectral noise

When σ_I is increased, we see highly detrimental effects for both ϵ_P and ϵ_Q, this can be explained by the well-known role that spectral cues play in localisation along the sagittal planes. Lateral error ϵ_L is also increased, although this effect is minimised due to the complementary nature between interaural timing and level differences. For yaw rotation, an increase on spectral noise has a much smaller effect on ϵ_Q. This again demonstrates how dynamic ITD as a function of head rotation serves as a strong cue to prevent quadrant errors. The spatial analysis (Fig. C.3b) shows that error increases are largest away from the median plane, this is likely because locations around the median plane are still partly supported by X_itd and $\mathbf{X}_+$, which are not or less severely affected by σ_I.

An increase in σ_S has no effect on ϵ_L, because knowledge of the source spectrum is irrelevant

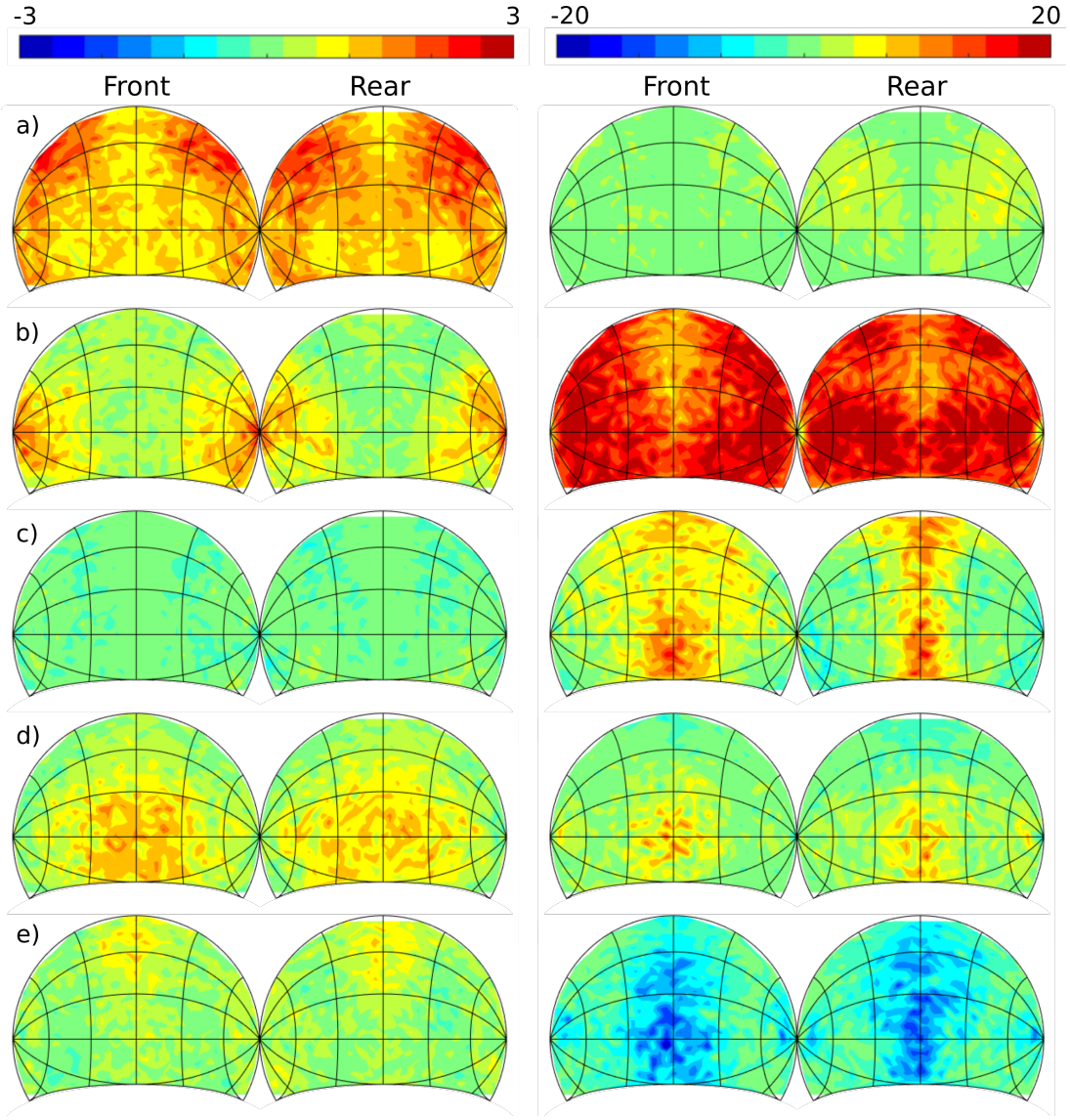

Fig. C.3 Model root mean square error difference between control values and separately varied model parameters, during static localisation. Left column: lateral error ϵ_L, right column: polar error ϵ_P. a) $2 \cdot \sigma_{itd}$, b) $2 \cdot \sigma_I$, c) $2 \cdot \sigma_S$, d) $\sigma_H = 10$, e) $\sigma_u = 2$. The results are plotted for 1527 target directions over the full sphere relative to the torso. Results were averaged over 7 subjects and 50 trials per subject per direction.

to lateral localisation. It also appears to be only slightly detrimental to ϵ_P and ϵ_Q. However, looking at Fig. C.3c, we see that specifically directions around the median plane are much more heavily influenced than others, which contrasts to the effects in C.3b. Indeed, on the median plane the listener can extract less information regarding the polar angle from $\mathbf{X}_-$ because of symmetry, and relies more on $\mathbf{X}_+$ (Reijniers et al., 2014). This requires knowledge of the sound-source, which becomes uncertain as σ_S increases.

C.5.4 Proprioceptive noise

A higher σ_H results in an increased error for all three metrics. This is not surprising, as one would expect an uncertainty of the orientation of the head and ears to correspond with a smearing of the direction estimate. Fig. C.3 shows that this smearing occurs more severely for sources close to straight ahead $(0, 0)$ and behind $(0, 180)$. This can be explained by the fact that σ_H is two-dimensional, i.e., it applies to the orientation of the head along the pitch and the yaw axis. Sources above the listener only suffer from the uncertainty in pitch and sources to the left or the right only suffer from the uncertainty in yaw. Whereas the directions at $(0, 0)$ and $(0, 180)$ are affected by both yaw and pitch uncertainty. In other words, this is an artifact of the present definition of the movement model, and empirical data would be required to determine if this adequately simulates true movements.

It may be counter-intuitive that an increase in noise on the motor control signal σ_u improves polar localisation (ϵ_P and ϵ_Q), but it can be easily explained by considering equation C.4. When δ_u is high, it means that the true head orientation $\theta_H(t_{i+1})$ deviates far from the previous orientation $\theta_H(t_i)$. If this deviation is large enough, then positive effects similar to those in the yaw rotation condition can be expected. This result raises the question about the validity of localisation experiments where subjects were instructed to remain still, as it is possible that a deviation from the instructed position accidentally provided additional acoustic cues.

C.6 Conclusion and Outlook

In this work we investigated the dynamic localisation model described in McLachlan et al. (2021) through a sensitivity analysis of the sensory noise parameters. The current parameters already provide much control and insight on the role of acoustic and proprioceptive information during sound-localisation, but the present Bayesian framework makes it easy to implement additional elements. To conclude, we will list a number of possible adjustments, based on psychoacoustic findings. Note that this list is intended to shed a light on the potential of the recursive Bayesian framework, but is by no means exhaustive.

First, the feature space may be reformulated to be more representative of the acoustic cues that humans utilise for sound localisation. For example, several studies found that the positive spectral gradient may be a more appropriate localization cue than the absolute spectral values in each frequency band (Baumgartner et al., 2014).

Second, lateral and polar estimation may be split into two separate processes. There is some neurological evidence that this may be the case (Ege et al., 2018), and previous work has shown that this split significantly affects the output of the Bayesian model (Peremans et al., 2022). One

way of splitting this process is by applying a Bayesian decision rule depending on the plane of localization, e.g., maximum a-posteriori for the lateral angle and random sampling for the polar angle.

Third, a non-uniform spatial prior can be implemented. Empirical findings suggest that a Gaussian distribution around the horizontal plane may better describe elevation estimation (Ege et al., 2018).

Finally, earlier work showed that the introduction of a pointing error to the Bayesian estimator successfully accounted for the deviance between model and experimental data (Baumgartner et al., 2014). It is important to note here, however, that this pointing error cannot be included arbitrarily. To prevent that this noise is simply added to account for any deviating results, its values should be carefully chosen and grounded in empirical evidence, e.g., a higher pointing error for sources behind the listener.

C.7 Acknowledgements

This research was supported by the Research Foundation Flanders (FWO) under Grant number G023619N and by the Agency for Innovation and Entrepreneurship (VLAIO).

PAPER D

Bayesian active sound localisation: to what extent do humans perform like an ideal-observer?

Glen McLachlan[1], Piotr Majdak[2], Jonas Reijniers[1], Michael Mihocic[2], Herbert Peremans[1]

[1] Department of Engineering Management, University of Antwerp, Belgium
[2] Acoustics Research Institute, Austrian Academy of Sciences, Austria

Abstract - Self-motion is an essential but often overlooked component of sound localisation. While the directional information of a source is implicitly contained in head-centred acoustic cues, that acoustic input needs to be continuously combined with proprioceptive information about the head orientation in order to decode these cues to a world-centred frame of reference. On top of that, the use of head movement significantly reduces ambiguities in the directional information provided by the incoming sound. In this work, we evaluate a Bayesian model that predicts dynamic sound localisation, by comparing its predictions to human performance measured in a behavioural sound-localisation experiment. Model parameters were set a-priori, based on results from various psychoacoustic and sensorimotor studies, i.e., without any post-hoc parameter fitting to behavioral results. In a spatial analysis, we evaluated the model's capability to predict spatial localisation responses. Further, we investigated specific effects of the stimulus duration, the spatial prior and sizes of various model uncertainties on the predictions. The spatial analysis revealed general agreement between the predictions and the actual behaviour. The altering of the model uncertainties and stimulus duration revealed a number of interesting effects providing new insights on modelling the human integration of acoustic and proprioceptive information in a localisation task.

Introduction

The acoustic cues acquired through head rotation are of crucial importance to spatial hearing. Not only do they improve sound externalisation (Brimijoin et al., 2013), they provide dynamic acoustic cues which contribute to sound localisation. First and foremost, they reduce ambiguities between the front and back (Wallach, 1940; Perrett and Noble, 1997a). Second, under certain conditions, they can improve the elevation estimation of a source through a relation between the rate of change in binaural cues and the amount of performed head rotation, a phenomenon termed the 'Wallach cue' (Wallach, 1940; Jiang et al., 2019). Dynamic acoustic cues become especially important when spectral cues are difficult to process, like in reverberant (Giguère and Abel, 1993) or virtual (McLachlan et al., 2023) environments, or when high-frequency cues are unavailable (Thurlow and Runge, 1967).

The movements that benefit sound localisation are not always voluntary. On the contrary, for both sensorimotor (Winter, 1995) and behavioural (Hadar et al., 1983, 1985) reasons, the human head is rarely completely still. Even when the only task given to subjects is to remain still, they continue to move by a small but measurable amount (Wersényi and Wilson, 2015). Despite this constant motion, we perceive the auditory world to be relatively stable. This suggests that there exists a mechanism that utilises positional information about the head to compensate for self-motion and converts the head-centred auditory cues into a stable, world-centred frame of reference (Brimijoin and Akeroyd, 2014). This notion is further supported by the fact that moving sound sources do not provide the same benefits to localisation as the head movements that would theoretically cause similar acoustic cues (Wightman and Kistler, 1999).

If we are to include these unavoidable dynamic effects in future studies of sound localisation, it is apparent that there is a need for a tool to better investigate or predict the effects of head movements in a reproducible manner. To this end, we previously proposed an ideal observer model for active sound localisation based on Bayesian inference, which can process dynamic cues obtained through self-motion. This model integrates acoustic and proprioceptive information over time to simulate sound localisation through self-motion (McLachlan et al., 2021). The model serves as a performance 'ceiling', given the available acoustic cues. It also provides a bottom-up approach to sound localisation, i.e., it lets one change the cues that are extracted from incoming sound and the way that they are utilised (e.g. by adjusting the spatial prior or increasing sensory noise), after which the effects on localisation performance can be tested.

In this paper, we investigated to what extent humans behave like an ideal observer during active sound localisation. We did this by comparing the output of a Bayesian model for active sound localisation to behavioural data over the full 2D sphere. First, we described the active sound

localisation model. This model continually collects auditory snapshots or 'looks', akin to the multiple looks model (Viemeister and Wakefield, 1991). Through recursive Bayesian estimation, these looks were accumulated over time, reducing sensory ambiguity. The use of snapshots means that dynamic cues were formed implicitly, i.e., the additional information obtained from head motion was obtained through a series of static looks. Note that this model controls head movement irrespective of the incoming sound and, hence, does not encompass 'closed-loop' processes such as triangulation or source tracking. Next, we described the localisation experiment and compared the results obtained here to the model data. Finally, we adjusted a subset of the parameters to investigate their effect on localisation performance. The present experiments focus on small head rotations ($10°$) along the yaw axis. Small head movements have been shown to comprise the majority of natural head movements (Kim et al., 2013a), and can be considered, if necessary, as a first step in a more complex movement framework. All experimental data, including the localisation model, were made publicly available in the Auditory Modeling Toolbox (AMT, Majdak et al., 2022).

Methods

Ethics statement

All subjects that participated in this study were adult volunteers. They were informed on the procedure and were free to withdraw at any time. They gave written informed consent before the experiment. The study applied the standard methodology of the Acoustics Research Institute (ARI) which has been approved by the Ethics Representatives of the ARI.

Template definition

The proposed model utilises a 'template-matching' procedure which requires a set of acoustic templates $\mathbf{T}_A$ that the observed information is compared to Middlebrooks (1992); Baumgartner et al. (2014); Reijniers et al. (2014). Each template in $\mathbf{T}_A$ contains the expected acoustic information from a specific direction. We assumed that $\mathbf{T}_A$ is the acoustic 'knowledge' that the brain has learned and stored over a lifetime of experience, and is thus signal-independent.

To compute $\mathbf{T}_A$, a set of acoustic features was extracted from the subject's head-related transfer functions (HRTFs) and the signals received at each ear. This process is identical to the feature extraction described in earlier work (Reijniers et al., 2014).

The ITD template T_{itd} was computed as the difference between times of arrival of the head-related impulse responses at each ear. The time of arrival was defined as the time it takes for the impulse response to reach a value 10 dB below its maxima. Each head-related impulse response

was low-pass filtered at a cutoff frequency of 3000 Hz before deriving the time of arrival. Then, the ITDs (in time units) were transformed into a scale of just-noticeable difference (JND) units, such that the error on the ITD was modeled as an additive instead of multiplicative factor (for further explanation, see Reijniers et al. (2014)).

Next, we consider the directional filters for the left $\mathbf{T}_L$ and right $\mathbf{T}_R$ ear separately. The head-related impulse responses were passed through a Gammatone filterbank with 32 channels in equivalent rectangular bandwidths, with centre frequencies ranging between 300 Hz and 15 kHz, as in Reijniers et al. (2014). These processed signals were half-wave rectified, low-pass filtered using five sequential first-order infinite impulse response filters with a cut-off frequency of 2000 Hz, and then transformed to a logarithmic domain (in dB). This stage simulates a simplified processing of the inner hair cell (Breebaart et al., 2001). Then, for each frequency channel, the root mean square of the signal was computed. Thus, $\mathbf{T}_L$ and $\mathbf{T}_R$ denote vectors with monaural spectral information in dB along the equivalent rectangular bandwidth channels.

Ultimately, the ITD and the monaural spectral vectors for both ears are combined into $\mathbf{T}_A$ which is a matrix containing the combined vectors per template source direction:

$$\mathbf{T}_A = [T_{itd}, \mathbf{T}_L, \mathbf{T}_R] \tag{D.1}$$

In this article, $\mathbf{T}_A$ consists of 2042 directions that are uniformly distributed over the sphere. These directions were obtained through spherical-harmonics interpolation of the measured HRTFs, involving Tikhonov regularisation to account for measured directions not covering the full sphere (Pollow et al., 2012).

Generative model

We assume that the listener wants to determine the source direction based on all prior information about the environment and all sensory information collected during the head movement. This prior and sensory information is combined into a posterior probability density function (PDF), from which finally a point estimate is retrieved.

To explain this process step-by-step, we first introduce the function of the spatial prior $p(\psi)$. Then we discuss how the likelihoods of the acoustic information L_A and proprioceptive information L_H are computed. Finally we combine these factors into the posterior PDF and obtain an estimate of the sound source direction from this distribution.

Spatial prior

In the Bayesian framework the probability of an occurring event may be affected by prior knowledge about the event. The spatial prior $p(\psi)$ quantifies the listener's a-priori assumptions about the source location before taking any sensory information into account. Polar estimations show a general bias towards the audio-visual horizon (Ege et al., 2018; Carlile et al., 1997). This can be modelled with a Gaussian spatial prior around the horizontal plane with a limited SD of about 12°. However, the best fitting SD of the prior seems to depend on the decision rule used.

The spatial prior is only one example of possible prior information available to a listener. Priors can be related to any variable, such as the number of sources (Rohe and Noppeney, 2015), the movement properties of the sound source (Senna et al., 2015) or its spectral content (Hofman and Van Opstal, 1998). In fact, the proposed model relies on the assumption that the source spectrum is unknown, but is derived from an ecologically valid prior.

Acoustic sensor model

The acoustic sensor model compares the stored template information $\mathbf{T}_A$ to a vector of acoustic features present in the observed sound signal, $\mathbf{y}_A$, which consists of the noiseless 'true' state of the acoustic information, $\mathbf{X}_A$, corrupted with noise due to uncertainties within the auditory system or caused by the environment:

$$\mathbf{y}_A = [y_{itd}, \mathbf{y}_L, \mathbf{y}_R] \tag{D.2}$$

with

$$\begin{aligned} y_{itd} &= X_{itd} + \delta_{itd} \\ \mathbf{y}_L &= \mathbf{X}_L - \hat{\mathbf{S}} + \delta_L + \delta_S \\ \mathbf{y}_R &= \mathbf{X}_R - \hat{\mathbf{S}} + \delta_R + \delta_S \end{aligned}$$

where δ_{itd} is the error on the ITD measurement with standard deviation σ_{itd}, δ_L and δ_R are the errors on the left and right monaural spectra measurements with covariance matrices $\mathbf{\Sigma}_L = \mathbf{\Sigma}_R = \sigma_I^2 \cdot \mathbf{I}$, respectively. Thus, σ_I represents the noise on the spectral measurements. Finally, $\hat{\mathbf{S}}$ is the mean expected source spectrum and δ_S is the error due to imperfect knowledge of the sound source with covariance matrix Σ_S (assuming a central process, this is the same at both ears). So, $\hat{S}$ and Σ_S define the observer's prior on the source spectrum:

$$P(S) = \mathcal{N}(\hat{S}, \Sigma_S) \tag{D.3}$$

i.e., the observer assumes that the source has spectrum $\hat{S}$ with an uncertainty which is contained by the source covariance matrix Σ_S.

With that, we define the full covariance matrix of the acoustic cues as:

$$\mathbf{\Sigma}_A = \begin{bmatrix} \sigma_{itd}^2 & 0 & 0 \\ 0 & \mathbf{\Sigma}_L + \mathbf{\Sigma}_S & \mathbf{\Sigma}_S \\ 0 & \mathbf{\Sigma}_S & \mathbf{\Sigma}_R + \mathbf{\Sigma}_S \end{bmatrix} \tag{D.4}$$

All error terms are assumed to be zero-mean Gaussian noise. Note that we assume the spectrum to be time-invariant and the source in the far field. We also assume here that each frequency channel has the same sensory noise of σ_I. Several studies have shown that the JND shows little dependence of signal frequency for sources louder than 50 dBA SPL (Jesteadt et al., 1977; Ozimek and Zwislocki, 1996).

We then consider the acoustic sensor model:

$$L_A(t_i) = p(\mathbf{y}_A(t_i) \mid \theta_H(t_i), \psi) \tag{D.5}$$

where $\mathbf{y}_A(t_i)$ and $\theta_H(t_i)$ are the observed acoustic information and the true head orientation, respectively, at time-step t_i, and ψ is the true sound source direction, which here is assumed to be independent of time.

The expression in Eq D.5 is calculated by computing the Mahalanobis distance between the measured acoustic cues $\mathbf{y_A}(t_i)$ and the set of acoustic cue templates $\mathbf{T_A}$ and covariance matrix $\mathbf{\Sigma}_A$. This is done at each sampled sound source direction, given the current head orientation $\theta_H(t_i)$.

Motor sensor model

The motor sensor model is defined as:

$$L_H(t_i) = p(\theta_H(t_i) \mid y_H(t_0 : t_i), u(t_0 : t_i)) \tag{D.6}$$

where $\theta_H(t_i)$ and $y_H(t_i)$ are the true and observed head orientations at each time step t_i, respectively, and u is the motor command signal, which is represented by the speed $\omega(t_i)$ of rotating the head around a given axis. These variables are defined as:

$$\begin{aligned} y_H(t_i) &= \theta_H(t_i) + \delta_H, \\ \theta_H(t_{i+1}) &= \theta_H(t_i) + u(t_i)\Delta t + \delta_u, \end{aligned} \tag{D.7}$$

The additive noise on both the movement equation and the sensor equation is again assumed to be zero-mean white Gaussian noise $\delta_u \sim \mathcal{N}(0, \sigma_u)$ and $\delta_H \sim \mathcal{N}(0, \sigma_H)$. Thus, σ_H describes the noise on the head orientation observation at each time step and σ_u describes the noise on the motor command that steers the head.

