PhD Thesis

# Spatial aspects of auditory salience

Zuzanna Podwinska

Supervisors:

Dr Bruno Fazenda, Prof. Bill Davies

University of Salford MANCHESTER

Acoustics Research Centre
School of Science, Engineering & Environment

September 2020

# Contents

# Acknowledgements

There are several people I would like to thank for their help and support throughout my PhD journey.

First and foremost, my supervisor, Bruno Fazenda and my co-supervisor, Bill Davies – for always being available to me, being encouraging and pushing me to do my best. I could always count on our meetings turning into long, interesting discussions which would often make me rethink my ideas and assumptions. I couldn't have asked for a better supervisory team.

Everyone in the Acoustics Research Centre, where I immediately felt welcome. In particular, I need to mention all my fellow postgraduate students who sat next to me in the fishbowl and G11 – the LaTeX masters, the stats experts, the cake providers, those who talked about research with me in the office, and of course, those who talked about life with me in the pub.

Everyone who has volunteered to participate in my experiments – this work would truly not be possible without them. And Lara Harris, for helping enormously with participant recruitment and room bookings.

Iwona Sobieraj, for the countless conversations about the joys and pains of pursuing a PhD, and for insisting that we should write a paper together.

My parents, for their continuous support and for always encouraging me to do what I believe is best for me, even if it means living 1000 miles away from them.

Last but certainly not least, James, for always being by my side, especially in the last months of writing this thesis. And Mitchie, for giving me an excuse not to stay in the office for too long, keeping me company when I worked from home, and the numerous times she saved my life from people simply walking past the house.

# Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation contains fewer than 100 000 words including appendices. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified below. Sections of work in this thesis which had been published at the time of writing are noted in the List of Publications.

## Collaborative work

- I was involved in the design of the experiment described in **Chapter 5** in collaboration with Prof. Bill Davies and Dr Anna Remington (University College London). Data collection for the experiment was undertaken by me and Beth Tobiansky. The analysis of data presented in Chapter 5 was performed by me.
- The combined salience and machine learning model described in **Chapter 11** was developed in collaboration with Dr Iwona Sobieraj, who designed and implemented the NMF part of the model, as well as the algorithm testing environment. My contribution was the implementation of the auditory salience model.

Zuzanna Podwinska

September 2020

# List of Publications

## Peer-reviewed conference papers

- Z. Podwinska, B. M. Fazenda and W. J. Davies (2019). 'Testing spatial aspects of auditory salience'. In: *Proceedings of the 25th International Conference on Auditory Display (ICAD 2019)*, pp. 191–198
  - This paper includes the results of experiments described in **Chapter 3** and **Chapter 4**.
- Z. Podwinska et al. (May 2019). 'Acoustic event detection from weakly labeled data using auditory salience'. In: *ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 41–45
  - The work described in **Chapter 11** and the model description in **Chapter 10** were published in this paper.

## Conference posters

- Z. Podwinska, B. M. Fazenda and W. J. Davies (2019). 'Modelling auditory salience of pattern regularity violations'. In: *The Predictive Brain Conference*, Marseille, France, 26-27 September 2019.
- Z. Podwinska, B. M. Fazenda and W. J. Davies (2017). 'Influence of source location and temporal structure on spatial auditory saliency'. In: *Acoustics '17*, Boston, Massachusetts, USA, 25-29 June 2017.

## Abstract

Models of auditory salience aim to predict which sounds attract people's attention, and their proposed applications range from soundscape design to machine listening systems and object-based broadcasting. A few different types of models have been proposed, but one of the areas where most of them still fall short is spatial aspects of sound – they usually operate on mono signals and do not consider spatial auditory scenes. Part of the reason why this is the case might be that the relationship between auditory salience and position of sound is still not clear. In addition, methods used to measure auditory salience vary greatly, and authors in the field do not always use the same definition of salience.

In Part I, this thesis aims to answer questions about the effect of spatial location of sound on auditory salience. This is done in four different experiments, which are based on previously published experimental methods but adapted to measure spatial effects. In general, the combined results of these experiments do not support the hypothesis that the spatial position of a sound alone influences how salient the sound is. However, they do show that unexpected changes in position might activate the deviance detection mechanism and therefore be salient. In addition, an experiment comparing three of the methods used reveals at least two dimensions of salience, which are measured by different methods to different extent. This emphasises the importance of carefully considering which experimental methods are used to measure auditory salience, and also providing a clear definition of what type of salience is of interest.

Part II demonstrates how spatial position of sound can be incorporated into an auditory salience model. The results of experiments described in this thesis support the idea that the basis of auditory salience is the violation of expectations. The surprise caused by a sudden change in sound position can therefore be modelled by a Kalman-filter-based deviance detection model, which predicts experimental data discussed above with good accuracy. Finally, an example is given of how an application of such a model can improve the performance of a machine learning algorithm for acoustic event detection.

# 1

# Introduction

## 1.1. Motivation

Perceptual modelling aims to mimic human responses to external stimuli such as images and sounds. Some models attempt to faithfully imitate physiology – for example, the working of a single neuron – but many take a more functional approach, where it is the final outcome of the system, or the general working mechanism that is being modelled. The models can span from a full, general model of auditory processing to more specialised models, which only tackle one aspect of it, such as sound quality or localisation. The latter group also includes auditory salience, which aims to detect sounds in the environment which are in some way *salient* to people. Such a model has many potential applications: for example, it could be used in object-oriented broadcasting as an additional layer of meta-data, providing information about which objects are worth prioritizing. It could also be useful in machine listening applications, for example, for scene analysis in humanoid robots (Schauerte et al. 2011) or soundscape quality assessment (Boes et al. 2018). It can also be applied to improve speech recognition and synthesis (Kalinli and Narayanan 2009) and it has even been suggested that information about salience can be incorporated

into audio data compression algorithms (Kakouros, Rasanen and Laine 2013).

Several auditory salience models have been developed in the last 15 years (starting with Kayser et al. 2005), but the field is not as developed as that of visual salience modelling. Part of the reason why auditory salience models are behind their visual counterparts is that measuring the salience of sounds is not straightforward. In fact, even the way salience is defined in studies is not always exactly the same – however, usually, sounds which automatically attract attention are considered salient. In vision, this is often assessed with eye-tracking, as people's eyes will automatically turn to important features in the environment. In hearing, however, determining which sounds are attended to is more difficult, as there is no clearly visible physical externalization of auditory attention. This leads researchers to develop and use various experimental methods. Some of the methods involve asking participants explicitly to mark salient events or compare salience between two sounds, and others are based on assumptions such as that detection in salient streams is easier. This lack of well-established, standard measurement methods makes the design and assessment of auditory salience models challenging.

One of the areas in which there is still room for improvement is spatial salience models. Most auditory salience models use a one-channel input and do not take spatial position of sound sources into account. Yet, spatial hearing has been a field of study with considerable interest for years, and the benefits of using acoustic signals from two ears are well known. Binaural hearing enhances stream segregation and improves speech intelligibility for sources separated in space. Spatial auditory salience has not been well studied directly either, at least not for locations all around the listener.

This thesis aims to address the question of the relationship between spatial location of sound and auditory salience. Does the absolute position of a sound around the listener influence how salient it is for this listener? Answering this question could

help create better salience models and enhance our understanding of the types of sounds which attract attention.

Four experiments were designed specifically to study whether spatial position of a sound influences its salience. Different methods are used that address this question from slightly different angles, some prioritizing low-level, automatic attention orienting, while others include other related phenomena, such as perceptual load and violation of expectations. Violations of expectations, in particular, tend to attract people's attention, so how expectations are built and what they are about is important for salience. Automatic attentional orienting is in fact related to a deviance detection mechanism in the brain. These processes can be also described in terms of prediction – what does not fit the prediction, is going to be salient. Some recent auditory salience models have successfully adopted these principles.

## 1.2. Outline of the thesis

The thesis is organised in two parts.

Part I focuses on measuring spatial auditory salience. **Chapter 2** gives an overview of the literature on auditory salience and methods used to measure it. It discusses the ways in which salience has been defined in the literature, and how different experimental methods relate to these definitions. The following four chapters describe experiments designed to study spatial auditory salience. **Chapter 3** describes an oddball detection experiment, in which participants detected a shortened inter-stimulus interval within two competing auditory streams. Their response times and accuracy are recorded to test auditory salience in a well-controlled setting. **Chapter 4** presents a more ecologically valid experiment in which participants are asked to report their attention in real-time. The experiment in **Chapter 5** tackles

auditory salience under perceptual load, and tests it in a dual-task scenario. **Chapter 6** describes a distraction experiment, in which implicit expectations about the distractors' sound type and spatial position are manipulated to elicit surprise. The data collected in these experiments include behavioural responses such as task accuracy and response times, as well as measurement of pupil dilation, which have been previously used to study salience. Finally, three of the methods used here are directly compared in a perceptual experiment described in **Chapter 7** to ascertain whether their outcomes correlate with each other.

Part II aims to apply results from Part I to improve models of auditory salience. **Chapter 9** provides a literature review of auditory salience models and a background on the computational and perceptual principles on which some of the more recent models are based, including prediction and deviance detection. Then, **Chapter 10** illustrates how spatial information can be added to such a model to predict some of the experimental results described before. An application of a prediction-based model is described in **Chapter 11**, where it is shown to improve an acoustic event detection algorithm.

Finally, **Chapter 13** provides a general discussion of both the experimental results in Part I and the modelling efforts in Part II, and how the experimental data can be used to inform the models.

## 1.3. Contributions

The work in this thesis has contributed the following knowledge to the field:

- Experimental data suggests that the absolute spatial location of a sound alone does not modulate its salience.
- Pupil dilation responses are sensitive to unexpected changes of spatial position

of a sound in a distraction experiment.

- Experimental methods used to measure auditory salience vary on two dimensions, indicating that they measure different aspects of salience.
- A model based on deviance detection can successfully predict pupil dilation responses to broken expectations about spatial location and type of sound.

# Part I.

# Measuring auditory salience

# 2

# Literature review

## 2.1. What is auditory salience?

The term "salience" (or "saliency") has become more widely used in the area of auditory modelling relatively recently, after Kayser's translation of a visual saliency map into the auditory domain (Kayser et al. 2005). The original visual saliency map proposed by Itti and Koch (2001) codes the global "conspicuity" of all locations in the visual field and emphasises "interesting or conspicuous' locations.

There is no single, universally accepted definition of auditory salience. Out of the publications which have discussed the topic over the last 15 years, not many have offered a clear definition of what exactly is meant by "salience". One of the few exceptions were Tordini, Bregman and Cooperstock (2015), who proposed that:

> A sound is salient, i.e., belongs to the foreground, when its selection in a complex scene is 'as easy' as its detection in isolation, i.e., over silence.

In a more recent publication, Tordini, Bregman and Cooperstock (2016) suggest there are in fact two types of salience: sensory and perceptual. Sensory salience describes

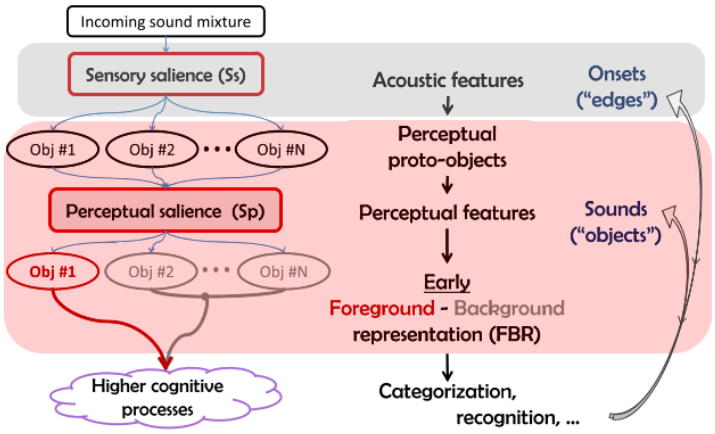



**Figure 2.1.:** An auditory perception framework reproduced from Tordini, Bregman and Cooperstock 2016. Note that this framework also includes top-down processes in the form of a feedback going from higher cognitive processes down to sounds and onsets.

how noticeable a sound event is in relation to its local neighbourhood, while perceptual salience operates on streams and describes how likely they are to become foreground. Thus, they work on short and long time scales respectively. This concept is illustrated in Figure 2.1.

Even though some authors writing about salience equate it with bottom-up attention or a cognitive mechanism (Slaney et al. 2012; Rodríguez-Hidalgo, Peláez-Moreno and Gallardo-Antolín 2017), most agree that salience is in fact a *property of sound*. Explanations of what exactly that property describes vary, but they seem to revolve around a few common points:

A salience is the **ability to attract attention** (or the likelihood that a sound will attract attention); salient sounds can be noticed without a conscious decision to focus on them (or even "without attention"), and people have difficulty ignoring them (e.g. Tsiami et al. 2016; Zhao et al. 2019)

B salience is the extent to which sounds (perceptually) **stand out** from the environment or their neighbours, how much they "pop out"; distinctiveness;

being easy to notice and detect, conspicuous (e.g. De Coensel and Botteldooren 2010; Liao et al. 2015; Tang and Cox 2018)

C salience is the "novelty and uniqueness, **deviating from the background**", how much a sound differs from its surroundings or the contrast between the sound and its surroundings; how much it deviates from regularities preceding it; its rarity in relation to the recent and long-term past (e.g. Tsuchida and Cottrell 2012; Tordini, Bregman and Cooperstock 2016)

D salience describes the sound's **importance and relevance**; salient sounds are informative and interesting (e.g. Botteldooren and De Coensel 2009; Rodríguez-Hidalgo, Peláez-Moreno and Gallardo-Antolín 2018)

The four points describe similar but not necessarily identical concepts. Arguably, A is what most auditory salience models aim to achieve – to predict which sounds will grab attention. It does not necessarily explain where auditory salience comes from (what does it take to attract attention?), but it can be useful for experimental paradigms, as long as one is able to detect when a sound has been attended to.

Point B also describes the perceptual effects of salient sounds – they stand out, are easy to notice and detect. This point relates to segregating sounds from their environment, or perceptual organisation. Of course, it is easy to see that A and B are related – sounds that stand out will often attract attention, and it is possible to interpret a lot of experimental paradigms through both lenses. In fact, some studies do mention more than one of the above points (e.g. Kayser et al. 2005; Huang and Elhilali 2017; Filipan et al. 2016a). However, this point is not the same as attention – for example, that a sound is easy to detect does not necessarily mean it will draw one's attention when one is not actively trying to detect it.

In some studies, salience is described not in perceptual or attentional terms, but by the properties of the sounds – as in point C. This includes mentions of salient sounds being different from the environment, deviating from patterns, being rare. While all of

these statements are in general true – these types of sounds tend to be salient – it seems to be more of a description of what makes a sound salient, rather than a definition of what salience is. Again, there is a close relationship between points B and C – sounds stand in the environment because they deviate in their features from the sounds preceding and surrounding them. The difference here is that the former refers to perception, whereas the latter – to stimulus features.

Finally, as in point D, some studies describe salience with words such as "important", "relevant", "informative" or "interesting". These relate to the idea that the brain monitors the environment and choses the parts which might potentially be important or relevant for further processing. However, these descriptors are all rather vague, and would each require further clarification.

How one decides to define auditory salience has implications for what type of method should be used to measure it. The following section will discuss some of the methods and show how they relate to the different definitions.

## 2.2. Testing auditory salience

There is no single agreed upon paradigm of testing auditory salience, and a variety of different methods have been proposed in the literature. This section presents behavioural testing methods grouped according to which definition of salience they fit best and explains some of the perceptual mechanisms behind each of the methods. Examples are also given of how physiological measurements can be used to infer auditory salience.

### 2.2.1. Human judgements

Perhaps the most straightforward way of testing whether a sound is salient is asking human subjects directly. For example, in an annotation task, Kim et al. (2014) asked participants to manually mark "interesting" sounds in a recording of a scene. Another type of experiment which involves human judgement is a comparison of two sounds (or scenes) in terms of their salience or "interestingness" (Kayser et al. 2005; Duangudom and Anderson 2007; Tsuchida and Cottrell 2012; Zhao et al. 2019). This type of experiment has the advantage of being able to sort test sounds from least to most salient. The downside is the subjectivity of the word "salient" or "interesting", which can have different meanings to different people. These types of experiments would most likely measure attention as defined by point D in the previous section.

### 2.2.2. Attention

Another type of experiment is based on the definition of salience being the ability to attract attention (point A in Section 2.1). Although attention has been extensively studied by philosophers, psychologists and neuroscientists for many years, it is not at all obvious how to define it. Since James (1890) wrote "every one knows what attention is"[1] , it has been described as a filter (Broadbent 1958), searchlight (Fritz et al. 2007), biased competition (Duncan 2006), and precision (Heilbron and Chait 2017). In general, it is a process or a group of processes, which prioritises some sensory inputs over others. It consists of bottom-up (involuntary) and top-down (voluntary)

[1]He continues: "It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others, and is a condition which has a real opposite in the confused, dazed, scatterbrained state which in French is called distraction, and Zerstreutheit in German."

processes.

Bottom-up processes cause an automatic attentional shift towards a salient event even when the person's focus is elsewhere – it is independent of the task that they might be performing. This attentional orienting might be brought about by different characteristics of sound, both low- and high-level. In general, attentional orienting is often caused by a violation of expectations built based on previous auditory inputs. It is important to stress that it does not depend on the frequency of occurrence of the sound as such or if it is novel, but rather on whether the sound matches the expectation (Parmentier et al. 2011; Vachon, Hughes and Jones 2012). For example, Nöstl, Marsh and Sörqvist (2012) demonstrated that the degree of attentional capture depends on how far the deviant sound is from the *expected* sound, not from the previous sound (local change). The brain is able to track complex and even abstract types of regularities and detects when input deviates from them. This predictive view of perception is discussed in more detail in Section 9.4.

Top-down selective attention can be consciously deployed by the listener to enhance perception of sound. Some studies indicate that this enhancement might in fact take place as low in the auditory system as the cochlea, by showing attentional effects on otoacoustic emissions (Giard et al. 1994; Maison, Micheyl and Collet 2001; Walsh, Pasanen and McFadden 2015), but other studies do not find this effect (Avan and Bonfils 1992; Michie et al. 1996; Timpe-Syverson and Decker 1999). It has, however, been repeatedly shown that attention enhances relevant sound representation in the brain (Alain, Arnott and Dyson 2014). Some also argue that the modulation goes beyond a simple gain-like enhancement and influences sharpening of relevant neural tuning curves (Kauramäki, Jääskeläinen and Sams 2007).

Attention can be directed to sound features such as frequency, timbre or location. For example, Kidd et al. (2005) found that providing listeners with a cue about the target speaker location significantly improved keyword identification compared to when no

cues were present. It has been found that both top-down and bottom-up orienting of attention to the position of a sound improved its localisation (Spence and Driver 1994). Also, Best et al. (2006) showed that attending to two spatially separated sources comes at a cost compared to streams in the same location, which suggests spatial attention might work as a "spotlight", similarly to vision. Even though it is possible to pay attention to features of sound such as pitch, there is strong indication, that in fact, attention operates on auditory objects (Shinn-Cunningham 2008). This would be consistent with the role of selective attention in vision, where it is believed to be object-based – e.g. Duncan (1984) found that it is easier to make judgments about two properties of the same object, than about properties of two different objects. Furthermore, Best et al. (2008) andBressler et al. (2014) have shown that selective auditory attention is enhanced by object continuity. This view is also supported by brain imaging studies, which have shown the same brain activity independently of which one of two features of an auditory object was attended (Zatorre, Mondor and Evans 1999).

Tracking auditory attention in real-time is not a straightforward task. Whereas in vision, tracking of automatic eye movements can be used, no such moving organ exists for hearing, and determining what a person is listening to is much more difficult. There have, however, been some attempts. For example, Huang and Elhilali (2017) use self-reporting, but avoid some of the ambiguity of other survey-based methods by asking participants explicitly to indicate where their attention is, and to do so in real time. This is somewhat analogous to gaze tracking in visual salience studies, but a less direct representation of the phenomenon, as it also involves conscious tracking of one's attention.

Attention tracking in audition can also be attempted with distraction experiments, in which it is assumed that salient sounds cause automatic attentional orienting away from the main task and therefore impair performance. However, there are at least two

things that need to be taken into account when using distraction experiments to measure attention.

First, it has been suggested in the duplex-mechanism theory of distraction (Hughes 2014) that there are in fact two separate processes that can cause distraction: attentional orienting and interference-by-process. The first type comes about when a sound draws attention away from the main task for a brief moment, causing slower responses. The second, on the other hand, arises when the sounds interfere with the type of brain processing which is needed for the particular task. It is usually demonstrated in a serial recall task, in which participants are asked to memorise the order of items presented to them visually. In this task, the subjects' response is significantly impaired by a sequence of tones or vowels which change from one to another – the so-called changing-state effect. However, if participants are asked to memorise visual items but not in any specific order, this effect disappears. Differences in pupil dilation responses also indicate that the two types of distraction are underpinned by different mechanisms (Marois, Marsh and Vachon 2019). These findings suggest that not all distraction can be attributed to attention, and any experimental methods need to take this into account.

Additionally, distraction effects might vary with the difficulty of the task and its perceptual demands. According to the load theory (Lavie 1995) the brain has a limited perceptual capacity, and until it reaches its limits, all sounds – task relevant or not – will be processed fully. This processing of all information until the capacity limit is reached is obligatory, therefore, if a task does not fill it, there will be some degree of automatic distraction from irrelevant input. However, once capacity is reached – for example, by a task with high perceptual load – no more input (whether auditory or visual) will be processed and no distraction will occur. However, Eltiti, Wallace and Fox (2005) argue that in fact distractor salience might be more important than perceptual load. In a visual experiment, they modulated distractor salience by making

it more or less similar to the target, and the target salience by making it larger than other, neutral items on the display. They argue that it is the decrease in distractor salience by increasing the number of items on a screen – as is often done in perceptual load experiments – that minimises distraction, rather than the lack of free perceptual capacity. Santangelo, Olivetti Belardinelli and Spence (2007) found a suppression of attentional spatial orienting to cued locations in high load conditions, compared to low load.

In general, measuring auditory attention is not a straightforward task. In addition, attention operates on many levels and it is not clear at which level we should be measuring it. Compare the automatic orienting in distraction with experiments relying on self-reporting of attention: some of the bottom-up attentional orienting might occur without reaching conscious awareness, or be brief enough that participants would not report it. However, for participants to report attending to a sound, they have to not only be aware of it, but also aware of their own attention, which is much more high-level.

### 2.2.3. Detection

Some studies rely on the detection definition of salience (point B). For example, participants might be asked to detect a sound in noise (Kayser et al. 2005) or to detect whether a sound clip contains a salient event (Kaya and Elhilali 2014). Another paradigm is based on oddball detection – pointing out a stimulus which is different from a series of standard, regular ones, often in the presence of competing streams. Response time and detection rate are the indication of stimulus salience (e.g. Tordini et al. 2013; Southwell et al. 2017).

A few conditions must be fulfilled for a sound to be successfully detected: first, it has to be audible – over the hearing threshold and not masked by other sounds; second, it

has to be separated from any background, or form a separate auditory object; finally, it has to be attended to, either consciously in a top-down manner, or through automatic, bottom-up attentional orienting.

There has been debate about the nature of the relationship between attention and auditory object formation. One view is that attention is strictly necessary for stream segregation. For example, Carlyon et al. (2001) played series of A-B-A tones, which over a span of several seconds tend to separate into two streams A and B. The participants would hear the tones in one ear, while their attention was directed to the other ear for the first 10 seconds. After that time, they would switch attention to the tones and asked for a stream segmentation judgement. If attention was not required for streaming, one would expect to see a 10-seconds-long streaming build-up on the tones causing the separation into A and B streams, but no indication of this was found.[2] Also, Shamma, Elhilali and Micheyl (2011) argue that attention facilitates stream segregation by binding together relevant features. Sussman et al. (2002) presented experimental participants with a repeating pattern of 5 tones (4 of the same frequency and one "deviant") and manipulated the listeners' attention with a task which required them to focus either on the tones' frequency, or on the pattern as a whole. They have shown that Mismatch Negativity (MMN), a brain potential associated with novel stimulus, is not evoked in response to the deviant tone when listeners follow tones as a pattern, as opposed to focusing on their frequency. This may be a proof of attention influencing grouping.

A different hypothesis is that stream segregation happens pre-attentively, and attention is only used to select which object becomes foreground (Bregman 1990). Indeed, there is proof that some information is processed in an unattended stream –

[2]Note that Macken et al. (2003) offer different explanations for the results of Carlyon et al. (2001), for example, that the auditory memory task might have influenced participants' pre-attentive processes.

for example, in a classic dichotic listening task, Moray (1959) showed that people will notice when their name appears in an unattended stream. The kind of high-level processing needed to recognise a name would not be possible without some kind of stream segregation. Macken et al. (2003) offer a paradigm for studying the role of attention in streaming, based on the disruptiveness of unattended sound, which they call the "irrelevant sound effect". In their experiments, the level of disruptiveness of task-irrelevant (unattended) tones in a visual memory task follows a pattern expected if streaming on those tones had taken place. Also, Masutomi et al. (2015) showed that segregation based on repetitions is not affected by attention. Interestingly, Deouell et al. (2007) also showed that people with unilateral neglect, who are not aware of any sounds on their left side, still experienced the scale illusion (Deutsch 1975), which relies on grouping of sounds from both ears.

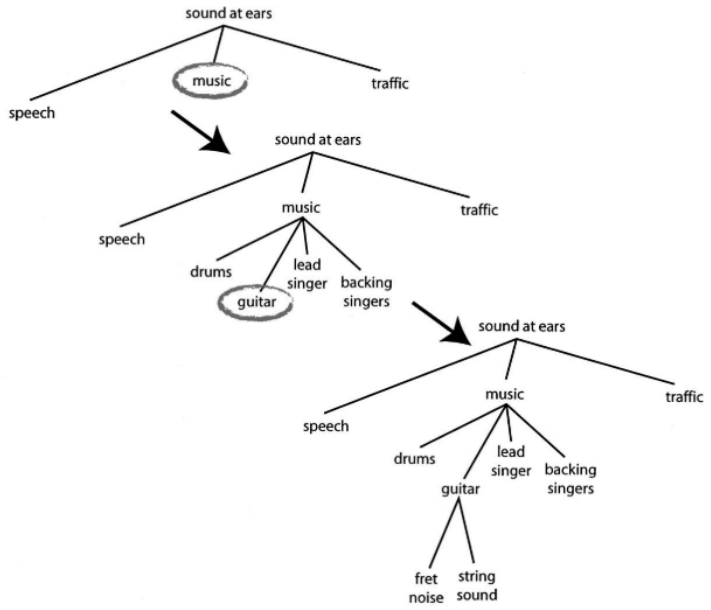



**Figure 2.2.:** The hierarchical decomposition model from Cusack et al. (2004). By aiding grouping, selective attention helps separate auditory scene into more objects.

Most likely, selective attention enhances stream segregation, particularly in complex environments, but is not necessary for it to happen (Shinn-Cunningham and Best

2015). Cusack et al. (2004) offer a hierarchical model of grouping, in which attention adds more detail to the representation of a scene – for example, it lets the listener divide a band into separate instruments (see figure 2.2). They replicated the work of Carlyon et al. (2001) in a series of detailed experiments, but noted that their results suggest that some basic streaming still takes place pre-attentively.

### 2.2.4. Physiological methods

Other than behavioural experiments, physiological measures can be used for determining the salience of sounds. For example, Liao et al. (2015) showed a connection between pupil dilation responses (PDR) and sound salience measured as subjective judgements. Indeed, there is some evidence that pupil dilation corresponds to the attentional orienting response (Marois et al. 2018). For example, Marois, Marsh and Vachon (2019) compared distraction caused by attentional orienting and interference-by-process, and found pupil dilation responses to the former, but not the latter. Liao et al. (2016) also showed that pupil dilation responses to deviant sounds are not modulated by top-down attention.

However, Zhao et al. (2019) did not find a relationship between pupil dilation responses and subjective salience ratings and suggested that pupil dilation might represent a later stage of salience processing. Also, Huang and Elhilali (2017) reported that although pupil dilation responses corresponded to changes in acoustic features of the stimuli, they did not *always* correspond to a behavioural response (self-reported attention).

In more general terms, pupil dilation is a response to changes in allocation of cognitive resources. In a literature review of 146 studies, Zekveld, Koelewijn and Kramer (2018) identified various external and internal factors which influence PDR to auditory stimuli, including automatic and intentional attention, increased task

demands, emotional valence, and an individual's hearing status. Pupillometry has been used to determine listening effort with focused attention (Koelewijn et al. 2015), and has been shown to respond to different levels of informational masking (Woodcock et al. 2019). Marois, Marsh and Vachon (2019) found an overall increase in pupil dilation to changing-state compared to "static" stimuli, which could indicate increased listening effort.

There is also evidence to suggest that auditory salience modulates inhibition of microsaccades – small, rapid eye movements (Zhao et al. 2019). Furthermore, Frith and Allen (1983) suggested ways in which skin conductance could be used to study the level and direction of attention. In addition to skin conductance response, Stekelenburg and Van Boxtel (2002) also showed inhibition of heart rate, respiration rate and depth, and electromyographic (EMG) activity of lower facial muscles in response to novel auditory stimuli.

The effects of involuntary, bottom-up attention can also be seen in event-related brain potentials (ERPs), specifically a negative N1, which is automatically evoked by novel sounds, and positive P300 (with P3a and P3b subcomponents), related to involuntary orienting towards a salient stimulus. Additionally, Mismatch Negativity (MMN) is an ERP difference wave between response to an oddball and regular auditory stimulus, and it is believed to reflect pre-attentive novelty detection. In contrast to P300, it is present even if the listener expects a deviant sound. It has been suggested that MMN and P300 represent two different time scales of auditory prediction and novelty detection (Wacongne et al. 2011). While MMN, present at about 200 ms, is a reaction to short-term novelty detection, P300, present at 250-400 ms after stimulus and requiring conscious attention (Chennu et al. 2013), represents unexpected changes in longer-term patterns.

## 2.3. Salient features

There currently is no single list of features which make a sound salient (Shinn-Cunningham and Best 2015). Most researchers agree that loudness is important for salience (Liao et al. 2015; Tordini, Bregman and Cooperstock 2016; Huang and Elhilali 2017). Tordini, Bregman and Cooperstock (2016) argue that after loudness, the most important salience features are tempo (faster patterns are more salient) and brightness ("darker" sounds are more salient). Kaya and Elhilali (2014) also showed effects of pitch and intensity on salience – interestingly, they found higher pitched sounds to be more salient. It has also been suggested that roughness may play a part, with rougher sounds being more salient (Zhao et al. 2019).

Although this thesis is only concerned with lower-level characteristics of sound, there certainly are higher level features and processes that influence salience. For example, a person's own name has the ability to attract attention even when they are focused on a task (Wood and Cowan 1995). Sounds associated with strong emotions can also have a larger attention-grabbing effect (Vuilleumier 2005). Deviations in sound category have also been shown to cause auditory distraction (Vachon, Marsh and Labonté 2019).

A recent study investigated brain responses to deviance in different features: timbre, pitch and intensity of notes in a melody, during a visual task in high and low load condition (Kaya, Huang and Elhilali 2020). They compared the same note when it matched and did not match the melody on one or more of the features (in and out of context) and found multiple interactions between the features.

### 2.3.1. Spatial salience

Although localisation of sounds has been thoroughly studied, not much is known about how the spatial position of a sound affects its salience, and auditory salience or

attention experiments with sounds positioned all around the listener are rare. Most studies of cross-modal spatial attention, for example, have presented stimuli in the frontal plane (Spence, Lee and Van der Stoep 2020).

There are other known spatial effects in auditory perception. For example, a right-ear advantage has been shown for speech stimuli, and there is debate whether it can be explained primarily by the specialisation of the left hemisphere in processing speech, or by attentional biases (Hiscock and Kinsbourne 2011). On the other hand, some have shown a left-ear disadvantage for non-speech irrelevant sound (meaning, sounds on the left cause greater distraction) to a task that involves serial recall (Hadlington, Bridges and Darby 2004) but not necessarily other memory-related tasks (Hadlington, Bridges and Beaman 2006), for distracting sounds with changing-state characteristics.

In a change detection experiment, moving target sounds originating on the left hand side (-60°) were detected faster than those originating +20° to the right (Peck et al. 2018). On the other hand, in an audio-visual distraction experiment, sounds on the right were more distracting than on the left (Corral and Escera 2008). An interesting rear-to-front cueing effect has also been observed – auditory stimuli on the side of a visual target enhanced responses to the target both when they were in front and rear. In other words, for example – sounds at 45° and 135° both caused attentional orienting to the right (Lee and Spence 2015).

Finally, it is known that *changes* in spatial location of the auditory stimulus can cause distraction (Chan, Merrifield and Spence 2005; Roeber, Widmann and Schröger 2003) – i.e. a sound coming from an unexpected location will be more distracting than one coming from an expected location. This has been shown by modifying spatial position (which was task-irrelevant) of a stimulus in a distraction experiment with an auditory task (Roeber, Widmann and Schröger 2003). Also, in a word recognition experiment, irrelevant words which changed location randomly between trials were more distracting than those coming from one location (Chan, Merrifield and Spence 2005).

This effect has also been shown with a visual task and auditory stimulus (Corral and Escera 2008) – in an even/odd classification task, a sound coming from an unexpected location was more distracting than one coming from an expected one. What is more, the effect seemed to increase with increased spatial separation.

## 2.4. Summary

This chapter provided a review of the literature on the measurement of auditory salience. No single definition of auditory salience exists, but the most common way to describe it is as the ability of a sound to attract attention. The experimental methods in the literature also vary, from surveys to methods based on detection and competing streams. In the following chapters, four different experimental approaches are described, each investigating auditory salience from a slightly different perspective.

# 3

# Automatic attentional orienting

## 3.1. Introduction

In this chapter, an experiment is described which was designed to test spatial auditory salience in a well-controlled experimental setting. The aim was to find out if any particular location is likely to automatically attract auditory attention than other locations.

## 3.2. Method

This experiment is based on the Segregation of Asynchronous Patterns (SOAP) paradigm (Tordini et al. 2013). The approach assumes that two perceived auditory streams will compete for attentional resources, and as a result one of them will become *foreground*, and the other will be *background*. If no arbitrary top-down effects are in place, a more salient stream will win the competition and be the foreground. The main assumption here is that it will be easier to detect changes in the foreground (more salient) stream.

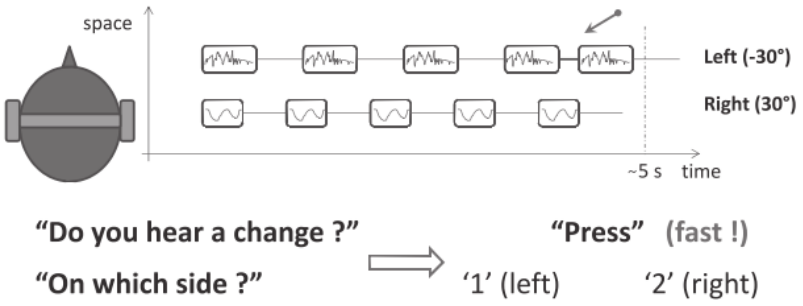



**Figure 3.1.:** The original SOAP paradigm, reproduced from Tordini et al. (2013).

In the original SOAP experiment, two sound patterns were presented to the left and right ears through headphones (see Figure 3.1). Both streams were patterns consisting of short birdsong excerpts separated by constant inter-stimulus interval (ISI). A crucial part of the design is to make sure that the two patterns are asynchronous, to avoid creating a rhythm which could be morphed into a single object. The participants' task was to detect a change in ISI in one of the streams, without being told which stream to attend. According to the SOAP framework, they should be statistically more likely to attend to, and detect changes in, the more salient stream.

The SOAP framework was extended in this experiment to include spatial effects. The participants were seated in an acoustically treated listening room, surrounded by six loudspeakers as in Figure 3.2. The stimuli were short noise bursts, either high- or low-pass filtered at 2 kHz. The sounds were designed so that there is no overlapping spectral content, to ensure easy stream segregation. Each pattern contained only one type of stimuli and lasted for 6 seconds. The regular inter-stimulus interval was 250 ms, and the shortened one – 80 ms. Instead of simply left and right, sound patterns arrived at the listener from 2 out of 6 locations around them. The participants were asked to detect a shortened ISI and indicate whether it occurred in the high or low frequency pattern.

To ensure asynchrony, one of the two patterns always included shorter stimuli than

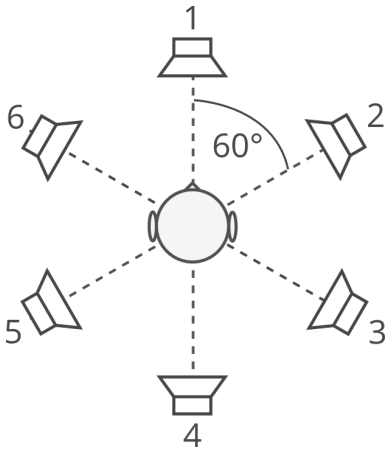



**Figure 3.2:** Loudspeaker set-up in the listening room.

the other (150 versus 200 ms). This resulted in one pattern sounding faster than the other (which is referred to here as fast tempo).

Independent variables were then: sound location (6 target stream location, each with 5 remaining background stream locations), frequency (high and low), and tempo (fast and slow). Each participant was exposed to all conditions in a full-factorial design, which resulted in 120 trials per person.

Before the main experiment, participants completed a short training session and a baseline test, where only one pattern was present at one time. Information from this baseline condition was later used to determine acceptance windows for each person (see Section 3.3).

19 volunteers took part in the experiment, all with self-reported normal hearing, average age 30.4, 4 female, 18 right-handed. Data collected included response time and accuracy.

## 3.3. Results

### 3.3.1. Response times

Time elapsed from the end of the shortened ISI to button press was recorded as response time (RT). Only correct responses were taken into account. RT distributions differed quite significantly between participants (see Figure 3.3).

Data was analysed with a Generalised Linear Mixed Model, using the lme4 package (Bates et al. 2015) in R (R Core Team 2019), with location, frequency, and tempo as fixed effects, and participant as random effect. A model including frequency-tempo and frequency-location interactions was used as it gave the best fit, based on the Akaike information criterion (Akaike 1974). GLMMs have a few advantages that are important in this case: they can deal with missing data (incorrect responses are not used, so there are no data points for them), they let the researcher specify response distribution, and can take into account baseline differences between groups (in this case: participants). Lo and Andrews (2015) argue that GLMMs are the preferred method of analysing reaction time data (rather than, e.g. tranformations to normality and ANOVA), and suggest an identity link function and either an inverse Gaussian or gamma distribution. In this analysis, an inverse Gaussian with an identity link was used and a similar method is used for analysing response time data throughout this thesis. The results are shown in Table 3.1.

The results indicate that there are significant interactions: frequency-tempo and frequency-location. A post-hoc analysis of contrasts with a Tukey p-value adjustment (using the *emmeans* package in R – Lenth 2019) shows that, for low frequency noise, there are no significant differences between locations. However, for high frequency stimuli, there are significant differences between some pairs of locations: back and front ($MD = -0.10$, $p = 0.001$), back and right-front ($MD = -0.11$, $p = 0.0003$), and

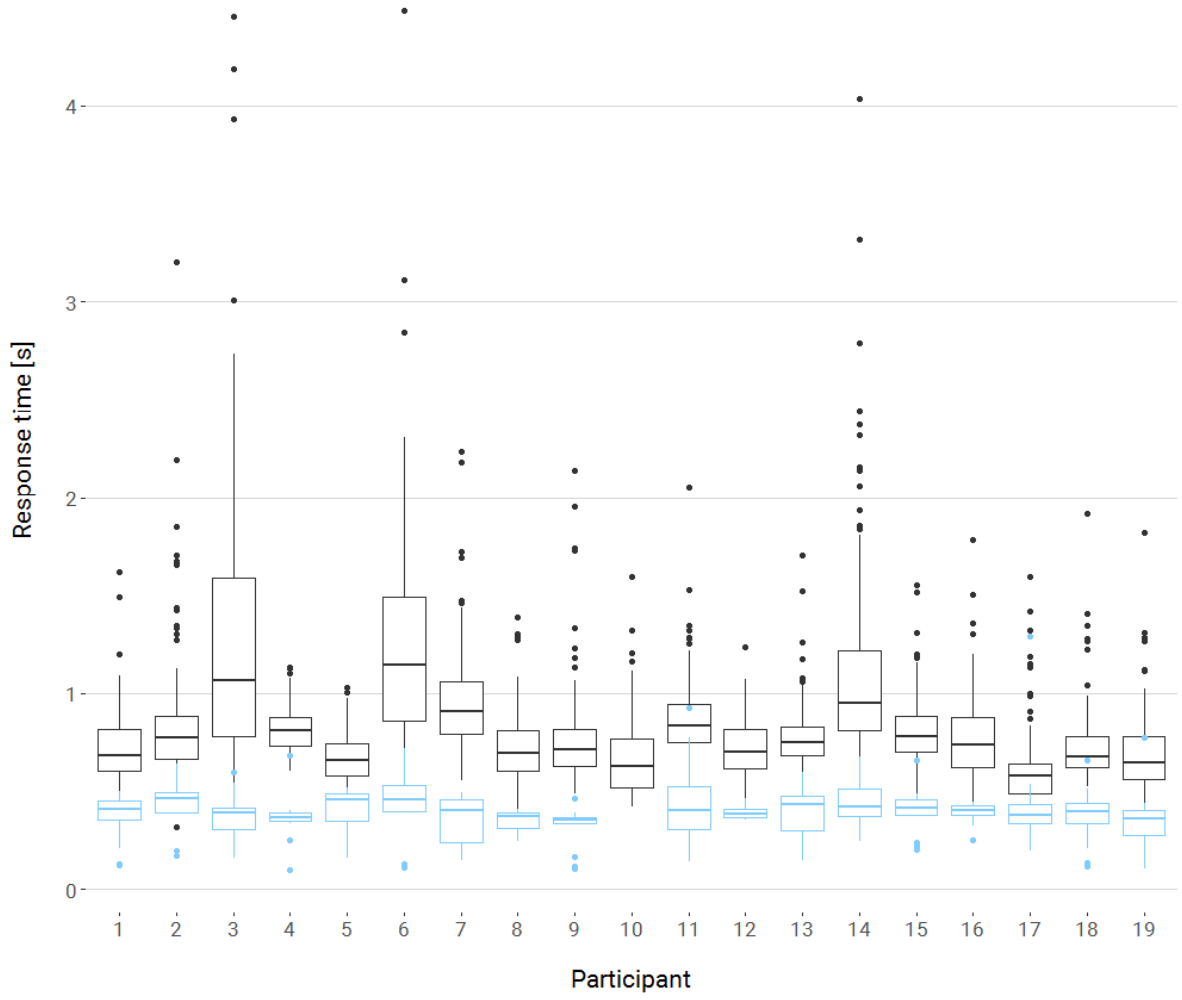



**Figure 3.3.:** Response times of individual participants. The box-plots show medians, 25th and 75th percentiles of response times. Black: main experiment, blue: baseline experiment (not available for participant number 10)

back and right-back ($MD = -0.09$, $p = 0.006$). Figure 3.4 shows estimated mean response times for locations vs frequency. For high frequency stimuli, response times were also on average 67 ms lower for fast compared to slow patterns ($p < 0.0001$). Figure 3.5 shows the interaction plot for frequency vs tempo.

As could be expected from the differences in average RT between participants, the

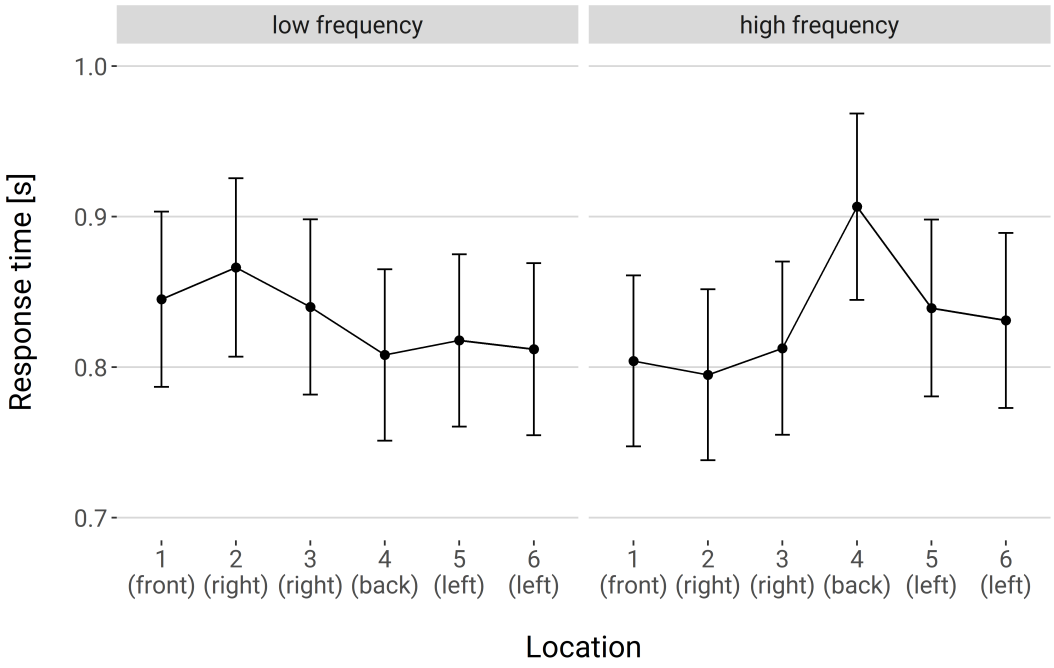



**Figure 3.4.:** Response time marginal means with 95% confidence intervals estimated from the model in Table 3.1, for different locations. Left: low frequency, right: high frequency.

random effect for participants varied quite significantly (standard deviation of about 100 ms). By including 'participant' as a random effect and allowing the intercept to vary across participants, each individual is in effect assigned a different baseline response time.

### 3.3.2. Accuracy

Accuracy data was binary: each response was either correct or incorrect. Following the analysis in Tordini et al. (2013), to discard late responses, a personalised acceptance window was calculated based on the baseline condition. The goal was to remove guesses and only consider correct responses where a participant was attending the target stream. In the experiment by Tordini et al. (2013), statistics of

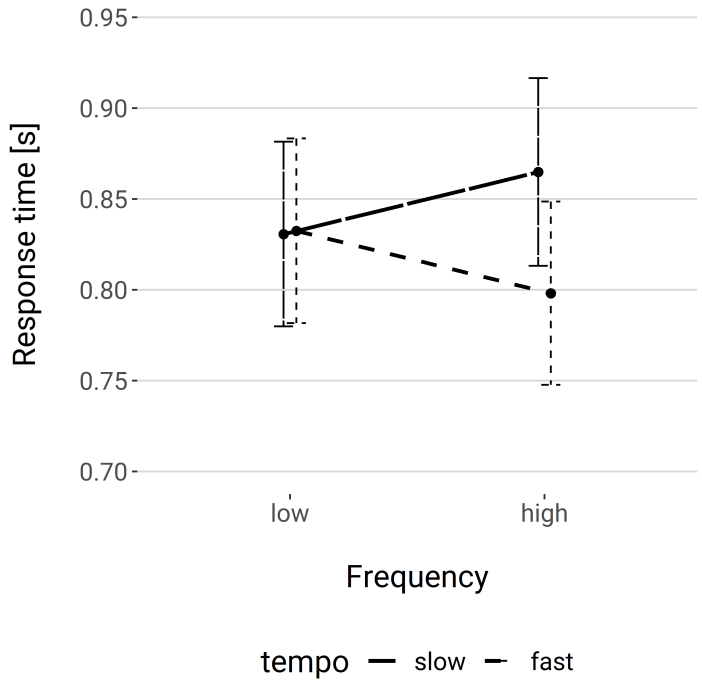



**Figure 3.5:** Response time marginal means with 95% confidence intervals estimated from the model in Table 3.1, for different noise frequency and tempo.

baseline response times (such as the 90th percentile) were used directly to set the upper limit for each participant. However, in the experiment described here, mean differences between baseline and main conditions were rather large, and for some participants response distributions from those two conditions did not even overlap (see Figure 3.3). This could be due to increased cognitive difficulty of the main task compared to the task by Tordini et al. (2013) – perhaps making high/low frequency judgements requires more decision time than the more natural left/right judgements.

However, the baseline data can still be used to discard late responses if those cognitive effects are nullified by aligning baseline and main time response distributions so that they medians are equal. The baseline condition should still be easier and have fewer late responses, because attention is always directed towards the target stream (which is an ideal condition). Therefore, responses within a time window corresponding to that of the baseline condition should indicate cases in which participants were actually attending the target stream.

Based on this, the baseline and main distributions were time aligned so that their medians were equal, and then the upper limit of the acceptance window was set at the 95th percentile of the baseline data (for an example see Fig. 3.6). All responses outside of this window were considered incorrect. This procedure caused on average 27% of each participant's correct responses to be marked as incorrect. For one participant baseline data was not available, so they were discarded from this analysis (leaving N=18).

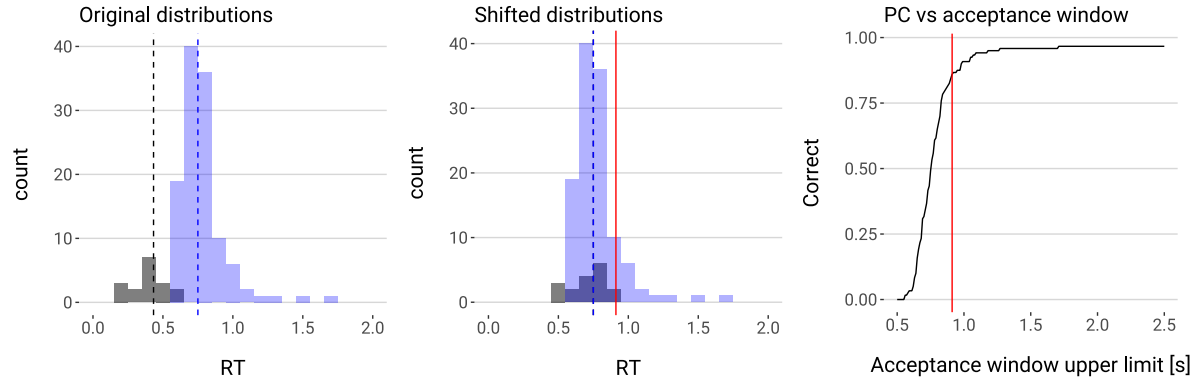



**Figure 3.6.:** Effect of applying acceptance window for participant 12.
Left: original baseline (grey) and main (purple) distributions of response times. Middle: shifted baseline distribution. Dashed lines represent distribution medians.
Right: influence of acceptance window length on the resulting proportion of correct responses. Red solid lines show the acceptance window upper limit used.

Again, a GLMM was fitted, in this case with a binomial distribution and logit link function, which is appropriate for a dichotomous response variable. Fixed and random effects in the model were the same as in RT analysis above. Results are similar to those obtained from analysis of reaction times, with two significant interactions: frequency/tempo and frequency/location (see Fig. 3.7).

Reflecting a result from response times, for high frequency sounds, participants were 2.3 times ($p < 0001$) more likely to be correct for fast rather than slow patterns, while

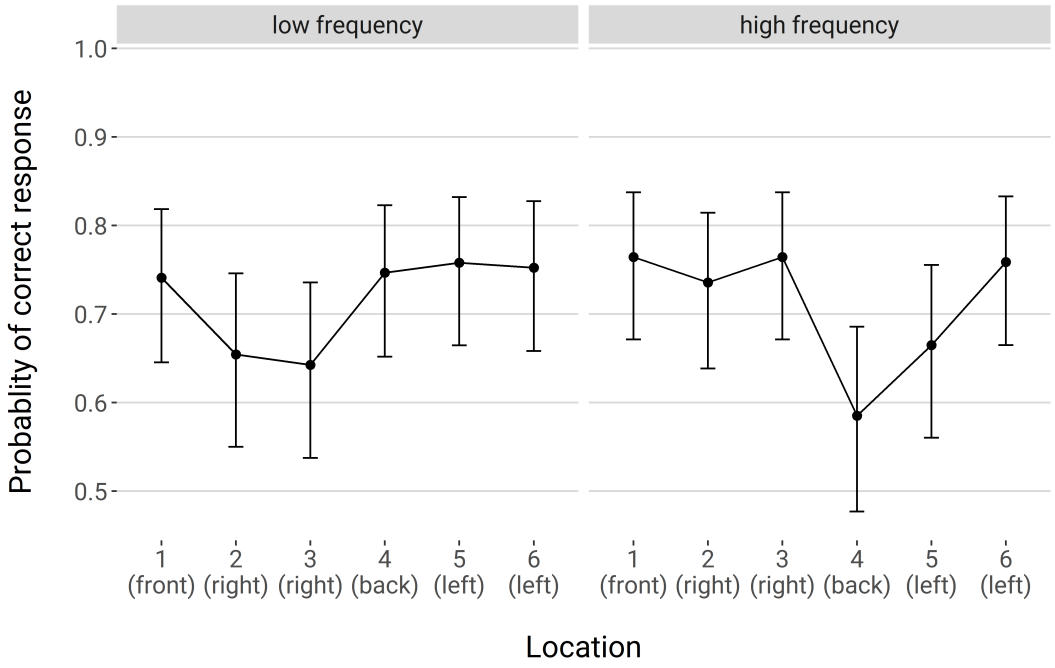



**Figure 3.7.:** Probabilities of correct response (with 95% confidence intervals) estimated from the model in Table 3.2, for different spatial locations and stimulus frequency.

for low frequency noise the difference was not statistically significant. Tordini, Bregman and Cooperstock (2016) also found that oddball inter-stimulus intervals were more often correctly detected in faster streams.