Assuming head orientation measurements to be independent of acoustic measurements, we show in McLachlan et al. (2021) that Eq D.6 can be reformulated. The dependency on all sensor readings and all head rotations executed so far can be expressed recursively as

$$L_H(t_i) = p(\theta_H(t_i) \mid y_H(t_0 : t_i), u(t_0 : t_i)) = p(\theta_H(t_i) \mid \hat{\theta}_H(t_i)) \tag{D.8}$$

with $\hat{\theta}_H(t_i) \sim \mathcal{N}(\mu_{\theta_H}(t_i), \sigma_{\theta_H}(t_i))$ the estimated head orientation updated at each step through a Kalman filter with:

$$\begin{aligned}
\mu_{\theta_H}(t_{i+1}) &= (1 - K) \cdot (\mu_{\theta_H}(t_i) + u(t_i)\Delta t) + K \cdot y_H(t_{i+1}), \\
\sigma^2_{\theta_H}(t_{i+1}) &= (1 - K) \cdot (\sigma^2_{\theta_H}(t_i) + \sigma^2_u)
\end{aligned}$$

and K the Kalman gain:

$$K = \frac{\sigma^2_{\theta_H}(t_i) + \sigma^2_u}{\sigma^2_{\theta_H}(t_i) + \sigma^2_u + \sigma^2_H}, \tag{D.9}$$

The expression in Eq D.8 is calculated by computing the Mahalanobis distance between given head orientation $\theta_H(t_i)$ and $\hat{\theta}_H(t_i)$.

At the initial time step t_0 we define: $\mu_{\theta_H}(t_0) = y_H(t_0)$ and $\sigma^2_{\theta_H}(t_0) = \sigma^2_H$.

Posterior computation

By marginalisation over all possible head orientations and using Bayes' theorem, we can combine the spatial prior and sensor model output to obtain the joint posterior PDF:

$$p_{t_i} = C \cdot p_{t_{i-1}} \times \int_{\theta_H} (L_H \times L_A) d\theta_H, \tag{D.10}$$

Turning to Bayesian terminology, p_{t_i} is the posterior PDF, $p_{t_{i-1}}$ is the prior PDF and the joint sensor model computes the likelihood. C is a normalisation constant. Note that the prior at time step t_i equals the posterior from time step t_{i-1}. At the initiation of the cumulative process, $p_{t_{i-1}} = p(\psi)$, which is the spatial prior. The detailed derivation of this equation is explained in McLachlan et al. (2021).

Fig D.1 illustrates how the posterior updates over time as more information arrives. Note that

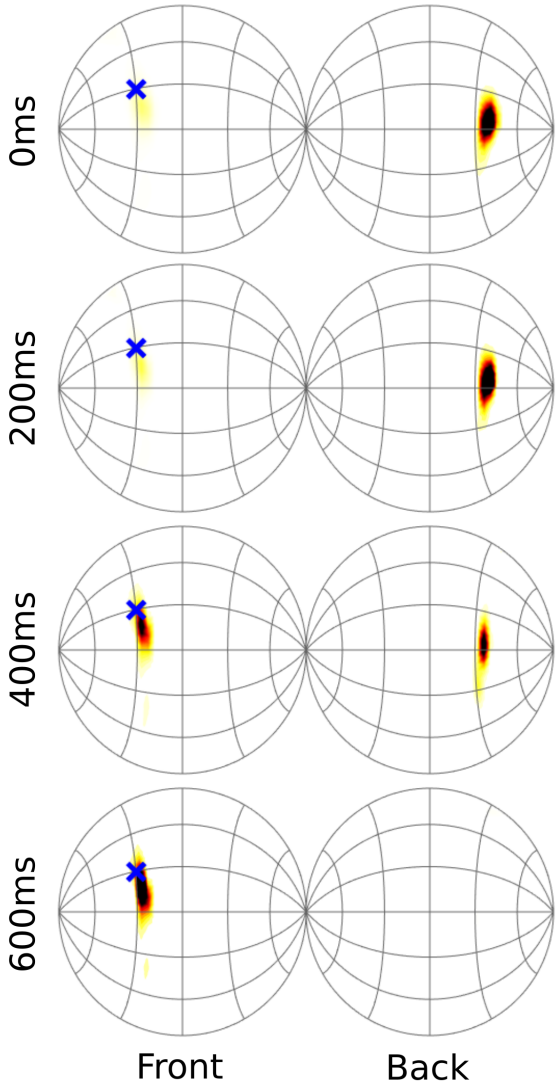

Fig. D.1 Example posterior distribution of source direction at different time steps during yaw rotation. Darker areas indicate higher probabilities. The blue 'x' is the true source direction.

in the numerical implementation of the model, the initial look contains all acoustic information, and the following looks only consider changes in the ITD cue to update the posterior. The reasoning for this is explained in the Methods.

Localisation decision rule

Eq D.10 returns a probability distribution over the sphere, i.e., a probability from a large but discrete set of source directions. The last step in the process is to obtain a point estimate from this posterior PDF. To do so, a decision rule must be defined. The present model uses the posterior matching strategy, where a weighted random sample is taken from the posterior PDF. However, it is easy to implement other strategies, such as the maximum a posteriori strategy, i.e., selecting

the location at the maximum of the posterior. It was found that localisation performance lies somewhere between the maximum a posteriori and posterior matching strategies (Ege et al., 2018).

Model parameters

General

The stimulus was a broadband time-invariant Gaussian white noise burst. The duration was the same as in the behavioural experiment, for both movement conditions. The simulated head movement was copied directly from the experimental head tracker data. In other words, for each simulated trial, the model executed the same head rotation as the subject did during the experiment.

The template $\mathbf{T_A}$ was listener-specific, i.e., derived from the individual's measured HRTFs. Earlier work suggests that the auditory system can detect changes (ITD (Bernstein et al., 2001), ILD (Brown and Tollin, 2016; Bernstein et al., 2001), spectrum (Hofman and Van Opstal, 1998; Martin and McAnally, 2008)) on a short time scale of about 5 ms. However, the full integration window of acoustic information for sound localisation appears to be more in the range of 100-200 ms (Zwislocki, 1969; Yabe et al., 1998). Furthermore, in localisation studies along the horizontal plane, the azimuth estimation reached best performance for stimulus durations of only 3 ms (Vliegen and Van Opstal, 2004). In studies along the vertical plane, a longer duration of 80 ms was required to reach the best performance in the elevation estimation (Hofman and Van Opstal, 1998). When head movements are allowed, a stimulus duration of approximately 100 ms seems to be required to provide a substantial benefit from the head movement (Macpherson, 2013). For the above reasons, the time step size Δt for the updating of the posterior was set to 100 ms. The effects of other step sizes are reported in the discussion.

The localisation task was simulated for 33 source directions and repeated 20 times without head movement, 20 times with a 10° rotation to the left, and 20 times to the right. This reflects the sound directions and repetitions used in the behavioral experiment. The simulations were repeated for each subject that participated in the behavioral experiment, then the results were pooled for further statistical analyses.

The spatial prior was assumed to emphasise directions around the horizon (Ege et al., 2018; Fischer and Peña, 2011). That is, for elevation, the prior has a mean around zero and a restricted variance σ_p. In Ege et al. (2018), the optimal value of σ_p was found to be around 11.5°. Initial simulations in the present model showed that this prior is too strong, as will be shown in the discussion. An explanation may be that this value was determined from a localisation exper-

iment that only included source directions in the elevation range of $[-35°, 35°]$ in the frontal hemisphere. For better comparability with the behavioural data, the model spatial prior was weakened to $30°$.

Acoustic information

In this implementation of the model, $\mathbf{y_L}$ and $\mathbf{y_R}$ were measured once, and y_{itd} was measured several times during stimulus presentation. The reason for this is twofold. First, it was found that dynamic spectral cues are not informative for sound localisation during small head rotations, which makes it unnecessary to take several measurements of the spectral information during head rotation (McLachlan et al., 2023). Second, the model's recursive estimation process relies on the assumption that the measurements are independent and identically distributed. This assumption does not hold for natural source spectra: we found the spectra of two subsequent segments of 100 ms for sources from the ESC-50 database (Piczak, 2015) to be highly correlated ($\rho \approx 0.8$).

The SD on the ITD measurement at each time step t, σ_{itd}, was set to 0.6 JND. The SD on the measurement of the spectral content, σ_I, was set to 3.5 dB. These values and units were derived earlier for the static localisation model in Reijniers et al. (2014). In the discussion, the effects of higher ITD measurement noise are reported.

The covariance matrix of the knowledge of the spectral content of the incoming sound source, Σ_S, i.e., the source prior, was derived from the ESC-50 database (Piczak, 2015), which is a collection of 2000 environmental audio recordings. Each of the sound files of the database were chopped up in intervals of 0.2s and, for each of these intervals, the source log-magnitude spectrum was expressed as function of the equivalent rectangular bandwidth centre frequencies. The resulting spectra were pooled in a single dataset, from which the average source spectrum $\hat{S}$ and the covariance matrix Σ_S were calculated.

Proprioceptive information

The SD on the measurement of the head orientation at each time step t, σ_H, and the SD on the motor command steering the head rotation, σ_u, were both initially set to $0°$. In other words, this assumed that the listener can perfectly estimate and control the head orientation. Human subjects are able to report motion and orientation perception with very high precision (Karnath et al., 1994; Diaz-Artiles and Karmali, 2021). In a seated position, the standard deviation of head rotation around the starting position (notated in our model as σ_u) is around $2°$ (Wersényi and Wilson, 2015). In the discussion the effects of higher motor noise are reported.

As with virtually all sensory systems, motor imprecision increases with stimulus magnitude (Todorov, 2005; Harris and Wolpert, 1998), i.e., noise increases with exerted force. However, motor noise (σ_H and σ_u) was assumed here to be additive for simplicity.

Theoretically, Eq D.10 requires marginalisation over all possible head rotations. However, with low proprioceptive noise, the probability distribution of head orientations based on the accumulated proprioceptive evidence will be near-zero for most orientations. Hence, a computational heuristic was used to only consider orientations within a range of $2\sigma_H$ from the true orientation, so that about 95% of the orientations were considered.

Acoustic measurements and behavioural experiment

Apparatus

The behavioural localisation experiment and acoustic HRTF measurements were conducted in a semi-anechoic room with 91 speakers (E301, KEF Inc.) distributed over the sphere within the elevation angles $-47°$ to $90°$. A head-mounted display (HMD, Oculus Rift, CV1, Meta Inc.) was used for the visual presentation of the virtual environment: a sphere with grid lines, which serve as anchor points to the subject's orientation in space. Although HMDs can minimally affect localisation performance (Gupta et al., 2018; Poirier-Quinot and Lawless, 2023), this decision was made as it was a higher precision error was found in darkness than when providing spatial information with an HMD (Majdak et al., 2010). Three infrared cameras were used for the tracking of the listener within the six degrees of freedom.

The experiment was controlled by a computer running a 64-bit Windows 10, equipped with an 8-core, 3.6-GHz CPU (i7-11700KF, Intel Inc.), 16 GB of RAM, and a graphic card with dedicated 8 GB of RAM (GeForce RTX 3070, NVIDIA Inc.). The experiment was controlled by the ExpSuite 1.1 application LocaDyn, version 0.9.7.

The tracking system provided a translation accuracy of below $1\ cm$ (Borrego et al., 2018) and a rotation accuracy of below $1°$ (for a similar tracking system (Monica and Aleotti, 2022)). The position and orientation of the subject's head were recorded for later analyses and to simulate using the model.

Subjects

Eight normal-hearing subjects (four female, four male) participated in the experiment. Their absolute hearing thresholds were within the average (± 1 standard deviation, SD) of the age-relevant norms (Corso, 1959; Stelmachowicz et al., 1989) within the frequency range from 0.125 to 12.5 kHz. The age range of the subjects was between 22 and 33 years.

Stimuli

Stimuli were always played over loudspeakers using virtual-base amplitude panning (Pulkki, 1997). Thirty-three source directions were distributed over the full sphere, which are visible in Fig D.3 and Fig D.4.

The acoustic stimulus used in this experiment was a wideband (20 to 20000 Hz) white noise burst, gated with a 10 ms cosine ramp. Each trial used the same noise realisation. The stimulus was gated off after 500 ms in the passive condition, and after 10° of head rotation for the active condition. For the latter, this means that the stimulus duration depended on the rotation velocity, with mean 634.9 ms and SD 335.7 ms.

Presentation level was measured to be 48 dBA SPL at the ear drum, with a ± 2.5 dB level roving range between trials.

Procedure

The localisation task procedure was identical to that of McLachlan et al. (2023). At the start of each trial, the subject kept the head still on the reference orientation at $(0^\circ, 0^\circ)$. The stimulus was then played and the subject remained still or initiated rotation depending on the movement condition. At the end of the stimulus, the subject pointed towards their perceived source direction with a hand-tracking device to provide their localisation estimate. No feedback was provided about performance during or after the trials.

In the condition labelled 'passive', the subject was instructed to keep the head still for the duration of the stimulus. For the condition labelled 'active', the subject was instructed to make a single-sided rotation (either to the left or to the right) as soon as they heard the stimulus onset. Half of the trials instructed a leftward rotation, the other half was rightward. The head rotation speed was unrestricted, but was monitored through the tracking system of the VR headset and recorded for analysis.

In total, the passive experiment consisted of 660 trials (33 directions and 20 repetitions) per subject. The active experiment consisted of 1320 trials per subject: 660 for a leftward rotation and 660 for a rightward rotation. The trials were divided into 6 blocks, with trials and blocks presented in random order. For passive localisation, trials which exceeded 2° of movement in any direction were excluded (225 omissions in total). For active localisation, trials that resulted in a total yaw rotation smaller than 7° or larger than 13°, or with a pitch rotation larger than 6° were omitted (648 omissions in total).

Subjects were trained before commencing the experiment. The training consisted of 300 trials

with the 500ms white noise burst, played from a direction randomly selected from a uniform distribution within the range of available directions. Subjects were not excluded based on their performance at the end of training.

Localisation metrics

There are many metrics available for sound localisation performance, this makes comparison between localisation studies difficult. It is, however, generally accepted that a distinction needs to be made between two types of errors (Carlile et al., 1997). The first type is the local error. Here we use the lateral-polar coordinate system (Morimoto and Aokata, 1984), (θ, ϕ), where $\theta \in [-90, 90]$ and $\phi \in (-180, 180]$, with $(\theta, \phi) = (0, 0)$ defined as straight ahead. The local error was expressed in root mean-squared error (RMSE) value of the lateral and polar errors. Polar RMSE was only considered in the range of $\pm 30^\circ$ lateral angle, and estimates in the wrong hemisphere (i.e., front-back and up-down confusions) were excluded from the local errors, following the definition by Middlebrooks (1999).

The second type of error is the reversal error, which generally is reported as a percentage, i.e., the rate of reversals in a given set of trials. The first reversal error considered was the quadrant error (QE) rate, which is defined as any polar error larger than 90°. Additionally, we used the front-back confusion (FBC) rate and the up-down confusion (UDC) rate , which are defined as any response crossing the frontal plane and the horizontal plane, respectively. This is the same definition as used by Carlile et al. (1997), and thus allows for a direct comparison between present and previous data. Note that this is a very coarse definition for the reversal error, as it confounds FBCs and UDCs with local errors near the frontal or horizontal plane, respectively.

Results

Global statistics

First, we present the global results, in order to compare to the existing literature. Table D.1 presents the localisation data of both the behavioural (B) experiment and the model (M) simulations (means and SDs averaged over the subjects), for passive (P) and active (A) conditions.

Local errors

The lateral errors in the behavioural results agreed with previous findings from similar experimental setups (Oldfield and Parker, 1984; Middlebrooks, 1999). However, the model results were small compared to the behavioural data. This may have been the result of σ_{itd} being set too small, or of the behavioural responses being confounded with a 'pointing' error. The effect

Table D.1 Averages and SDs of behavioural (B) and modelled (M) localisation performance in the passive (P) and active (A) conditions. The performance is represented as the lateral and polar RMSE (in degrees), QE, FBC, and UDC rates (in %). Means and SDs were computed over eight (virtual) subjects. For comparison, the results from previous work are reported too (Middlebrooks, 1999). N.R.: not reported.

Condition	L. RMSE (deg)	P. RMSE (deg)	QE (%)	FBC (%)	UDC (%)
BP (Middlebrooks, 1999)	10.6 ± 2.0	22.7 ± 5.1	4.6 ± 5.9	N.R.	N.R.
BP	8.1 ± 1.4	20.0 ± 2.4	7.9 ± 4.5	10.7 ± 6.3	6.7 ± 4.2
BA	8.2 ± 1.8	18.9 ± 5.0	0.7 ± 1.1	1.5 ± 1.2	4.2 ± 4.1
MP	2.7 ± 0.5	17.9 ± 2.5	7.4 ± 2.0	4.0 ± 1.3	3.1 ± 0.5
MA	2.7 ± 0.2	15.3 ± 2.2	2.0 ± 1.1	1.1 ± 0.3	1.8 ± 0.7

of ITD noise on the model performance is tested below.

The polar errors of the behavioural results were consistent with previous findings (Middlebrooks, 1999; McLachlan et al., 2023). Like the lateral error, the mean polar error of the simulated trials was lower than that of the behavioural data. However, the difference here was smaller.

No decrease was seen in polar error in condition BA. This is evidence against the 'Wallach cue' (Perrett and Noble, 1997a), and agrees with previous findings (McLachlan et al., 2023). However, condition MA did show a decrease in polar error. Although this improvement is still small, it is an indicator that the Wallach cue may be theoretically informative, as the model was able to obtain elevation information from yaw movement. The reason why this isn't seen in humans could be due to motor noise. This is investigated below.

Interestingly, the SD of the polar RMSE increased in condition BA, even though the mean did not change much. This suggests that the effects of head movement may be subject-dependent, e.g., motor noise during motion may be higher for some individuals.

Reversal errors

QE rates in conditions BP and MP agreed with previous work (Middlebrooks, 1999). Furthermore, the near-complete removal of QEs in conditions BA and MA also confirms the consensus that head rotation resolves all reversal errors (McLachlan et al., 2023).

The FBC rate in condition MP was notably lower than in condition BP. Looking at earlier studies, FBC rates of normal hearing listeners were closer to $3 - 6\%$ (Carlile et al., 1997; Fischer et al., 2020; Makous and Middlebrooks, 1990), although the errors were highly subject-

dependent. This means that the high FBC rate in this study is somewhat anomalous. Hence, the cause for the discrepancy here seems to lie in the behavioural results, not in the model predictions.

There was also a slight decrease in the UDC rate. This reduction may have been caused by the improved polar estimation obtained from the Wallach cue. It is possible that the rotation made during some trials contained a significant roll-component, which helps distinguish between the lower and upper hemispheres (McLachlan et al., 2021). However, the tracker data shows that overall roll rotation was small, with the mean absolute roll $0.82°$ and the SD $0.51°$.

The high SD in the reversal errors shows that this metric is highly subject dependent. The SDs of reversal errors between model 'virtual listeners' were small compared to the behavioural results. This is not surprising, as the individual differences between subjects are likely not fully explained by the individual HRTFs. The same noise parameters were used for each individual, while it is likely that they differ per individual (Barumerli et al., 2023). Furthermore, higher level processes such as listening strategies (Thurlow et al., 1967) or attention (Klatt et al., 2018) will also be a cause for individual differences.

Spatial analysis

Following the methods of visualisation of previous work, the localisation responses were modelled as elliptical Kent distributions (Leong and Carlile, 1998; Carlile et al., 1997). The centroids visualise the bias, i.e., the mean vector, of all responses for one source direction. The ellipsoid outlines visualise the equal probability contours of the distribution of responses. The major and minor axes of the ellipsoid are two SDs in length and represent the first two orthogonal 'principal components' of the dataset that account for the maximum amount of variance in the data. Fig D.2 illustrates the Kent distribution of all responses for the frontal direction of condition BP.

Fig D.3 visualises the behavioural and model localisation results per source direction on the sphere around the listener. Quadrant errors were excluded.

Response bias

The direction of the centroids of condition BP are similar to those found in (Carlile et al., 1997). More specifically, the centroids show a bias towards the audio-visual horizon and towards the interaural axis, i.e. the left and right ear. This supports the already strong evidence for a spatial prior on the horizon (Ege et al., 2018). Several other studies have shown that human sound localisation displays a peripheral bias that increases with eccentricity (Oldfield and Parker, 1984;

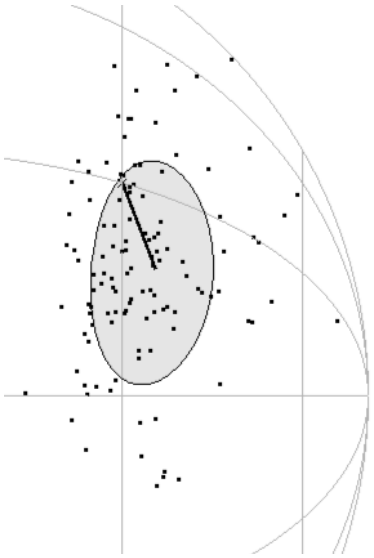

Fig. D.2 Centroid and Kent distribution for condition BP and source direction $(30°, 30°)$. Black dots are the individual subject responses, from which the Kent distribution was calculated.