Finally, further analysis of the location-frequency interaction reveals that, for high frequency noise, the rear location was significantly less likely to get a correct response than 4 other locations: front ($OR = 2.3$, $p = 0.007$), right-front ($OR = 1.97$, $p = 0.045$), right-back ($OR = 2.3$, $p = 0.007$), left/front ($OR = 2.23$, $p = 0.010$).

| Fixed effects | Est. [s] | SE | t value | p-value | |
|---|---|---|---|---|---|
| (Intercept) | 0.844 | 0.031 | 27.67 | < 0.0001 | *** |
| Location 2 | 0.021 | 0.026 | 0.83 | 0.408 | |
| Location 3 | −0.005 | 0.025 | −0.20 | 0.839 | |
| Location 4 | −0.037 | 0.024 | −1.53 | 0.125 | |
| Location 5 | −0.027 | 0.024 | −1.12 | 0.261 | |
| Location 6 | −0.033 | 0.024 | −1.37 | 0.171 | |
| Tempo (fast) | 0.002 | 0.014 | 0.13 | 0.898 | |
| Frequency (high) | −0.007 | 0.026 | −0.25 | 0.801 | |
| Tempo:Frequency | −0.007 | 0.020 | −3.40 | 0.0007 | *** |
| Location2:Frequency | −0.030 | 0.034 | −0.88 | 0.378 | |
| Location3:Frequency | 0.014 | 0.034 | 0.40 | 0.693 | |
| Location4:Frequency | 0.139 | 0.036 | 3.91 | < 0.0001 | *** |
| Location5:Frequency | 0.062 | 0.034 | 1.82 | 0.069 | |
| Location6:Frequency | 0.060 | 0.034 | 1.77 | 0.078 | |
| Random effect: Participant | | | | | |
| Number of groups | | | 19 | | |
| Standard deviation | | | 0.104 | | |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

**Table 3.1.:** Results of a GLMM model on response time data (link function: identity, family: inverse Gaussian). Formula used in the model: RT $\sim$ 1 + Location + Tempo + Frequency + Tempo$*$Frequency + Location$*$Frequency + (1 | Participant)

| Fixed Effects | Est. | SE | Z value | p-value | |
|---|---|---|---|---|---|
| (Intercept) | 1.05 | 0.241 | 4.36 | < 0.0001 | *** |
| Location 2 | −0.41 | 0.235 | −1.76 | 0.079 | |
| Location 3 | −0.47 | 0.235 | −1.98 | 0.048 | * |
| Location 4 | 0.03 | 0.243 | 0.12 | 0.903 | |
| Location 5 | 0.09 | 0.245 | 0.37 | 0.713 | |
| Location 6 | 0.06 | 0.244 | 0.25 | 0.807 | |
| Tempo (fast) | −0.00002 | 0.138 | 0.00 | 1.000 | |
| Frequency(high) | −0.28 | 0.263 | −1.08 | 0.280 | |
| Tempo:Frequency | 0.82 | 0.197 | 4.14 | < 0.0001 | *** |
| Location2:Frequency | 0.26 | 0.341 | 0.76 | 0.446 | |
| Location3:Frequency | 0.47 | 0.343 | 1.35 | 0.176 | |
| Location4:Frequency | −0.86 | 0.341 | −2.53 | 0.011 | * |
| Location5:Frequency | −0.58 | 0.344 | −1.69 | 0.091 | |
| Location6:Frequency | −0.09 | 0.349 | −0.26 | 0.794 | |
| Random effect: Participant | | | | | |
| Number of groups | | 18 | | | |
| Standard deviation | | 0.655 | | | |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

**Table 3.2.:** Mixed effects generalized linear regression results on accuracy data (distribution: binomial, link function: logit). The estimates shown are on the log scale. Formula used in the model: Correct ∼ 1 + Location + Tempo + Frequency + Tempo∗Frequency + Location∗Frequency + (1 | Participant).

## 3.4. Summary

The experiment described in this chapter tested auditory salience in an oddball detection paradigm, in which two streams of repeating stimuli were used. High frequency sounds were found to be less accurately detected when they were behind the listener, than when they were in front. An interaction was also found between stimulus frequency and pattern tempo. These results are discussed in light of existing research in Chapter 13.

# 4

# Higher level attention

## 4.1. Introduction

One of the shortcomings of the experiment in Chapter 3 was that the stimuli were simple, synthetic sounds. Although this allowed for straightforward manipulation of the sound, it could be argued that the perception of and responses to those stimuli do not accurately represent everyday listening situations. The goal of the second experiment was to test spatial salience in a more ecologically valid scenario.

## 4.2. Method

The experimental procedure was inspired by Huang and Elhilali (2017), who tested salience of sound events in two competing scenes. The participants heard one scene in each ear, and were asked to continuously indicate which one they were focusing on. For that, they used a mouse and a visual interface like in Figure 4.1.

A similar procedure was used here, but with stimuli arriving from different locations

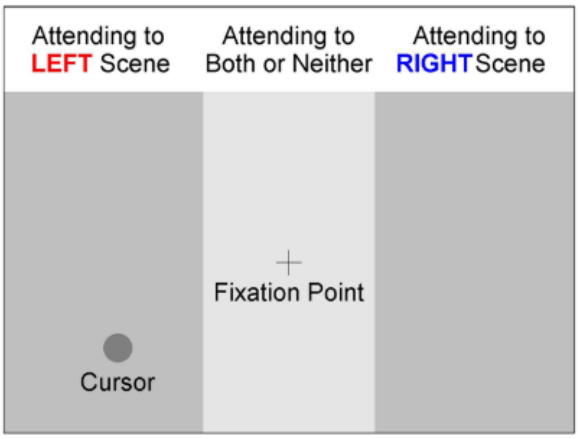



**Figure 4.1:** Graphical interface used in the experiment of Huang and Elhilali (2017). Participants were asked to move their mouse to the left or right area, depending on which scene they were attending to.

all around the listener instead of just left and right. Additionally, it can be argued that the situation would be more realistic if competition for attention was between sound *events*, rather than full scenes, presented dichotically. Therefore, different locations in this experiment did not correspond to different scenes, but rather to events. Similarly to Huang and Elhilali (2017), the participants were asked to indicate, in real time, to which location in the scene their attention was directed. To do that, they used a joystick, and no visual display was provided, partly to avoid forcing participants to focus their attention on a display in front of them. Participants were allowed to move their heads slightly, but were reminded to indicate the location of the sound in relation to the room, rather the direction they were facing.

The experiment by Huang and Elhilali (2017) used recordings of different types of existing sound scenes. However, using recordings of full scenes would make manipulation of experimental variables difficult, so here the scenes were designed from individual sounds instead. They consisted of a steady background and two types of events: distractors and targets. The experiment checked how often participants paid attention to targets, while responses to distractors were not analysed (they were effectively treated as part of the background). The position in time of distractors was randomised but was the same for all participants. The position of targets was randomised for each participant separately, in an attempt to average out

any interactions between specific distractors and targets.

The experiment was a full-factorial repeated-measures design with the following independent variables:

- loudness (2 levels)
- spectral centroid (2 levels)
- location (4 levels)
- semantic category (3 levels)
- background (2 levels)

This results in 96 different conditions. Because habituation to a particular sound might make it less salient (as it is less surprising), it was crucial not to use the same stimulus more than once. For this reason, 96 different sound events were used as targets.

Because this design relies on accurate localisation of targets, a baseline experiment was conducted directly after the main experiment, with the same target stimuli and the same reproduction method, but with no background or distractors. The participants were asked to indicate which direction each target was coming from, as soon as they heard it, and to return to the centre after the sound was over. This allowed collection of baseline data which indicated individual localisation accuracy.

### 4.2.1. Reproduction system

The stimuli were reproduced over a 2nd order ambisonic system, using the Higher Order Ambisonic Library Matlab toolbox (Politis 2016). The reproduction system was 8 loudspeakers arranged in an octagon, at ear-level (see Fig. 4.3). The background was not ambisonic but rather an 8-channel signal sent directly to the loudspeakers. All sounds were reproduced with 44100 Hz sampling frequency.

### 4.2.2. Target sounds

Targets were short clips from recordings of real-world sounds (from Font, Roma and Serra (2013), *BBC Sound Effects Library* (2018) and Xeno-canto (n.d.)), on average 3 seconds long. Time spacing between consecutive stimuli varied randomly from 2 to 4 seconds. The stimuli belonged to three different semantic categories, which were determined based on the soundscape taxonomy established in a sorting experiment by Bones, Cox and Davies (2018). The categories were: *nature* (subcategory: *animals*), *people* (subcategory: *voices*, which did not include speech), and *manmade* (subcategory: *industrial*). All target sound clips used in this experiment are listed in Appendix A.

The spectral centroid represented an objective measure of the perceived brightness of the sound, and was calculated as:

$$SC = \frac{\sum_{n=1}^{N} f(n)Y(n)}{\sum_{n=1}^{N} Y(n)} \tag{4.1}$$

where $Y(n)$ is the amplitude of the $n^{th}$ bin of the spectrum, and $f(n)$ is the centre frequency of that bin. To avoid any artefacts that come with filtering, and the risks of unnatural sounding stimuli, sound spectra were not manipulated directly. Instead, sound events were chosen so that their spectral centroid falls within one of two groups: 1000-2500 Hz or 4000-5500 Hz. Recordings were chosen not to have significant background noise that could influence the value of the spectral centroid.

The short-term loudness of sound was calculated using the Dynamic Loudness Model (Chalupper and Fastl 2002) available through the PsySound3 toolbox in Matlab (Cabrera et al. 2008). Rennies, Verhey and Fastl (2010) found this model equally good as the loudness model for time-varying sounds by Glasberg and Moore (2002), and 'slightly better' for spectrally varying sounds. It is also significantly more time efficient. As an indication of the loudness of each sound, the maximum of time-smoothed short-term loudness was used (STL window = 2 ms, smoothing window = 100 ms). Sound level was manipulated to create two levels with loudness

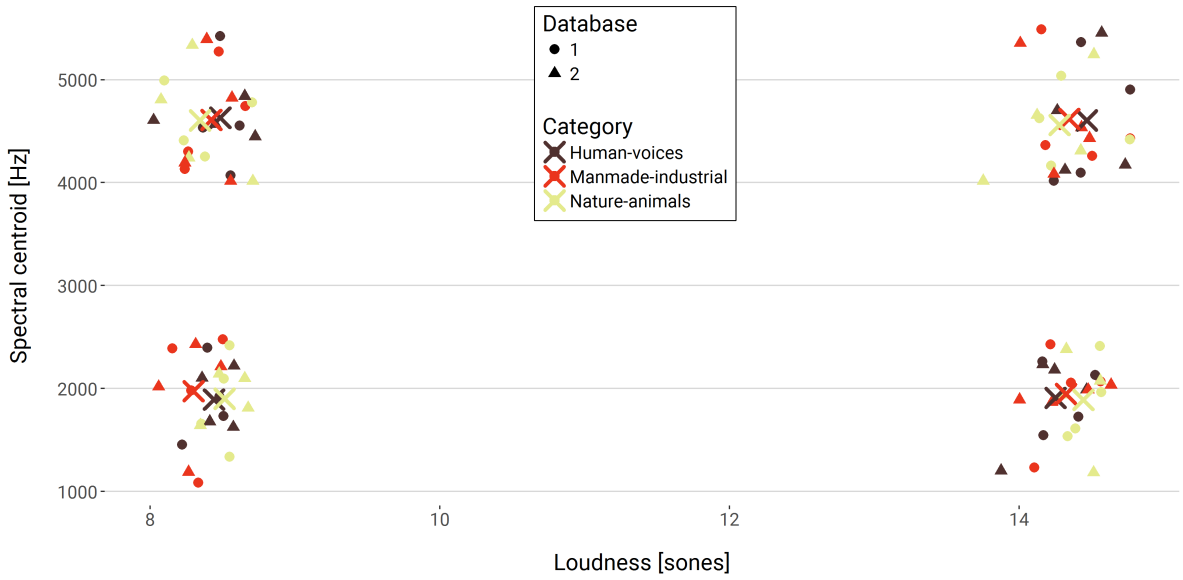
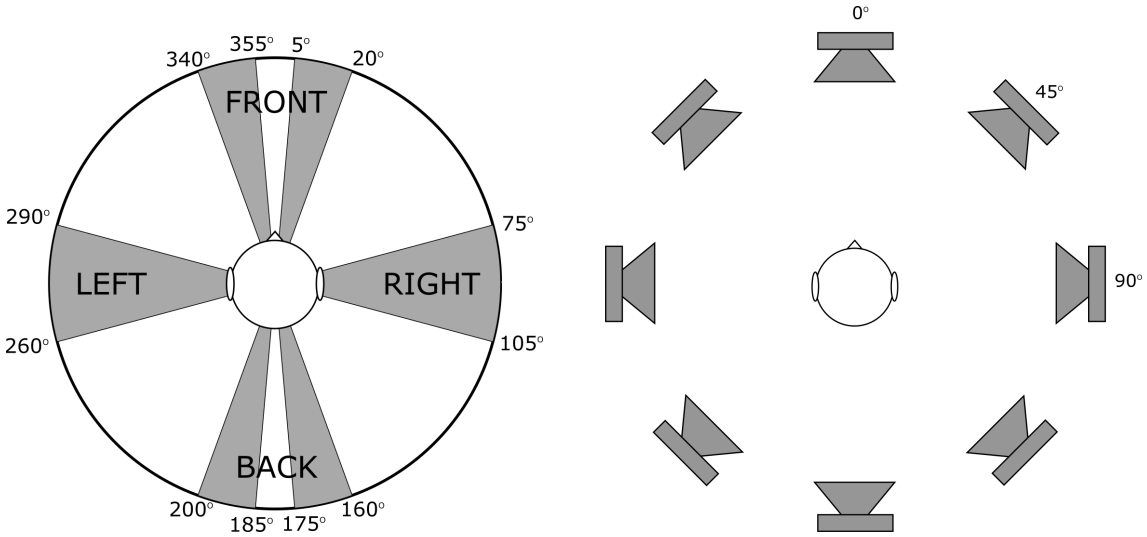



**Figure 4.2.:** Stimuli used in the experiment. Database corresponds to two stimuli groups, used with different backgrounds. Colors indicate one of the three semantic categories. Recordings were chosen to fall within the two spectral centroid levels, and then their loudness was manipulated, while keeping the pairs of brightness groups as similar as possible.

means 8.4 and 14.4 sones, and standard deviation of 0.2 sones. These two levels correspond to the loudness of a 1kHz tone at about 70 and 78 dB SPL. Each sound was assigned to either one of the two levels in a way that minimised mean and variance differences between brightness levels. Fig. 4.2 shows all targets on the loudness-brightness spectrum.

Targets were positioned in one of four 30° areas (cones in Figure 4.3a) around the listener: front, back, right and left. The exact location of stimuli varied randomly within these areas. The choice of cone width was guided by a trade-off: on one hand, it would be best to avoid the borders between areas (e.g. 45° front/right border), where small localisation errors would be more problematic. On the other hand, from the perspective of scene realism, the cones should be wide enough so that the targets do not always appear at the exact same location. Additionally, 10° cones around the

front and back locations were excluded (see Figure 4.3a). The location of each target was determined randomly for each participant, while keeping the number of targets in each area equal. Elevation was always the same, at approximately ear level.




**(a)** Grey areas show cones where target sounds were localised

**(b)** Loudspeaker positions; distractors and background were played from all loudspeakers.

**Figure 4.3.:** Target locations and experimental setup.

### 4.2.3. Scenes

Participants were presented with two different sound scenes, each about 5 minutes long, each with a different background sound and distracting events. Targets were divided into 2 balanced groups (this is represented by different shapes in Figure 4.2) and each group was played over one of the backgrounds. The 2 targets/backgrounds combinations, as well as the order of the scenes, were randomised between participants.

In the first scene ("speech"), the background was steady babble noise with distracting

louder speech excerpts, recorded by Al Noori, Duncan and Li (2017). The speech was present in all 8 loudspeaker channels, with an equal number of male and female speakers in each channel. The speech clips were 5 seconds long, with on average 2 seconds of silence in-between, in each channel. Most of the time, there was more than one talker present at the same time, but never in the same channel. The speech was in 9 different languages and participants were asked about their knowledge of these languages in a questionnaire after the test (12 reported no knowledge of these languages whatsoever, 5 – knowing a few words in one of them, 1 – knowing a few words in 6 of them; no one reported knowing any of the languages well). The speech was originally recorded with 16000 Hz sampling frequency. For reference, the spectrum of the "speech" background is shown in figure 4.4.

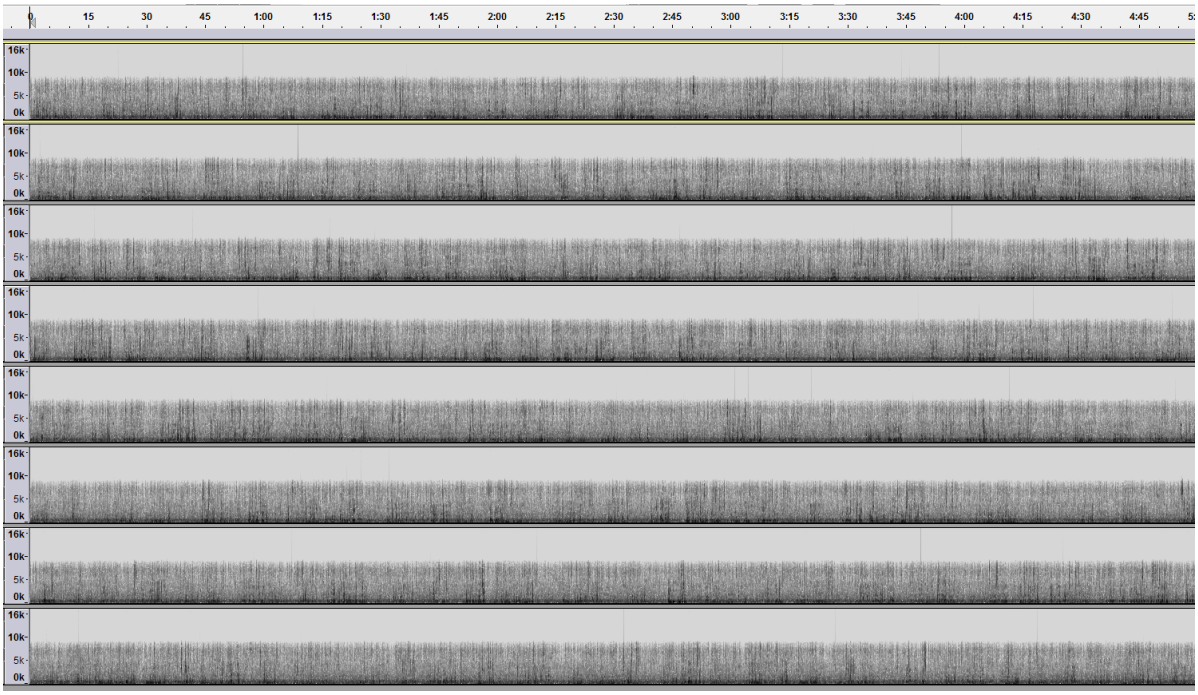



**Figure 4.4.:** Spectrograms of the "speech" background. Each row represents one channel. The horizontal axis represents time in seconds, and the the vertical axis shows frequency in Hz.

The other scene ("nature") had a steady wind sound as background, and distracting sound events from the semantic category "nature", but different subcategories than

the targets: 48 were sounds of insects, 32 – leaves and branches, and 16 – water, all spread evenly across all 8 loudspeaker channels. These distractors were distributed over the background in a similar manner as the target events, with one or two distractors present at any given time, and 2-4 s breaks in-between. Some, but not all distractors overlapped with targets. Average background loudness was 4.3 sones, and average distractor loudness was 11 sones. The spectral range of distractors was quite wide, ranging from 780 Hz to 13600 Hz (median: 4838 Hz). Figure 4.5 shows the spectrum of this background.

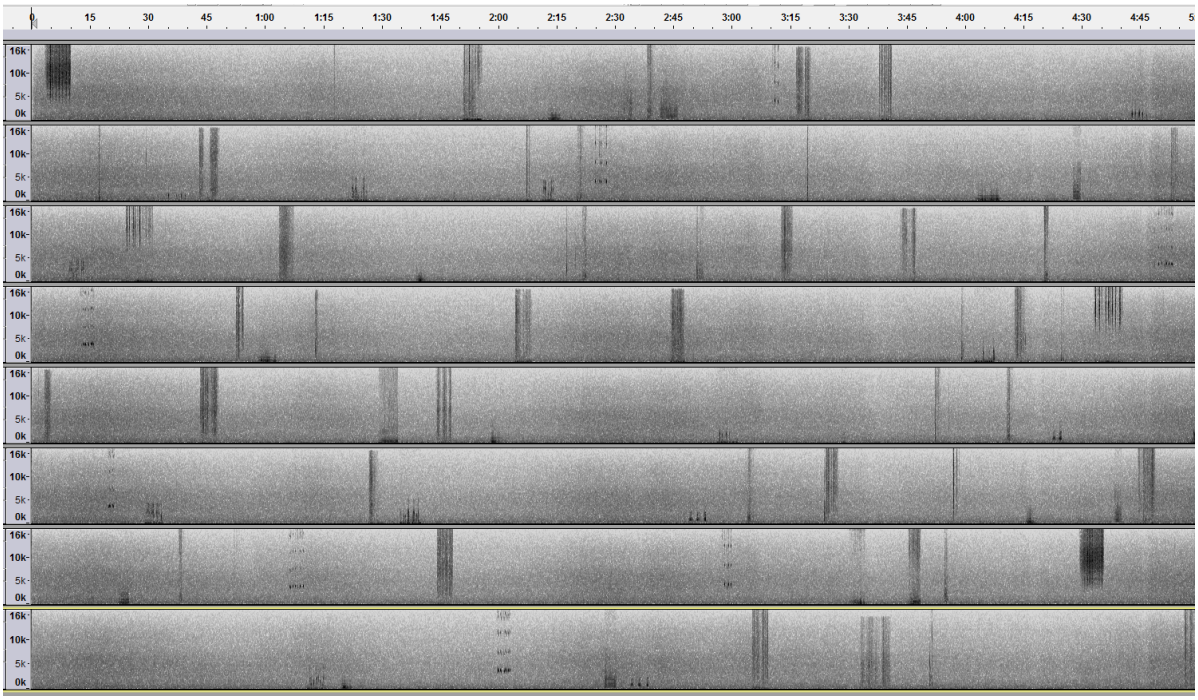



**Figure 4.5.:** Spectrograms of the “nature” background. Each row represents one channel. The horizontal axis represents time in seconds, and the the vertical axis shows frequency in Hz.

## 4.3. Results

15 volunteers took part in the experiment, 8 male and 7 female, mean age = 28.3, 13 right-handed and 2 left-handed.

### 4.3.1. Data preprocessing

Figure 4.6 shows an example of raw data collected from the joystick movements of one of the participants in the baseline experiment.

A target event was considered attended to (a "hit") if, within a certain time window (acceptance window), the joystick was in the quadrant of the event. Thus, two things needed to be decided: limits of the acceptance window and the size of each quadrant. Both were determined from the baseline experiment.

No participants responded within the first 400 ms of any event, so this value was chosen as the lower limit of the acceptance window. We assume this to be the minimum time required for the cognitive and motor functions necessary to give a response in this setting. The upper limit of the window was set to 2 s, with which all participants were very close to their best localisation performance. A longer window could overlap with subsequent targets, and a shorter one would miss a larger portion of the attended events.

The joystick area was divided into quadrants, each including one of the areas where targets were present, and also allowing for localisation errors around these areas (analysis quadrants were 90° wide, while target areas – only 30°). Because participants were instructed to keep the joystick in the centre if they were unsure what they were listening to, this area had to be removed from analysis. Analysis of joystick movements in the baseline experiment showed that the result is not very sensitive to

the size of the central area (until it becomes close to the size of the whole joystick area). Figure 4.6 shows the chosen centre area and response quadrants.

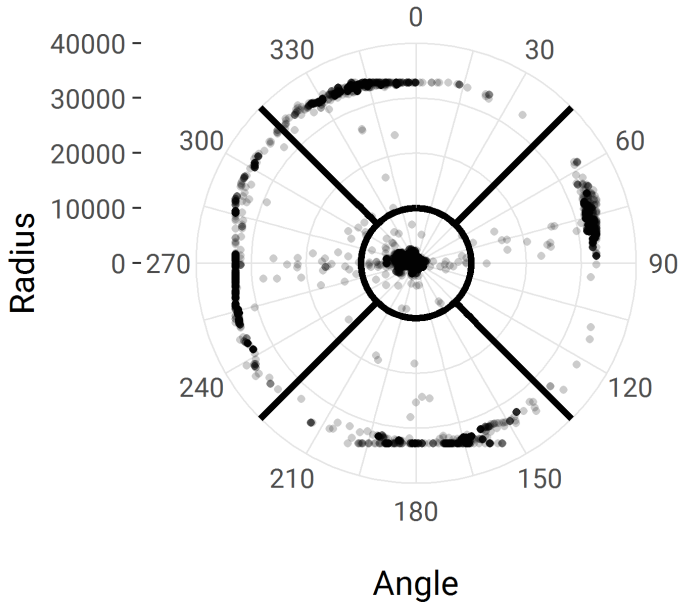



**Figure 4.6:** Raw joystick movement data for one of the baseline experiment participants. Each dot is one joystick position sampled at regular time intervals. The dots are partly transparent, so the darker the region, the more data points there are. Solid black lines show how the space was divided into quadrants and the centre area.

### 4.3.2. Localisation errors

Average localisation accuracy in the baseline experiment varied from 68% to 100% between participants, indicating that, despite removing direct front and back locations from playback, localisation errors were still an issue. This accuracy was different for different sound locations, on average: 79% for the front, 81% for the back, and 99% for left and right. As was expected, the main difficulty lied in localising sounds positioned in the front and back, while sounds on the left and right were localised almost perfectly. The data were analysed with Generalised Linear Mixed Model (GLMM) regression, with a binomial distribution and logit link (a mixed logistic regression). Using *participant* as a random variable made it possible to analyse experimental conditions isolated from differences between participants. The statistical

model confirmed that none of the other factors (loudness, brightness, category) had an effect on localisation accuracy – see Table 4.1.

| Fixed effects | Est. | SE | Z | p-value | |
|---|---|---|---|---|---|
| (Intercept) | 1.88 | 0.37 | 5.02 | <0.0001 | *** |
| Channel - right | 4.09 | 0.72 | 5.67 | <0.0001 | *** |
| Channel - back | 0.16 | 0.20 | 0.81 | 0.416 | |
| Channel - left | 4.79 | 1.01 | 4.76 | <0.0001 | *** |
| Loudness - loud | −0.14 | 0.20 | −0.70 | 0.483 | |
| Brightness - high | 0.02 | 0.20 | 0.10 | 0.920 | |
| Category - manmade | −0.24 | 0.24 | −0.99 | 0.322 | |
| Category - nature | −0.24 | 0.24 | −0.99 | 0.322 | |
| Background - nature | −0.14 | 0.20 | −0.70 | 0.483 | |
| Random effect: Participant | | | | | |
| Number of groups | | | 15 | | |
| Standard deviation | | | 0.98 | | |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

**Table 4.1.:** Mixed effects generalized linear regression results on localisation accuracy data in the baseline experiment (distribution: binomial, link function: logit). The estimates shown are on the log scale. Formula used in the model: Correct ∼ 1 + Channel + Loudness + Brightness + Category + Background + (1 | Participant).

These localisation errors will likely influence the main experiment responses as well. The following section discusses how these these errors could be disentangled from effects of attention and distraction.

### 4.3.3. Main experiment

The total percentage of target sounds attended varied among participants, with an average of 64% and a standard deviation of 10%.

To study the effects of experimental variables on the hit/miss responses, data from the baseline and main experiments were pooled together, forming a new variable in the analysis – experiment type. By looking at interactions between the experiment type and other variables, it can be seen if adding distracting sounds – in other words, introducing attentional effects – had an effect on any of these variables.

Again, a Generalised Linear Mixed Model (logit link, binomial distribution) was fitted with *participant* as a random effect, and 2-way interactions between the experiment type and the other independent variables (loudness, brightness, location, category and background type). The results are shown in Table 4.3. Wald tests reveal significant interaction effects between experiment type and loudness, and between experiment type and location.

Analysis of contrasts confirms that participants were 1.7 times more likely to attend to loud than to quiet targets in the main experiment ($p < 0.0001$), while no effect is observed in the baseline. This is to be expected, as louder sounds will be more salient, but loudness should not affect localisation.

Comparison of contrasts between different locations shows the same significant differences for main and baseline experiments: front/right, front/left, back/right, back/left. These differences appear to be mainly due to localisation errors. All of these effects, however, are smaller for the main experiment than the baseline. The effect of experiment type on responses to different locations can be seen on Figure 4.7. Clearly, the 'hit rate' in the main experiment is generally lower than in the baseline, because in the former, participants were not asked to attend to target sounds and there were distractors. The general trend looks similar in both experiments, with more 'hits' to the sounds on the right and left, and fewer for front and back.

To see if there were any interactions between independent variables, the main experiment data was analysed separately from the baseline data. A GLMM model

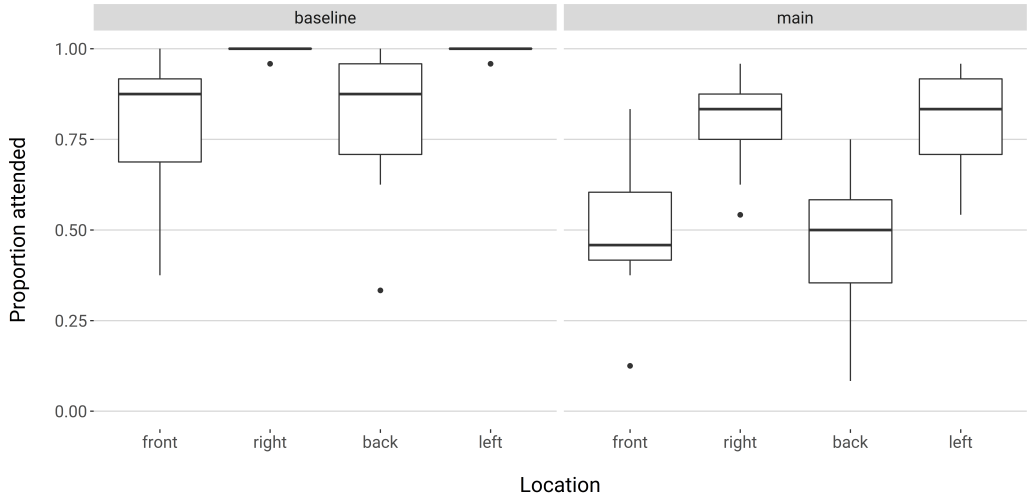



**Figure 4.7.:** Responses to sounds in different positions for the baseline localisation experiment (left panel) and the main experiment (right panel). Boxplots show the medians, 25th and 75th percentiles of hit scores calculated for a particular condition and for each participant.

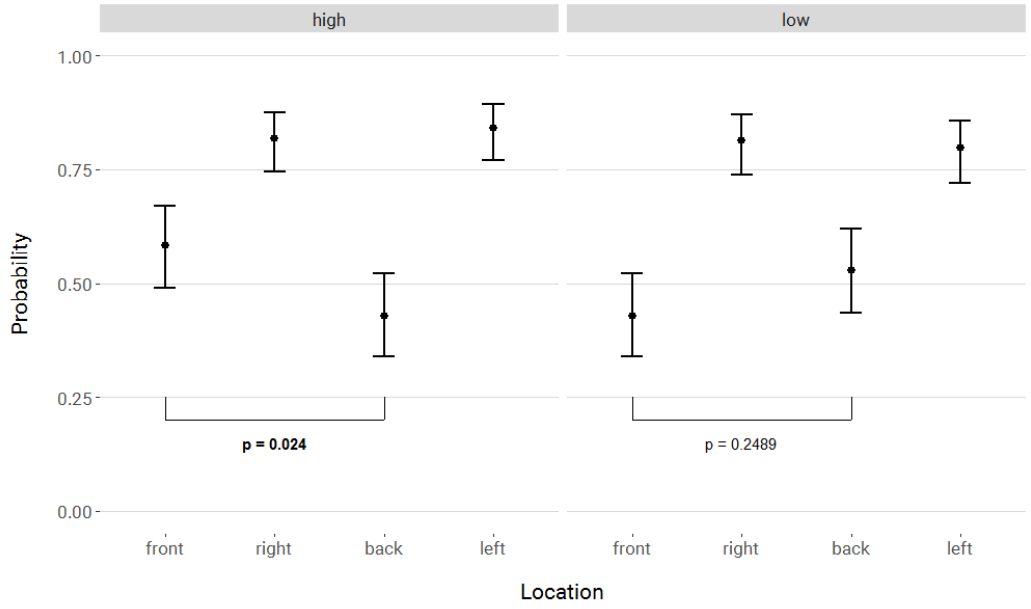



**Figure 4.8.:** Probability of attending to sounds in different positions in the main experiment estimated from the model in Table 4.2, split by brightness of the sound. Error bars show 95% confidence intervals. Based on model in Table 4.2.

with the best fit based on the Akaike information criterion (AIC) included one interaction: location/brightness (see Table 4.2). The model indicated that brightness significantly changed responses to front and back locations. Analysis of contrasts shows that in the main experiment, although no significant differences were found for low brightness targets in front and back, there is a significant difference between high brightness targets presented in front and back locations, with sounds in front being more salient – see Figure 4.8.

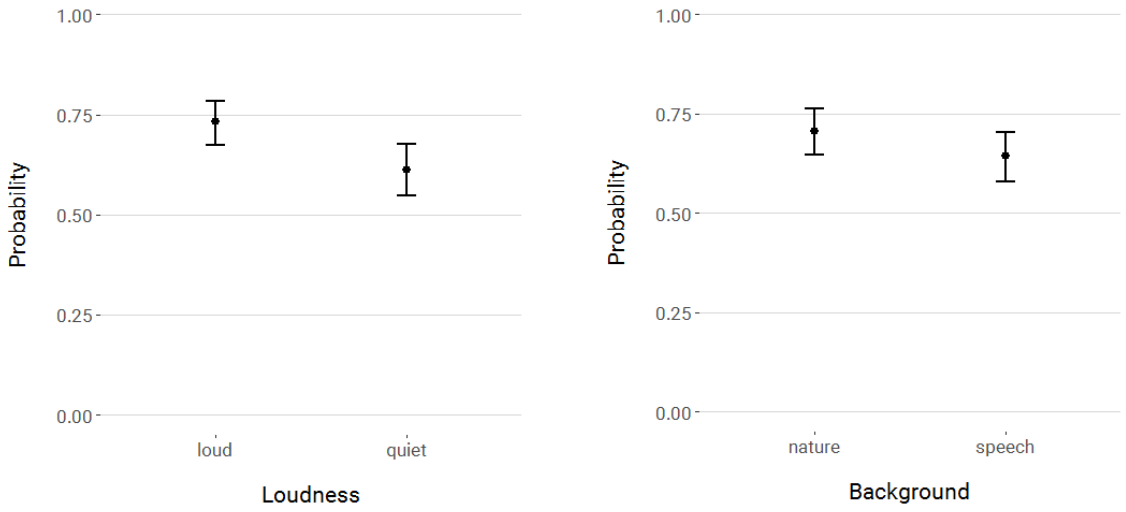
| Fixed effects | Est. | SE | Z | p-value | |
|---|---|---|---|---|---|
| (Intercept) | −0.89 | 0.23 | −3.89 | <0.0001 | |
| Location - right | 1.76 | 0.25 | 7.12 | <0.0001 | *** |
| Location - back | 0.41 | 0.22 | 1.85 | 0.064 | |
| Location - left | 1.66 | 0.24 | 6.81 | <0.0001 | *** |
| Brightness - high | 0.62 | 0.22 | 2.83 | 0.005 | ** |
| Loudness - loud | 0.55 | 0.12 | 4.54 | <0.0001 | *** |
| Category - manmade | 0.32 | 0.15 | 2.14 | 0.033 | * |
| Category - nature | 0.22 | 0.15 | 1.47 | 0.142 | |
| Background - nature | 0.29 | 0.12 | 2.41 | 0.016 | * |
| Location-right:Brightness | −0.59 | 0.35 | −1.68 | 0.094 | |
| Location-back:Brightness | −1.03 | 0.31 | −3.31 | 0.001 | ** |
| Location-left:Brightness | −0.32 | 0.35 | −0.92 | 0.357 | |
| Random effect: Participant | | | | | |
| Number of groups | | | 15 | | |
| Standard deviation | | | 0.43 | | |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

**Table 4.2.:** Coefficient estimates, standard errors, Z statistics and p-values of the interactions in a GLMM model (binomial distribution, logit link function) fitted with main experiment data. The estimates shown are on the log scale. Model formula: Correct ∼ 1 + Channel + Brightness + Loudness + Category + Background + Channel∗Brightness + (1 | Participant).

The model also confirms a significant main effect of loudness ($p < 0.0001$), and

suggests that there is a significant effect of background type, with higher probability of attending to targets in the *nature* background. This is perhaps not surprising, as compared to *speech*, the *nature* background was more sparse, with less activity from sound sources. Figure 4.9 shows both of these effects.




**(a)** Effect of loudness

**(b)** Effect of background type

**Figure 4.9.:** Probability of attending to target sounds in the main experiment, based on model in Table 4.2. Error bars show 95% confidence intervals.

There are significant differences between sound categories, with the manmade/industrial category being more likely to be attended to than the muman/voices category. Note, however, that after a Tukey p-value correction, none of the pairwise contrasts between the three categories are statistically significant. A difference between categories could point to an influence of semantic meaning on salience. However, it is worth keeping in mind that, while the targets were balanced on the loudness and brightness scales, there might be other acoustic properties of the sounds which vary between the categories.

| Fixed effects | Est. | SE | Z | p-value | |
|---|---|---|---|---|---|
| (Intercept) | 1.66 | 0.29 | 5.38 | < 0.0001 | |
| Channel - right | 3.91 | 0.72 | 5.42 | < 0.0001 | *** |
| Channel - back | 0.15 | 0.19 | 0.77 | 0.444 | |
| Channel - left | 4.61 | 1.01 | 4.56 | < 0.0001 | *** |
| Loudness - loud | −0.12 | 0.19 | −0.66 | 0.509 | |
| Brightness - high | 0.02 | 0.19 | 0.09 | 0.925 | |
| Category - manmade | −0.22 | 0.23 | −0.93 | 0.351 | |
| Category - nature | −0.22 | 0.23 | −0.93 | 0.351 | |
| Background - nature | 0.12 | 0.19 | 0.66 | 0.509 | |
| Experiment - main | −2.02 | 0.31 | −6.50 | < 0.0001 | *** |
| Experiment:Background | 0.17 | 0.22 | 0.74 | 0.458 | |
| Category-manmade:Experiment | 0.53 | 0.28 | 1.93 | 0.054 | |
| Category-nature:Experiment | 0.43 | 0.28 | 1.57 | 0.116 | |
| Brightness:Experiment | 0.11 | 0.22 | 0.51 | 0.613 | |
| Loudness:Experiment | 0.68 | 0.22 | 3.01 | 0.003 | ** |
| Location-right:Experiment | −2.43 | 0.74 | −3.27 | 0.001 | ** |
| Location-back:Experiment | −0.25 | 0.25 | −1.03 | 0.302 | |
| Location-left:Experiment | −3.11 | 1.03 | −3.03 | 0.002 | ** |
| Random effect: Participant | | | | | |
| Number of groups | | | 15 | | |
| Standard deviation | | | 0.51 | | |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

**Table 4.3.:** Coefficient estimates, standard errors, Z statistics and p-values of the interactions in a GLMM model (binomial distribution, logit link function) including 'experiment' as a variable (main vs baseline). Note that the main purpose of this model is to show how the main experiment interacted with other variables. The estimates shown are on the log scale. Model formula: Hit $\sim$ 1 + Channel + Loudness + Brightness + Category + Background + Experiment + Channel$*$Experiment + Loudness$*$Experiment + Brightness$*$Experiment + Category$*$Experiment + Background$*$Experiment + (1 | Participant)

## 4.4. Summary

The experiment described in this chapter tested auditory salience under a natural listening situation, in which participants reported their attention in real time. Loud sounds and sounds in the category *industrial* were found to be more salient. An interaction was also found between brightness of a sound and its spatial location. Further discussion of these results can be found in Chapter 13.

# 5

# Salience with perceptual load

## 5.1. Introduction

As discussed in Chapter 2, the salience of an auditory stimulus can interact with the perceptual load of the listener. When the load is low, it is likely that the majority of sounds, even if not very salient, will be noticed. On the other hand, when the load is high, and the listener has little free perceptual capacity, only the most salient stimuli will emerge. Therefore, it is possible that some salience differences are difficult to detect at low load levels. The following chapter investigates if any spatial salience effects arise with sufficiently high perceptual load.

This chapter describes a dual-task experiment in which participants were asked to perform two simultaneous tasks, prioritizing one of them (the primary task), while responses to the secondary task gave an indication of how salient the stimuli involved in the secondary task were. The higher stimulus salience, the easier it should be to detect it, even while engaged in the primary task.

Duangudom and Anderson (2013) proposed using an auditory dual-task experiment to determine the salience of the secondary stimulus, which was varied by changing

the strength of an amplitude modulation of the sound. In their case, the primary task involved counting low tones in a stream of high and low ones, and the secondary task was a detection task among interferers. While these authors varied the salience of the stimulus, here, it is the difficulty of the primary task that was varied, which influenced perceptual load under which participants are. It is expected that increasing perceptual load will make the target stimulus more difficult to detect. The two research questions asked in this chapter are:

1. Does the secondary target position influence how likely it is to be detected in a dual-task experiment, regardless of perceptual load?
2. Does increasing perceptual load lead to some differences in salience emerging between different spatial positions?

## 5.2. Method

The experiment was based on the method of Remington and Fairnie (2017), but changed to better reflect natural listening conditions. It is an auditory dual-task paradigm, where the primary task is to identify which one of the two known sounds is present in a scene (potentially among other irrelevant sounds), and the secondary task is a detection of a known stimulus.

The primary target stimuli were recordings of a motorcycle and a lorry, and one of them (but only one) was present in all of the scenes. The participants completed a training session before the experiment to make sure they could consistently differentiate between the two targets. Perceptual load was manipulated by adding task-irrelevant sounds: a plane, a drill, a dog barking, a bus and a train. There were 4 perceptual load conditions: 1, 2, 4 or 6 simultaneous sounds (including the primary target sound). All of these sounds were 3 s long and loudness-equalised.

In addition to this, 50% of the trials included a distinctive sound of an ice-cream van melody, which served as the secondary target. The level of this secondary target was adjusted individually between a few different sound levels (below or at the level of other sounds), so that it was the lowest sound level at which their performance for the secondary task only was at least 90%. This allowed for a level of normalisation to individual ability to detect the secondary target.

Scenes were reproduced over a 2nd order Ambisonic system, and the stimuli were placed around the listener as shown in Fig. 5.1, and on the same vertical level. Note that none of the sounds, including the secondary target, ever spatially overlapped with other sounds.

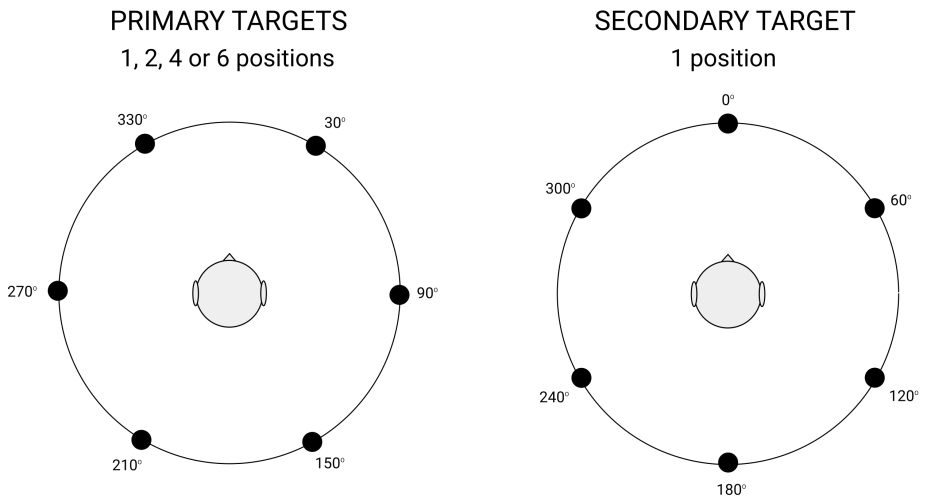



**Figure 5.1.:** Spatial positions of primary and secondary targets. If there were fewer than 6 primary sounds present, their locations were chosen randomly.

Trials were presented in blocks, and each block only contained trials from one perceptual load level. The order of the blocks was randomised and reverse counterbalanced (e.g. if a participant heard the block with 2 sounds first, followed by 1, 6 and 4, they then heard them these blocks again in reverse order: 4, 6, 1, 2).

The participants sat in an acoustically treated room, surrounded by loudspeakers, and

in front of a screen and a keyboard. On each trial, a 3 s long sound scene was presented, and a message on the screen instructed the participant to respond, as quickly as possible, if they heard a motorcycle or a lorry. They responded by pressing one of two buttons on a keyboard (each of the two buttons had a sticker with an image of either a motorcycle or a lorry). As soon as they pressed the button, the sound clip was stopped. Then, the second screen was shown, asking whether an ice-cream van sound was present or not. The screen stayed visible until the participant responded with a button press, at which point the experiment continued with the next trial. Participants were informed in advance of the two tasks, and had a chance to practice each one separately and both together. They were instructed that the primary task was more important and they should prioritize it. Finally, after the main experimental part, each participant performed a short "control" experiment, in which they listened to the same scenes, but were only performing the detection task (ice-cream van sound).

In addition to gathering behavioural responses and response times, pupil dilation responses were also measured. Introducing a physiological measure can offer a more direct way of measuring responses to sounds, which bypasses the need for conscious behavioural responses from participants. Pupil dilation responses have also been established as a measure of cognitive effort, so they could help to confirm the effect of adding sound sources on perceptual load.

25 volunteers took part in the experiment, mean age 24.8 (ranging between 18 and 46). Pupil dilation was measured for 15 participants.

## 5.3. Results

### 5.3.1. Behavioural data – increasing set size

First, it was important to confirm the assumption that increasing the number of sounds in a scene (i.e. increasing the *set size* variable) would cause an increase in participants' perceptual load. This can be done by analysing how increasing the *set size* changed task performance in the primary task.

Fig. 5.2 shows the response times and the proportion of correct responses in the primary task. There is a clear increase in response times and a decrease in the proportion of correctly identifying the primary target, which suggests that increasing the number of sounds in a scene did increase perceptual load for participants. This is confirmed by a Generalised Linear Mixed Model (GLMM), with response times (in milliseconds) as a dependent variable, *participant* as a random variable, and *set size* as an independent discrete variable with 4 levels, as shown in Table 5.1. An analysis of contrasts with a Tukey p-value correction reveals that response times for all pairs of set sizes were significantly different from each other (all $p < 0.0001$). A mixed-effects logistic regression with correct primary task response as the outcome variable (correct/incorrect), *participant* as a random variable, and *set size* as a discrete independent variable, shows a very similar pattern (see also Table 5.1). Analysis of contrasts, with a Tukey adjustment, shows that task performance for all pairs of set sizes, except between set size 1 and 2, was significantly different ($p < 0.0001$). Based on this, it can be concluded that increasing the number of sounds in a scene caused a significant, gradual increase in perceptual load, manifested as increased task difficulty.

Fig. 5.3 (in black) shows participants' performance in the secondary task. Each data point in the box plots represents the proportion of correct responses in the specific set size for one participant (therefore the variance represents variability between

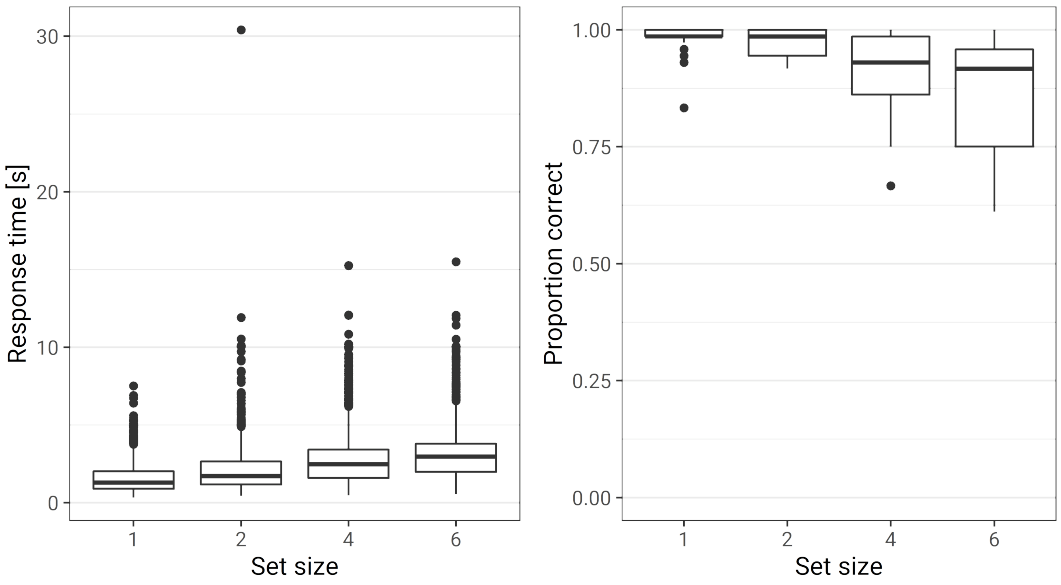



**Figure 5.2.:** Response times and proportion of correctly identified targets in the primary task.

participants). Similarly to the primary task, there is a steady decline in the secondary task performance as perceptual load increases.

To ensure that the perceptual load effects, and not just energetic masking influenced secondary target detection, secondary (detection) task performance in the main experiment was compared to the control experiment, where participants only performed the detection task. As can be seen in Fig. 5.3 (in grey), proportion correct for the detection-only trials shows a slight decline from size 2 onwards, suggesting that there may have been some masking effects for the two larger set sizes. A mixed-effects logistic regression with *participant* as a random effect, *set size* and *experiment* (detection-only/dual-task) as independent variables, and the detection task response outcome (correct/incorrect) shows a significant interaction between the two independent variables (see Table 5.2). This confirms that introducing the primary task significantly decreased secondary (detection) task performance, and therefore that a significant portion of this *set size* effect can be attributed to effects other than energetic masking.

| | Dependent variable: | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Proportion Correct | | | | RT [ms] | | | |
| Fixed effects | Est. | SE | z value | | Est. | SE | t value | |
| (Intercept) | 4.37 | 0.27 | 16.33 | *** | 2197.6 | 14.5 | 151.54 | *** |
| Set size 2 | −0.48 | 0.22 | −2.18 | * | 363.2 | 11.6 | 31.26 | *** |
| Set size 4 | −1.65 | 0.19 | −8.54 | *** | 962.7 | 12.3 | 78.38 | *** |
| Set size 6 | −2.19 | 0.19 | −11.66 | *** | 1403.7 | 11.7 | 120.52 | *** |
| Random effect: Participant | | | | | | | | |
| Number of groups | 25 | | | | 25 | | | |
| Standard deviation | 0.99 | | | | 229.2 | | | |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

**Table 5.1.:** GLMM regression results for two outcome variables of the primary task: proportion correct (binomial distribution, logit link function) and response times (inverse gaussian distribution, identity link function). The estimates of the binomial model shown are on the log scale. Set size was treated here as a categorical variable. Model formulas (identical except for the dependent variable): PC/RT $\sim$ 1 + SetSize + (1|Participant).

### 5.3.2. Behavioural data − target position

Fig. 5.4 shows the proportion of correct detections of the secondary target for different target positions. In order to analyse the effect of the position of the secondary target, *target position*, which is in effect a circular variable, was transformed into its sine and cosine components, representing the target's positioning on the left/right and front/back axis, respectively (see Fig. 5.5). Thanks to this, the target position, which is a circular variable, can be incorporated into the statistical models in a more meaningful way than if it were 6 separate categorical levels. It should also make the interpretation of any interactions much more straightforward.

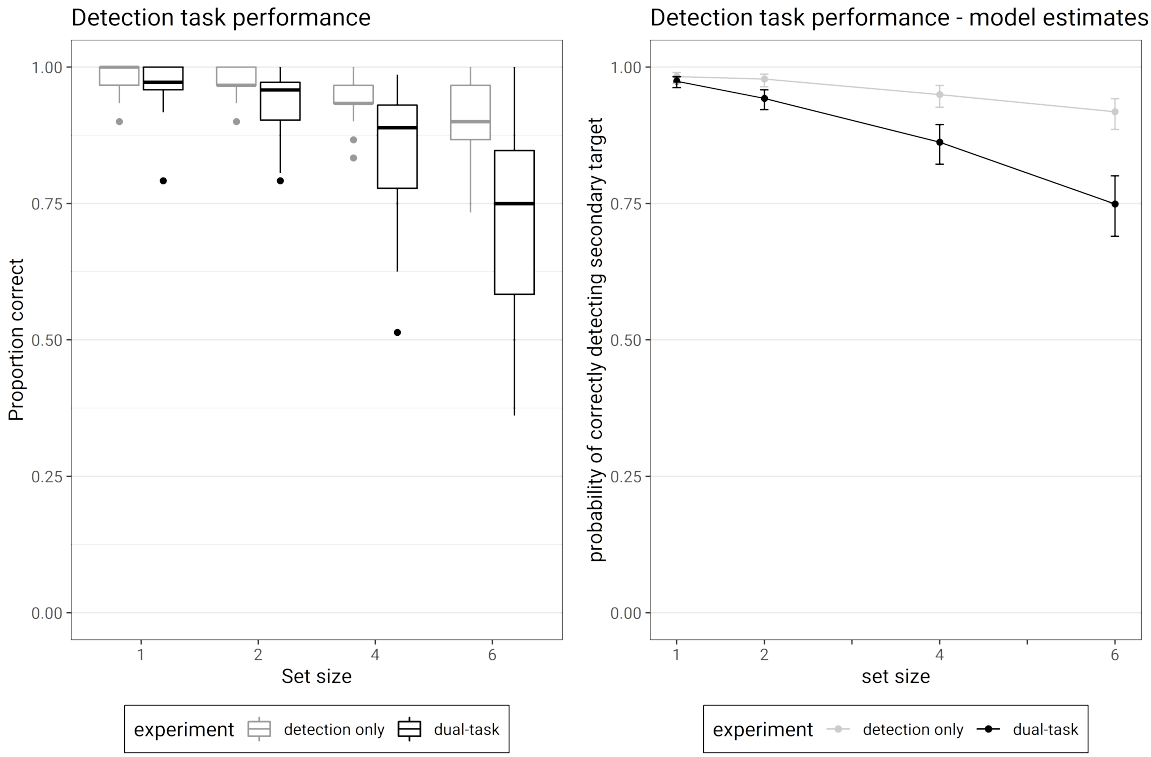



**Figure 5.3.:** Secondary task performance in the control (detection task only) and main (dual-task) experiment. Left pane shows the summary of task performance for each participant, and the right pane the overall estimated probability of detecting the secondary target, with 95% confidence intervals. Although the control test shows a slight decline, the decrease in the dual-task condition is significantly larger.