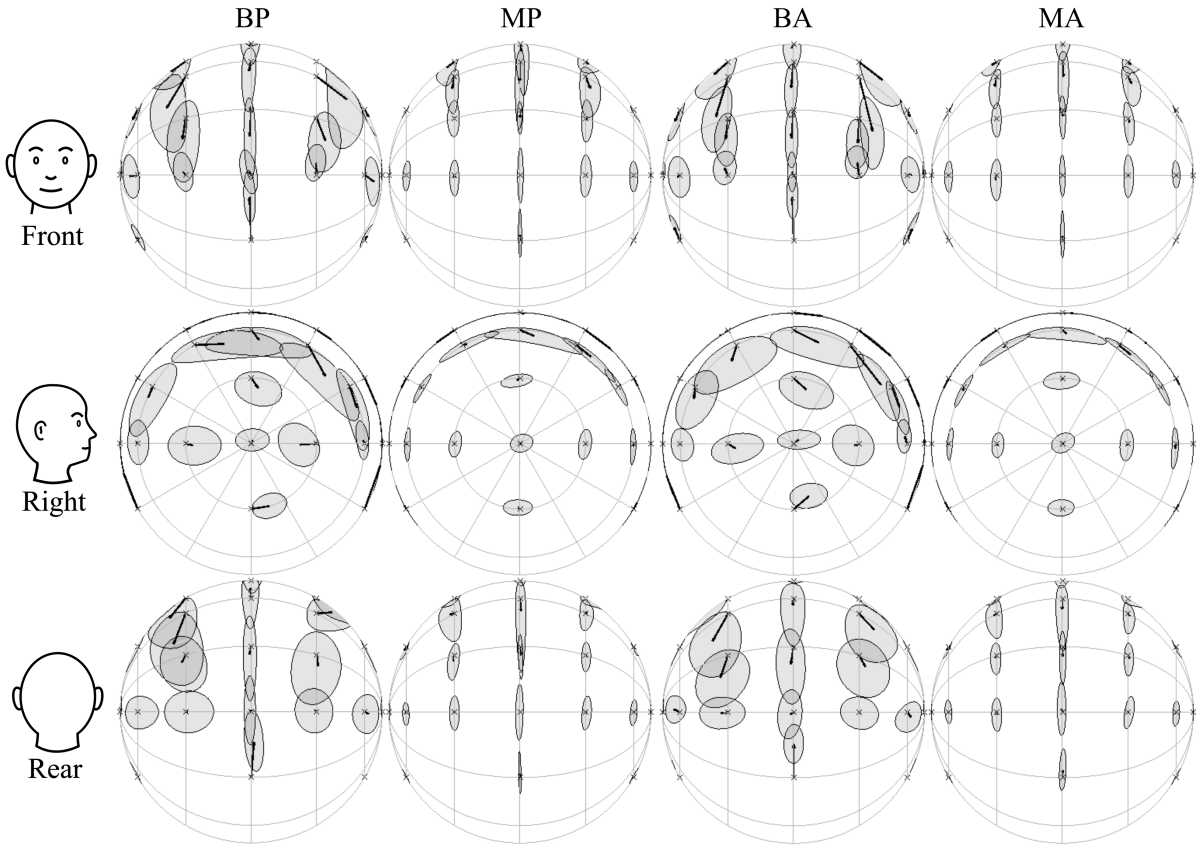

Fig. D.3 Centroids and Kent distributions of behavioural (B) and modelled (M) responses in the passive (P) and active (A) conditions, averaged over eight subjects. The rows show the same data viewed towards the front, the right, and the back of the head. Quadrant errors were excluded.

Lewald and Ehrenstein, 1998; Odegaard et al., 2015).

In condition BA, the bias towards the horizon seemed to increase further. This can be explained if we assume an increase in uncertainty on the orientation of the head, or perhaps on the spectral cues during motion. Due to this increase in sensory noise, the relative strength of the spatial prior towards the horizon would increase. The biases did not change between conditions MP and MA. Possibly, this effect was not seen in the model because the head orientation was assumed to be perfectly known. Below we investigate the influence of increased proprioceptive noise.

Two explanations for spatial biases have been proposed previously: 1) a Bayesian approach describes a pre-existing spatial prior for certain locations that is added to the sensory information, 2) alternatively, responses may be pulled to certain directions due to compressions and expansions in the sensory representations of auditory (and visual) space (Odegaard et al., 2015; Brimijoin, 2018). Fig D.3 shows that the selected spatial prior for the model results in vertical biases that are similar to the behavioural data. This is evidence that the spatial biases can, at least in part, be explained by a prior.

Response variance

Similar to the Kent distributions in Carlile et al. (1997), condition BA shows larger distributions for sources at higher elevations, and for sources in the rear hemisphere. The response spread is also highest along the polar dimension, more specifically, along the cones of confusion (Blauert, 1997).

The spread in model responses also followed the cones of confusion. On the median plane the response distributions appear fairly similar. However, the response spread for sources at higher lateral angles was noticeably smaller than in the behavioural responses, especially for the lateral responses. This was to be expected from the lower local errors that were seen earlier.

The biggest difference was seen for sources in the rear. For conditions MP and MA, the response distributions for sources in the rear were nearly identical to those in front. On the contrary, the behavioural data contained a much larger spread in the rear, especially along the lateral dimension. This suggests that the increase in variance, i.e., decrease in precision, found in the rear hemisphere of the behavioural results cannot be (fully) attributed to a lower spatial resolution in acoustic cues. Instead, a response or 'pointing' error may have been responsible. Even if the auditory system can perfectly estimate a source location, the action of pointing in a direction as a response may introduce an additional error. It is reasonable to hypothesise that sources behind the listener are more difficult to point towards consistently. Pointing errors have been modelled previously (Baumgartner et al., 2014; Barumerli et al., 2023), though only as

a simple Gaussian noise source, which does not take into account the source direction. More research is required to accurately model the spatial dependence of a response error.

There were no large differences in Kent distributions between passive and active conditions, neither for the behavioural data nor for the model. One exception is the source position directly above the listener in the model predictions. In condition MP, a much higher polar spread is predicted than in condition BP. This can be explained by the Gaussian shape of the spatial prior, which affects sources at higher elevations more heavily than those around the horizon. Interestingly, this spread isn't visible in condition MA, which means that head rotation significantly improved estimation of this source direction and outweighed the spatial prior. This is another indicator of the available Wallach cue in the model simulations. This suggests that either perfect knowledge of the head orientation or lower noise on the ITD made head rotation more informative for the model than in the behavioural experiments.

Quadrant errors

The spatial distribution of QEs was visualised in Fig D.4. Condition BP reveals that QE rates were more common in the rear hemisphere than in the front: 59.4% and 40.6%, respectively. Most notable are the source directions directly in front of the listener, which showed nearly no QEs at all. Condition MP also shows more QEs in the back than in the front: 74.6% and 25.4%. This suggests that acoustic information accounts for most of the quadrant errors found. However, as was found in the Kent distributions, the source positions above the listener showed very different results between conditions BP and MP, this was caused by the spatial prior towards the horizon.

Previous studies have shown that the quantity and spatial distribution of reversal errors were highly subject-dependent. For example, one study found an even higher percentage of QEs to be located in the rear hemisphere than in the present study (Middlebrooks, 1999). In another study, only two subjects showed a significant majority of reversal errors in the rear, while two showed a majority in the front, and two showed an equal distribution (Makous and Middlebrooks, 1990). Similarly, this study contained three subjects with a majority in the back, one with a majority in the front, and four with no clear preference. For the localisation of low-pass stimuli, most confusions happened for sources in the front, note that this may be because sources from behind undergo more filtering than sources from the front (Carlile et al., 1999).

From the present findings and the available literature, it is apparent that reversal errors involve a complex process that differs between individuals.

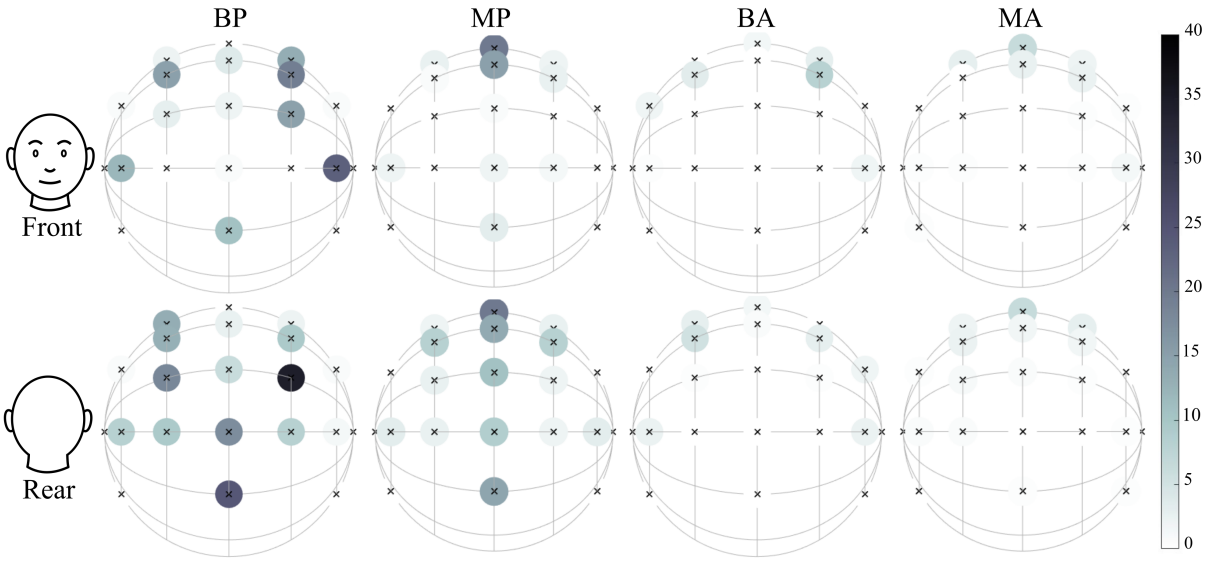

Fig. D.4 Quadrant error rates [%] of behavioural (B) and modelled (M) responses in the passive (P) and active (A) conditions, averaged over eight subjects. The rows show the same data viewed towards the front and the back of the head.

Effect of ITD noise

Table D.1 shows that the lateral RMSE was too small compared to the behavioural results. Here, we tested new values of σ_{itd} to investigate whether this parameter is the cause of the discrepancy. The results were plotted in Fig D.5.

The lateral error increased monotonically with σ_{itd}. However, even for $\sigma_{itd} = 3°$, the errors were still lower than the behavioural results. This shows that ITD noise alone cannot account for the discrepancy in lateral error. As expected, polar errors remained mostly unaffected in the passive condition, as the static ITD contains little to no information on the polar angle. However, the reduction in polar error in the active condition (due to the Wallach cue) was only visible for $\sigma_{itd} < 1.2°$. Minimising σ_{itd} led to a complete removal of QEs in the active condition, whereas maximising it made the passive and active conditions near-identical. Together, the results suggest that a low noise on the ITD cue is essential to utilise the dynamic cue that resolves reversal errors when moving the head. They also show that the value of $\sigma_{itd} = 0.6$ that was derived from a previous experiment is a plausible value (Reijniers et al., 2014).

Effect of the proprioceptive noise

For the initial simulations, the head orientation was assumed to be perfectly known. Here we investigated the effects of an increased uncertainty on the head orientation. The simulations were rerun with different values for σ_H and σ_u. The results were plotted in Fig. D.6.

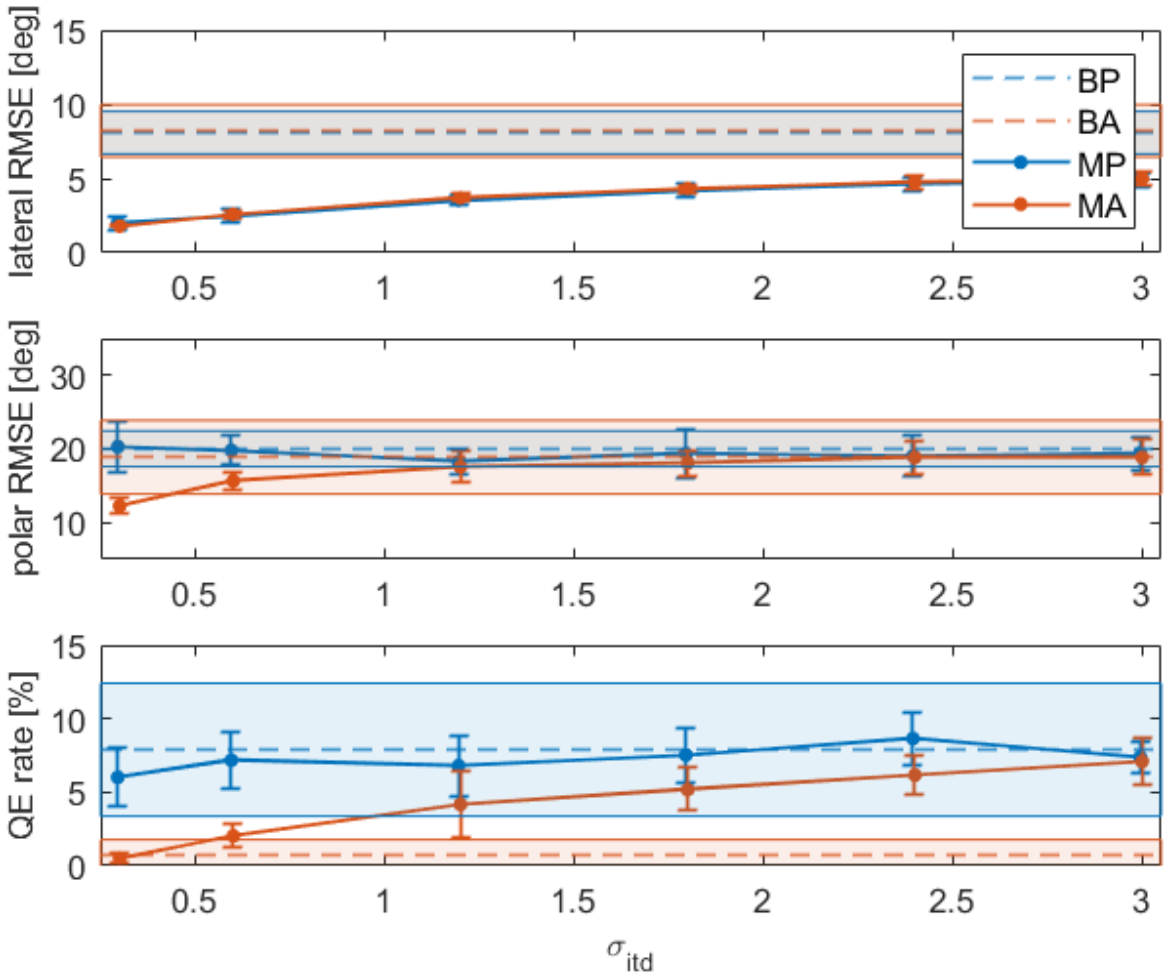

Fig. D.5 Lateral RMSE, polar RMSE and QE rate of the modelled data as a function of σ_{itd} (in units of JND). Blue markers are passive results, orange markers are active results. The markers and the error bars represent the mean and standard deviation over the eight modelled subjects. For reference, the dashed lines and the coloured areas show the behavioural means and standard deviations over the eight subjects, respectively.

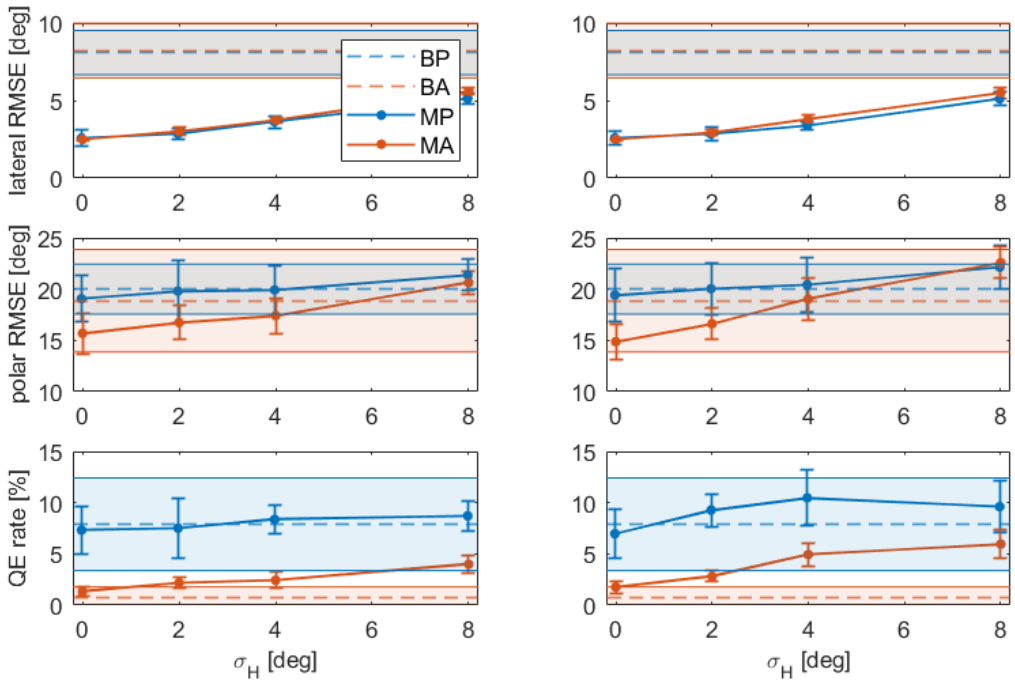

Fig. D.6 Lateral RMSE, polar RMSE and QE rate of the modelled data as a function head orientation measurement noise σ_H, with head control noise $\sigma_u = 0°$ (left column) and $\sigma_u = 8°$. Blue markers are passive results, orange markers are active results. The markers and the error bars represent the mean and standard deviation over the eight modelled subjects. For reference, the dashed lines and the coloured areas show the behavioural means and standard deviations over the eight subjects, respectively.

The results show that a larger value for σ_H increased the lateral error, similar to σ_{itd}. There appeared to be no interaction effect with σ_u.

Polar error monotonically increased with σ_H, though the active condition seemed to suffer slightly more. In other words, the difference in polar error between condition MP and MA became smaller as σ_H increased. This effect was even stronger with a high σ_u. This suggests that uncertainty on the head position is the reason why the Wallach cue cannot be used by human listeners.

For $\sigma_u = 0°$, the QE rate was affected slightly for higher values of σ_H, notably less than when σ_{itd} was increased. For $\sigma_u = 8°$, the effects were larger, and even condition MP seemed to suffer more QEs.

Together, this leads to the conclusion that the uncertainty on the proprioception measurement of the head orientation, σ_H, can account for several (though not all) of the differences found between the behavioural data and the model output, but that the noise on the execution of the control signal, σ_u needs to be low.

Note that σ_H and σ_u will likely be higher in the active condition than in the passive condition, as motor noise is multiplicative (Todorov, 2005).

Effect of the spatial prior

The bias vectors in the behavioural results appeared slightly stronger than for the simulations with $\sigma_p = 30$. This implies that a stronger spatial prior may be necessary. There are many different possible spatial prior shapes, e.g. a Laplace distribution (Gerven et al., 2009), or prioritising the front or high lateral angles (Garcia et al., 2017). In this study, the analysis was restricted to the horizontal prior. Fig. D.7 shows the elevation gain (Ege et al., 2018) (i.e., the slope of a linear regression between responses and true directions) and QE rate for different values of σ_p. As we are interested in the 'pull' towards the horizon, the elevation gain is a more appropriate indicator than the polar RMSE.

The plot suggests that the correct prior σ_p lies between $20°$ and $30°$, where the gain and QE rates are closest to the behavioural results. Thus, $\sigma_p = 30$ was again a fairly good estimate, though the large standard deviations between subjects in the behavioural data suggest that the strength of the prior may be subject-dependent. As σ_p increases, the prior approaches a uniform distribution and the relative weight of the sensory information will increase. The plot shows that, as a result, the responses approach an elevation gain of 1. This suggests that the bias present in human responses is not due to acoustic factors, but indeed due to a prior towards the

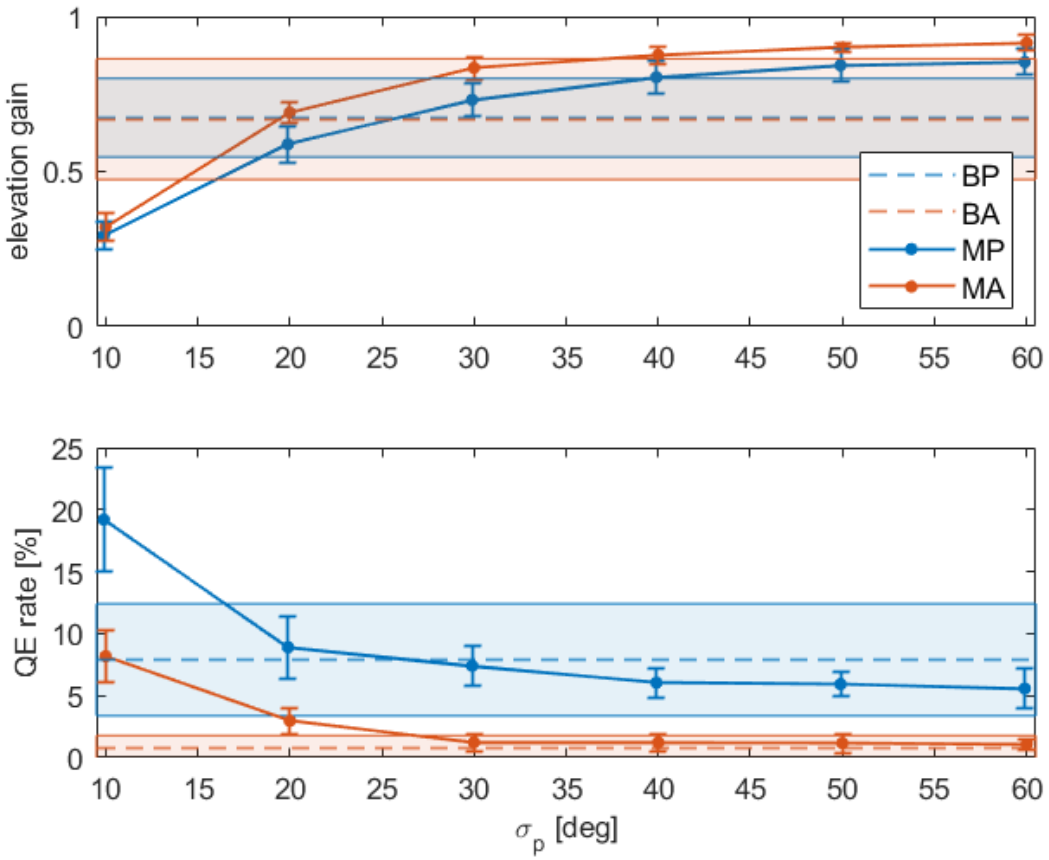

Fig. D.7 Elevation gain (Ege et al., 2018) and QE rate of the modelled data as a function of σ_p (in degrees). Blue markers are passive results, orange markers are dynamic results. The markers and the error bars represent the mean and standard deviation over the eight modelled subjects. For reference, the dashed lines and the coloured areas show the behavioural means and standard deviations over the eight subjects, respectively.

horizon.