A mixed-effects logistic regression model was used with the response to the secondary task (correct/incorrect) as the dependent variable, *participant* as a random variable, and independent variables: *set size*, *sin(position)* and *cos(position)*. Finally, 2-way interactions between *set size* and the two positional variables, *sin(position)* and *cos(position)*, were included, to see if any positional effects are modulated by perceptual load. The results of this model are shown in Table 5.3. [1]

[1]Note that this statistical analysis only uses the trials with the secondary target present, because for other trials "target position" is meaningless. This means there are only true positives and false negatives – one can measure sensitivity but not specificity.

| Fixed effects | Estimate | 95% CI | SE | z value | p-value | |
|---|---|---|---|---|---|---|
| (Intercept) | 4.36 | [ 3.87, 4.88] | 0.26 | 16.95 | <0.001 | *** |
| Set size | −0.33 | [−0.42,−0.24] | 0.05 | −7.15 | <0.001 | *** |
| Experiment – main | −0.56 | [−1.04,−0.11] | 0.24 | −2.38 | 0.017 | * |
| Set size:Experiment | −0.13 | [−0.23,−0.03] | 0.05 | −2.59 | 0.010 | ** |
| Random effect: Participant | | | | | | |
| Number of groups | | | 25 | | | |
| Standard deviation | | | 0.70 | | | |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

**Table 5.2.:** Results of a GLMM model (binomial distribution, logit link function) on the secondary task performance. Estimates are on the log scale. Set size was modelled here as a continuous variable. Model formula: SecondaryCorrect $\sim$ 1 + SetSize + Experiment + SetSize $*$ Experiment + (1 | Participant)

The model shows a significant interaction between *set size* and *sin(position)*, but not *set size* and *cos(position)*. This suggests that increasing perceptual load modulated target detection on the left-right axis, but not front-back. More specifically, a Johnson-Neyman analysis (Johnson and Neyman 1936) reveals that the slope of *sin(position)* is significant for *set size* values smaller than 2.6 and larger than 7.8. At set sizes 1 and 2, therefore, participants detected the target more often when it was on the right, but this effect disappeared at set sizes 4 and 6. For the lowest set size, target on the right was about 2 times more likely to be detected as target on the left, and for *set size* $= 2$, this ratio was 1.7. Fig. 5.6 shows the Johnson-Neyman plot and estimated probabilities for all tested set sizes and a range of *sin(position)* values.

As mentioned before, there was no significant interaction between *set size* and *cos(position)*. However, because a regression model with interactions does not show main effects for the variables involved in interactions – but rather their conditional

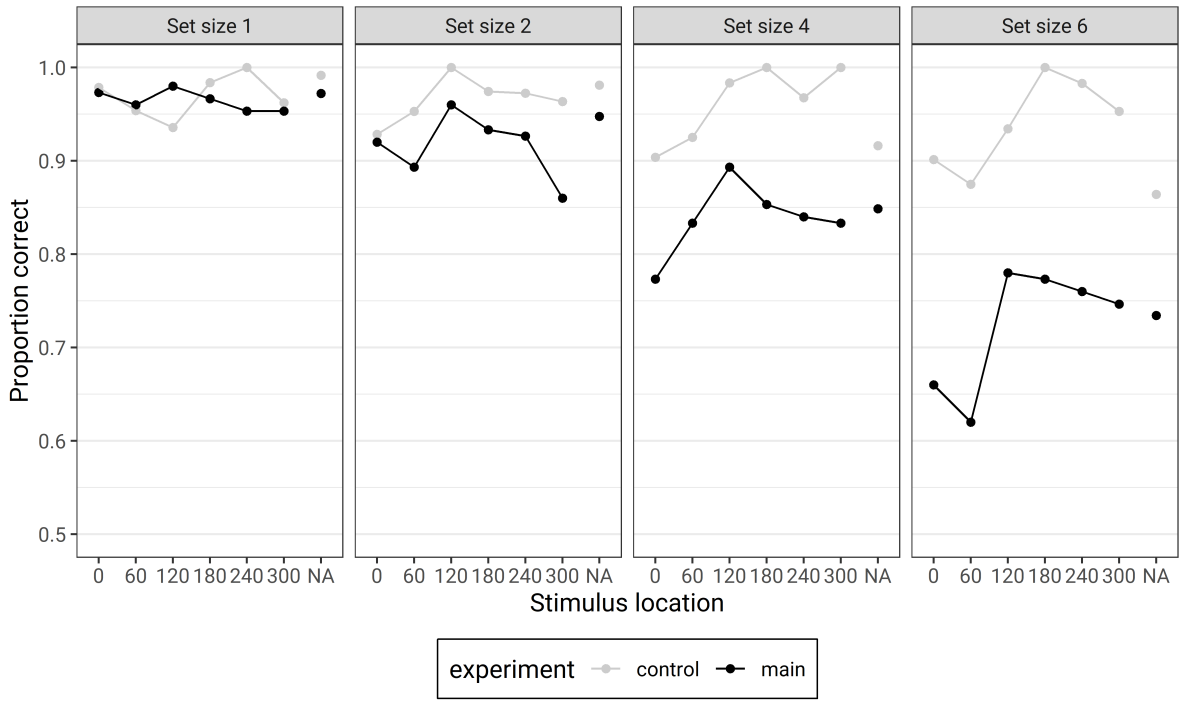



**Figure 5.4.:** Proportion of correctly detecting if secondary target is present or not, for different set sizes and secondary target positions (NA – trials with no secondary target).

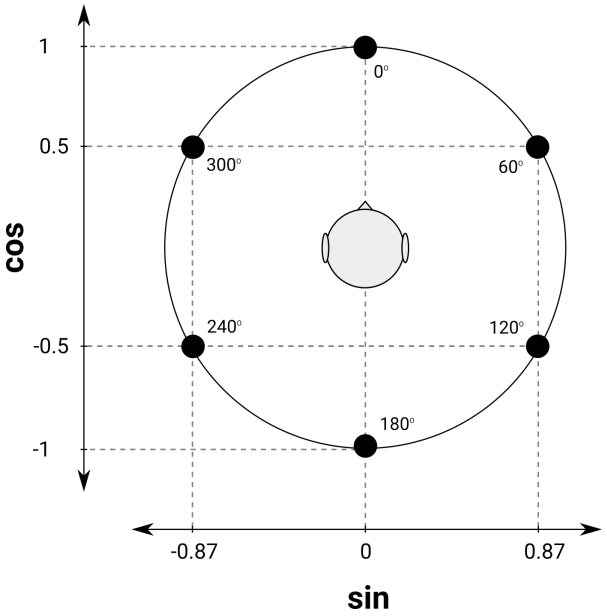



**Figure 5.5:** Positions of the secondary target and their corresponding sine and cosine values. Positive sine values indicate positions to the right of centre, and negative – to the left. Positive cosine values represent positions in front of the listener, negative – in the rear.

| | Interaction model | | | | Main effects model | | | |
|---|---|---|---|---|---|---|---|---|
| Fixed effects | Est. | SE | z-value | | Est. | SE | z-value | |
| (Intercept) | 3.82 | 0.25 | 15.46 | *** | 3.83 | 0.25 | 15.54 | *** |
| SetSize | −0.45 | 0.03 | −15.21 | *** | −0.45 | 0.03 | −15.40 | *** |
| sin(Position) | 0.45 | 0.19 | 2.41 | * | 0.45 | 0.19 | 2.41 | * |
| cos(Position) | −0.19 | 0.19 | −0.99 | | −0.33 | 0.07 | −4.41 | *** |
| sin(Position):SetSize | −0.10 | 0.04 | −2.43 | * | −0.10 | 0.04 | −2.43 | * |
| cos(Position):SetSize | −0.03 | 0.04 | −0.84 | | | | | |
| Random effect: Participant | | | | | | | | |
| Number of groups | 25 | | | | 25 | | | |
| Standard deviation | 0.996 | | | | 0.995 | | | |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

**Table 5.3.:** Results of GLMM models (binomial distribution, logit link function) with and without the interaction term between cos(Position) and the set size variable. The dependent variable in both was correct/incorrect response in the secondary task, estimates are on the log scale. Note that in the main effects model cos(position) represents a main effect, while in the interaction model, it shows a conditional effect (when $SetSize = 0$). Interaction model formula: SecondaryCorrect $\sim$ 1 + sin(Position) + cos(Position) + SetSize + sin(Position)$*$SetSize + cos(Position)$*$SetSize + (1 | Participant). Main effects model formula: SecondaryCorrect $\sim$ 1 + sin(Position) + cos(Position) + SetSize + sin(Position)$*$SetSize + (1 | Participant).

effects, here: when set $size = 0$ – it is not possible based on this model alone to determine whether there is a main effect of *cos(position)*. Therefore, another logistic regression model is fitted, identical but without the non-significant interaction (also shown in Table 5.3). The result indeed shows a significant main effect of *cos(position)*, which suggests that in general, the secondary target behind the participant was 1.9 times more likely to be detected than the target in front of them.

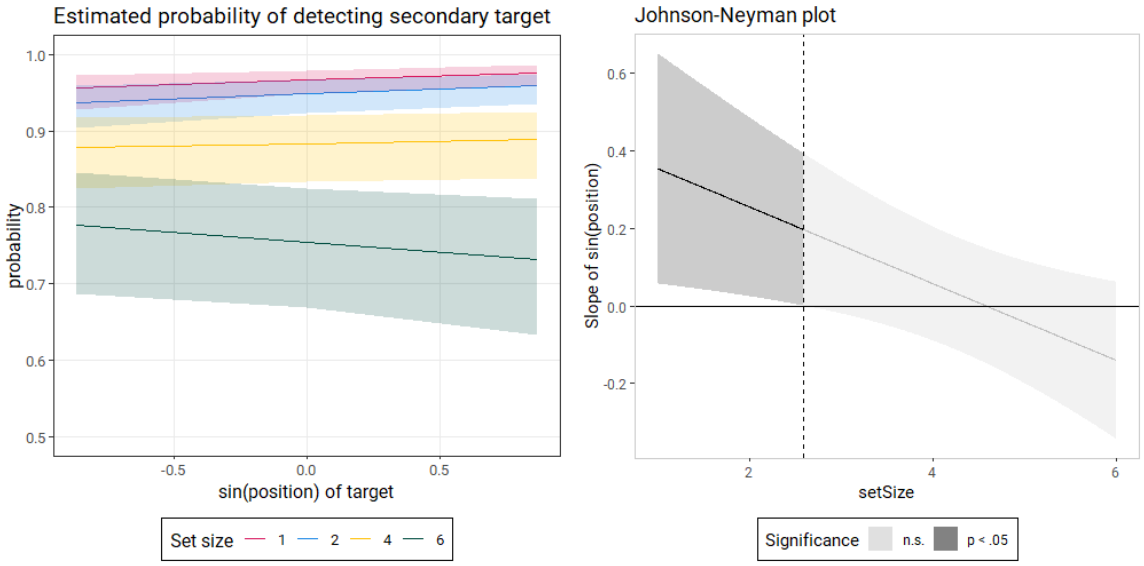



**Figure 5.6.:** Left: estimated probabilities of detecting the secondary target for different set sizes and target positions on the left-right axis (-1 is most to the left, and 1 is most to the right). Right: a Jonhson-Neyman plot for the interaction between *set size* and *sin(position)*. *Sin(position)* has a statistically significant positive slope for set sizes below 2.6. Shaded areas show 95% confidence intervals.

To summarise, these results indicate that in this dual-task experiment, the secondary target was easier to detect when it was behind the listener than when it was in front, and this effect was not modulated by perceptual load. In addition, at the two lowest load levels, participants were more likely to detect sounds on the right than sounds on the left, and this effect disappeared with increased perceptual load.

### 5.3.3. Pupillometry

Pupil dilation responses (PDR) was recorded using the Pupil Labs eye-tracking headset and software (Kassner, Patera and Bulling 2014). Before a statistical analysis, the raw recorded pupil dilation data were pre-processed following a method described in Kret and Sjak-Shie (2019). The procedure was as follows:

1. Remove values with confidence $< 0.6$. Each measurement point recorded by Pupil Labs has a confidence value, which represents the quality of the measurement.
2. Remove values outside of the valid range of pupil sizes: smaller than 1.5 mm and larger than 9 mm.
3. Detect blinks based on a speed filter and remove corresponding data points.
4. Detect outliers via residuals analysis and remove corresponding data points.
5. Resample data to 1000 Hz and interpolate missing data points, where a group of missing values is not larger than 500 ms. For larger groups, set these data points to invalid.
6. Smooth data out by low-pass filtering at 4 Hz.
7. If data were recorded for both eyes, choose the eye with a larger fraction of valid data points.
8. Then, cut pupil dilation recordings into segments, which start at the beginning of each trial (sound scene) and last 4 s.
9. For each trial, a baseline pupil dilation level was calculated as the average dilation for 200 ms before the trial start.
10. Set trials with more than 30% of invalid data points as not valid and removed from the dataset.
11. Finally, calculate mean and maximum dilation for each valid segment.

Fig. 5.7 shows the average PDR for all 4 set sizes, and the shaded area shows the time segment over which statistics (mean, maximum) were calculated.

A linear mixed-effects model with *set size* as independent variable (see Table 5.4) shows that, compared to set size 1, there is an increase in neither peak nor mean PDR for set size 2, however, set sizes 4 and 6 cause a significantly larger PDR. An analysis of contrasts shows that there is no significant difference between mean PDR for set sizes 4 and 6 ($p = 0.652$), however there might be a difference between peak PDR for

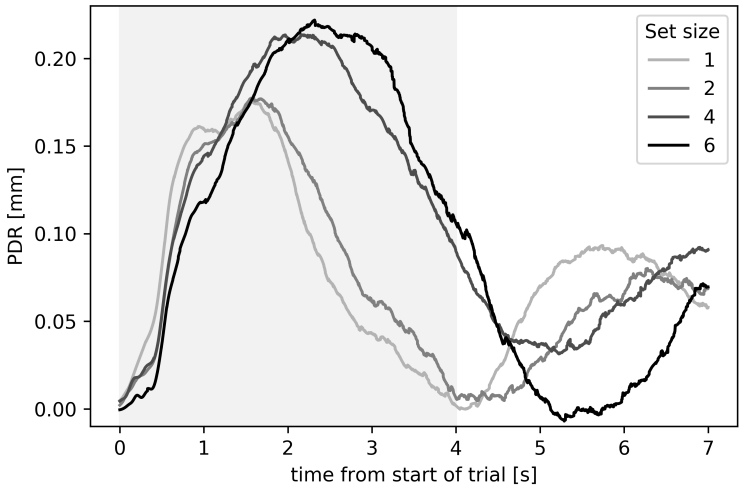



**Figure 5.7:** Averaged pupil dilation responses for all participants and different set sizes, aligned to start of trial. Shaded area shows the segment over which mean and peak PDR were calculated.

these set sizes ($p = 0.039$). Generally, there seems to be an increase in PDR with increasing set size, but it's not as linear and gradual as observed in behavioural performance metrics.

As in the previous section, an analysis of target position on PDR was performed. Mixed-effects linear models with *cos(position)* and *sin(position)* on mean and peak PDR show that there is no interaction between the positional variables and any of the set sizes (see Table 5.5). There is also no main effect of either *sin(position)* or *cos(position)* on mean or max PDR. [2]

In this experiment, there was no effect of target sound position on pupil dilation responses, regardless of perceptual load.

[2]Note that because of the evident lack of a linear relationship between set size and PDR, set size is coded in these models as a discrete variable with 4 levels.

| | Dependent variable: | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean PDR | | | | Peak PDR | | | |
| Fixed effects | Est. [mm] | SE | z-value | | Est. [mm] | SE | z-value | |
| (Intercept) | 0.800 | 0.045 | 17.64 | | 0.981 | 0.061 | 15.97 | |
| Dilation baseline | −0.191 | 0.008 | −23.16 | *** | −0.153 | 0.011 | −13.36 | *** |
| Set size 2 | 0.002 | 0.010 | 0.17 | | 0.005 | 0.014 | 0.37 | |
| Set size 4 | 0.063 | 0.010 | 6.05 | *** | 0.083 | 0.014 | 5.80 | *** |
| Set size 6 | 0.050 | 0.010 | 4.85 | *** | 0.044 | 0.014 | 3.07 | ** |
| Random effect: Participant | | | | | | | | |
| Number of groups | | 15 | | | | 15 | | |
| Standard deviation | | 0.13 | | | | 0.17 | | |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

**Table 5.4.:** Results of two Linear Mixed Models, with mean PDR (left column) and peak PDR (right column) as outcome variable. Model formulas: PDR $\sim$ 1 + PupilBaseline + SetSize + (1|Participant).

| | Interaction model | | | | Main effects model | | | |
|---|---|---|---|---|---|---|---|---|
| Fixed effects | Est. [mm] | SE | z-value | | Est. [mm] | SE | z-value | |
| (Intercept) | 0.801 | 0.053 | 15.01 | *** | 0.799 | 0.05 | 15.01 | *** |
| Dilation baseline | −0.187 | 0.011 | −16.61 | *** | −0.186 | 0.011 | −16.61 | *** |
| SetSize 2 | −0.017 | 0.014 | −1.19 | | −0.017 | 0.014 | −1.18 | |
| SetSize 4 | 0.048 | 0.014 | 3.34 | *** | 0.048 | 0.014 | 3.34 | *** |
| SetSize 6 | 0.039 | 0.015 | 2.69 | ** | 0.039 | 0.015 | 2.69 | ** |
| sin(Position) | −0.004 | 0.014 | −0.27 | | −0.005 | 0.007 | −0.65 | |
| cos(Position) | 0.0002 | 0.014 | 0.01 | | −0.0002 | 0.007 | −0.03 | |
| sin(Position):SetSize 2 | −0.005 | 0.020 | −0.25 | | | | | |
| sin(Position):SetSize 4 | −0.005 | 0.020 | −0.27 | | | | | |
| sin(Position):SetSize 6 | 0.007 | 0.020 | 0.33 | | | | | |
| cos(Position):SetSize 2 | −0.011 | 0.020 | −0.57 | | | | | |
| cos(Position):SetSize 4 | 0.013 | 0.020 | 0.65 | | | | | |
| cos(Position):SetSize 6 | −0.003 | 0.020 | −0.17 | | | | | |
| Random effect: Participant | | | | | | | | |
| Number of groups | | 15 | | | | 15 | | |
| Standard deviation | | 0.12 | | | | 0.12 | | |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

**Table 5.5.:** Results of two mixed linear models, in which the outcome variable was mean PDR, with and without interaction terms with the set size variable. Note that in the main effects model sin(position) and cos(position) represent main effects, while in the interaction model, they show conditional effects (when $SetSize = 0$). Interaction model formula: MeanPDR $\sim$ 1 + PDRBaseline + SetSize + sin(Position) + cos(Position) + sin(Position)$*$SetSize + cos(Position)$*$SetSize + (1 | participant). Main effects model formula: MeanPDR $\sim$ 1 + Baseline + SetSize + sin(Position) + cos(Position) + (1 | participant)

## 5.4. Summary

This chapter described a dual-task auditory experiment designed to test spatial auditory salience under different levels of perceptual load, which was manipulated by changing the difficulty of the primary task. Behavioural responses to the secondary task revealed small, but statistically significant effects of spatial position of the target. In particular, an effect which was not diminished by perceptual load, and therefore could perhaps be attributed to stimulus salience, indicated that target sounds positioned behind the listener were about 2 times more likely to be detected than those in front.

These results are discussed in Chapter 13.

# 6

# Context and expectations

## 6.1. Introduction

As discussed in Chapter 2, distraction happens – at least partly – when one's attention is involuntarily drawn away from a task by an irrelevant stimulus. In other words, auditory deviants cause attentional orienting away from the primary object of focus, which causes distraction. Therefore, if salience is defined as the ability to draw attention, it can be assumed that sounds which cause more distraction are more salient.

A classic distraction experiment involves subjecting participants to a stream of sounds, most of which are repeating and predictable – *standards* – and a small percentage are different and unexpected – *deviants*. Distraction by deviant sounds has been demonstrated in visual, audio-visual and auditory experiments. Usually, the distraction is measured as prolonged response times on a task. Deviant sounds also elicit brain responses such as MMN (deviance detection) and P3a (attention switch). For example, in a distraction experiment with an auditory task, deviant frequency tones elicited both MMN and P3a, whereas MMN was present even when the task

was ignored (deviance detection worked but there was not attention switching) (Schröger, Giard and Wolff 2000).

These types of paradigms can be described as creating expectations for the listener (*standard sound*) and breaking them (*deviant sound*), which causes automatic attentional orienting, manifesting as distraction. These expectations can be about different characteristics of the stimulus. Originally, most distraction experiments were based on a visual task and auditory distractor. For example, a visual classification test (even/odd) with auditory distractors (tones or noise) (Parmentier et al. 2011), or a visual recognition test (left and right arrows) and repeating tone patterns as distractors (Nöstl, Marsh and Sörqvist 2012).

Schröger and Wolff (1998) proposed a fully auditory paradigm for measuring distraction, where two different dimensions of the same sound serve as target and distractor. In their experiment, the target was length of the sounds (short or long), and the distraction was caused by unexpected changes in pitch. This paradigm was later adapted to spatial location instead of pitch (Roeber, Widmann and Schröger 2003), so that the standard sound was always in the same location (left or right of the listener), and occasional deviants in the other one. Indeed, when tones appeared in an unexpected location, this caused distraction. Also, in a word recognition experiment, irrelevant words which changed location randomly between trials were more distracting than those coming from one location (Chan, Merrifield and Spence 2005).

This chapter describes an experiment which was designed to test spatial auditory distraction in a more natural listening scenario. To do this, rather than short repeating sounds, a continuous stream of a natural sound is used, which means that expectations are built not over the whole experiment – like in most of the studies described above – but over shorter periods of up to 10 seconds. After the expectation has been created, a "distracting" sound is played, which either agrees with or violates it in terms of its spatial location (see Fig. 6.1 for an illustration of a single trial).

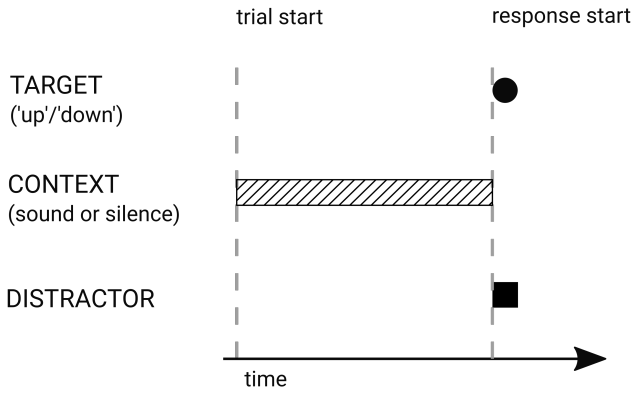



**Figure 6.1:** Building blocks of a single trial: context sound or a period of silence, followed by a distractor sound played simultaneously with the target speech.

If a sound suddenly changes its position, two things might happen in the listener's brain: 1) the spatial continuity of that particular sound stream breaks, 2) a sound is recognised to have appeared in an unexpected location (against spatial expectation). Both of these can be thought of as surprising, hence distracting. Therefore, this experiment attempts to answer the following questions:

1. Can the spatial distraction effect be measured with more natural stimuli and shorter expectation build-up?
2. If attentional orienting occurs, is it due to breaking stream continuity or a sound appearing in an unexpected location, or both? If both, how do these effects compare?

## 6.2. Method

In the experiment, participants were sitting in a room surrounded by loudspeakers, and were instructed to perform a simple task based on auditory stimuli. They heard a voice saying either "up" or "down" and were asked to press either the upward or the downward arrow respectively.

In some cases, a distracting sound was played at the same time as the target speech.

The distractor was played simultaneously with the target, rather than just before like in similar visual-task-based experiments, in order to avoid temporal cueing effects. Having a cue about when the target arrives could improve performance and cancel out distraction effects compared to the baseline no-distraction condition.

In some of the distracting scenarios, the distractor and target were preceded by a few (between 3.7 and 11.1) seconds of "context" sound. The context could either match or not match the distractor on two variables: spatial location and type of sound.

There were six conditions altogether:

- **A** no distraction
- **B** silence followed by a single distracting sound
- **C1** context sound followed by a distracting sound, where the context matches the distractor in **location and type**
- **C2** context sound followed by a distracting sound, where the context matches the distractor in **location but not type**
- **D1** context sound followed by a distracting sound, where the context matches the distractor in **type but not location**
- **D2** context sound followed by a distracting sound, where the context does not match the distractor in neither **location nor type**

Figure 6.2 shows example scenarios for each of the six conditions.

The sounds were spatially distributed in four positions around the listener: 45° (front-right), 135° (back-right), 225° (back-left) and 315° (front-left). Crucially, the context and distracting sounds were never in the same spatial position as the target. This was done to minimise any energetic masking effects and facilitate auditory stream separation.

Each trial started with a 1 s-long narrow-band noise burst, always in the same spatial

position as the target. The purpose of the noise was to a) mark the beginning of a new trial, and b) guide participants' attention to where the target was going to be, in order to minimise the cost of initial switching of attention and the search for the target stream (see Kidd et al. (2005)).

The context stimuli were bird songs, chosen at random from a list of 37 recordings, their length varying from 3.7 to 11.1 seconds (mean: 7 s). The matching distractors (conditions C1, D1) were 200 ms long excerpts of bird song. Each context sound had a corresponding matching distractor, so that they sounded like one continuous recording. The non-matching distractors (conditions C2, D2) were 200 ms non-bird sounds picked randomly out of a list of 32 (e.g. dog barks, car horn, water drop). The distractor in condition B could be either from the bird, or the non-bird list. A full list of all recordings used and their sources is available in Appendix B.

Experimental design was full factorial, 6 conditions x 24 possible spatial configurations of target, context and distractor.

Participants were also asked to wear the Pupil Labs eye-tracking headset (Kassner, Patera and Bulling 2014), which measured dilation of their pupils as they were performing the task. As discussed in Chapter 2, pupil dilation responses have been associated with unexpected sounds and automatic attentional orienting, which can help determine the salience of each experimental condition.

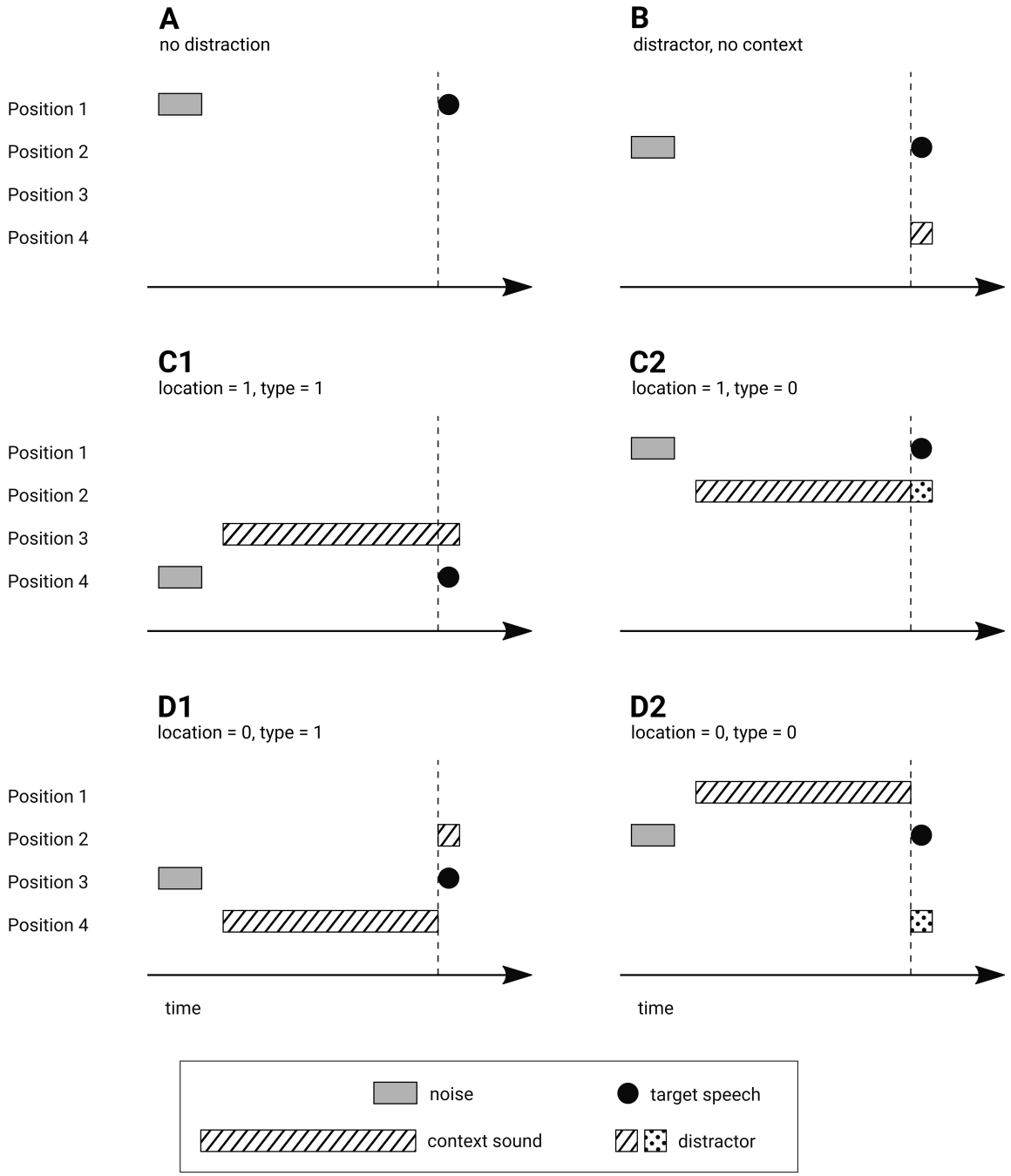



**Figure 6.2.:** Examples of context/distractor/target combinations for each of the experimental conditions. Note that each panel only shows one possible combination of the stimuli positions – in the experiment, the positions were balanced and radomised.

## 6.3. Results

36 volunteers took part in the test, mean age 26 (min: 18, max: 60). Two participants were over 50 years old, but their responses did not seem to differ in any significant way from other participants (see Fig. 6.3).

The data was analysed in two main ways, to investigate:

- effect of condition – to see if introducing the different types of stimuli causes distraction,
- effects of location and type of sound on distraction.

The first one is a straightforward comparison between the six conditions. In the second, the data from conditions A and B is discarded, because they did not include the context sound. For the remaining data, two variables are considered, *Location* and *Type* (of sound), which can be either 1 (change from context to distractor) or 0 (no change). Each of the four conditions – C1, C2, D1, D2 – can be assigned a distinct combination of variables location and type. Specifically:

- C1: Location = 0, Type = 0
- C2: Location = 0, Type = 1
- D1: Location = 1, Type = 0
- D2: Location = 1, Type = 1

For an illustration, see Fig. 6.2.

### 6.3.1. Behavioural responses

The mean proportion of correct responses was 99%, with only two participants responding correctly in fewer than 96% of trials. Response times were recorded as the

time from the beginning of the target word to button press. The average response time across participants was 850 ms with standard deviation of 110 ms. Only one participant had a significantly larger average response time than the others (and 85% accuracy). Their low accuracy likely comes from exceeding the 2s period during which responses were collected after each trial. A decision was made to exclude the outlier participant from further analysis, as such long response times suggest they might not have been following the instructions to respond as quickly as possible. Figure 6.3 shows all participants response times and accuracies, with the shaded areas marking 2 standard deviations from the mean on each axis. Fig. 6.4 shows individual response time distributions for the remaining 35 participants.

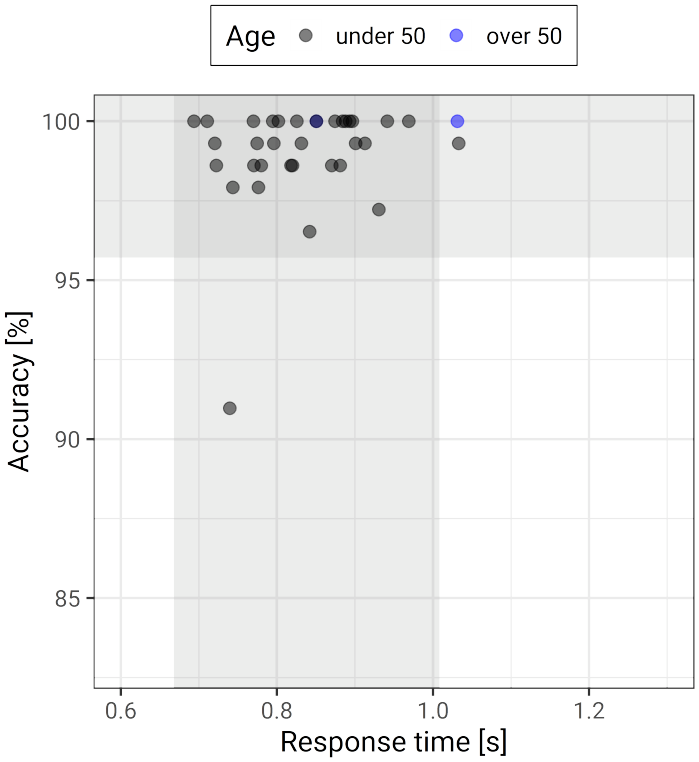



**Figure 6.3:** Mean response times and accuracy in the test – each dot represents a participant. Shaded areas show 2 standard deviations from the mean on each axis. Blue dots are participants over 50 years old.

For all response times analyses in this section, generalised linear mixed models (GLMM) were used, with the inverse Gaussian distribution and identity link function, and *participant* as a random variable.

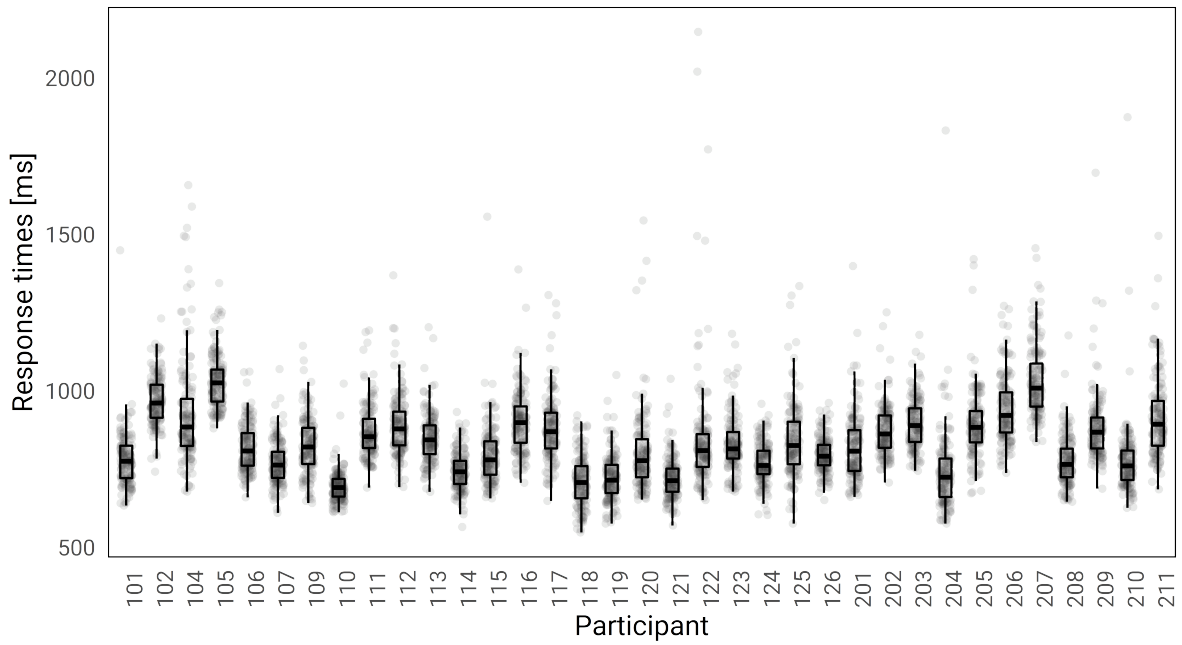



**Figure 6.4.:** Individual participants' response times. Boxplots show the median, 25th and 75th percentile of each participants' response times. Dots show the individual responses (the darker the are, the more responses there are).

The results show that introducing the various distractors had a significant effect on the response times. Results of a GLMM model with context length and different conditions as predictors is shown in Table 6.1 and Fig. 6.5. Compared to condition A (no distraction), conditions C2 (different sound), D1 (different location), and D2 (different sound and location) caused a significant increase in response times. Somewhat surprisingly, condition B (single distractor) was not significantly different from A. Condition C1 (context sound, no violated expectations) seems to have a similarly non-significant effect as B (no context sound, violated expectations).

In addition, the length of time between the beginning of the trial and the target (*context length*) had a significant effect on the responses. With each 1 s increase in context length, response times decreased by approximately 3 ms. To find out if this effect was present only for context sounds, or also when the target was preceded by

| Fixed effects | Est. [ms] | SE | t value | p-value | |
|---|---|---|---|---|---|
| (Intercept) | 871.89 | 30.43 | 28.65 | < 0.0001 | *** |
| Context length [s] | −3.00 | 0.63 | −4.77 | < 0.0001 | *** |
| **Condition B** | 7.42 | 4.41 | 1.68 | 0.093 | |
| **Condition C1** | 8.20 | 4.40 | 1.86 | 0.062 | |
| **Condition C2** | 9.60 | 4.37 | 2.20 | 0.028 | * |
| **Condition D1** | 10.24 | 4.38 | 2.33 | 0.020 | * |
| **Condition D2** | 12.39 | 4.39 | 2.82 | 0.005 | ** |
| Random effect: Participant | | | | | |
| Number of groups | | 35 | | | |
| Standard deviation | | 33.50 | | | |

* p < 0.05, ** p < 0.01, *** p < 0.001

**Table 6.1.:** GLMM results on response time data (identity link, inverse Gaussian distribution). Model formula: RT ∼ 1 + ContextLength + Condition + (1|Participant).

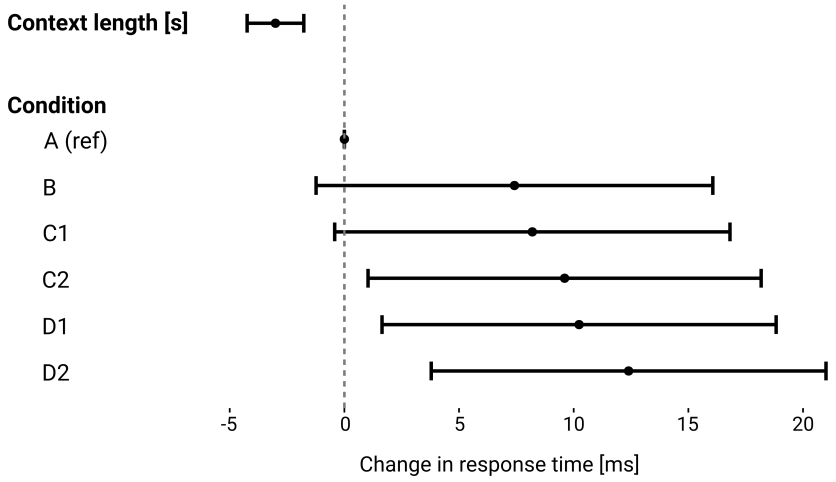



**Figure 6.5:** Estimated effects of different variables on response times and their 95% confidence intervals based on the model in Table 6.1.

silence, a GLMM regression model with an *ContextLength*ContextType* interaction was analysed, where *ContextType* was either 'sound' (conditions C1, C2, D1, D2) or "silence" (conditions A and B), and the outcome variable was response time (in

milliseconds). The interaction was not significant ($Est. = 0.4$, $p = 0.78$), suggesting that the effect of "context length" was independent of whether an actual context sound was present or not.

To test if there were significant effects of *Type* and *Location* on response times, as described in the previous section, a GLMM was analysed with these two variables as main effects, and one with an interaction between them. Neither the interaction ($p = 0.935$), nor the main effects (location: $p = 0.439$, type: $p = 0.670$) were statistically significant.

Finally, GLMM regression is used to investigate if there are any differences between responses to different spatial position of the distractor. The model's dependent variable is response times (in ms), and independent variables *context length* and the four possible distractor spatial positions. A post-hoc analysis of contrasts found no significant differences between any of the positions (see Tables 6.2 and 6.3).

| | Est. [ms] | SE | t value | p-value | |
|---|---|---|---|---|---|
| (Intercept) | 878.68 | 22.05 | 39.85 | <0.0001 | *** |
| ContextLength [s] | −3.04 | 0.62 | −4.91 | <0.0001 | *** |
| Position: front-right | 0.78 | 3.64 | 0.21 | 0.83 | |
| Position: back-right | 4.19 | 3.63 | 1.16 | 0.25 | |
| Position: back-left | 0.70 | 3.62 | 0.19 | 0.85 | |
| Random effect: Participant | | | | | |
| Number of groups | | 35 | | | |
| Standard deviation | | 33.53 | | | |

* p < 0.05, ** p < 0.01, *** p < 0.001

**Table 6.2.:** Mixed-effects generalised linear regression results with response time as the dependent variable and distractor spatial position as one of the independent variables. Model formula: RT ∼ 1 + ContextLength + Position + (1 | Participant).

| | front-left | front-right | back-right | back-left |
|---|---|---|---|---|
| front-left | | 1.00 | 0.66 | 1.00 |
| front-right | | | 0.78 | 1.00 |
| back-right | | | | 0.77 |

**Table 6.3.:** P-values from a post-hoc comparison of response times for different distractor locations. P-value adjustment: Tukey method for comparing a family of 4 estimates.

### 6.3.2. Pupil dilation

First, pupil dilation data was pre-processed following the same procedure as in Chapter 5, with code adapted from Kret and Sjak-Shie (2019), with the exception of how dilation segments were selected. Here, the pupil dilation recordings were cut into segments between 500 ms and 2 s after the onset of the target speech.

Only participants with at least 50% valid trials were included in the pupil dilation analysis. It is difficult to determine the exact reason for missing data in each case, but most likely the majority of it was due to a poor fit of the headset. In this case, there is no reason to believe these missing data points are not random and in any way correlated with the experimental variables. A mixed-effects logistic regression model confirms that the *Condition* variable was not able to predict whether a trial is valid or not (compared to condition A, B: $p = 0.59$, C1: $p = 0.93$, C2: $p = 0.53$, D1: $p = 0.46$, D2: $p = 0.65$; none of the post-hoc pairwise contrasts were statistically significant).

Similarly to response times, first, the effects of the *Condition* variable on PDR are analysed, using a mixed-effects linear model with *Participant* as a random variable. Pupil dilation *baseline* and *context length* are also included as independent variables. The results are summarised in Table 6.4 and Fig. 6.6

The results for mean and peak PDR reveal a very similar pattern. In the remainder of

| | Dependent variable: | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Peak PDR | | | | Mean PDR | | | |
| Fixed effects | Est. [mm] | SE | t value | | Est. [mm] | SE | t value | |
| (Intercept) | 0.93 | 0.07 | 13.80 | *** | 0.70 | 0.05 | 12.93 | *** |
| Baseline | −0.11 | 0.01 | −11.25 | *** | -0.11 | 0.01 | −13.24 | *** |
| Context Length | −0.01 | 0.003 | −2.02 | * | -0.01 | 0.003 | −2.36 | * |
| **Condition B** | 0.08 | 0.02 | 3.80 | *** | 0.07 | 0.02 | 3.55 | *** |
| **Condition C1** | 0.003 | 0.02 | 0.14 | | 0.02 | 0.02 | 0.87 | |
| **Condition C2** | 0.08 | 0.02 | 3.82 | *** | 0.08 | 0.02 | 4.03 | *** |
| **Condition D1** | 0.04 | 0.02 | 2.01 | * | 0.04 | 0.02 | 2.11 | * |
| **Condition D2** | 0.08 | 0.02 | 3.83 | *** | 0.07 | 0.02 | 3.89 | *** |
| Random effect: Participant | | | | | | | | |
| Number of groups | | 19 | | | | 19 | | |
| Standard deviation | | 0.19 | | | | 0.14 | | |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

**Table 6.4.:** Results of mixed-effects linear regression on peak and mean pupil dilation responses. Model formulas: PDR $\sim$ 1 + DilationBaseline + ContextLength + Condition + (1 | Participant).

this chapter, only mean PDR results will be discussed. Pupil dilation baseline is a significant predictor: the larger the baseline, the smaller the dilation during the trial. Context length is also significant, with increased length related to smaller PDR.

Changing the experimental condition had an effect on pupil dilation. More specifically, conditions B, C2, D1, and D2 were significantly different from the control condition A, and C1 was not. A post-hoc analysis of contrasts (with Tukey p-value adjustment) reveals statistically significant differences between conditions A and B ($MD = -0.068$, $p = 0.005$), A and C2 ($MD = -0.077$, $p < 0.001$), A and D2 ($MD = -0.074$, $p = 0.001$), C1 and C2 ($MD = -0.060$, $p = 0.02$), and C1 and D2

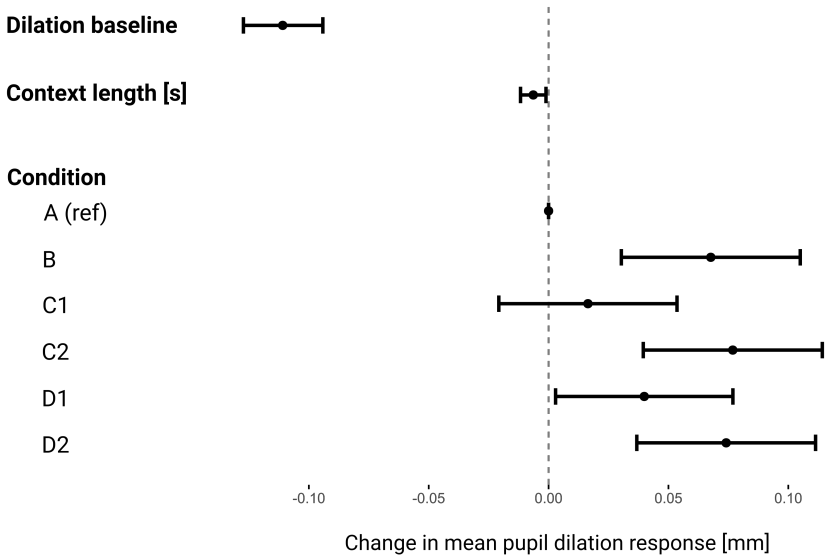



**Figure 6.6:** Estimated effects of mean PDR from a mixed-effects linear regression model (Table 6.4) and corresponding 95% confidence intervals.

($MD = -0.058, p = 0.03$).

Next, a mixed-effects model with *Location* and *Type* shows no significant interaction between the two variables. When analysing main effects, there is a significant effect of *Type* on mean PDR, but not of *Location* – see Table 6.5. Fig. 6.7 shows the estimated mean PDR for both variables. An analysis of peak PDR shows the same pattern, with no significant interaction ($p = 0.139$), no main effect of *Location* ($p = 0.127$), but a main effect of *Type* ($p < 0.001$).

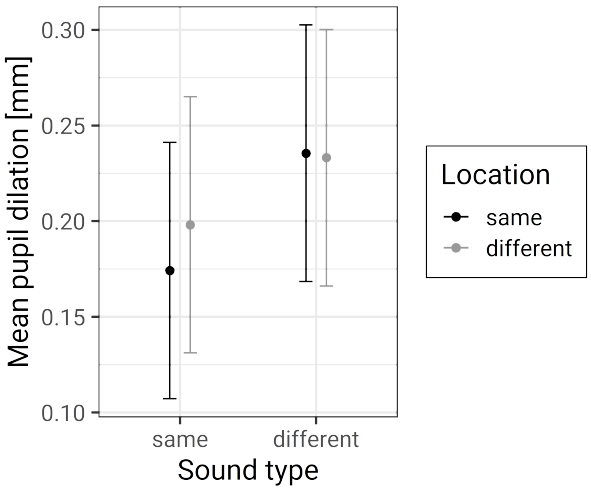



**Figure 6.7:** Estimated mean pupil dilation responses for conditions with the same/changing location and the same/changing sound type, based on the interaction model in Table 6.5. The model shows a significant effect of sound type but not location, and no interaction.

| | Main effects model | | | | Interaction model | | | |
|---|---|---|---|---|---|---|---|---|
| Fixed effects | Est. [mm] | SE | t value | | Est. [mm] | SE | t value | |
| (Intercept) | 0.60 | 0.05 | 11.04 | *** | 0.59 | 0.05 | 10.86 | *** |
| Baseline | −0.09 | 0.01 | −9.87 | *** | −0.09 | 0.01 | −9.87 | *** |
| ContextLength | −0.004 | 0.003 | −1.31 | | −0.004 | 0.003 | −1.31 | |
| **Location** | 0.01 | 0.01 | 0.95 | | 0.02 | 0.02 | 1.47 | |
| **Type** | 0.05 | 0.01 | 4.15 | *** | 0.06 | 0.02 | 3.73 | *** |
| **Location:Type** | | | | | −0.03 | 0.02 | −1.13 | |
| Random effect: Participant | | | | | | | | |
| Number of groups | | 19 | | | | 19 | | |
| Standard deviation | | 0.132 | | | | 0.131 | | |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

**Table 6.5.:** Effects of Location and Type on mean PDR. The model on the right shows that an interaction between these two variables is not significant, however, it does not show main effects (but rather, conditional effects). The model on the left shows main effects of Location (not significant) and Type (significant). Main effects model formula: PDR $\sim$ 1 + DilationBaseline + ContextLength + Location + Type + (1 | Participant), interaction model formula: PDR $\sim$ 1 + DilationBaseline + ContextLength + Location + Type + Location$*$Type + (1 | Participant)

Finally, similarly to behavioural responses, a linear mixed-effects model with *dilation baseline*, *context length*, and *spatial position* as independent variables and mean PDR as dependent variable found no significant differences between the different spatial positions (see Tables 6.6 and 6.7).

| | Est. [mm] | SE | t value | p-value | |
|---|---|---|---|---|---|
| (Intercept) | 0.738 | 0.054 | 13.55 | <0.0001 | *** |
| Baseline | −0.111 | 0.008 | −13.19 | <0.0001 | *** |
| ContextLength | −0.006 | 0.003 | −2.31 | 0.02 | * |
| Position: front-right | 0.023 | 0.016 | 1.47 | 0.14 | |
| Position: back-right | 0.008 | 0.016 | 0.51 | 0.61 | |
| Position: back-left | 0.020 | 0.016 | 1.28 | 0.20 | |
| Random effect: Participant | | | | | |
| Number of groups | | 19 | | | |
| Standard deviation | | 0.141 | | | |

* p < 0.05, ** p < 0.01, *** p < 0.001

**Table 6.6.:** Mixed-effects linear regression results with average pupil dilation as the dependent variable and distractor spatial position as one of the independent variables. Model formula: MeanPDR ∼ 1 + PupilBaseline + ContextLength + Position + (1 | Participant).

| | front-left | front-right | back-right | back-left |
|---|---|---|---|---|
| front-left | | 0.46 | 0.96 | 0.58 |
| front-right | | | 0.78 | 1.00 |
| back-right | | | | 0.87 |

**Table 6.7.:** P-values from a post-hoc comparison of mean pupil dilation different distractor locations. P-value adjustment: tukey method for comparing a family of 4 estimates.

## 6.4. Summary

The experiment described in this chapter tested the effects of breaking expectations about spatial location and type of sound. Even though introducing distracting sounds did increase response times compared to a no-distraction condition, no effects of changing location or type of sound on behavioural responses were found. However, introducing a new sound to the scene did cause significantly larger pupil dilation responses than when the same sound continued throughout the scene. No main effect of changing spatial location of sound was found. However, a condition in which spatial continuity of a stream was broken, showed a small increase in pupil dilation compared to the control condition.

A discussion of these results can be found in Chapter 13.

7

# Comparison of methods

## 7.1. Introduction

In the previous chapters, perceptual experiments designed to test the effects of location of sound on auditory salience were described. As there is no standard testing method for auditory salience, four different methods were used in this thesis, each of which addressed the question from a slightly different perspective.

In Chapter 2, different ways in which salience has been defined in the literature were discussed. Out of the four most common ways, three could be used as a basis for experiments: salience in terms of attention ("the ability to grab attention"), detection ("standing out") and relevance (or importance). Experiments reported in Chapters 4 and 6 are both based on attention. The experimental method used in Chapter 3 was described as based on detection in the original paper, but could also be explained in terms of attracting attention. The experiment in Chapter 5 was based on detection.