However, there remains a problem in the distribution of the errors. First, it was noted that the prior affects sources at higher elevations more heavily than those around the horizon, resulting in an unrealistically high spread in responses and high QE rate for sources above the listener (see Fig. D.3 and Fig. D.4, condition MP). There also remains the lateral bias that is unaccounted for with the spatial prior tested in this study. These discrepancies could mean that humans have an additional auditory spatial bias upwards and towards higher lateral angles. Alternatively, as stated earlier, it is possible that responses are pulled or 'snap' to certain directions due to the sensory representations of space (Garcia et al., 2017; Odegaard et al., 2015). The latter explanation would be an example of non-ideal observer behaviour.

Effect of the time step

To investigate the effect of the time step size Δt, the model was rerun with different update rates. Note that each simulated trial always contained at least two time steps at the start and at the end of the stimulus duration, to make dynamic cues available. The results are shown in Fig. D.8.

Generally, it seems that a smaller Δt makes localisation performance slightly more accurate, this is the result of more looks being available. This improvement is mostly visible in the polar RMSE of condition MA, which shows that the benefit of the Wallach cue becomes more prevalent if more ITD looks are allowed. Unsurprisingly, a similar effect was seen when the noise on ITD looks was kept low. For the other metrics, the improvement performance is surprisingly small. Regarding the QE rate, it appears that two ITD looks (one at the start and one at the end of rotation) were sufficient to prevent most errors, and that any look in between is somewhat redundant. Finally, there appeared an unexpected dip in the QE rate of condition MP at $\Delta_t = 50ms$. A repetition of the model simulations revealed that this was a statistical anomaly.

Conclusions

This article introduced a Bayesian ideal observer model that enables a bottom-up investigation of human performance in the task of active sound localisation. In order to investigate to what extent humans perform as ideal observers, the model output was compared to behavioral results obtained in a free-field localisation experiment.

With parameters selected a priori, i.e., without the use of any post-hoc fit to the behavioral

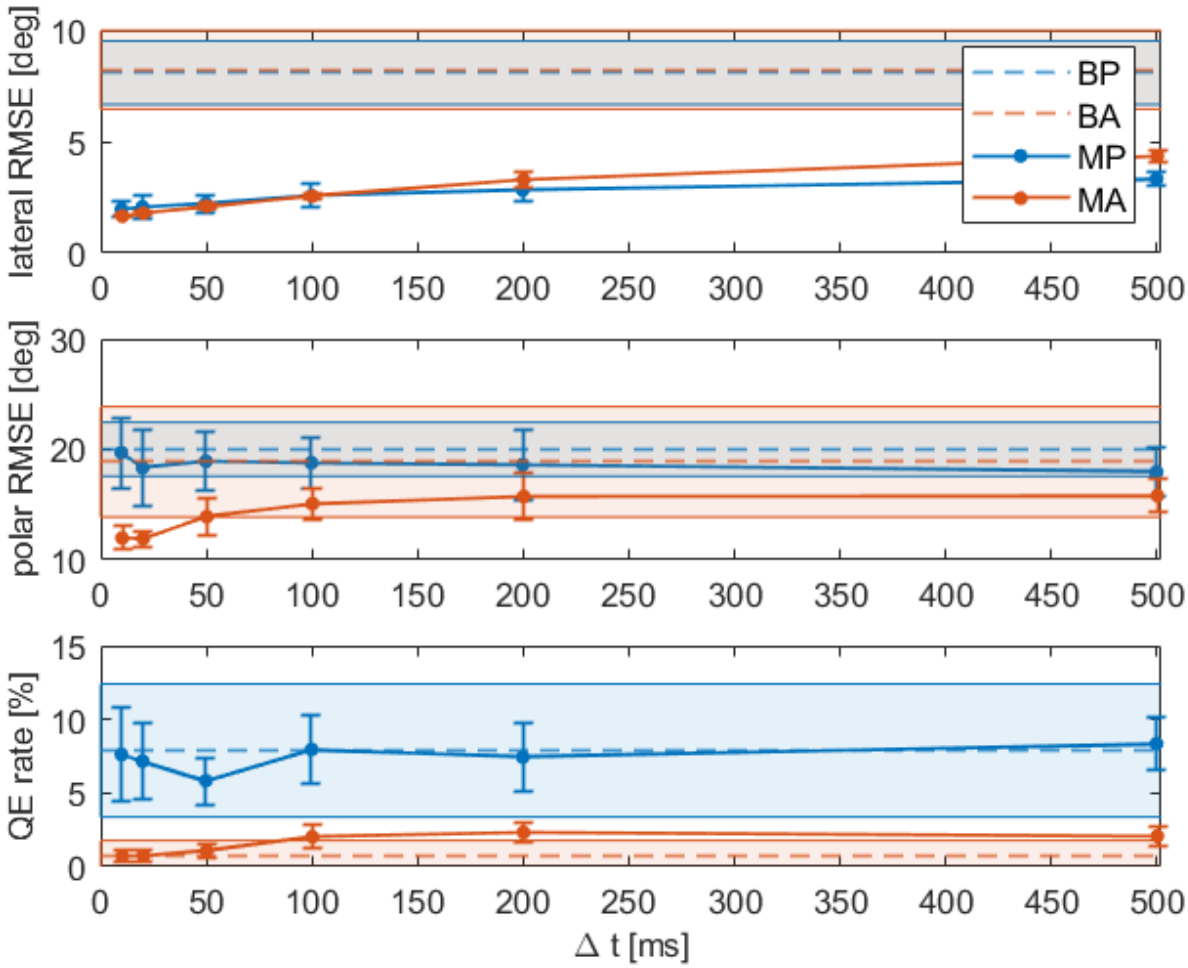

Fig. D.8 Lateral RMSE, polar RMSE, and QE rates of the modelled data as a function of time step size Δt. The symbols show the averages and the error bars represent ± 1 SDs over the (virtual) subjects. For reference, the horizontal dashed lines show the behavioural data.

data, the model predicted and explained the human performance in a general sense. This is an encouraging finding, supporting the hypothesis that model parameters can be derived a-priori based on general behavioural experiments. Furthermore, as the model only processed changes in ITD, it also confirms the earlier finding that humans do not utilise dynamic spectral cues for localisation, at least during small head rotations (McLachlan et al., 2023).

In a more detailed spatial analysis, the model predictions deviated from the behavioral data. The largest differences were found for sources to the rear and above the listener. We investigated in detail the conditions where the model agreed or deviated from the behavioural data by studying the effects of several model parameters on the predictions. The ITD noise parameter revealed that an ITD cue alone does not account for the lateral RMSE and QE rate found in behavioural data. The uncertainty on the head orientation had a significant effect and was able to partially explain the behavioural data.

The discrepancies we found between the model predictions and behavioural data are important for future investigations. First, there is a high variance in behavioral responses for sources behind the listener, whereas the model showed little difference between the front and rear hemisphere. This indicates that these errors result from non-acoustic factors, such as the pointing error. Second, human listeners showed a response bias towards larger lateral angles, which was not seen in the model predictions. The origins of this discrepancy remain an open question, with a lateral spatial prior or a stretched sensory representation of space as potential candidates. Third, the model predictions showed a higher response variance and QE rates for sources placed above the listener. In the model, this might be an effect of the Gaussian spatial prior towards the horizon, which might not fully reflect the spatial prior of the human listeners, or even point to a more complex mechanisms at play.

Our framework can be applied in the future to a variety of phenomena that have been identified in previous studies on active sound localisation, such as the improvement of elevation perception with yaw movements for low-pass stimuli (Perrett and Noble, 1997a), elimination of FBCs for low-pass stimuli (Macpherson, 2013), and the relative weight of dynamic ILD and ITD in the localisation process (Pöntynen and Salminen, 2019).

Acknowledgements

The authors would like to thank Philip Leong for providing the MATLAB code to the ‘Spak’ spherical data processing tool. Part of this code was rewritten for the purpose of this work and integrated into the AMT. This research was supported by the Research Foundation Flanders (FWO, under grant number G023619N), the Agency for Innovation and Entrepreneurship

(VLAIO, grant number HBC.2019.2191), and the European Union (project “SONICOM”, grant number 101017743, RIA action of Horizon 2020).

PAPER E

Head rotations follow those of a truncated Fick gimbal during an auditory guided visual search task

Glen McLachlan[1], Pedro Lladó[2], Herbert Peremans[1]

[1] Department of Engineering Management, University of Antwerp, Belgium
[2] Acoustics Lab, Department of Signal Processing and Acoustics, Aalto University, Finland

Abstract - Recent interest in dynamic sound localisation models has created a need to better understand the head movements made by humans. Previous studies have shown that static head positions and small oscillations of the head obey Donders' law: for each facing direction there is one unique three-dimensional orientation. It is unclear whether this same constraint applies to audiovisual localisation, where head movement is unrestricted and subjects may rotate their heads depending on the available auditory information. In an auditory guided visual search task, human subjects were instructed to localise an audiovisual target within a field of visual distractors in the frontal hemisphere. During this task, head and torso movements were monitored using a motion capture system. Head rotations were found to follow Donders' law during search tasks. Individual differences were present in the amount of roll that subjects deployed, though there was no statistically significant improvement in model performance when including these individual differences in a gimbal model. The roll component of head rotation could therefore be predicted with a truncated Fick gimbal, which consists of a pitch axis nested within a yaw axis. This led to a reduction from three to two degrees of freedom when modelling head movement during localisation tasks.

E.1 Introduction

Sound localisation performance improves when listeners are able to move their heads during stimulus presentation. Changes in interaural cues as a result of head movement help resolve front-back confusions and can, in certain situations, improve elevation estimation (Wallach, 1940; Thurlow and Runge, 1967; Perrett and Noble, 1997b; Kato et al., 2003; Iwaya et al., 2003; McAnally and Martin, 2014; McLachlan et al., 2023).

Traditionally, experiments and models of human sound localisation have mainly focused on passive localisation, where the head remains stationary. However, new available technologies in head tracking and virtual reality have led to a growing interest in understanding how head movements affect sound localisation performance (Carlile and Leung, 2016). A few models exist that consider the dynamic position of the head during sound localisation (McLachlan et al., 2021; Lladó et al., 2024), though this remains a relatively new area of study and introduces a number of new challenges in auditory modelling. First, it raises the question of how acoustic and sensorimotor information are combined. Second, the incoming information needs to be integrated over time. Third, a separate movement model must be defined to simulate head rotation. In this manuscript we focus on addressing the latter challenge.

Models of head motion have been proposed in the past with various levels of complexity. There are models that consider rotation (Kunin et al., 2007; Ghosh and Wijayasinghe, 2012), translation (Medendorp et al., 1998) or acceleration patterns (Zangemeister et al., 1981a). Movement also seems to depend highly on the individual (Kim et al., 2013a). Furthermore, experiments that also investigated unrestricted and natural head movement found that movement is task-dependent (Ceylan et al., 2000; Kunin et al., 2007), so a model for one task may not be applicable to another. In the context of sound localisation, it is possible that humans move their heads to optimise the acoustic information that they receive, which may lead to its own unique head movement behaviour Brimijoin et al. (2010); Hendrikse et al. (2022).

The arguments above make it difficult, though important, to find general rules of head motion that may simplify a movement model. A common example of such simplifications is the assumption that humans generally move their heads according to Donders' law, i.e., the head does not make use of all three degrees of freedom when rotating (Donders, 1847; Von Helmholtz, 1867; Listing, 1905). Rather, for any direction of the head, its rotation around its direction is unique. In other words, the amount of roll (i.e. torsion) is not controlled separately, instead it is a function of the amount of yaw (i.e. horizontal) and pitch (i.e. vertical) rotation (Medendorp et al., 1998; Ghosh and Wijayasinghe, 2012). Note that humans are anatomically capable of executing independent roll rotations, but that Donders' law implies that, in practice, humans do

not utilise this additional degree of freedom.

The goal of the present study is to investigate head movement behaviour by collecting head (and torso) motion data during an auditory-guided visual search task. More specifically, we wish to answer two subsequent questions: 1) do the head movements in this task follow Donders' law and 2) how can we best model this? The results of this study can then be used to integrate a more realistic movement model into existing sound localisation models that process self-motion.

E.2 Methodology

E.2.1 Coordinate and rotation conventions in three dimensions

All axes and rotations described in this paper follow the right-hand rule and are expressed in the global coordinate system, with the axes fixed to the world. The positive x-axis points forwards, with positive roll rotations towards the right shoulder. The positive y-axis points to the left, with positive pitch rotations downwards. The positive z-axis points upwards, with positive yaw rotations towards the left. The coordinate system and the directions of rotation are illustrated in Fig. E.1.

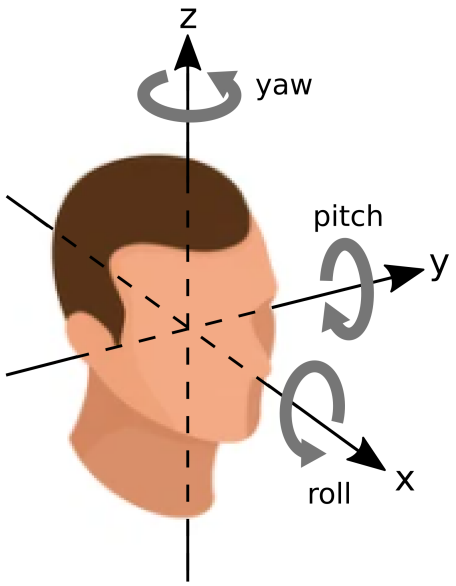

Fig. E.1 Visualisation of three-dimensional axes and corresponding positive rotations.

The orientation of an object in a three-dimensional space with respect to a fixed coordinate system can be described by a set of three rotations around three principal axes, the Euler angles.

If θ is the rotation angle around a single axis, then rotation over the spatial x axis, i.e., roll, is

defined as:

$$R_x(\theta_x) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & cos(\theta_x) & -sin(\theta_x) \\ 0 & sin(\theta_x) & cos(\theta_x) \end{pmatrix}, \tag{E.1}$$

Rotation over the spatial y axis, i.e., pitch, is defined as:

$$R_y(\theta_y) = \begin{pmatrix} cos(\theta_y) & 0 & sin(\theta_y) \\ 0 & 1 & 0 \\ -sin(\theta_y) & 0 & cos(\theta_y) \end{pmatrix}, \tag{E.2}$$

Rotation over the spatial z axis, i.e., yaw, is defined as:

$$R_z(\theta_z) = \begin{pmatrix} cos(\theta_z) & -sin(\theta_z) & 0 \\ sin(\theta_z) & cos(\theta_z) & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{E.3}$$

The rotation matrix to obtain the final orientation can be expressed by a successive multiplication of the three single-axis rotation operators. This can be done in any order, but this will result in different final orientations, so it is also important to define the order of rotations. In this paper we follow the Fick convention, which defines R as a rotation over the x-axis, then the y-axis and then the z-axis using world-fixed rotation axes.

$$R = R_z * R_y * R_x, \tag{E.4}$$

We can also express rotation matrix R as a vector $\mathbf{r}$, which can bring the object from the reference orientation to the orientation of interest by just one rotation around a single axis. This representation has been used in several studies investigating head and eye movements (Haslwanter, 1995; Medendorp et al., 1998; Kunin et al., 2007).

Rotation vector $\mathbf{r}$ with components $(r_x, r_y, r_z)^T$ is defined as (Haslwanter, 1995):

$$\mathbf{r} = \frac{1}{1 + R_{11} + R_{22} + R_{33}} * \begin{pmatrix} R_{32} - R_{23} \\ R_{13} - R_{31} \\ R_{21} - R_{12} \end{pmatrix}, \tag{E.5}$$

where R_{ij} are the indices of rotation matrix R, where i is the row and j is the column. The length of $\mathbf{r}$ is a function of the angle of rotation α around the vector $\mid \mathbf{r} \mid = tan(\alpha/2)$ and the direction of the rotation axis coincides with that of $\mathbf{r}$, again using the right hand rule for the sense of rotation.

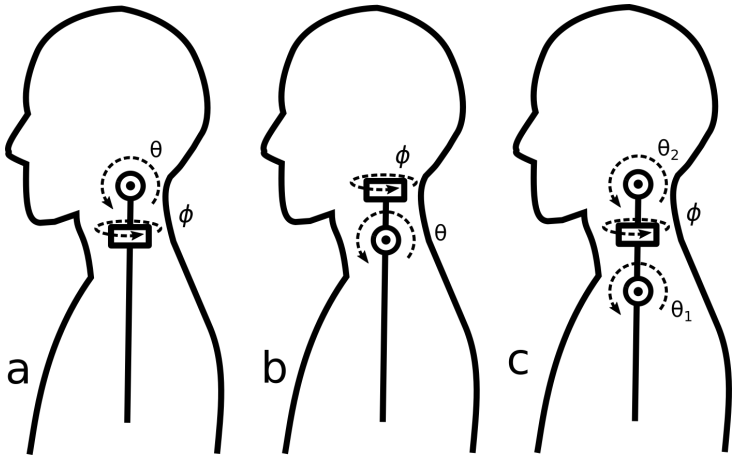

Fig. E.2 a) Truncated Fick gimbal, b) Truncated Helmholtz gimbal and c) k-gimbal model, where θ denotes a pitch rotation and ϕ denotes a yaw rotation. Note that the axes were visually separated for clarity; the models do not undergo any translation. The upper rotation axes are nested within the bottom axes, e.g., the truncated Fick gimbal is a pitch axis nested within a yaw axis.

E.2.2 Models for head rotation

According to Donders' law, the rotation vectors of a head rotation in any direction should fit on a second-order twisted surface (Tweed and Vilis, 1990):

$$r_x = a_1 + a_2 r_y + a_3 r_z + a_4 r_y^2 + a_5 r_y r_z + a_6 r_z^2, \tag{E.6}$$

where a_1 adjusts the surface offset, a_2 and a_3 adjust the surface orientation, a_4 and a_6 yield a parabolic curvature and a_5 allows for the surface to twist.

By examining the roll, pitch and yaw components of the rotation vector, it was inferred that the constraints on the head during natural movement resemble a specific form of Donders' law, following the rotations of a truncated Fick gimbal (with no roll axis) (Glenn and Vilis, 1992; Ceylan et al., 2000; Kunin et al., 2007). Here the yaw axis of rotation is fixed relative to the trunk and can change the direction of the supported pitch axis (Radau et al., 1994). Thus, the roll component of the rotation vector (r_x) in such a system is not controlled independently, but depends on the eccentricity of oblique facing directions. The truncated Fick gimbal is illustrated in Fig. E.2a.

Kunin et al. (2007) further extended the truncated Fick gimbal system by utilising a coefficient, k, which reflects the ratio of the angles by which the first and second pitch axes rotate: $k\theta_2 = (1-k)\theta_1$. This system is referred to as the k-gimbal model (see Fig. E.2c). This model provides a mathematical and biomechanical explanation for the types of roll rotation made as a result of yaw and pitch rotation, resulting in the following relationship (Kunin et al., 2007):

$$r_x = r_z \cdot \tan(2k \cdot \tan^{-1} r_y - \tan^{-1} r_y) + r_{x0}, \tag{E.7}$$

where r_{x0} is the amount of roll offset. If $k = 0$, then the model represents a truncated Fick gimbal, if $k = 1$, then it represents a truncated Helmholtz gimbal (see Fig. E.2).

E.2.3 Apparatus

The subjective evaluation took place in the multichannel anechoic chamber 'Wilska' at Aalto University's Acoustics Lab (Espoo, Finland). A total of twenty Genelec 8331A coaxial loudspeakers were arranged in the frontal hemisphere, positioned 2.04 meters from the center (see Figure E.3). These loudspeakers were placed at four distinct elevations. On the horizontal plane, seven loudspeakers were evenly spaced with an angular separation of 30° in azimuth. On the planes with ±30° in elevation, five loudspeakers were evenly spaced with an angular separation of 45°. On the plane with an elevation of 60°, three loudspeakers were evenly spaced with an angular separation of 90°. In one of the experimental conditions, sources at azimuth $90°$ and $-90°$ were excluded to prevent unnecessary neck strain for the subjects (represented in Figure E.3 as empty circles). See Section E.2.4 for detailed information about the experimental conditions.

A 2x2 LED matrix (with a 15 mm center-to-center spacing) was installed directly in front of each loudspeaker, serving as a visual target for the search task. The target loudspeaker's LED matrix always displayed an even number of illuminated red LEDs (either two or four, randomly determined for each trial). Non-target loudspeakers displayed an odd number of illuminated LEDs (one or three, randomly determined for each loudspeaker). The LED system was controlled by an Arduino UNO WiFi Rev2, interfaced with Max 8 via serial communication. An auditory stimulus synchronized with the visual target was emitted from the target loudspeaker to assist in locating the visual target. This sound stimulus consisted of pink noise with an onset ramp of 10 ms, as described in Method 3 of ANSI/ASA S3.71(ANSI/ASA, 2019), set at an A-weighted level of 65 dB SPL measured at the subject's position. The stimulus continued until the subject responded. The room lights were dimmed to aid the visual search task."

The head and torso movements were tracked using the Motive software and six OptiTrack Prime 13W cameras at a sampling rate of 100 Hz. The origin of the world coordinates was calibrated at floor level beneath the chair in which the subjects were seated. The origin of the head tracker was calibrated for each individual separately, at the center of the head and slightly above eye level. Reflective markers were located on a hat and the torso of the subjects to track head and torso movements respectively, which were grouped as rigid bodies in Motive to analyse these two body parts separately.

To control the participants' field of view in the experiment, pinhole goggles were constructed by taping off standard protective construction goggles. This left only a small rectangular aperture through which a single target could be seen at any given time, with a maximum width of $30°$.