The Chapter 5 experiment showed results not entirely consistent with the other experiments, and it is possible that the different results come from it being based on a different definition of salience. However, even though all three attention-based

experiments showed a lack of a spatial effect, this cannot really be used to claim that they all measure exactly the same phenomenon. In fact, there are reasons to think there are differences in what type of "salience" each of them tests. For one, they function on different levels of attention. The method used in Chapter 6 (distraction) relies on attentional orienting, which is low-level, automatic, and could potentially even be subconscious. On the other side of the spectrum, in the Chapter 4 experiment (self-reporting), the sounds are picked up not only if they cause attentional orienting, but also if they are consciously noticed, and cause the participant to point towards them. This is a much higher level process. Additionally, although no instructions about what to listen to were given in the *self-reporting* experiment, it allowed for possible effects of top-down attention and personal preference. In the *distraction* experiment, the participants' attention was focused on the speech, so the effects of top-down attention should not play a role.

A more direct comparison of these methods, which could determine if they do in fact measure the same type of salience, would be useful. To address this issue, an experiment which compares the three attention-based methods was conducted.

## 7.2. Method

The experiment consists of three parts, each employing one of the methods used in the experiments described in Chapters 3, 4 and 6. The same stimuli were tested with each method to allow for direct comparisons of salience metrics produced by each method.

The stimuli were a subset of the short sounds used by Zhao et al. (2019), who conducted a large-scale online survey, asking participants to compare salience between two sounds. On this basis, they were able to sort 18 sounds from the least to most salient. Out of these, 10 were selected for this study, spread over the salience

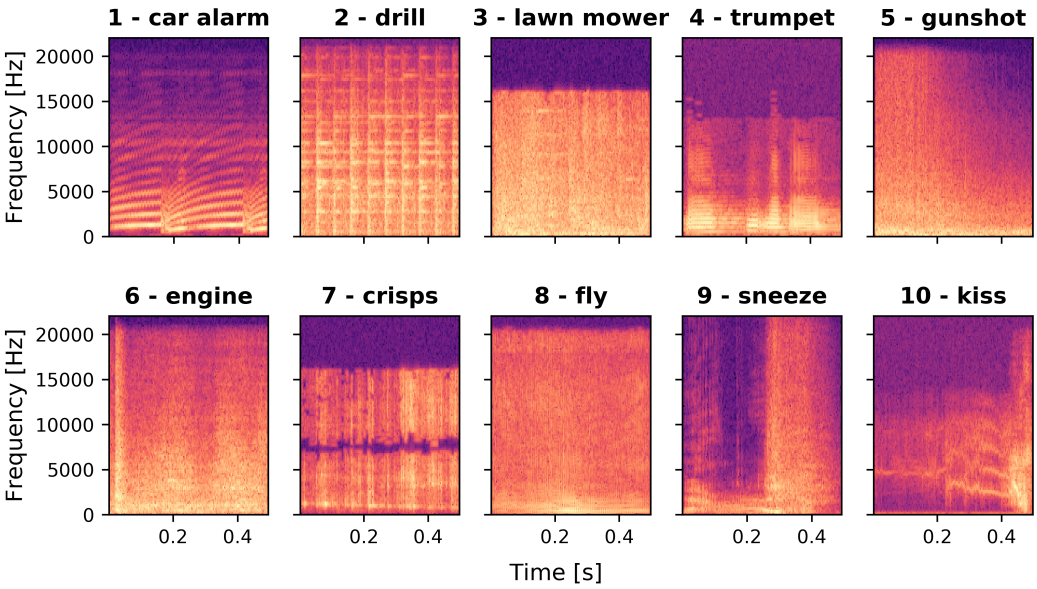


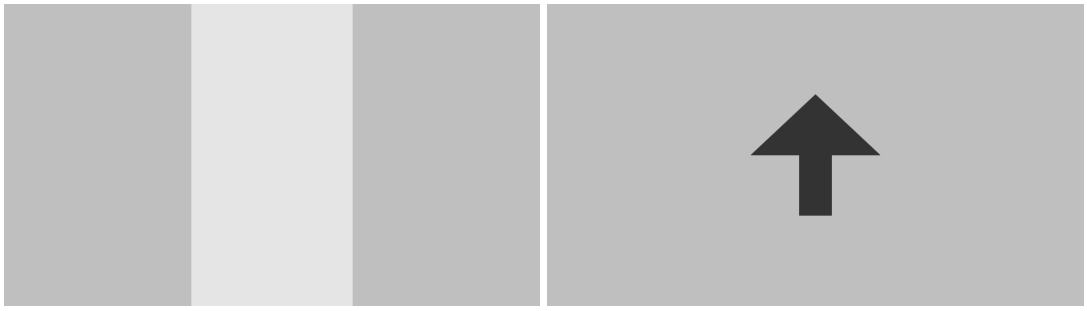

**Figure 7.1.:** Spectrograms of the stimuli used in the comparison experiment.

scale. Their spectrograms are shown in Fig. 7.1. The sounds were all 500 ms long and RMS equalised.

In this experiment the more basic, non-spatial versions of each method were used, and the test was conducted using headphones. This allowed the methods to be simplified and avoid any additional effects brought about by spatialisation. The three parts of the experiment are described below.

**The first part**, real-time self-reporting, was done in a very similar way to the original experiment by Huang and Elhilali (2017), and corresponds to the experiment in Chapter 4. Participants heard one scene in their left ear and one scene in their right ear, and were asked to indicate – in real time – which one they were listening to. They did this by moving a mouse cursor on a screen, which was divided into 3 areas: left for the left scene, right for the right scene, and the middle for none, both or undecided (see Fig. 7.2a). Four sound scenes were picked from the ones used by Huang and Elhilali (2017) (trimmed to 30 seconds) and up to 3 of the 10 stimuli were picked

randomly to be inserted into each scene. To spread the stimuli over the duration of the scene, each one was placed in one of three 10-second ranges, and the position within the ranges was randomised. Participant's mouse movements during the scene presentation was collected.



**(a)** Part 1 – self-reporting **(b)** Part 2 – distraction

**Figure 7.2.:** Graphical interfaces for the first two experimental parts. There was no visual interface for the third part (oddball).

**The second part**, distraction (Chapter 6), was a visual task with auditory distractors. On each trial, an image of an arrow pointing either up or down was displayed (see Fig. 7.2b). The task was to press the up or down arrow on the keyboard in response to the visual stimulus as quickly as possible. Simultaneously with the arrow, a sound was played: the standard was a 500 ms long 440 Hz tone (90% of trials), and the deviant distractors were the stimuli described above (10% of trials). Each distractor stimulus was played twice, which gives a total of 20 distractors randomly inserted among 180 standards. Response times to the stimuli were measured as a proxy for salience (the larger response times – the larger the distraction – the more salient the sound).

**The third part**, SOAP, was again very similar to the original by Tordini et al. (2013) and corresponds to the experiment in Chapter 3. The regular stimuli were 500 ms long, and shortened versions (378 ms) of the stimuli were made using Audacity, in order to ensure asynchrony (sounds were shortened without affecting pitch). Regular

inter-stimulus interval was 250 ms, and the task was to detect a shortened inter-stimulus interval (80 ms) and indicate whether it was in the right or left ear, by pressing left or right arrow on the keyboard. Response accuracy (the larger, the more salient the sound) and response times (the larger, the less salient) were measured.

The three parts were all presented in one sitting (about 30 minutes in total), always in the same order, as described above: self-reporting, distraction, oddball detection. The particular order was based on trying to avoid participants' inhibition to the stimuli as much as possible. It was assumed that it was most crucial for the self-reporting experiments that the participants had not heard the sounds before. The oddball experiment was placed last, because its method involved multiple repetitions of the stimuli.

## 7.3. Results and discussion

### 7.3.1. Data pre-processing

In the first part, to calculate a salience score from mouse movements for each stimulus, first, scores were assigned to each measurement point over the duration of the stimulus, so that score of 1 means the scene was attended (cursor in the matching area), $-1$ – the opposite scene was attended, 0 – neither (cursor in the middle). Then, the scores were summed over the duration of the stimulus (500 ms) and the score was divided by the number of measurement points in the 500 ms range, so that it is in the range between $-1$ and 1. Additionally, a hit rate was calculated – each stimulus was assigned a 1 when it was reported and 0 when it was not, then these were averaged for each stimulus. While the hit rate metric only measured how often a stimulus was attended, the "score" could also give an indication of the average period of time over which it was attended.

For part 2, response times (RT) were used. All RTs were above 200 ms. Correct response rate was at least 96% for all participants, on average 98%.

For part 3, response times and correct response rate (hit rate) were recorded. Because response times in the SOAP paradigm are the only metric used here which is expected to have a *negative* correlation with salience, to facilitate comparison with other metrics, these response times were transformed to *response speed* ($speed = 1/RT$). The average hit rate was 67% and varied between participants from 49% to 88%.

Response times generally tend to show significant differences between participants. Each participant's median response times in part 2 (distraction method) were compared with their response times and hit rate in part 3 (oddball method). Both of these show clear correlations, with Pearson's coefficients of 0.77 and -0.75 respectively (see Fig. 7.3). This means that a participant with generally lower scores and longer RTs in the oddball detection method, tended to also have long RTs in the distraction method. To avoid this effect obscuring the differences between stimuli, response times and speed were normalised by subtracting an individual's median response.

11 volunteers participated in the test. Fig. 7.4 shows the metrics obtained from each of the methods for each stimulus. The stimuli are sorted from most to least salient according to Zhao et al. (2019).

### 7.3.2. Statistical analysis

To summarise, 5 metrics of salience were obtained from the three methods described above: self-reporting score, self-reporting hit rate, distraction response time, oddball detection hit rate and oddball detection speed. All of these are expected to show higher values for more salient sounds. To allow for comparison between these metrics, they were aggregated per stimulus. For response times and speed, the median was

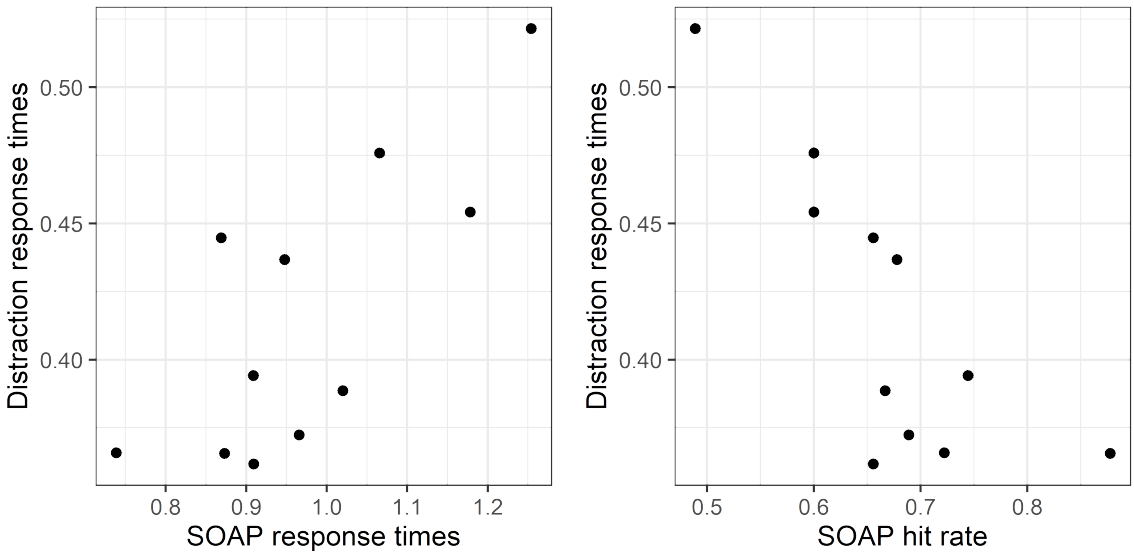
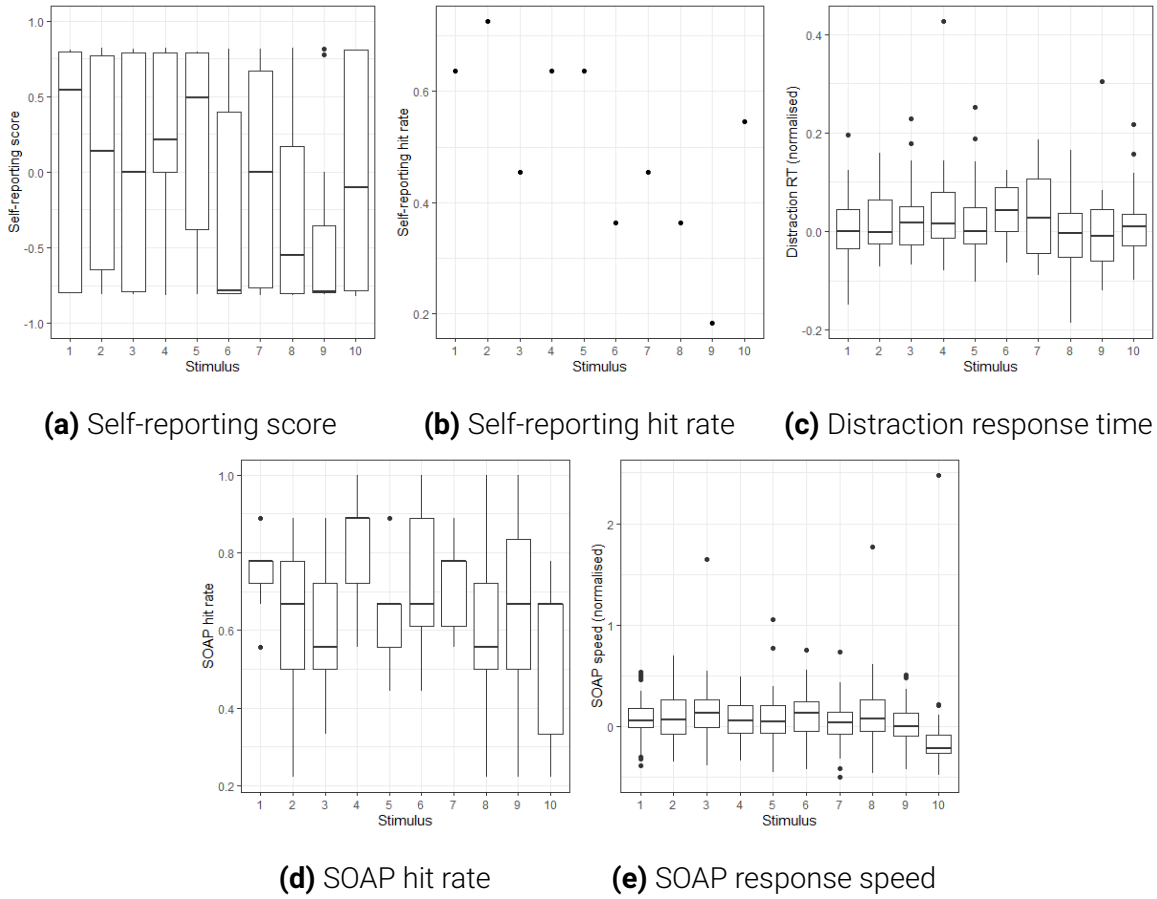



**Figure 7.3.:** Individual average RTs and hit rates. Pearson's correlation coefficient for Distraction and SOAP RTs: 0.77, for Distraction RT and SOAP hit rate: $-0.75$.

calculated, and the mean for hit rates and scores. Figure 7.5 shows how the metrics correlate with each other and with the salience scores from Zhao et al. (2019).

Perhaps unsurprisingly, the self-reporting score and hit rate – calculated from the same mouse movement data – show a very high correlation. There is also reasonably high positive correlation between the distraction response times and the oddball hit rate, and between the oddball hit rate and response speed – although the latter perhaps smaller than expected, given that they are both obtained from the same experimental method. Additionally, both the self-reporting metrics and the SOAP response speed seem to correlate with the salience scores from Zhao et al. (2019). The weakest correlation appears to be between the self-reporting metrics and oddball metrics. In general, the correlation matrix suggests two clusters of metrics: 1) the self-reporting metrics, 2) the oddball and distraction metrics, with the Zhao et al. (2019) score showing some correlation with both of these clusters.

To further investigate how the different metrics of auditory salience interact, a




**(a)** Self-reporting score **(b)** Self-reporting hit rate **(c)** Distraction response time

**(d)** SOAP hit rate **(e)** SOAP response speed

**Figure 7.4.:** Scores obtained in the 3 experimental parts for each stimulus. Boxplots show the median, 25th and 75th percentile of scores calculated per participant.

Principal Component Analysis (PCA) was performed, with data again aggregated per stimulus. The Kaiser-Meyer-Olkin measure of sampling adequacy for the dataset was 0.6 and Bartlett's test for sphericity p-value = 0.003. These indicate that the dataset can be considered for factor analysis.

The analysis was performed using the FactoMineR package in R (Lê, Josse and Husson 2008), and the obtained components are summarised in Table 7.1. Based on the eigenvalues, the first two components were chosen for further exploration, which cumulatively explain 80% of the variance in the data. To facilitate interpretation of

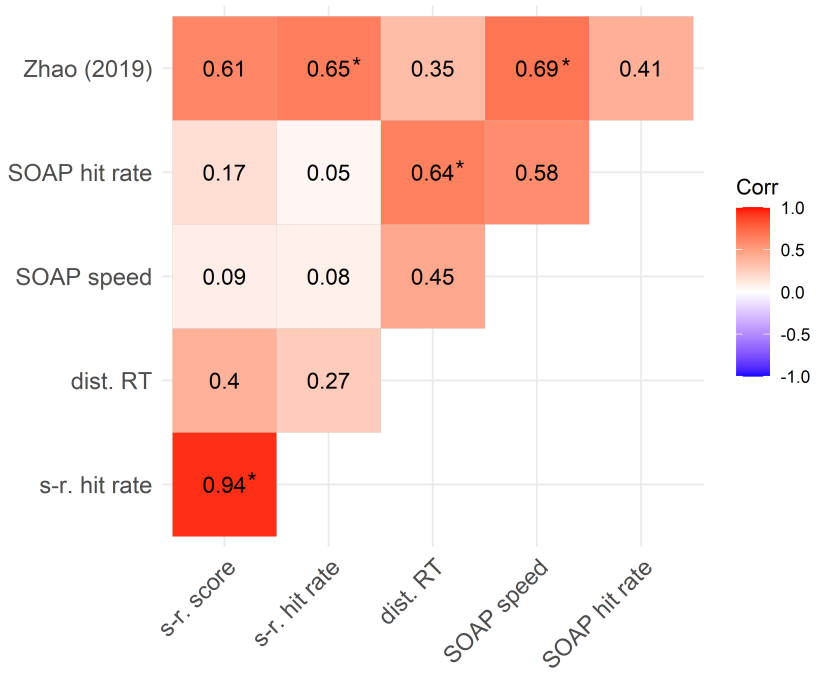



**Figure 7.5:** Matrix of Pearson's correlation coefficients between all 5 salience metrics and the scores from Zhao et al. (2019). Asterisks show correlation coefficients with p-value < 0.05.

dimensions in the data, a varimax rotation of the two components was performed . Fig. 7.6 shows the two resulting dimensions before and after rotation.

| | Comp1 | Comp2 | Comp3 | Comp4 | Comp5 | Comp6 |
|---|---|---|---|---|---|---|
| Eigenvalue | 3.16 | 1.63 | 0.78 | 0.31 | 0.08 | 0.04 |
| Percentage of variance | 52.71 | 27.18 | 12.92 | 5.17 | 1.40 | 0.62 |
| Cumulative % of variance | 52.71 | 79.89 | 92.80 | 97.98 | 99.38 | 100.00 |

**Table 7.1.:** Dimensions found in the Principal Component Analysis.

What can be seen in the plots of the variables is that the metrics obtained from the self-reporting and oddball experiments seem to be orthogonal on the first two dimensions. The self-reporting metrics load almost exclusively on the first component, while the oddball metrics load mainly on the second. The distraction response times

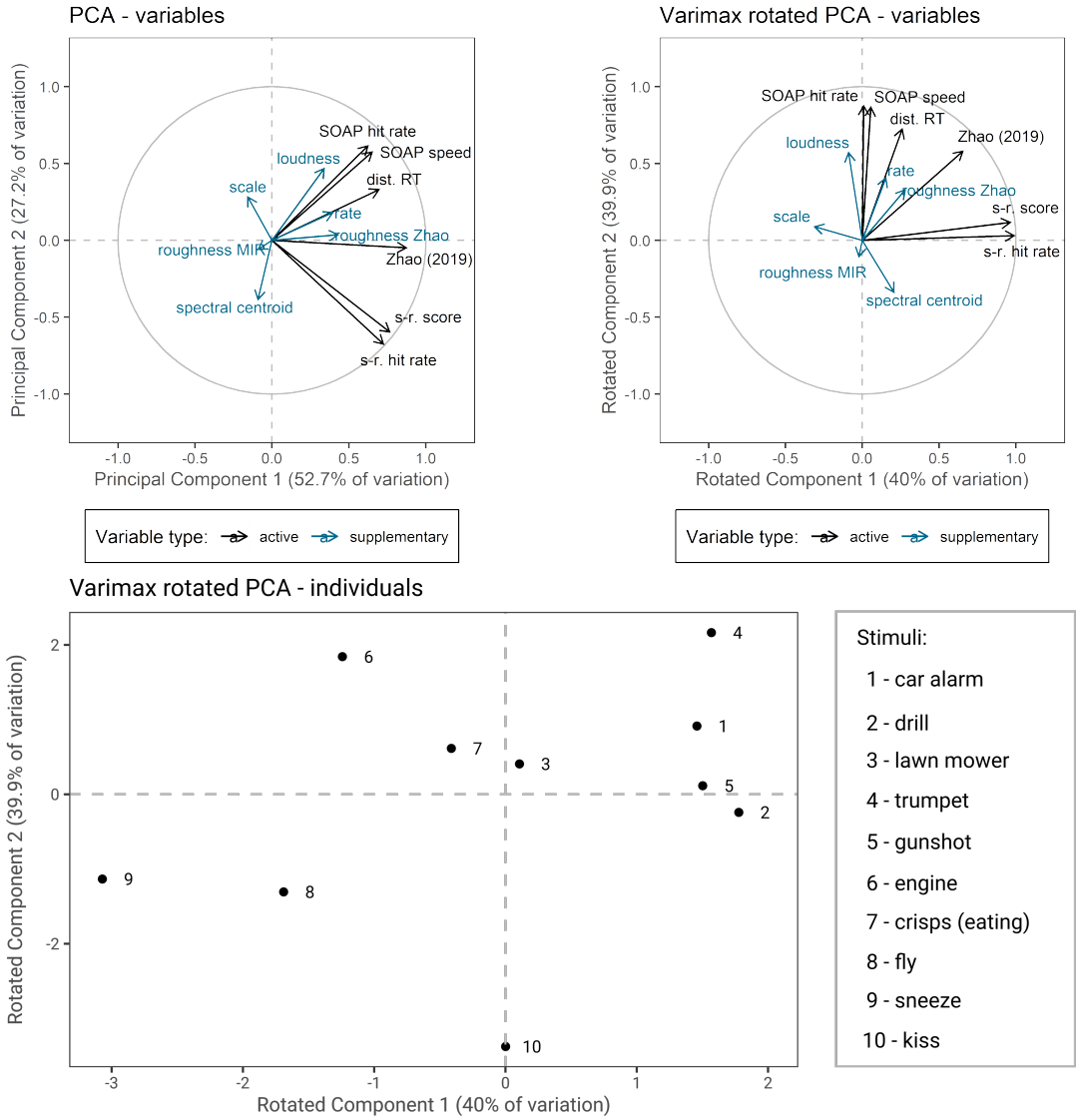



**Figure 7.6.:** Results of a PCA analysis on the tested metrics. Top panel shows the active variables (salience metrics) and supplementary variables (acoustic features) on the first two components. The plot on the left shows the original principal components, while the plot on the right shows the two components after a varimax rotation. The bottom plot shows the individuals (stimuli) on the rotated components. The number of the stimulus corresponds to the Zhao et al. (2019) salience score, with 1 being the most, and 10 the least salient.

also load more strongly on the second dimension than the first. This suggests that there are important differences in the salience metrics obtained from the experimental methods tested here. A possible interpretation of the two rotated components is that the first one represents higher-level, conscious 'seek' processes, while the second – a lower-level, autonomous reaction. Interestingly, the Zhao et al. (2019) scores seem to be in the middle, loading onto both components to a similar extent. Therefore, this metric might in fact be interpreted as a sum of the self-reporting and oddball and distraction scores, perhaps measuring a combination of both high-level and autonomous attentional processes.

The bottom panel of Fig. 7.6 shows the position of each of the stimuli on the two rotated components. Note that the number assigned to the stimuli corresponds to the Zhao et al. (2019) salience score, with 1 being the most, and 10 the least salient of the stimuli according to that scale. While sounds 8 and 9 are low on both dimensions, stimulus 10 is very low on the second dimension (dominated by the lower-level experimental methods) but does not seem to contribute to the first (high-level) one at all. On the other hand, the most salient stimulus according to Zhao et al. (2019) ("car alarm") contributes to both dimensions, but the second most salient ("drill") is loaded mostly on the first, dominated by self-reporting. The sound which is most salient on both dimensions seems to be stimulus 4 ("trumpet"). It is worth noting that while the stimuli used here span a range on the salience metric of Zhao et al. (2019), they might all be salient to some extent, as they are all distinct sound events which have been selected for an experiment studying exactly that. In addition, the dataset does not include any anchor points, such as sounds which would not be expected to cause any attentional response, therefore it would be difficult to judge salience of these stimuli in absolute terms.

To provide additional information about the dimensions found, the following acoustic features which have been associated with auditory salience were calculated for each

stimulus:

1. **loudness**, which has been shown multiple times to be a strong salience feature (Kim et al. 2014; Liao et al. 2015; Huang and Elhilali 2017); it was obtained from the model for varying sounds (Glasberg and Moore 2002) implemented in the Loudness Toolbox (Genesis 2009);
2. **brightness**, represented by the spectral centroid, has also been suggested as a strong feature (Tordini, Bregman and Cooperstock 2016), calculated here with Matlab's *spectralCentroid* function;
3. **roughness** has also been associated with high salience (Zhao et al. 2019; Arnal et al. 2015); roughness was extracted using the MIR Toolbox (Lartillot and Toiviainen 2007), which computes dissonance between peaks of the spectrum; furthermore, roughness values reported by Zhao et al. (2019) were added, who obtained them from Modulation Power Spectra (MPS), as a contribution of high frequency temporal modulations in the MPS;
4. **rate** and **scale**, as measures of temporal and spectral modulations respectively, are included after Huang and Elhilali (2017), who found them to be important in explaining events attended by participants; they were calculated using the *NSL Auditory-Cortical Matlab Toolbox* (2008).

All features have been calculated in Matlab, over short time-windows and averaged per stimulus. These features were added as supplementary variables and are also shown in Fig. 7.6 (in blue).

The feature which shows the highest correlation with any of the dimensions is loudness, which loads highly on the second component, suggesting it mainly affects lower-level salience mechanisms. It also seems highly correlated with the oddball hit rate and response speed. In the original experiment by Tordini, Bregman and Cooperstock (2016), calculated loudness was the second most correlated with oddball responses, even after initial perceptual loudness matching of the stimuli.