E.2.4 Experimental conditions

Three movement conditions were tested, each intended to incentivise different modes of movement. In all conditions subjects were seated in a fixed chair with arm rests, hence the subjects' hips were fixed. In the first condition, subjects were instructed not to move the torso. We refer to this as the no torso condition (NT). In the second condition, the subjects also could not move the torso and, in addition, they wore pinhole goggles that blocked their peripheral vision. It was checked that the subject could not see more than one loudspeaker at a time with the pinhole goggles before the experiment started. This forced them to fully rotate their heads towards a source to see it, as opposed to merely rotating to bring it into their field of view. We refer to this as the pinhole goggles condition (PG). In this PG condition, sources at azimuth $90°$ and $-90°$ were excluded, to prevent unnecessary neck strain for the subjects. In the third condition, the subjects were free to move however they preferred from a seated position, including head and torso movements. We refer to this as the free condition (F). No further instructions were given on how the head or body should be moved.

E.2.5 Subjects

During a recruitment period from May 10^{th} to May 19^{th}, 2023, seventeen caucasian subjects (three female, fourteen male) were recruited from the staff at Aalto University for the experiment. Subjects provided written informed consent, and reported normal hearing and no recent neck injuries. The age range of the subjects was between 19 and 34 years. There was no financial compensation for participating in the experiment.

E.2.6 Experimental design

The localisation task conducted here was inspired by the experimental design used in Bolia et al. (1999), Simpson et al. (2005), and Lladó et al. (2024), and was adopted by the standard "Methods for Measuring the Effect of Head-worn Devices on Directional Sound Localization in the Horizontal Plane" (ANSI/ASA, 2019). The subject held two buttons, which were used to answer how many LEDs were activated at the target. Left (blue) indicated *two* and right (red) indicated *four*.

At the start of each trial, the subjects oriented their body and head straight forward facing the LED board at ($0°$ azimuth, $0°$ elevation). A new trial was initiated by pressing the two buttons

at the same time. The LEDs were activated on all loudspeakers. Simultaneously, an acoustic stimulus was played from the target direction. The subject then searched the field in order to localise the target. Once the target was identified, the subject had to press the appropriate button, according to the correct number of LEDs activated. The sound stimulus was presented until a response button was pressed. The head tracking for each trial began when the stimulus was presented and was terminated when the subject pressed the response button.

The order of conditions and source directions was randomised independently for each subject, and only one movement condition was tested within a single block of trials. Each source direction was presented a total of 5 times per block.

This experiment was performed under the ethical approval for listening experimentation by the Research Ethics Committee of Aalto University.

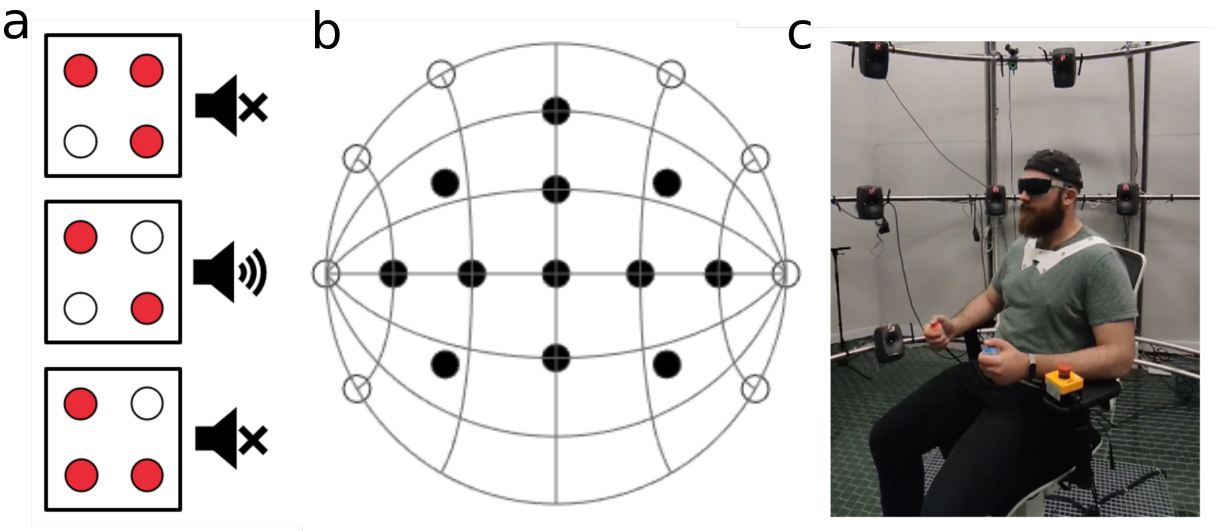

Fig. E.3 a) Example of target and two distractor LED clusters. b) Source direction distribution in the frontal hemisphere, used during the listening experiments. Empty circles were excluded in condition PG. c) Photo of experimental setup, including pinhole goggles and head and shoulder reflectors for tracking.

E.2.7 Post-processing and statistical analyses

The tracker data was rotated and translated per subject by the median starting position of that subject, so that the resulting median starting orientation and translation of the centre of the head was at zero along all world-centred axes for each subject.

Trials were omitted if they had unrealistic roll values larger than $50°$, durations that exceeded 5s or, in the case of NT and PG, had torso rotations that exceeded $5°$. This resulted in $1 + 2 + 166$ total omissions, respectively, out of 4165 trials. This means that the majority of omissions were

due to forbidden torso rotations. For the calculation of gimbal scores and k values, data points were only considered for rotations larger than 6°.

As a statistical measure of the goodness-of-fit of the twisted surfaces to the obtained data, the R^2 was reported. The statistical significance between performance of the truncated Fick gimbal model and alternative model implementations was tested using two-tailed t-tests.

E.3 Results

E.3.1 Maximum rotations

We calculated the medians, upper and lower quartiles and ranges of the minimum and maximum rotations made over all trials. These results are plotted in Fig. E.4, separated for the world-centred yaw, pitch and roll axes and for all test conditions.

Unsurprisingly, the largest rotations were made around the yaw axis. The maximum yaw rotation was the same for NT and PG, even though the maximum direction of the stimuli were not the same (see Fig. E.3). When localising sound, subjects will often rotate their heads toward the presented sounds, though they do not necessarily fully turn their heads to face the stimulus (Thurlow et al., 1967). The field of view then appeared to be responsible for a discrepancy of about 30° between the orientation of the head and the stimulus orientation.

Pitch rotation was asymmetrical: extension (negative pitch) reached larger angles than flexion (positive pitch). This is likely caused by the distribution of the stimuli, which omitted the lowest elevations (see Fig. E.3). The maximum rotations for pitch were larger in condition PG. Again this is most likely because the restricted field of view forces the subject to (almost) fully rotate towards each source. In condition NT subjects made almost no use of flexion.

The lowest rotations were seen around the roll axis. However, with maximum rotations up to 20°, the roll axis is still important to consider for potential acoustic cues, which can already be informative for rotations smaller than 10° (McLachlan et al., 2021).

There was very little difference between conditions NT and F. The maximum ranges were slightly larger for condition F, as were the lower and upper quartiles. However, the lower and upper quartiles of torso rotations were found to be smaller than 2°, indicating that very little use was made of the extra allowed freedom of movement. Thus, for the remainder of the manuscript we will focus on the results of conditions NT and PG, where no torso movements were allowed.

Together, these general results show that, without torso movement, the maximum ranges of

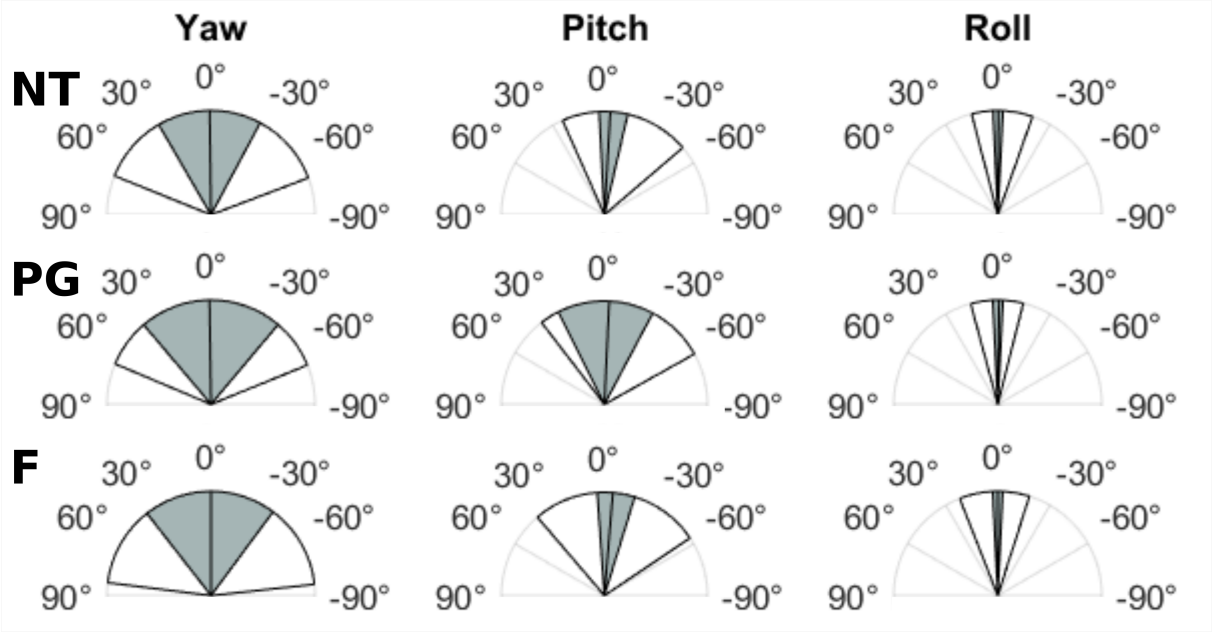

Fig. E.4 Median, upper and lower quartiles and ranges of maximum rotations made over all trials. Results are separated for the world-centred yaw, pitch and roll axes, for movement conditions NT, PG and F. Note that negative pitch is an upwards rotation.

head rotation are around $60°$ for yaw and pitch rotation, and $20°$ for roll rotation.

Note that the obtained data contained significant variances in starting positions between subjects. This is a result of individual differences in what is considered a comfortable 'resting' position for the head. For example, some subjects consistently rested their head at $10°$ or $20°$ below the calibrated zero-pitch orientation. Note that the plots in Fig. E.4 show values after the median initial orientation over all trials was set to zero.

E.3.2 Rotation trajectories

The azimuth trajectories over time of all trials performed by all subjects were plotted in Fig. E.5a. The same trajectories were plotted on a sphere in Fig. E.5b, which provides pitch rotation information instead of time information. For both conditions, the target that elicited the maximum rotation size was selected, this was source direction $(90, 60)$ for condition NT and $(45, 30)$ for condition PG. The source directions were indicated with a red cross.

The shapes of the rotation trajectories were similar between subjects. In the time domain, trajectories consistently followed a sigmoid curve, with the maximum velocity halfway through the rotation. This pattern has also been observed in other studies (Zangemeister et al., 1981b; Land, 2004). Trajectories in condition PG had more consistent end points than when full vision was available. This is because the subjects were forced to fully rotate towards the location of

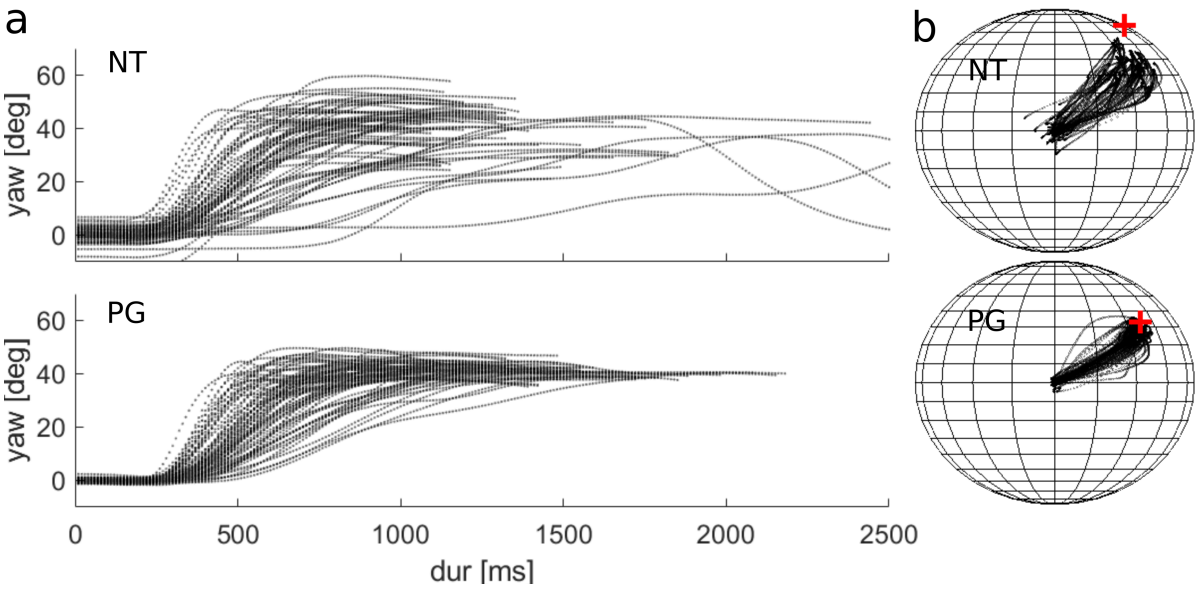

Fig. E.5 Trajectory plots of the head, for all trials towards the same target direction (NT: top row, $(90, 60)$; PG: bottom row, $(45, 30)$). Distance between dots indicates a higher velocity. a) Yaw trajectories plotted against time. b) Trajectories plotted on a sphere. The red cross indicates the source direction.

the target in order to see it, this adds a factor of consistency between subjects. In condition NT, any end point that contains the target within the field of view is acceptable. The start points were more consistent for the same reason: it was easier to align the head to a straight-forward direction, because the target at $(0, 0)$ was only visible at this exact orientation.

E.3.3 Twisted surfaces

To test the viability of Donders' law, we fitted a surface to the rotation vectors of the data of each subject with a nonlinear least-squares solver in MATLAB `lsqcurvefit`. The procedure minimised the residual error ϵ for the fitted twisted surface:

$$r_x = a_1 + a_2 r_y + a_3 r_z + a_4 r_y^2 + a_5 r_y r_z + a_6 r_z^2 + \epsilon, \tag{E.8}$$

Fig. E.6 shows the trajectories of the rotation vectors for each trial and the fitted twisted surfaces to those trajectories, for the subjects with the highest value, the median and the lowest value of a_5, respectively. Values for a_5 and R^2 were reported in the bottom right corner for each subject. The top row shows the results for condition NT ($a_5 = -0.50, -0.96, -1.80$), the bottom row shows the same results for condition PG ($a_5 = -0.41, -0.77, -1.14$).

Tab. E.1 presents the means and standard deviations of the fitted twisted surface parameters for conditions NT and PG, alongside the results from an earlier study on head rotation, where

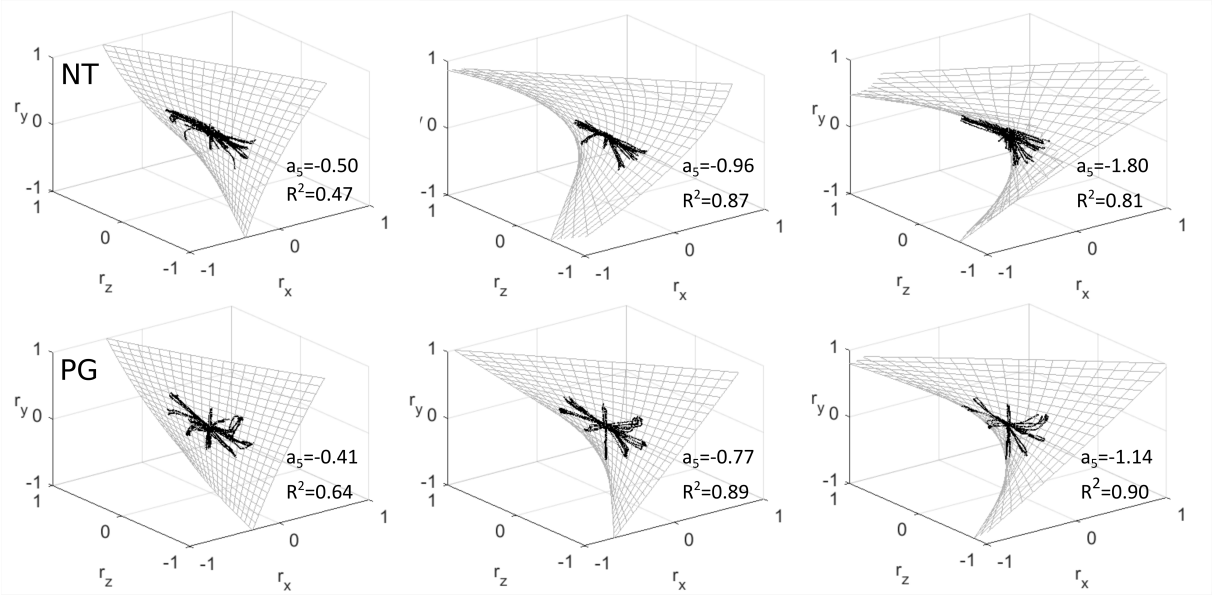

Fig. E.6 Fitted second-order surfaces for subjects with the lowest, median and highest absolute twist score (a_5), respectively. For condition NT (top row) and PG (bottom row). The thick lines are the individual trajectories of the rotation vectors per trial. The respective a_5 and R^2 values were reported in the bottom right corner of each plot.

	NT	PG	Reference (Medendorp et al., 1998)
a_1	0.000 ± 0.005	0.000 ± 0.005	0.001 ± 0.002
a_2	-0.019 ± 0.052	-0.018 ± 0.044	-0.004 ± 0.024
a_3	-0.003 ± 0.075	0.013 ± 0.051	-0.005 ± 0.064
a_4	-0.105 ± 0.215	-0.012 ± 0.056	0.072 ± 0.071
a_5	-1.062 ± 0.321	-0.794 ± 0.171	-0.732 ± 0.268
a_6	-0.017 ± 0.072	-0.004 ± 0.051	-0.026 ± 0.064

Table E.1 Mean and standard deviation of fitted values for parameter space $\mathbf{a}$ of second-order twisted surfaces. Standard deviations were computed between subjects.

subjects were instructed to rotate the head from a central position towards twelve targets on a circle in front of them, arranged like the hours on the face of a clock (Medendorp et al., 1998). The mean and standard deviation of a_5 (NT:-1.062 ± 0.321, PG:-0.794 ± 0.171) were similar to those found in the reference study (-0.732 ± 0.268).

As a metric for the goodness-of-fit of the twisted surfaces we used the coefficient of determination, R^2. In Fig. E.7 we report R^2 scores of the fitted twisted surfaces for conditions NT and PG. This was computed first with all parameters included and followed by individually excluding each of the six parameters. E.g. for R_1^2: $a_1 = 0$, for R_2^2: $a_2 = 0$.

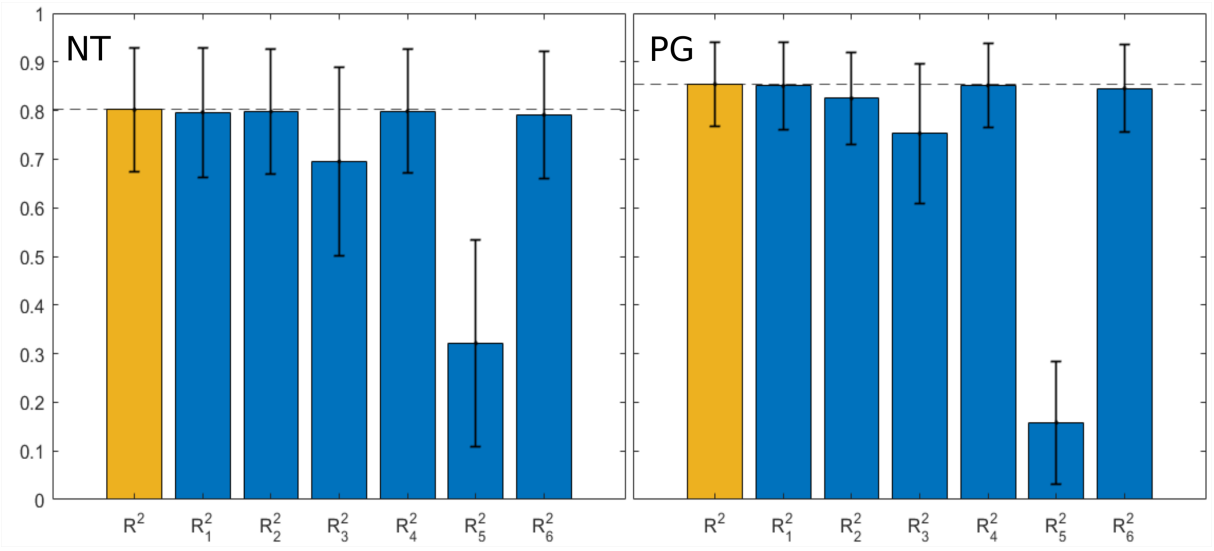

Fig. E.7 Mean and standard deviation of R^2 of second-order surfaces fitted to each subject. The yellow bar indicates the R^2 value for the twisted surface with all six available parameters. The blue bars indicate the R^2 value for the surface with one parameter excluded. E.g. for R^2_1: $a_1 = 0$, for R^2_2: $a_2 = 0$.

E.3.4 Gimbal scores

To quantify the twist in the surface fitted to the rotation trajectories, we can compute the gimbal score, G, which describes the dependence of the roll component of the rotation vector on the yaw and pitch components of the same vector as follows (Ceylan et al., 2000; Kunin et al., 2007):

$$G = \frac{r_x}{r_y r_z} \tag{E.9}$$

The roll component (r_x) of each trial was plotted against the pitch-yaw product ($r_y r_z$), for conditions NT and PG in Fig. E.8. This was done for the same subjects as in Fig. E.6.

The gimbal score was then estimated by using linear regression on the plots of r_x versus the product $r_y r_z$, as was done in an earlier study (Kunin et al., 2007). Note that the assumption of a linear relationship between r_x and $r_y r_z$ implies that only a_1 and a_5 affect the twisted surface fit.

Linear regressions were computed for extension and flexion separately, as it was found earlier that positive pitch rotations can lead to different behaviour than negative rotations (Kunin et al., 2007). The gimbal score defined in Eq. E.9 was reported for both signs of pitch for easier comparison with previous work.

The mean gimbal scores during extension were close to -1. During flexion, the slopes were somewhat steeper in condition NT and flatter in condition PG (NT extension: -0.988 ± 0.435, NT flexion: -1.217 ± 1.474, PG extension: -0.928 ± 0.358, PG flexion: -0.684 ± 0.254).

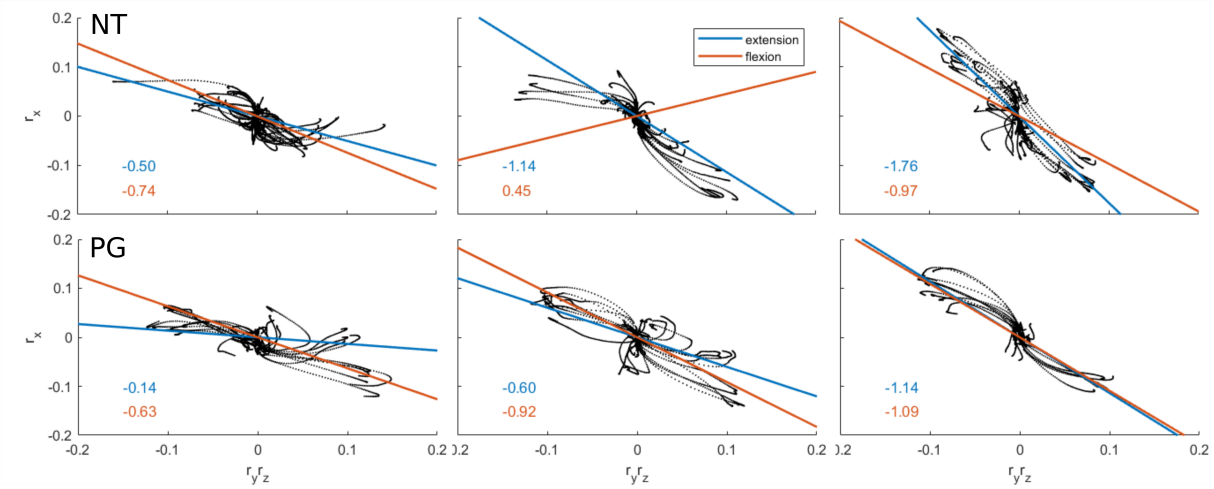

Fig. E.8 Roll component of rotation vector (r_x) plotted against the pitch-yaw product ($r_y r_z$), for conditions NT (top row) and PG (bottom row) and for subjects with the lowest, median and highest absolute twist score (a_5), respectively. Linear regressions were computed for extension and flexion separately. The numbers in the bottom left corner of each plot indicate the gimbal score G from Eq. E.9.

Note that the standard deviation in NT flexion is very high, the reason for this is discussed in Section E.4.2.

E.3.5 k-gimbal model

We tested five different implementations of the k-gimbal model (Eq. E.7) for prediction of the roll component of the rotation vectors (r_x). The first implementation assumed zero roll ($k = 0.5$). The second assumed a truncated Fick gimbal ($k = 0$). Third, we fitted a single k-value to the trials combined over all subjects. This led to $k = -0.04$ for NT and $k = 0.10$ for PG. Fourth, a unique k value was fitted for each subject. Finally, a separate k was computed for extension and flexion, leading to two values of k per subject. The k values were obtained by minimising r_{x0} in Eq. E.7 using the nonlinear simplex method (`fminsearch` in MATLAB).

In Fig. E.9, we plot, per model approach, the standard deviation of the error between the true and estimated r_x in degrees, also referred to as the torsional thickness σ (Medendorp et al., 1998), defined as:

$$\sigma = std(2 * atan(\hat{r}_x - r_x)), \tag{E.10}$$

where $\hat{r}_x$ is the value of r_x predicted by the fitted twisted surface in Eq. E.6.

Statistical significance between the Fick gimbal (k_0) and the other implementations of the k-gimbal model was tested using two-tailed t-tests. Statistical significance was only found for the zero-roll model ($k_{0.5}$) for both conditions NT and PG.

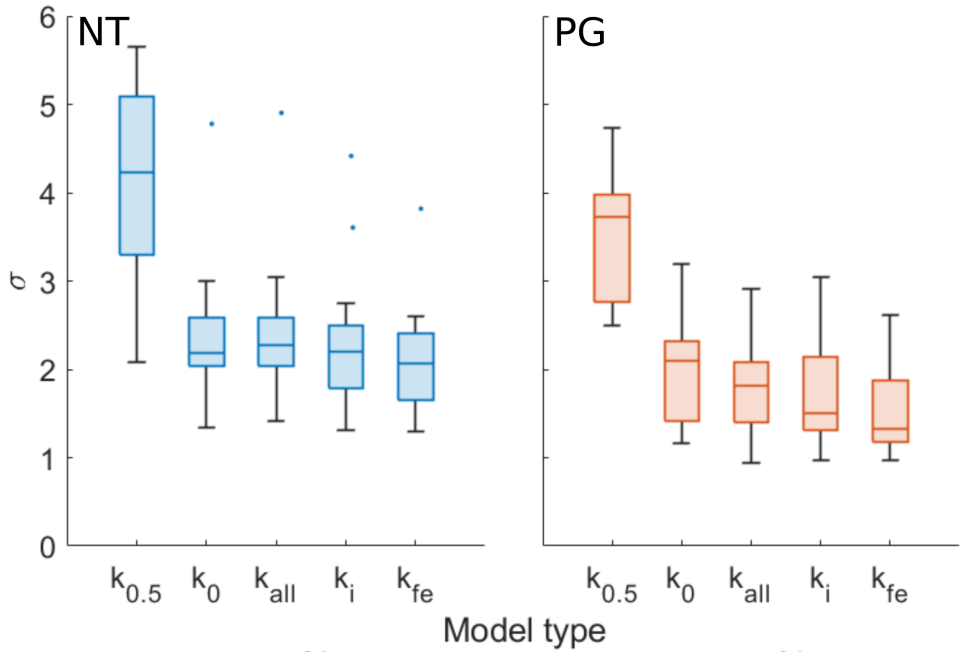

Fig. E.9 Torsional thickness σ (in degrees) of the k-gimbal model, for 5 different approaches of setting the k value. Left window: NT, right window: PG. $k_{0.5}$: Zero-roll, k_0: Fick gimbal, k_{all}: single value fitted to all subjects, k_i: k value fitted individually for each subject, k_{fe}: two k values fitted for extension and flexion, separately for each individual. Each boxplot shows the statistics of all 17 subjects (median, first and third quartiles, minima and maxima, outliers).

E.4 Discussion

E.4.1 Rotations adhere to Donders' law

The R^2 values in Fig. E.7 support the notion that head movements generally adhered to Donders' law; in condition NT the mean R^2 was 0.80, in condition PG it was 0.85. Furthermore, from the mean values in Tab. E.1 and the R^2 values in Fig. E.7, we can conclude that the majority of the roll data was explained by the twist score, a_5. When a_1, a_2, a_4 or a_6 was omitted from the equation, there was almost no effect on R^2. The only other factor that appeared to have a small effect was a_3, which suggests that there was a small linear relationship between roll and yaw, regardless of the pitch position. This may, however, be a result of the resting head position of some of the subjects. If the zero-pitch orientation of a subject does not properly align the anatomical yaw axis of rotation with the world z-axis, then any yaw rotation will inherently have a roll component, resulting again from an $r_y r_z$ interaction. The median thickness scores of 2° in Fig. E.9 are in accordance with the scores found in Ceylan et al. (2000). This suggests that adherence to Donders' law was not influenced by the additional available auditory information in this study. A standard deviation of 2° is small, but may still be significant, given that the upper and lower quadrants of roll were very small (see Fig. E.4). For general modelling purposes, however, Donders' law appears to be an appropriate assumption.

E.4.2 Roll is a linear function of pitch and yaw

The high dependence on the twist score, a_5, shows that roll was almost strictly a function of the pitch/yaw product. Looking at Fig. E.8, we see an almost linear relationship between r_x and $r_y r_z$. We also see that the fitted linear regressions nearly pass through the origin, indicating that the relationship can be modelled as a slope, omitting the offset.

The majority of the gimbal scores (i.e., the slopes of the linear regressions) amongst all subject were found to be negative. This supports the characterisation of head rotations as a Fick gimbal as opposed to a Helmholtz gimbal (Glenn and Vilis, 1992). There were some exceptions for flexion in condition NT, though the linear regressions in these instances are not very meaningful due to the general lack of flexion performed, as could also be seen in Fig. E.4. The sign and magnitudes of the gimbal scores are in general accord with previous studies, which reported values between 0 and -2 (Ceylan et al., 2000; Kunin et al., 2007). Gimbal scores did differ substantially between subjects, which means that there were differences in the amount of roll that they deployed during head rotations.

Interestingly, only a few subjects showed a large difference in slope of the linear regression between extension and flexion, given that the magnitude of flexion was large enough to correctly fit the regression (see Fig. Roll component of rotation (r_x) plotted against the pitch-yaw product ($r_y r_z$), for each subject in condition NT. Linear regressions were computed for extension and flexion separately. The numbers in the bottom left corner of each plot indicate the gimbal score G from Eq. E.9.). For those subjects, it does seem necessary to differentiate between two different gimbal models depending on the sign of pitch.

E.4.3 Rotations can be approximated by a truncated Fick gimbal

Different models were tested to include more individualised data. However, Fig. E.9 presents the somewhat surprising result that none of the individualisation approaches of the k-gimbal model significantly improved the model performance. No statistical significance between the different approaches of obtaining the k value means that, while individual differences in the twist score do exist, the truncated Fick gimbal (i.e. $k = 0$) serves as a good approximation for the roll component in the head rotations of the average subject. Thus, the differences in gimbal scores between subjects were not large enough to make full individualisation necessary. This is also confirmed by the k values found when fitting over all subjects, which were close to 0 ($k = -0.04$ for NT, $k = 0.10$ for PG). Note that there are a number of subjects where using an individualised k value does improve modelling performance, but that the majority does not appear to need this.

E.4.4 Pinhole goggles decrease variability in rotation

There were some observable differences between the NT and PG conditions. First of all, Fig. E.6 and E.7 show that the a_5 values were lower in condition PG, whereas the R^2 values were higher. This means that pinhole goggles decreased the amount of roll that subjects deployed, while they also made the head movements adhere better to Donders' law. Similarly, it was found earlier that pinhole goggles constrain the head rotation vectors to a plane (i.e., Listing's law) (Ceylan et al., 2000). However, the results still could not be well predicted by a planar surface (concluded from the very low R^2 score when a_5 was omitted in Fig. E.7). In Tab. E.1 we also see that the standard deviation of all fitted parameters decreased substantially in condition PG, which means that the pinhole goggles caused the inter-subject differences to decrease. One explanation could be that most of the head movement variability between subjects is caused by the extra degrees of freedom from the eyes. Restricting this would then make the subjects behave more consistently. However, the standard deviations in Medendorp et al. (1998) were similarly small, while the subjects did not have any restricted vision here. Alternatively, the lower standard deviation may be a result of more restricted start and end points, as condition PG forced subjects to fully rotate their heads towards the visual targets (see Fig. E.5). In condition NT, any end point that contains the target within the field of view is acceptable. This appears to be a reasonable explanation, as the subjects in Medendorp et al. (1998) were required to accurately point their head towards each target before moving to the next, which makes the task performed more similar to condition PG. This is further supported by the twist scores of Medendorp et al. (1998), which are most similar to those of condition PG. Interestingly, the R^2 in condition PG is also lower than in condition NT when a_5 is omitted. This means that, with pinhole goggles, the twist score describes an even larger percentage of the data. Finally, Fig. E.9 shows that thickness scores were consistently lower in condition PG.

E.4.5 Auditory cues keep head rotation behaviour consistent.

Due to the restricted vision of condition PG, it was expected that the subjects would depend more on auditory cues to localise the correct target, which in turn might lead to different head movement strategies. From the data collected in conditions NT and PG, this did not appear to be the case, as the trajectories (as seen in Fig. E.5) did not differ significantly except for the lower deviation in start and end points, as explained in the section above. From this we can conclude that the restricted vision did not fundamentally alter rotation behaviour, likely because the available auditory cues kept head movement consistent.

E.4.6 Limitations

As the aim of the present study was head rotation behaviour, we did not extensively investigate head translation. The absolute maximum translations for the x, y and z-axes were substantial. Condition NT: 15.4 cm, 9.5 cm and 4.9 cm; condition PG: 19.3 cm, 10.4 cm and 8.3 cm; condition F: 20.8 cm, 14.6 cm and 5.3 cm. It is difficult to infer how much of this translation was due to the definition of the origin of the coordinate system, which was defined slightly above eye level and not on the exact axis of rotation of the head.

A second limitation is that no condition was tested with solely visual information. In the future, it would be interesting to compare rotation behaviour to the present study when no auditory information is available.

E.5 Conclusion

In this study, the head movements of seventeen subjects were tracked during an auditory-guided visual search tasks. By fitting a second-order twisted surface to the rotation vectors over all trials, it was shown that Donders' law was generally met, i.e., humans used only two out of three available rotational degrees of freedom. There were inter-subject differences in the gimbal score, i.e., the rotation angle around r_x as a function of rotation angles around r_y and r_z. Surprisingly, these differences did not lead to a statistically significant improvement when the k-gimbal model, which was introduced in previous work (Kunin et al., 2007), was tested with individualised k-values. This leads to the conclusion that the measured head rotations could be described by the k-gimbal model with $k = 0$. This is equivalent to a truncated Fick gimbal, which consists of a pitch axis nested within a yaw axis. Finally, subjects were more alike in their movements when wearing pinhole goggles. This was likely because the restricted vision enforced clear start and end points of each rotation, resulting in a new set of more consistent—though potentially less natural—head rotations amongst subjects.

Taken together, subjects generally used similar search strategies following Donders' law to identify an audiovisual target. With restricted vision and, therefore, an increased dependency on auditory cues, the executed rotations did not change fundamentally. From this we can conclude that visual conditions had little influence on movement behaviour because of the available acoustic cues.

Future research might consider the head movements of hearing impaired individuals, as they have been found to show more unpredictable movement behaviour Brimijoin et al. (2010); Hendrikse et al. (2022). A second avenue for further investigation involves the neural control mechanisms responsible for Donders' law. The fact that head movements follow the same

motor rules, whether auditory information is available or not, can provide insights into the neural substrates of the control of head movements Crawford et al. (2003) and of the processing of head-motion related sensory feedback McLachlan et al. (2021).

Acknowledgements

The authors express their gratitude to Ville Pulkki, who gave approval for the research visit at Aalto University.

E.6 Supplementary figures

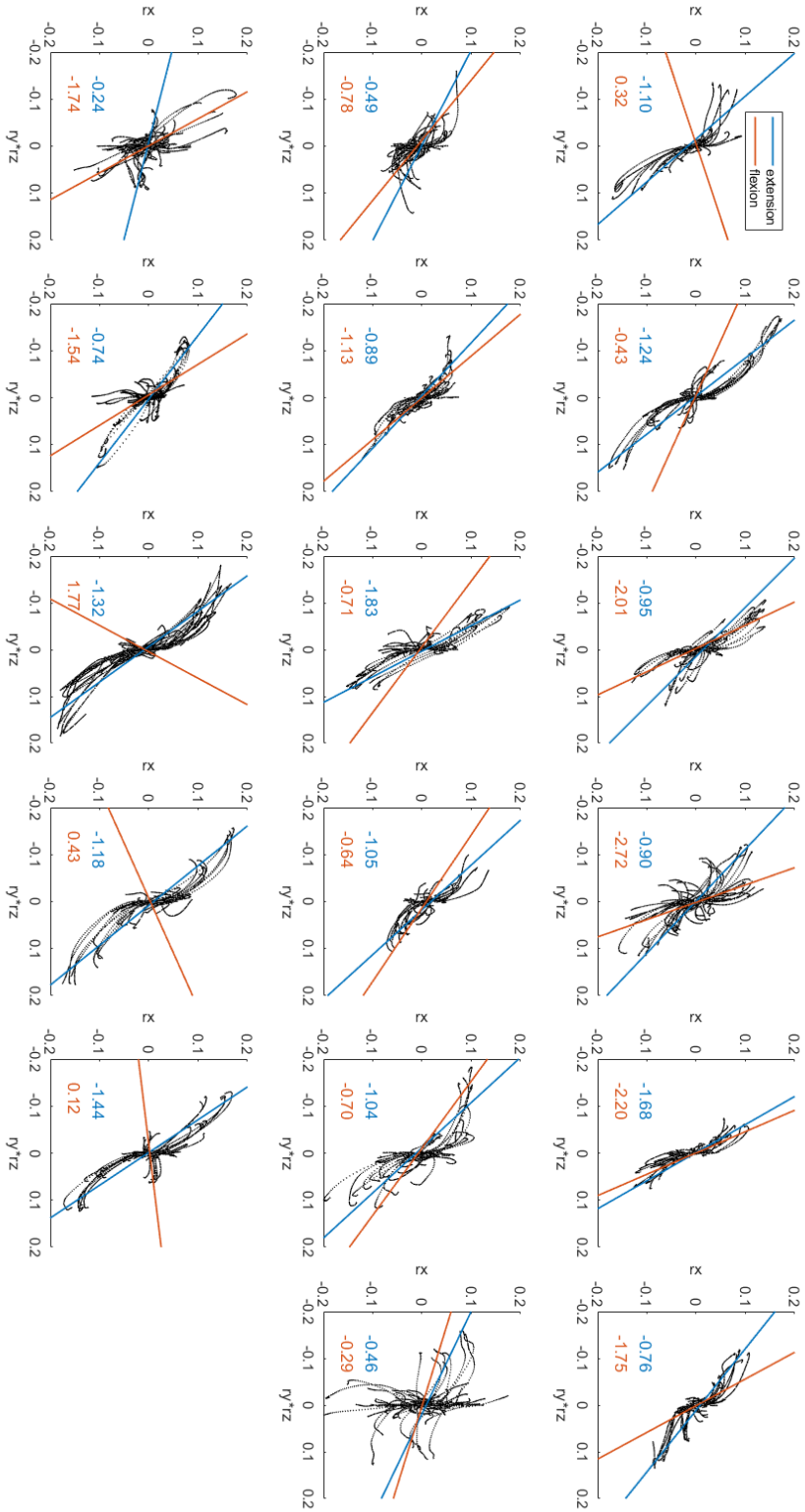

Fig. SE.1 Roll component of rotation (r_x) plotted against the pitch-yaw product ($r_y r_z$), for each subject in condition NT. Linear regressions were computed for extension and flexion separately. The numbers in the bottom left corner of each plot indicate the gimbal score G from Eq. E.9.

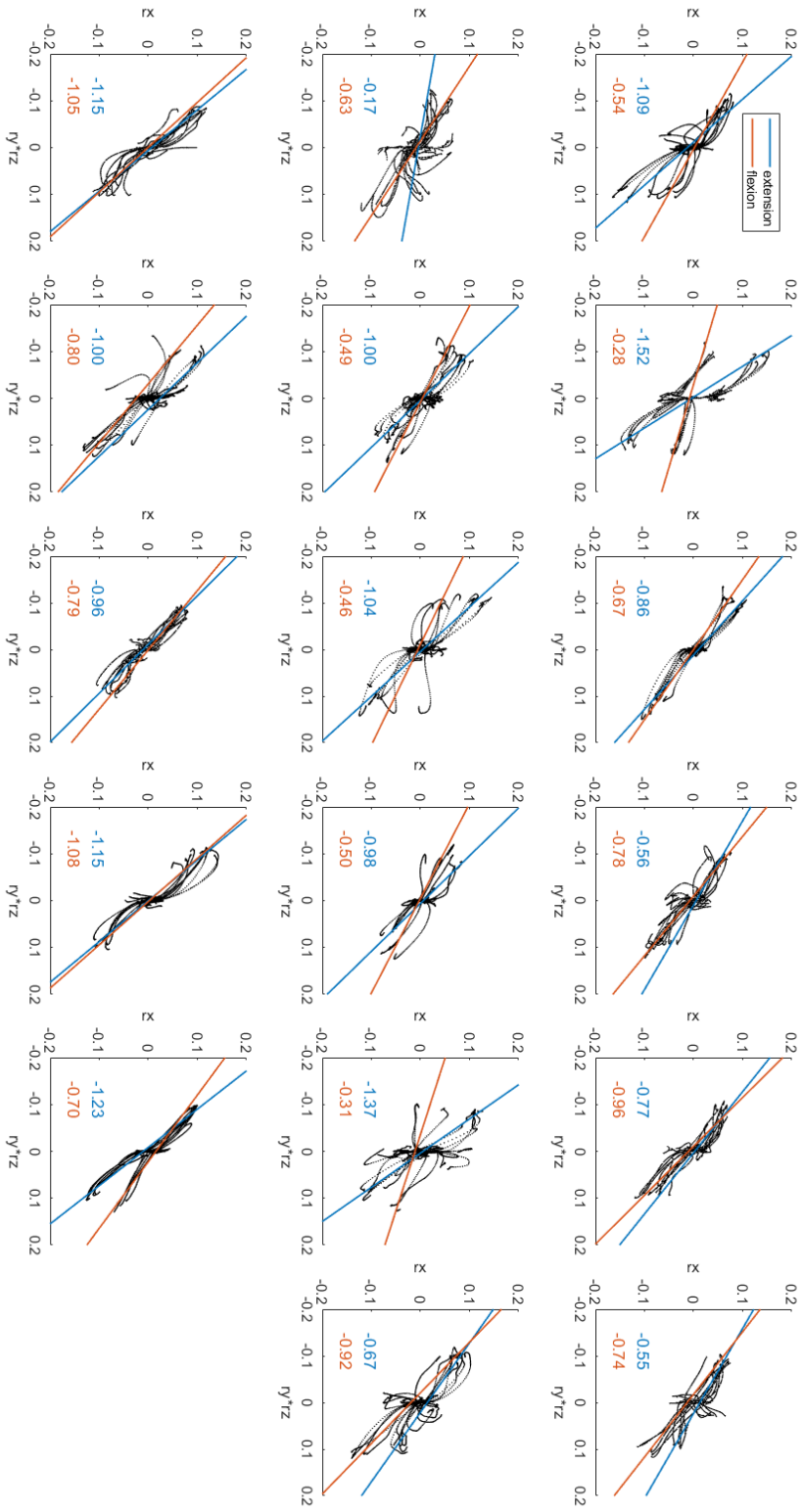

Fig. SE.2 Roll component of rotation (r_x) plotted against the pitch-yaw product ($r_y r_z$), for each subject in condition PG. Linear regressions were computed for extension and flexion separately. The numbers in the bottom left corner of each plot indicate the gimbal score G from Eq. E.9.

REFERENCES

P. Aarabi. The fusion of distributed microphone arrays for sound localization. *EURASIP Journal on Advances in Signal Processing*, 2003(4):1–10, 2003.

D. Alais and D. Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current biology*, 14(3):257–262, 2004.

Y. A. Al'tman, I. Kudryavtseva, and E. Radionova. The pattern of response of the inferior colliculus of the cat during the movement of a sound source. *Neuroscience and behavioral physiology*, 15(4):318–324, 1985.

D. E. Angelaki, Y. Gu, and G. C. DeAngelis. Multisensory integration: psychophysics, neurophysiology, and computation. *Current opinion in neurobiology*, 19(4):452–458, 2009.

ANSI/ASA. Methods for measuring the effect of head-worn devices on directional sound localization in the horizontal plane. *ANSI/ASA S3.71, 2019 Edition*, 2019.

T. Ashby, T. Brookes, and R. Mason. Towards a head-movement-aware spatial localisation model: Elevation. In *21st International Congress on Sound and Vibration 2014, ICSV 2014*, volume 4, pages 2808–2815, 2014.

P. Avan, F. Giraudet, and B. Büki. Importance of binaural hearing. *Audiology and Neurotology*, 20(Suppl. 1):3–6, 2015.

L. Bahl, J. Cocke, F. Jelinek, and J. Raviv. Optimal decoding of linear codes for minimizing symbol error rate (corresp.). *IEEE Transactions on information theory*, 20(2):284–287, 1974.

H. Bahu, T. Carpentier, M. Noisternig, and O. Warusfel. Comparison of different egocentric pointing methods for 3d sound localization experiments. *Acta acustica united with Acustica*, 102(1):107–118, 2016.

D. Barber, A. T. Cemgil, and S. Chiappa. *Bayesian time series models*. Cambridge University Press, 2011.

M. Barnett-Cowan and L. R. Harris. Temporal processing of active and passive head movement. *Experimental brain research*, 214(1):27–35, 2011.

R. Barumerli, P. Majdak, J. Reijniers, R. Baumgartner, M. Geronazzo, and F. Avanzini. Predicting directional sound-localization of human listeners in both horizontal and vertical dimensions. In *Audio Engineering Society Convention 148*. Audio Engineering Society, 2020.

R. Barumerli, P. Majdak, M. Geronazzo, D. Meijer, F. Avanzini, and R. Baumgartner. A bayesian model for human directional localization of broadband static sound sources. *Acta Acustica*, 7:12, 2023.

R. Bassett and J. Deride. Maximum a posteriori estimators as a limit of bayes estimators. *Mathematical Programming*, 174:129–144, 2019.

P. W. Battaglia, R. A. Jacobs, and R. N. Aslin. Bayesian integration of visual and auditory signals for spatial localization. *Josa a*, 20(7):1391–1397, 2003a.

P. W. Battaglia, R. A. Jacobs, and R. N. Aslin. Bayesian integration of visual and auditory signals for spatial localization. *J. Opt. Soc. Am. A*, 20(7):1391–1397, Jul 2003b.

C. Baumann, C. Rogers, and F. Massen. Dynamic binaural sound localization based on variations of interaural time delays and system rotations. *The Journal of the Acoustical Society of America*, 138(2):635–650, 2015.

R. Baumgartner, P. Majdak, and B. Laback. Modeling sound-source localization in sagittal planes for human listeners. *The Journal of the Acoustical Society of America*, 136(2):791–802, 2014.

R. Baumgartner, D. K. Reed, B. Tóth, V. Best, P. Majdak, H. S. Colburn, and B. Shinn-Cunningham. Asymmetries in behavioral and neural responses to spectral cues demonstrate the generality of auditory looming bias. *Proceedings of the National Academy of Sciences*, 114 (36):9743–9748, 2017.

A. Becher, J. Angerer, and T. Grauschopf. Novel approach to measure motion-to-photon and mouth-to-ear latency in distributed virtual reality systems. *arXiv preprint arXiv:1809.06320*, 2018.

D. R. Begault, E. M. Wenzel, and M. R. Anderson. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *Journal of the Audio Engineering Society*, 49(10):904–916, 2001.

U. Beierholm, L. Shams, K. Koerding, et al. Comparing bayesian models for multisensory cue combination without mandatory integration. *Advances in neural information processing systems*, 20, 2007.

U. R. Beierholm, S. R. Quartz, and L. Shams. Bayesian priors are encoded independently from likelihoods in human multisensory perception. *Journal of vision*, 9(5):23–23, 2009.

B. Bernschütz. A spherical far field hrir/hrtf compilation of the neumann ku 100. In *Proceedings of the 40th Italian (AIA) annual conference on acoustics and the 39th German annual conference on acoustics (DAGA) conference on acoustics*, page 29. AIA/DAGA, 2013.

L. R. Bernstein, C. Trahiotis, M. A. Akeroyd, and K. Hartung. Sensitivity to brief changes of interaural time and interaural intensity. *The Journal of the Acoustical Society of America*, 109 (4):1604–1615, 2001.

V. Best, D. Brungart, S. Carlile, C. Jin, E. Macpherson, R. Martin, K. McAnally, A. Sabin, and B. Simpson. A meta-analysis of localization errors made in the anechoic free field. In *Principles and applications of spatial hearing*, pages 14–23. World Scientific, 2011.

J. Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

J. Blauert and J. Braasch, editors. *The Technology of Binaural Understanding*. Modern Acoustics and Signal Processing. Springer International Publishing, 2020. ISBN 978-3-030-00385-2.

R. S. Bolia, W. R. D'Angelo, and R. L. McKinley. Aurally aided visual search in three-dimensional space. *Human factors*, 41(4):664–669, 1999.

A. Borrego, J. Latorre, M. Alcañiz, and R. Llorens. Comparison of oculus rift and htc vive: feasibility for virtual reality-based exploration, navigation, exergaming, and rehabilitation. *Games for health journal*, 7(3):151–156, 2018.

L. Boucher, A. Lee, Y. E. Cohen, and H. C. Hughes. Ocular tracking as a measure of auditory motion perception. *Journal of Physiology-Paris*, 98(1-3):235–248, 2004.

J. Braasch. Localization in the presence of a distracter and reverberation in the frontal horizontal plane: Ii. model algorithms. *Acta Acustica united with Acustica*, 88(6):956–969, 2002.

J. Breebaart, S. Van De Par, and A. Kohlrausch. Binaural processing model based on contralateral inhibition. i. model structure. *The Journal of the Acoustical Society of America*, 110(2): 1074–1088, 2001.

A. S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

W. O. Brimijoin. Angle-dependent distortions in the perceptual topology of acoustic space. *Trends in Hearing*, 22:2331216518775568, 2018.

W. O. Brimijoin and M. A. Akeroyd. The role of head movements and signal spectrum in an auditory front/back illusion. *i-Perception*, 3(3):179–182, 2012.

W. O. Brimijoin and M. A. Akeroyd. The moving minimum audible angle is smaller during self motion than during source motion. *Frontiers in neuroscience*, 8:273, 2014.

W. O. Brimijoin, D. McShefferty, and M. A. Akeroyd. Auditory and visual orienting responses in listeners with and without hearing-impairment. *The Journal of the Acoustical Society of America*, 127(6):3678–3688, 2010.

W. O. Brimijoin, A. W. Boyd, and M. A. Akeroyd. The contribution of head movement to the externalization and internalization of sounds. *PloS one*, 8(12):e83068, 2013.

A. D. Brown and D. J. Tollin. Slow temporal integration enables robust neural coding and perception of a cue to sound source location. *Journal of Neuroscience*, 36(38):9908–9921, 2016.

H. H. Bülthoff and H. A. Mallot. Integration of stereo, shading and texture. In *11th European Conference on Visual Perception (ECVP 1988)*, pages 119–146. Wiley, 1990.

J. Burger. Front-back discrimination of the hearing systems. *Acta Acustica united with Acustica*, 8(5):301–302, 1958.

J. A. G.-U. Calvo, M. M. van Wanrooij, and A. J. Van Opstal. Adaptive response behavior in the pursuit of unpredictably moving sounds. *Eneuro*, 8(3), 2021.

S. Carlile and V. Best. Discrimination of sound source velocity in human listeners. *The Journal of the Acoustical Society of America*, 111(2):1026–1035, 2002.

S. Carlile and J. Leung. The perception of auditory motion. *Trends in hearing*, 20: 2331216516644254, 2016.

S. Carlile, P. Leong, and S. Hyams. The nature and distribution of errors in sound localization by human listeners. *Hearing research*, 114(1-2):179–196, 1997.

S. Carlile, S. Delaney, and A. Corderoy. The localisation of spectrally restricted sounds by human listeners. *Hearing research*, 128(1-2):175–189, 1999.

S. Carlile, K. Balachandar, and H. Kelly. Accommodating to new ears: the effects of sensory and sensory-motor feedback. *The Journal of the Acoustical Society of America*, 135(4):2002–2011, 2014.

M. Ceylan, D. Henriques, D. Tweed, and J. Crawford. Task-dependent constraints in motor control: pinhole goggles make the head move like an eye. *Journal of Neuroscience*, 20(7): 2719–2730, 2000.

Y. E. Cohen and E. I. Knudsen. Maps versus clusters: different representations of auditory space in the midbrain and forebrain. *Trends in neurosciences*, 22(3):128–135, 1999.

M. Colombo and P. Seriès. Bayes in the brain—on bayesian modelling in neuroscience. *The British journal for the philosophy of science*, 2012.

M. Cooke, Y.-C. Lu, Y. Lu, and R. Horaud. Active hearing, active speaking. In *ISAAR 2007-International Symposium on Auditory and Audiological Research*, pages 33–46, 2007.

J. Cooper, S. Carlile, and D. Alais. Distortions of auditory space during rapid head turns. *Experimental brain research*, 191(2):209–219, 2008.

J. F. Corso. Age and sex differences in pure-tone thresholds. *The journal of the Acoustical Society of America*, 31(4):498–507, 1959.

W. Cox and B. J. Fischer. Optimal prediction of moving sound source direction in the owl. *PLoS computational biology*, 11(7):e1004360, 2015.

J. Crawford, J. Martinez-Trujillo, and E. Klier. Neural control of three-dimensional eye and head movements. *Current opinion in neurobiology*, 13(6):655–662, 2003.

E. De Boer. Auditory time constants: a paradox? In *Time Resolution in Auditory Systems: Proceedings of the 11th Danavox Symposium on Hearing Gamle Avernæs, Denmark, August 28–31, 1984*, pages 141–158. Springer, 1985.

A. Diaz-Artiles and F. Karmali. Vestibular precision at the level of perception, eye movements, posture, and neurons. *Neuroscience*, 468:282–320, 2021.

F. Donders. Beitrag zur lehre von den bewegungen des menschlichen auges. *Hollandische Beitrenge zu den anatomischen und physiologichen Wissenschaffen*, 1:105–145, 1847.

R. Ege, A. J. Van Opstal, and M. M. Van Wanrooij. Accuracy-precision trade-off in human sound localisation. *Scientific reports*, 8(1):1–12, 2018.

R. Ege, A. J. Van Opstal, and M. M. Van Wanrooij. Perceived target range shapes human sound-localization behavior. *eneuro*, 6(2), 2019.

M. O. Ernst. A bayesian view on multimodal cue integration. *Human body perception from the inside out*, 131:105–131, 2006.

M. O. Ernst. Learning to integrate arbitrary signals from vision and touch. *Journal of vision*, 7 (5):7–7, 2007.

M. O. Ernst and M. S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

M. O. Ernst and H. H. Bülthoff. Merging the senses into a robust percept. *Trends in cognitive sciences*, 8(4):162–169, 2004.

M. A. R. Ferreira and H. Lee. *Multiscale Modeling: A Bayesian Perspective*. Springer Series in Statistics. Springer-Verlag, New York, 2007. ISBN 978-0-387-70897-3.

B. J. Fischer and J. L. Peña. Owl's behavior and neural representation predicted by bayesian inference. *Nature neuroscience*, 14(8):1061–1066, 2011.

T. Fischer, M. Caversaccio, and W. Wimmer. A front-back confusion metric in horizontal sound localization: The fbc score. In *ACM Symposium on Applied Perception 2020*, pages 1–5, 2020.

E. Fosler-Lussier. Markov models and hidden markov models: A brief tutorial. *International Computer Science Institute*, 1998.

T. C. Freeman, J. Leung, E. Wufong, E. Orchard-Mills, S. Carlile, and D. Alais. Discrimination contours for moving sounds reveal duration and distance cues dominate auditory speed perception. *PloS one*, 9(7):e102864, 2014.

T. C. Freeman, J. F. Culling, M. A. Akeroyd, and W. O. Brimijoin. Auditory compensation for head rotation is incomplete. *Journal of experimental psychology: human perception and performance*, 43(2):371, 2017.

J. H. Fuller. Head movement propensity. *Experimental Brain Research*, 92(1):152–164, 1992.

S. E. Garcia, P. R. Jones, G. S. Rubin, and M. Nardini. Auditory localisation biases increase with sensory uncertainty. *Scientific Reports*, 7(1):40567, 2017.

D. Genzel, U. Firzlaff, L. Wiegrebe, and P. R. MacNeilage. Dependence of auditory spatial updating on vestibular, proprioceptive, and efference copy signals. *Journal of neurophysiology*, 116(2):765–775, 2016.

D. Genzel, M. Schutte, W. O. Brimijoin, P. R. MacNeilage, and L. Wiegrebe. Psychophysical evidence for auditory motion parallax. *Proceedings of the National Academy of Sciences*, 115 (16):4264–4269, 2018.

G. M. Gerken, V. K. Bhat, and M. Hutchison-Clutter. Auditory temporal integration and the power function model. *The Journal of the Acoustical Society of America*, 88(2):767–778, 1990.

M. Gerven, B. Cseke, R. Oostenveld, and T. Heskes. Bayesian source localization with the multivariate laplace prior. *Advances in neural information processing systems*, 22, 2009.

B. K. Ghosh and I. B. Wijayasinghe. Dynamics of human head and eye rotations under donders' constraint. *IEEE transactions on automatic control*, 57(10):2478–2489, 2012.

C. Giguère and S. M. Abel. Sound localization: Effects of reverberation time, speaker array, stimulus frequency, and stimulus rise/decay. *The Journal of the Acoustical Society of America*, 94(2):769–776, 1993.

B. Glenn and T. Vilis. Violations of listing's law after large eye and head gaze shifts. *Journal of Neurophysiology*, 68(1):309–318, 1992.

H. Goossens and A. Van Opstal. Influence of head position on the spatial representation of acoustic targets. *Journal of neurophysiology*, 81(6):2720–2736, 1999.

H. H. Goossens and A. J. Van Opstal. Human eye-head coordination in two dimensions under different sensorimotor conditions. *Experimental Brain Research*, 114(3):542–560, 1997.

J. A. Grange and J. F. Culling. The benefit of head orientation to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 139(2):703–712, 2016.

Y. Gu, D. E. Angelaki, and G. C. DeAngelis. Neural correlates of multisensory cue integration in macaque mstd. *Nature neuroscience*, 11(10):1201–1210, 2008.

R. Gupta, R. Ranjan, J. He, and G. Woon-Seng. Investigation of effect of vr/ar headgear on head related transfer functions for natural listening. In *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2018.

U. Hadar, T. J. Steiner, E. Grant, and F. C. Rose. Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2(1-2):35–46, 1983.

U. Hadar, T. J. Steiner, and F. Clifford Rose. Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9:214–228, 1985.

D. A. Hambrook, M. Ilievski, M. Mosadeghzad, and M. Tata. A bayesian computational basis for auditory selective attention using head rotation and the interaural time-difference cue. *PloS one*, 12(10):e0186104, 2017.

C. M. Harris and D. M. Wolpert. Signal-dependent noise determines motor planning. *Nature*, 394(6695):780–784, 1998.

T. Haslwanter. Mathematics of three-dimensional eye rotations. *Vision research*, 35(12):1727–1739, 1995.

J. Hebrank and D. Wright. Spectral cues used in the localization of sound sources on the median plane. *The Journal of the Acoustical Society of America*, 56(6):1829–1834, 1974.

H. B. Helbig and M. O. Ernst. Optimal integration of shape information from vision and touch. *Experimental brain research*, 179:595–606, 2007.

M. M. Hendrikse, T. Eichler, V. Hohmann, and G. Grimm. Self-motion with hearing impairment and (directional) hearing aids. *Trends in Hearing*, 26:23312165221078707, 2022.

T. Hirahara, D. Kojima, D. Morikawa, and P. Mokhtari. The effect of head rotation on monaural sound-image localization in the horizontal plane. *Applied Acoustics*, 178:108008, 2021.

P. M. Hofman and A. J. Van Opstal. Spectro-temporal factors in two-dimensional human sound localization. *The Journal of the Acoustical Society of America*, 103(5):2634–2648, 1998.

P. M. Hofman and A. J. Van Opstal. Bayesian reconstruction of sound localization cues from responses to random spectra. *Biological cybernetics*, 86(4):305–316, 2002.

A. Honda, K. Ohba, Y. Iwaya, and Y. Suzuki. Detection of sound image movement during horizontal head rotation. *i-Perception*, 7(5):2041669516669614, 2016.

T. E. Hudson, L. T. Maloney, and M. S. Landy. Movement planning with probabilistic target information. *Journal of neurophysiology*, 98(5):3034–3046, 2007.

N. J. Ingham, H. C. Hart, and D. McAlpine. Spatial receptive fields of inferior colliculus neurons to auditory apparent motion in free field. *Journal of Neurophysiology*, 85(1):23–33, 2001.

Y. Iwaya, Y. Suzuki, and D. Kimura. Effects of head movement on front-back error in sound localization. *Acoustical science and technology*, 24(5):322–324, 2003.

R. A. Jacobs. Optimal integration of texture and motion cues to depth. *Vision research*, 39(21): 3621–3629, 1999.

C. Jenny, P. Majdak, and C. Reuter. Sofalizer for unity 1.x, 2022. URL `https://github.com/sofacoustics/SOFAlizer-for-Unity`.

W. Jesteadt, C. C. Wier, and D. M. Green. Intensity discrimination as a function of frequency and sensation level. *The Journal of the acoustical society of America*, 61(1):169–177, 1977.

J. Jiang, B. Xie, H. Mai, L. Liu, K. Yi, and C. Zhang. The role of dynamic cue in auditory vertical localisation. *Applied Acoustics*, 146:398–408, 2019.

H.-O. Karnath, D. Sievering, and M. Fetter. The interactive contribution of neck muscle proprioception and vestibular stimulation to subjective "straight ahead" orientation in man. *Experimental Brain Research*, 101(1):140–146, 1994.

M. Kato, H. Uematsu, M. Kashino, and T. Hirahara. The effect of head motion on the accuracy of sound localization. *Acoustical science and technology*, 24(5):315–317, 2003.

C. Kayser and L. Shams. Multisensory causal inference in the brain. *PLoS biology*, 13(2): e1002075, 2015.

D. Kersten, P. Mamassian, and A. Yuille. Object perception as bayesian inference. *Annu. Rev. Psychol.*, 55:271–304, 2004.

C. Kim, R. Mason, and T. Brookes. Head movements made by listeners in experimental and real-life listening activities. *Journal of the Audio Engineering Society*, 61:425–438, 2013a.

J. Kim, M. Barnett-Cowan, and E. A. Macpherson. Integration of auditory input with vestibular and neck proprioceptive information in the interpretation of dynamic sound localization cues. In *Proceedings of meetings on acoustics ICA2013*, volume 19, page 050142. Acoustical Society of America, 2013b.

R. B. King and S. R. Oldfield. The impact of signal bandwidth on auditory localization: Implications for the design of three-dimensional audio displays. *Human factors*, 39(2):287–295, 1997.

D. J. Kistler and F. L. Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *J Acoust Soc Am*, 91(3):1637–47, Mar. 1992.

L.-I. Klatt, S. Getzmann, E. Wascher, and D. Schneider. The contribution of selective spatial attention to sound detection and sound localization: Evidence from event-related potentials and lateralized alpha oscillations. *Biological Psychology*, 138:133–145, 2018.

D. C. Knill and A. Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.

K. P. Körding and D. M. Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427 (6971):244–247, 2004.

K. P. Körding, U. Beierholm, W. J. Ma, S. Quartz, J. B. Tenenbaum, and L. Shams. Causal inference in multisensory perception. *PLoS one*, 2(9):e943, 2007.

A. Kothig, M. Ilievski, L. Grasse, F. Rea, and M. Tata. A bayesian system for noise-robust binaural sound localisation for humanoid robots. In *2019 IEEE International Symposium on Robotic and Sensors Environments (ROSE)*, pages 1–7. IEEE, 2019.

J. Kreitewolf, J. Lewald, and S. Getzmann. Effect of attention on cortical processing of sound motion: an eeg study. *NeuroImage*, 54(3):2340–2349, 2011.

M. Kumon and S. Uozumi. Binaural localization for a mobile sound source. *Journal of biomechanical science and engineering*, 6(1):26–39, 2011.

M. Kunin, Y. Osaki, B. Cohen, and T. Raphan. Rotation axes of the head during positioning, head shaking, and locomotion. *Journal of neurophysiology*, 98(5):3095–3108, 2007.

M. F. Land. The coordination of rotations of the eyes, head and trunk in saccadic turns produced in natural situations. *Experimental brain research*, 159(2):151–160, 2004.

M. S. Landy, L. T. Maloney, E. B. Johnston, and M. Young. Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision research*, 35(3):389–412, 1995.

E. H. Langendijk and A. W. Bronkhorst. Contribution of spectral cues to human sound localization. *The Journal of the Acoustical Society of America*, 112(4):1583–1596, 2002.

P. Leong and S. Carlile. Methods for spherical data analysis and visualization. *Journal of neuroscience methods*, 80(2):191–200, 1998.

J. Leung, D. Alais, and S. Carlile. Compression of auditory space during rapid head turns. *Proceedings of the National Academy of Sciences*, 105(17):6492–6497, 2008.

J. Lewald and W. H. Ehrenstein. The effect of eye position on auditory lateralization. *Experimental brain research*, 108(3):473–485, 1996.

J. Lewald and W. H. Ehrenstein. Auditory-visual spatial integration: a new psychophysical approach using laser pointing to acoustic targets. *The Journal of the Acoustical Society of America*, 104(3):1586–1597, 1998.

J. Lewald and H.-O. Karnath. Vestibular influence on human auditory space perception. *Journal of Neurophysiology*, 84(2):1107–1111, 2000.

H. Li. A brief tutorial on recursive estimation with examples from intelligent vehicle applications (part i): Basic spirit and utilities. *HAL*, 2014. URL `https://hal.science/hal-01023525v2`.

J. B. Listing. *Beitrag zur physiologischen Optik*. Number 147. W. Engelmann, 1905.

P. Lladó, P. Hyvärinen, and V. Pulkki. The impact of head-worn devices in an auditory-aided visual search task. *The Journal of the Acoustical Society of America*, 155(4):2460–2469, 2024.

P. Lladó, R. Barumerli, R. Baumgartner, and P. Majdak. Predicting the effect of headphones on the time to localize a target in an auditory-guided visual search task. *Frontiers in Virtual Reality*, 5, 2024.

N. Loveless, S. Levänen, V. Jousmäki, M. Sams, and R. Hari. Temporal integration in auditory sensory memory: neuromagnetic evidence. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 100(3):220–228, 1996.

Y.-C. Lu and M. Cooke. Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners. *Speech Communication*, 53(5):622–642, 2011.

R. C. Luo and C.-C. Chang. Multisensor fusion and integration: A review on approaches and its applications in mechatronics. *IEEE Transactions on Industrial Informatics*, 8(1):49–60, 2011.

R. A. Lutfi and W. Wang. Correlational analysis of acoustic cues for the discrimination of auditory motion. *The Journal of the Acoustical Society of America*, 106(2):919–928, 1999.

N. Ma, T. May, and G. J. Brown. Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2444–2453, 2017.

W. J. Ma and M. Rahmati. Towards a neural implementation of causal inference in cue combination. *Multisensory research*, 26(1-2):159–176, 2013.

E. A. Macpherson. A computer model of binaural localization for stereo imaging measurement. *Journal of the Audio Engineering Society*, 39(9):604–622, 1991.

E. A. Macpherson. Head motion, spectral cues, and wallach's'principle of least displacement'in sound localization. In *Principles and applications of spatial hearing*, pages 103–120. World Scientific, 2011.

E. A. Macpherson. Cue weighting and vestibular mediation of temporal dynamics in sound localization via head rotation. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, page 050131. Acoustical Society of America, 2013.

E. A. Macpherson and J. C. Middlebrooks. Listener weighting of cues for lateral angle: the duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America*, 111(5):2219–2236, 2002.

P. Majdak and M. Mihocic. Expsuite. http://sourceforge.net/projects/expsuite/, 2022.

P. Majdak, M. J. Goupell, and B. Laback. 3-D localization of virtual sound sources: effects of visual environment, pointing method, and training. *Attention Perception and Psychophysics*, 72 (2):454–69, Feb. 2010.

P. Majdak, M. J. Goupell, and B. Laback. Two-dimensional localization of virtual sound sources in cochlear-implant listeners. *Ear and hearing*, 32(2):198–208, 2011.

P. Majdak, R. Baumgartner, and B. Laback. Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization. *Frontiers in Psychology*, 5: 319, 2014. ISSN 1664-1078.

P. Majdak, R. Baumgartner, and C. Jenny. Formation of three-dimensional auditory space. In J. Blauert and J. Braasch, editors, *The Technology of Binaural Understanding*. Springer, Berlin, Heidelberg, 2020. ISBN 978-3-030-00385-2.

P. Majdak, C. Hollomey, and R. Baumgartner. Amt 1. x: A toolbox for reproducible research in auditory modeling. *Acta Acustica*, 6:19, 2022. URL https://www.amtoolbox.org/.

J. C. Makous and J. C. Middlebrooks. Two-dimensional sound localization by human listeners. *The journal of the Acoustical Society of America*, 87(5):2188–2200, 1990.

C. Mark, C. Metzner, L. Lautscham, P. L. Strissel, R. Strick, and B. Fabry. Bayesian model selection for complex dynamic systems. *Nature Communications*, 9(1):1803, May 2018. ISSN 2041-1723. Number: 1 Publisher: Nature Publishing Group.

W. Martens, S. Sakamoto, L. Miranda, and D. Cabrera. Dominance of head-motion-coupled directional cues over other cues during walking depends upon source spectrum. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, page 050129. Acoustical Society of America, 2013.

R. L. Martin and K. I. McAnally. Spectral integration time of the auditory localisation system. *Hearing research*, 238(1-2):118–123, 2008.

T. May, S. Van De Par, and A. Kohlrausch. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Transactions on audio, speech, and language processing*, 19(1):1–13, 2010.

T. May, N. Ma, and G. J. Brown. Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2679–2683. IEEE, 2015.

D. McAlpine, D. Jiang, T. M. Shackleton, and A. R. Palmer. Responses of neurons in the inferior colliculus to dynamic interaural phase cues: evidence for a mechanism of binaural adaptation. *Journal of Neurophysiology*, 83(3):1356–1365, 2000.

K. I. McAnally and R. L. Martin. Sound localization with head movement: implications for 3-d audio displays. *Frontiers in neuroscience*, 8:210, 2014.

G. McLachlan, P. Majdak, J. Reijniers, and H. Peremans. Towards modelling active sound localisation based on bayesian inference in a static environment. *Acta Acustica*, 5:45, 2021.

G. McLachlan, P. Majdak, J. Reijniers, M. Mihocic, and H. Peremans. Dynamic spectral cues do not affect human sound localization during small head movements. *Frontiers in neuroscience.-Lausanne*, 17:1–10, 2023.

W. Medendorp, B. Melis, C. Gielen, and J. V. Gisbergen. Off-centric rotation axes in natural head movements: implications for vestibular reafference and kinematic redundancy. *Journal of Neurophysiology*, 79(4):2025–2039, 1998.

C. Mendonça, P. Mandelli, and V. Pulkki. Modeling the perception of audiovisual distance: Bayesian causal inference and other models. *PloS one*, 11(12):e0165391, 2016.

J. C. Middlebrooks. Narrow-band sound localization related to external ear acoustics. *The Journal of the Acoustical Society of America*, 92(5):2607–2624, 1992.

J. C. Middlebrooks. Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *The Journal of the Acoustical Society of America*, 106(3): 1493–1510, 1999.

J. C. Middlebrooks. Sound localization. *Handbook of clinical neurology*, 129:99–116, 2015.

A. W. Mills. On the minimum audible angle. *The Journal of the Acoustical Society of America*, 30(4):237–246, 1958.

R. Monica and J. Aleotti. Evaluation of the oculus rift s tracking system in room scale virtual reality. *Virtual Reality*, 26(4):1335–1345, 2022.

D. Morikawa, Y. Toyoda, and T. Hirahara. Head movement during horizontal and median sound localization experiments in which head-rotation is allowed. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, page 050141. Acoustical Society of America, 2013.

M. Morimoto and H. Aokata. Localization cues of sound sources in the upper hemisphere. *Journal of the Acoustical Society of Japan (E)*, 5(3):165–173, 1984.

D. Muir and J. Field. Newborn infants orient to sounds. *Child development*, pages 431–436, 1979.

K. Müller. *here: A Simpler Way to Find Your Files*, 2020. URL `https://CRAN.R-project.org/package=here`. R package version 1.0.1.

J. Nix and V. Hohmann. Sound source localization in real sound fields based on empirical statistics of interaural parameters. *J. Acoust. Soc. Am*, 119(1):463–479, 2006.

B. Odegaard, D. R. Wozny, and L. Shams. Biases in visual, auditory, and audiovisual perception of space. *PLoS computational biology*, 11(12):e1004649, 2015.

S. R. Oldfield and S. P. Parker. Acuity of sound localisation: a topography of auditory space. i. normal hearing conditions. *Perception*, 13(5):581–600, 1984.

E. Ozimek and J. J. Zwislocki. Relationships of intensity discrimination to sensation and loudness levels: Dependence on sound frequency. *The Journal of the Acoustical Society of America*, 100(5):3304–3320, 1996.

C. V. Parise, K. Knorre, and M. O. Ernst. Natural auditory scene statistics shapes human spatial hearing. *Proceedings of the National Academy of Sciences*, 111(16):6104–6108, 2014.

M. T. Pastore, S. J. Natale, C. Clayton, M. F. Dorman, W. A. Yost, and Y. Zhou. Effects of head movements on sound-source localization in single-sided deaf patients with their cochlear implant on versus off. *Ear and hearing*, 41(6):1660–1674, 2020.

R. Pavão, E. S. Sussman, B. J. Fischer, and J. L. Peña. Natural itd statistics predict human auditory spatial perception. *eLife*, 9:e51927, oct 2020. ISSN 2050-084X.

H. Peremans, G. McLachlan, P. Majdak, and J. Reijniers. Ideal versus non-ideal observer models for sound localization. In *Proceedings of the 24th International Congress on Acoustics*, 2022.

S. Perrett and W. Noble. The contribution of head motion cues to localization of low-pass noise. *Perception and psychophysics*, 59(7):1018–1026, 1997a.

S. Perrett and W. Noble. The effect of head rotations on vertical plane sound localization. *The Journal of the Acoustical Society of America*, 102(4):2325–2332, 1997b.

K. J. Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.

Z. Pizlo. Perception viewed as an inverse problem. *Vision research*, 41(24):3145–3161, 2001.

D. Poirier-Quinot and M. S. Lawless. Impact of wearing a head-mounted display on localization accuracy of real sound sources. *Acta Acustica*, 7:3, 2023.

G. D. Pollak. Circuits for processing dynamic interaural intensity disparities in the inferior colliculus. *Hearing research*, 288(1-2):47–57, 2012.

M. Pollow, K.-V. Nguyen, O. Warusfel, T. Carpentier, M. Müller-Trapet, M. Vorländer, and M. Noisternig. Calculation of head-related transfer functions for arbitrary field points using spherical harmonics decomposition. *Acta acustica united with Acustica*, 98(1):72–82, 2012.

H. Pöntynen and N. H. Salminen. Resolving front-back ambiguity with head rotation: The role of level dynamics. *Hearing research*, 377:196–207, 2019.

H. Pöntynen, O. Santala, and V. Pulkki. Conflicting dynamic and spectral directional cues form separate auditory images. In *Audio Engineering Society Convention 140*. Audio Engineering Society, 2016.

A. Portello, G. Bustamante, P. Danès, J. Piat, and J. Manhes. Active localization of an intermittent sound source from a moving binaural sensor. In *European Acoustics Association Forum Acusticum*, page 12p, 2014.

M. Puckette et al. Pure data: another integrated computer music environment. *Proceedings of the second intercollege computer music concerts*, pages 37–41, 1996.

V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of the audio engineering society*, 45(6):456–466, 1997.

P. Radau, D. Tweed, and T. Vilis. Three-dimensional eye, head, and chest orientations after large gaze shifts and the underlying neural strategies. *Journal of Neurophysiology*, 72(6):2840–2852, 1994.

L. Rayleigh. Xii. on our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74):214–232, 1907.

J. Reijniers, D. Vanderelst, C. Jin, S. Carlile, and H. Peremans. An ideal-observer model of human sound localization. *Biological cybernetics*, 108(2):169–181, 2014.

L. A. Reiss and E. D. Young. Spectral edge sensitivity in neural circuits of the dorsal cochlear nucleus. *Journal of Neuroscience*, 25(14):3680–3691, 2005.

T. Rohe and U. Noppeney. Sensory reliability shapes perceptual inference via two mechanisms. *Journal of vision*, 15(5):22–22, 2015.

K. Saberi and D. R. Perrott. Minimum audible movement angles as a function of sound source trajectory. *The Journal of the Acoustical Society of America*, 88(6):2639–2644, 1990.

E. Schechtman, T. Shrem, and L. Y. Deouell. Spatial Localization of Auditory Stimuli in Human Auditory Cortex is Based on Both Head-Independent and Head-Centered Coordinate Systems. *Journal of Neuroscience*, 32(39):13501–13509, Sept. 2012. ISSN 0270-6474, 1529-2401.

C. Schymura, T. Walther, D. Kolossa, N. Ma, and G. J. Brown. Binaural sound source localisation using a bayesian-network-based blackboard system and hypothesis-driven feedback. In *Fourm Acusticum*. European Acoustics Association, 2014.

C. Schymura, F. Winter, D. Kolossa, and S. Spors. Binaural sound source localisation and tracking using a dynamic spherical head model. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

R. L. Seilheimer, A. Rosenberg, and D. E. Angelaki. Models and processes of multisensory cue combination. *Current opinion in neurobiology*, 25:38–46, 2014.

I. Senna, C. V. Parise, and M. O. Ernst. Hearing in slow-motion: Humans underestimate the speed of moving sounds. *Scientific reports*, 5(1):1–5, 2015.

L. Shams and U. R. Beierholm. Causal inference in perception. *Trends in cognitive sciences*, 14(9):425–432, 2010.

L. Shams, W. J. Ma, and U. Beierholm. Sound-induced flash illusion as an optimal percept. *Neuroreport*, 16(17):1923–1927, 2005.

E. A. Shaw. Transformation of sound pressure level from the free field to the eardrum in the horizontal plane. *The Journal of the Acoustical Society of America*, 56(6):1848–1861, 1974.

B. G. Shinn-Cunningham, S. Santarelli, and N. Kopco. Tori of confusion: Binaural localization cues for sources within reach of a listener. *The Journal of the Acoustical Society of America*, 107(3):1627–1636, 2000.

N. Shinozaki, H. Yabe, Y. Sato, T. Hiruma, T. Sutoh, T. Matsuoka, and S. Kaneko. Spectrotemporal window of integration of auditory information in the human brain. *Cognitive Brain Research*, 17(3):563–571, 2003.

B. D. Simpson, R. S. Bolia, R. L. McKinley, and D. S. Brungart. The impact of hearing protection on sound localization and orienting behavior. *Human Factors*, 47(1):188–198, 2005.

P. G. Stelmachowicz, K. A. Beauchaine, A. Kalberer, and W. Jesteadt. Normative thresholds in the 8-to 20-khz range as a function of age. *The Journal of the Acoustical Society of America*, 86(4):1384–1391, 1989.

E. Stengård and R. Van den Berg. Imperfect bayesian inference in visual perception. *PLoS computational biology*, 15(4):e1006465, 2019.

M. K. Stern and J. H. Johnson. Just noticeable difference. *The Corsini Encyclopedia of Psychology*, pages 1–2, 2010.

A. A. Stocker and E. P. Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nature neuroscience*, 9(4):578–585, 2006.

S. Särkkä. *Bayesian Filtering and Smoothing*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, Cambridge, 2013. ISBN 978-1-107-03065-7.

X. Teng, X. Tian, and D. Poeppel. Testing multi-scale processing in the auditory system. *Scientific Reports*, 6(1):34390, Oct. 2016. ISSN 2045-2322. Number: 1 Publisher: Nature Publishing Group.

W. R. Thurlow and P. S. Runge. Effect of induced head movements on localization of direction of sounds. *The Journal of the Acoustical Society of America*, 42(2):480–488, 1967.

W. R. Thurlow, J. W. Mangels, and P. S. Runge. Head movements during sound localization. *The Journal of the Acoustical society of America*, 42(2):489–493, 1967.

J. Tobias. *Foundations of modern auditory theory*. Elsevier, 2012.

E. Todorov. Stochastic optimal control and estimation methods adapted to the noise characteristics of the sensorimotor system. *Neural computation*, 17(5):1084–1108, 2005.

D. Tweed and T. Vilis. Geometric relations of eye position and velocity vectors during saccades. *Vision research*, 30(1):111–127, 1990.

M. Ursino, A. Crisafulli, G. Di Pellegrino, E. Magosso, and C. Cuppini. Development of a bayesian estimator for audio-visual integration: a neurocomputational study. *Frontiers in computational neuroscience*, 11:89, 2017.

J.-M. Valin, F. Michaud, and J. Rouat. Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems*, 55 (3):216–228, 2007.

D. C. Van Barneveld and A. John Van Opstal. Eye position determines audiovestibular integration during whole-body rotation. *European Journal of Neuroscience*, 31(5):920–930, 2010.

K. van der Heijden, J. P. Rauschecker, E. Formisano, G. Valente, and B. de Gelder. Active sound localization sharpens spatial tuning in human primary auditory cortex. *Journal of Neuroscience*, 38(40):8574–8587, 2018.

M. van der Heijden and P. X. Joris. Interaural correlation fails to account for detection in a classic binaural task: Dynamic itds dominate n0sπ detection. *Journal of the Association for Research in Otolaryngology*, 11(1):113–131, 2010.

M. M. Van Wanrooij, P. Bremen, and A. John Van Opstal. Acquired prior knowledge modulates audiovisual integration. *European Journal of Neuroscience*, 31(10):1763–1771, 2010.

I. Viaud-Delmon and O. Warusfel. From ear to body: the auditory-motor loop in spatial cognition. *Frontiers in Neuroscience*, 8, 2014. ISSN 1662-453X. Publisher: Frontiers.

N. F. Viemeister and G. H. Wakefield. Temporal integration and multiple looks. *The Journal of the Acoustical Society of America*, 90(2):858–865, 1991.

J. Vliegen and A. J. Van Opstal. The influence of duration and level on human sound localization. *The Journal of the Acoustical Society of America*, 115(4):1705–1713, 2004.

J. Vliegen, T. J. Van Grootel, and A. J. Van Opstal. Dynamic sound localization during rapid eye-head gaze shifts. *Journal of Neuroscience*, 24(42):9291–9302, 2004.

H. Von Helmholtz. *Handbuch der physiologischen Optik: mit 213 in den Text eingedruckten Holzschnitten und 11 Tafeln*, volume 9. Voss, 1867.

H. Wagner and T. Takahashi. Influence of temporal cues on acoustic motion-direction sensitivity of auditory neurons in the owl. *Journal of Neurophysiology*, 68(6):2063–2076, 1992.

H. Wallach. The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology*, 27(4):339, 1940.

E. A. Wan, R. Van Der Merwe, and S. Haykin. The unscented kalman filter. *Kalman filtering and neural networks*, 5(2007):221–280, 2001.

X.-X. Wei and A. A. Stocker. A bayesian observer model constrained by efficient coding can explain'anti-bayesian'percepts. *Nature neuroscience*, 18(10):1509–1517, 2015.

Y. Weiss, E. P. Simoncelli, and E. H. Adelson. Motion illusions as optimal percepts. *Nature neuroscience*, 5(6):598–604, 2002.

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*, 94(1): 111–123, 1993.

G. Wersényi and J. Wilson. Evaluation of head movements in short-term measurements and recordings with human subjects using head-tracking sensors. *Acta Technica Jaurinensis*, 8(3): 218–229, 2015.

F. L. Wightman and D. J. Kistler. Monaural sound localization revisited. *The Journal of the Acoustical Society of America*, 101(2):1050–1063, 1997.

F. L. Wightman and D. J. Kistler. Resolution of front–back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America*, 105(5):2841–2853, 1999.

V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Korner. A probabilistic model for binaural sound localization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36 (5):982–994, 2006.

D. A. Winter. Human balance and posture control during standing and walking. *Gait and posture*, 3(4):193–214, 1995.

H. Yabe, M. Tervaniemi, J. Sinkkonen, M. Huotilainen, R. J. Ilmoniemi, and R. Näätänen. Temporal window of integration of auditory information in the human brain. *Psychophysiology*, 35(5):615–619, 1998.

W. A. Yost, X. Zhong, and A. Najam. Judging sound rotation when listeners and sounds rotate: Sound source localization is a multisystem process. *The Journal of the Acoustical Society of America*, 138(5):3293–3310, Nov. 2015. ISSN 0001-4966. Publisher: Acoustical Society of America.

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*, 91(3):409–420, 2005.

W. Zangemeister, S. Lehman, and L. Stark. Simulation of head movement trajectories: model and fit to main sequence. *Biological Cybernetics*, 41(1):19–32, 1981a.

W. H. Zangemeister, A. Jones, and L. Stark. Dynamics of head movement trajectories: main sequence relationship. *Experimental neurology*, 71(1):76–91, 1981b.

H. Ziegelwanger and P. Majdak. Modeling the direction-continuous time-of-arrival in head-related transfer functions. *The Journal of the Acoustical Society of America*, 135(3):1278–1293, 2014.

B. Zonooz, E. Arani, and A. J. Van Opstal. Learning to localise weakly-informative sound spectra with and without feedback. *Scientific reports*, 8(1):1–14, 2018.

B. Zonooz, E. Arani, K. P. Körding, P. R. Aalbers, T. Celikel, and A. J. Van Opstal. Spectral weighting underlies perceived sound elevation. *Scientific reports*, 9(1):1–12, 2019.

J. J. Zwislocki. Temporal summation of loudness: An analysis. *The Journal of the Acoustical Society of America*, 46(2B):431–441, 1969.