Rate, a measure of temporal modulation, also seems to correlate, however less strongly, with the second component, and with the distraction response times. Spectral centroid shows a negative correlation with the second component, suggesting that the oddball detection metrics were higher for sounds with low spectral centroid. This is in agreement with what Tordini, Bregman and Cooperstock (2015) found with the oddball detection SOAP paradigm, although the experiment in Chapter 3, where a similar method was used, did not show a significant effect of the spectral centroid. Roughness calculated from the Modulation Power Spectrum correlates well with the salience metric from Zhao et al. (2019), as was reported in their paper. However, roughness measured by averaging the dissonance between spectrum peaks does not correlate with either of the dimensions.

In general, the features tested here seem to correlate more with the second component, which supports the idea of it representing lower-level salience processes. In other words, the automatic attention-grabbing properties are more related to low-level acoustic features, while the higher-level switching of attention towards a sound of interest might rely on other, higher level characteristics, such as semantic meaning or emotional connotations.

Naturally, the selection of features presented here is in no way complete in the sense of the information available to the brain when analysing the auditory environment. Also, the selection of stimuli is not a representative sample in terms of the extent of these features.

## 7.4. Conclusions

The lack of a broadly agreed-upon and standardised method of testing auditory salience is still a barrier on the way to developing new auditory salience models. Here,

the three methods used in previous chapters were compared: oddball detection, self-reporting and distraction, and a previously published salience score based on a large-scale ratings-based survey.

What emerges from this comparison is that metrics derived from these methods do not all correlate with each other. In fact, no correlation at all was found between the self-reporting and oddball detection metrics, and very low correlation between self-reporting and distraction.

A PCA analysis suggests that two independent components underlie these metrics: one higher-level, more conscious, measured by the self-reporting method, and the other lower-level, more automatic, which is represented by the oddball detection and distraction methods. Interestingly, the salience scores from Zhao et al. (2019) seem to fall between the two, seemingly capturing both aspects of salience.

# 8

# Summary

Part I of this thesis discussed methods of measuring auditory salience and what effect spatial position of sound has on salience. Four experiments conducted to study these effects are summarised in Table 8.1. An oddball detection experiment described in Chapter 3 tested low-level auditory salience, with short noise bursts as stimuli. Oddball inter-stimulus-intervals in high frequency noise patterns were less likely to be detected if the pattern was behind the listener. The experiment described in Chapter 4 tested auditory salience in a more ecologically valid listening scenario. Participants heard recordings of sound events and reported their attention in real-time. Similarly as in the oddball experiment, sounds with high brightness were attended to less when they were behind the participants. The effects of perceptual load were evaluated in a dual-task detection experiment described in Chapter 5. The results suggest a small advantage for detecting sounds to the right in low perceptual load conditions, and behind the listener independent of perceptual load. The experiment in Chapter 6 studied effects of expectations of type and location of sound on auditory salience. Violations of spatial expectations were shown to elicit pupil dilation responses – however the effect was smaller than for violations of expectations about sound type. All of the results above are discussed in light of current research in Chapter 13.

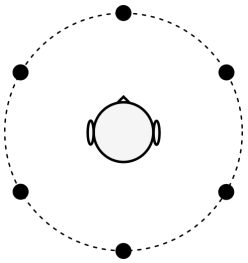
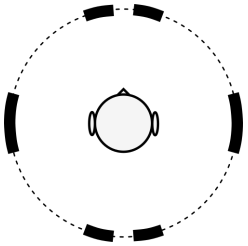
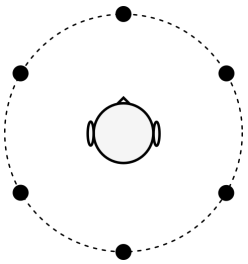
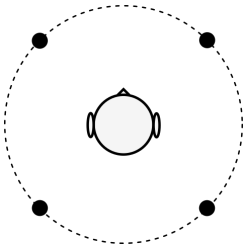
| Ch. | Experimental method | Stimuli | Tested positions | Main result |
|---|---|---|---|---|
| 3 | oddball detection (Segregation of Asynchronous Patterns) | high- and low-pass filtered noise bursts |  | slower detection in the back for high-pass filtered noise bursts |
| 4 | real-time attention tracking | environmental sound recordings |  | high spectral centroid sounds in the back attended to less than those in the front |
| 5 | detection in a secondary task | environmental sound recording (ice-cream van) |  | target easier to detect when behind, compare to the front |
| 6 | distraction in an auditory classification task | environmental sound recordings |  | breaking spatial continuity causes a pupil dilation response |

**Table 8.1.:** Summary of experiments described in Chapters 3 to 6.

Finally, the comparison experiment in Chapter 7 inspected results obtained from three of the methods in response to the same stimuli, and compared them to a previously published salience score. A PCA revealed at least two different dimensions of salience on which these metrics are: lower-level automatic salience, which correlates with acoustic features of sound, and higher-level salience, likely influenced by other characteristics, such as meaning and emotions.

# Part II.

# Modelling auditory salience

# 9

# Literature review

## 9.1. Introduction

The goal of auditory salience modelling is to predict which sounds will attract the attention of the listener, and to what extent.

There are a few different ways in which such a model can be built. For example, the input to the model might be a recording of a sound scene, or multiple recordings of auditory objects present in a scene. The output might be a salience score over time or a single score for each object, indicating how attention-grabbing it is compared to other objects (see Figure 9.1). In this chapter, auditory salience models published so far will be reviewed, and perceptual principles, which can provide a baseline for salience modelling, will be discussed.

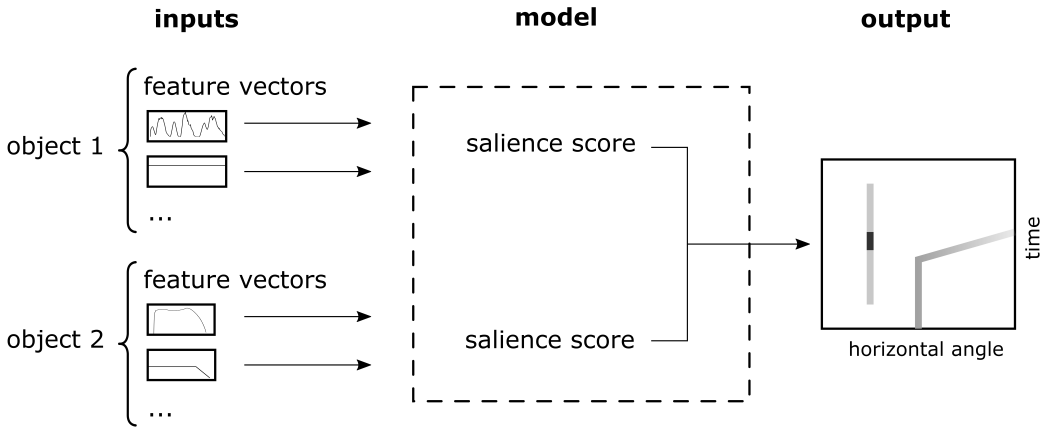



**Figure 9.1.:** Example auditory salience model structure.

## 9.2. Review of models

First attempts to model auditory salience were directly inspired by the concept of a visual salience map (Koch and Ullman 1985; Itti, Koch and Niebur 1998). Kayser et al. (2005) used this framework in audition by performing a spectro-temporal analysis of a sound signal and effectively treating the resulting spectrogram as an image. From it, three salience features – intensity, temporal structure and frequency structure – were calculated at four different spectrotemporal scales. Maps representing different scales were combined and normalised. Finally, maps from all three features were combined together to create a single salience map. This process is illustrated in Figure 9.2.

This approach was a simple adaptation of a known visual model to the auditory domain, yet it failed to take into account the differences between auditory and visual domains, such as the temporal nature of sound. Sound is one-directional and with short- and long-term dependencies which are crucial to perception, including the perception of salience. For example, Kaya and Elhilali (2017) suggest that a sudden start of loud noise could potentially be more salient than when the reverse happens – a sudden silence after a loud noise. In other words, the way sound scene has been building up to a particular point in time will influence the expectation of future sound

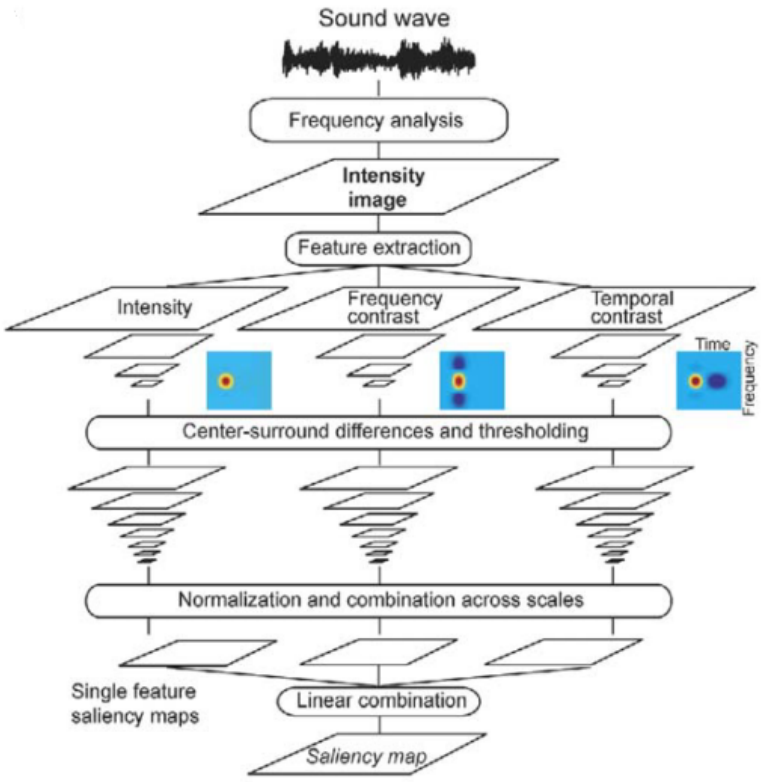



**Figure 9.2.:** Step-by-step computation of an auditory salience map. From Kayser et al. 2005.

events. Moreover, a spectrogram, as a representation of sound intensity in time and frequency, does not contain phase information. Filipan et al. (2016b) demonstrated that a phase modification distinctively audible to human listeners was not visible on a spectrogram and hence not picked up by a classic salience map algorithm.

Despite its limitations, Kayser's model was an inspiration for other researchers who modified it in various ways: added more relevant auditory features [1] (Kalinli and Narayanan 2007) and a biologically-inspired cortical model (Duangudom and Anderson 2007), or used a one-dimensional temporal salience score instead of a two-dimensional map (Kaya and Elhilali 2012).

[1] These were pitch and orientation features, which simulate responses to moving ripples – dynamic broadband sounds, with a modulated spectrum.

A different approach to modelling salience is to calculate it from signal statistics. Tsuchida and Cottrell (2012) adapted a visual SUN model (Salience Using Natural Statistics) and based their model on short- and long-term (life-long) sound statistics, which serve as prior information about the characteristics of natural sounds.

Schauerte et al. (2011) introduced a model based on auditory Bayesian surprise, similar to the visual model by Itti and Baldi (2009). It calculates surprise as the Kullback-Leibler divergence between prior and posterior probabilities of the incoming signal's frequency spectrum. The probabilities were initially modelled as Gaussian distributions, however, the authors later modified their approach by using Gamma distributions instead (Schauerte and Stiefelhagen 2013). More recently, Rodríguez-Hidalgo, Peláez-Moreno and Gallardo-Antolín (2018) proposed a salience model based on Bayesian *log-surprise* calculated over multiple time scales. After determining the Kullback-Leibler divergence for (Gaussian) signal distributions at two consecutive time steps, they use the logarithm of it as the salience score. The time scales, or memory spans, dictate the time windows over which distribution means and variances are calculated. Salience scores from 5 time scales ranging from 680 ms to 3.4 s are then combined into one score. They tested their model in an acoustic detection task and concluded it outperforms baseline models, including a saliency map (Kalinli and Narayanan 2007) and log-surprise without the memory aspect.

Kaya and Elhilali (2014) proposed a predictive-coding model in the form of a Kalman filter as a deviance detector (see Figure 9.3). The model also includes interactions between features, and it was the first model not to be based on a vision equivalent. Another biologically-inspired class of models is based on neural networks (Wrigley and Brown 2004; Boes et al. 2012).

Attention models which go beyond salience usually also include a stream competition mechanism which is influenced by both top-down and bottom-up effects, often with

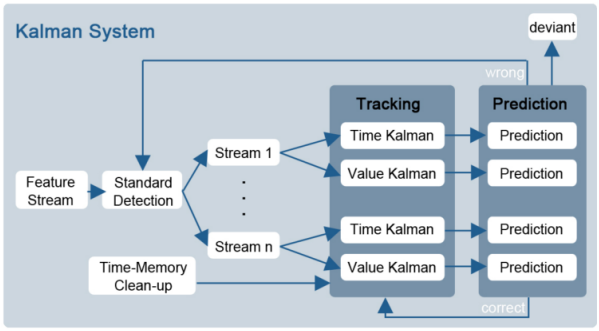



**Figure 9.3.:** Principle of a deviance detector based on predictive coding, from Kaya and Elhilali 2013

inhibition-of-return, which slowly decreases the salience of an event or stream over time (De Coensel and Botteldooren 2008; De Coensel and Botteldooren 2010; Boes et al. 2012). Other approaches model top-down attention as a form of weighting: for example, of neural network parameters (Wrigley and Brown 2004) or of features (Kalinli and Narayanan 2009). Also, Carlin and Elhilali (2015) proposed a model for attentional modulation of spectro-temporal receptive fields, which can operate both on features and objects. Various types of auditory salience models are summarised in Table 9.1.

Almost all auditory salience models only use one-channel recordings of a scene or one-channel signals representing each object separately. Wrigley and Brown (2004) take two-channel input, but they only use it to extract binaural pitch which helps with harmonic grouping, and no spatial information is taken into account. They also allowed for allocation of attention to each ear separately, but not to a specific location. The models which used separate signals for each object (De Coensel and Botteldooren 2008; De Coensel and Botteldooren 2010) do not take location of the objects into account.

| Authors | BU principle | TD principle | Features | Input | Output |
|---|---|---|---|---|---|
| De Coensel and Botteldooren 2008 | PD | attention switching mechanism | - | object signals | level of attention per object |
| Kayser et al. 2005 | SM | - | intensity, frequency contrast, temporal contrast | 1-ch signal | spectrotemporal salience map |
| Duangudom and Anderson 2007 | SM | - | output of a cortical model (global energy, temporal and spectral modulations) | 1-ch signal | spectrotemporal salience map |
| De Coensel and Botteldooren 2010 | SM | competition mechanism | intensity, spectral contrast, temporal contrast | object signals | 0/1 switch per object |
| Kaya and Elhilali 2012 | SM | - | envelope, spectrogram, rate, bandwidth, pitch | 1-ch signal | temporal salience score |
| Kalinli and Narayanan 2009 | SM-based gist features | task-dependent bias | intensity, frequency and temporal contrast, FM slope, pitch variations | 1-ch signal | prominent syllable detection |
| Tsuchida and Cottrell 2012 | P/B | - | cochleagrams (reduced to 2-3 dimensions with PCA) | 1-ch signal | spectrotemporal salience map |
| Schauerte et al. 2011; Schauerte and Stiefelhagen 2013 | P/B | - | spectrogram | 1-ch signal | salience score |
| Rodríguez-Hidalgo, Peláez-Moreno and Gallardo-Antolín 2018 | P/B | - | spectrogram | 1-ch signal | salience score |
| Kaya and Elhilali 2014 | predictive coding | - | envelope, harmonicity, spectrogram, bandwidth, temporal modulation | 1-ch signal | temporal salience likelihood |
| Wrigley and Brown 2004 | NN | weighting | pitch | 2-ch signal | spectrotemporal attentional stream |
| Boes et al. 2012 | NN | competition mechanism | intensity, time contrast, frequency contrast | 1-ch signal | neuron activity |
| Kim et al. 2014 | linear filter | - | loudness | 1-ch signal | salience detection |

**Table 9.1.:** Summary of models. Bottom-up principle types: PD – time domain peak detection, SM – salience map, P/B – probabilistic and/or Bayesian, NN – neural networks

## 9.3. Incorporating spatial information

The results described in Part I suggest that spatial location as such might not play a big role in determining salience of sounds. However, it is still worth keeping track of it, as it can be susceptible to the same expectation rules which affect other features. How can spatial location then be incorporated into auditory models?

Because the field of visual attention modelling is more developed, it is worth looking at which parts of it could be applied to auditory attention, or how to translate principles and ideas from one into the other. To do this, one needs to consider similarities and differences between the two modalities. Of special interest here would be finding a visual equivalent of the role spatial location plays in auditory processing.

One of the main differences between vision and audition is how space and time are processed. In vision, spatial information is derived directly from retinal image – visual information such as colour and intensity is in its essence spatially coded. This is in contrast to audition, where basic coding is tonotopic, and spatial information needs to be derived from the difference between signals incoming to left and right ear.

What analogies can then be drawn between visual and auditory dimensions? Kubovy (2017), in his Theory of Indispensable attributes (TIA), compared visual location to auditory frequency. He defines indispensable attributes as those that are necessary for perception of multiple objects, and argues that in vision space is such an attribute, while in audition it is frequency. He goes further to claim that "attention is allocated to pitch, not to location", which negates the existence of spatial auditory attention entirely. TIA has faced some criticism from the scientific community (e.g. Handel 1988). A study performed by Neuhoff (2003) has shown that changes in pitch are not at all necessary for formation of auditory objects, which contradicts the idea of frequency as an indispensable attribute. Also, as mentioned in Chapter 2, spatial cues

increase performance in an auditory task, which indicates the ability to direct auditory attention to space.

Perhaps the most interesting analogy for this research has been proposed by Shamma (2001), who described how some architectures and computational algorithms used in vision may also be used to explain certain auditory phenomena. One of the proposed analogies is between binaural azimuthal localisation and binocular depth perception. They both operate on the basis of comparing input received by two sensory organs (ears or eyes), and by doing so add an extra dimension to our perception. In fact, as the author notices, algorithms designed to determine interaural time differences and spatial disparities between binocular images are fundamentally identical.

In light of this theory, it would be worth considering different ways in which depth has been incorporated into visual salience models. In fact, most of them concentrate on 2-dimensional vision, similarly to how their auditory counterparts focus on non-spatial hearing. Existing 3D visual salience models have been divided by Wang et al. (2013) into:

- depth-weighting models, which assume that salience differs with depth, so each spot on a 2D salience map is assigned a different weighting depending on corresponding depth,
- depth-salience models, which calculate depth salience maps based on depth-related features and combine them with 2D salience maps,
- stereo-vision models, biologically plausible, which incorporate stereoscopic perception processing directly.

Figure 9.4 illustrates this classification.

In audition, the analogy to the first type of models would be assigning each horizontal location a different salience weighting. In light of the results of the experiments carried out in the previous chapters, this strategy might not be feasible, as there is no

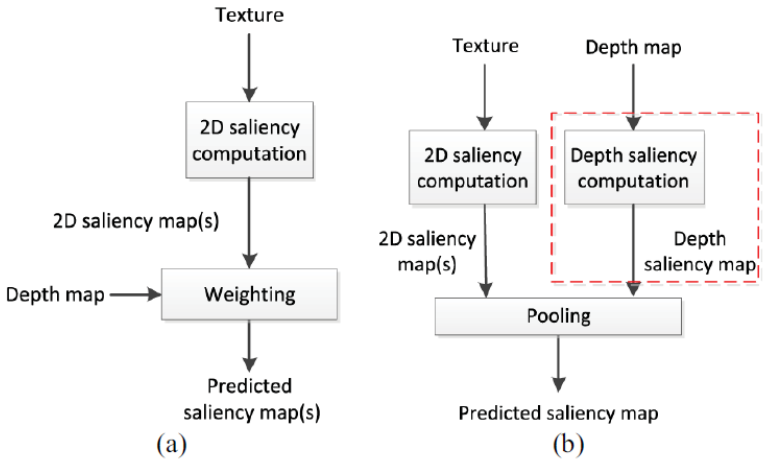



**Figure 9.4.:** Comparison of depth-weighting and depth-salience models, from Wang et al. 2013

indication of different weights for different locations. An analogy to the second type of visual models would require calculating auditory salience features derived from spatial properties of sound. These could for example be unexpected changes in sound location or a form of spatial contrast. Biologically inspired attempts could incorporate binaural localization models (such as Lindemann 1986; Breebaart, Par and Kohlrausch 2001; Dietz, Ewert and Hohmann 2011) or some form of extension of neural networks like in Bruce and Tsotsos (2005).

This thesis will propose a way in which spatial information can be added to a deviance detection-based auditory salience model as a new feature. There is evidence that models based on prediction and Bayesian principles might correspond to how the brain processes information. These principles are reviewed below.

## 9.4. Perceptual principles

Let us now consider attention in a probabilistic framework of perception – how the brain uses Bayesian principles, prediction and summary statistics – as it can serve as a

basis for auditory salience models.

### 9.4.1. Bayesian brain

Many cognitive and perceptual processes can be thought of in terms of inference – for example, given the light received by the eyes, what is the object in front of the observer (or how fast is it moving)? Note that in most cases, sensory input is just one type of data available. Another source of information is the knowledge the observer has about the world, or has gathered in recent observations (e.g. what objects are likely to be present in the particular situation, or basic laws of physics which govern movement). This problem might have more than one possible answer and the most likely one has to be selected. A framework that can describe these kinds of processes well is Bayesian inference, and it has in fact been argued that humans often do behave like optimal Bayesian observers (Knill and Pouget 2004).

Bayesian inference determines the probability of a hypothesis based on incoming data. In Bayes' theorem, this would be the probability of a certain hypothesis A given the event (data) B happened, which is related to a) the probability of acquiring data B under hypothesis A (the likelihood), and b) the probability of hypothesis A itself (the prior).

Take an example of determining the source of incoming sound. Auditory input was received by a listener and the listener would like to know whether this sound was produced by a tiger. With Bayes' theorem, this probability could be determined as follows:

$$P(tiger|sound) = \frac{P(sound|tiger)P(tiger)}{P(sound)} \tag{9.1}$$

So, the probability that the listener heard a tiger, given the received auditory input, is equal to how likely this input is to have come from a tiger, times the probability that there would be a tiger around to produce that sound. The denominator is a

normalising factor and does not depend on the hypothesis. This equation could be evaluated over many different potential sources (hypotheses) to find the most likely one. This naturally implies that the brain operates on probabilities rather than deterministic values, something Knill and Pouget (2004) call a Bayesian coding hypothesis.

Elhilali (2013) has proposed Bayesian inference as a framework which unifies bottom-up and top-down effects on sound perception. In this framework, bottom-up effects are built up from the auditory input (data), while top-down processes such as attention are the priors. Wolmetz and Elhilali (2016) presented a Bayesian model of auditory scene analysis which distinguishes between attention-driven (top-down) and context-driven (bottom-up) priors, which they tested with behavioural data obtained from a cued detection task with informative and uninformative cues. In a similar manner, Whiteley and Sahani (2012) proposed a model of perception where attention is not strictly a prior, but acts in a similar way on the inference (the classic prior is still there and can be influenced by the data). In this model, it is assumed that the perceptual representation is an approximation of the sensory input and the role of attention is to locally refine this approximation in the area of interest (notice the similarity with the hierarchical model of grouping by Cusack et al. (2004) and the evidence for attention tuning perception, described in Chapter 2).

### 9.4.2. Predictive coding

In 1999, Rao and Ballard proposed that the visual cortex is organised in terms of predictive coding: with predictions propagating down the neural hierarchy, and precision errors propagating upwards, as feedback. This idea has proven quite influential in neuroscience and has since gathered substantial evidence (Friston 2018).

Prediction error, the difference between prediction and sensory input, can be used in

Bayesian inference to refine priors (Aitchison and Lengyel 2017). In essence, at each point in time, first a prediction is obtained based on past input and priors (i.e. what is the likely next input given a hypothesis about the latent source of the input), then that prediction and current input are compared to get prediction error, and finally this error is used to correct the prior (probability of the hypothesis). Aitchison and Lengyel (2017) argue however that predictive coding has the potential to serve other purposes in the brain as well, and should be seen as a general "neural motif".

Heilbron and Chait (2017) reviewed evidence for predictive coding in the auditory cortex, examining animal and human brain imaging studies. They conclude that a generative (predictive) nature of the auditory cortex is likely, and neural activity seems to represent prediction error. For example, they point to evidence from responses to omission, which show brain activity recorded in response to sounds *omitted* in a sequence. These occur only when subjects do not expect the gap, and are very similar to responses to actual stimuli. Heilbron and Chait (2017) also argue that suppression of responses to repetitive stimuli might actually be caused by improved prediction rather than adaptation or neural fatigue. Brain imaging provides evidence that predictive coding is hierarchical (Wacongne et al. 2011; Heilbron and Chait 2017). Attention is likely to act on all these different levels (time scales).

In the predictive coding framework attention is sometimes seen as the process of optimising precision (Friston 2009) – weighting of the prediction error in order to enhance response to the stimuli of interest. While on the bottom-up level, precision is based on input reliability, top-down it can be directed towards a specific input (as expectation of reliability or higher confidence in the attended input). Chennu et al. (2013) investigated how top-down attention and expectation interacts with different levels of bottom-up stimuli-based predictability. They concluded that attention to sound enhanced local prediction error, but expectation about the patterns attenuated it. However, Heilbron and Chait (2017) concluded that the empirical evidence for a

relationship between attention and predictive precision is not very strong in audition.

### 9.4.3. Summary statistics

The probabilistic view of perception fits well with the evidence that the brain stores statistical information about sound, particularly the evidence coming from research on sound textures.

McDermott, Schemitsch and Simoncelli (2013) conducted a texture discrimination experiment to study if summary statistics are used in the perception of artificially generated auditory textures. They found that when discriminating between excerpts from two different textures (with different statistical properties), increasing the excerpt's length enhanced performance, as expected, because more information was available. However, for differentiating between two exemplars of the same texture, increasing the length actually worsened performance. This is consistent with the idea of the brain storing summary statistics – as more input comes in, the statistics of the same type of texture will converge to very similar values. They propose that there are two parallel processes: 1) sequential encoding of input, and 2) summary statistics calculated over time. As the sequential buffer is replaced with more recent input, only the statistical representation remains available. Interestingly, they note that it is possible that both processes are in fact the same process but on two ends of a spectrum of temporal scales.

McWalter and McDermott (2018) further investigated this accumulation of statistical information. They presented participants with two generated textures: one with a change of signal statistics ("step" texture) and one with constant statistics ("morph"), and asked which one of these was more similar to a reference texture, presented immediately afterwards. The key here was the direction of the change in the step and at which point it occurred: assuming the perceptual temporal integration window is

longer than the second part of the step, the first part will influence (bias) the similarity ratings. They showed that the beginning of the step did indeed bias the ratings in the predicted direction, even though participants were instructed to make judgements based on the *endpoint* of the step texture. The effect was bigger for more variable textures, suggesting that the brain integrates signal statistics over time windows which vary with texture variability. A possible explanation is that different statistics have different integration windows (e.g. slow vs fast modulations). Additionally, they found that breaking texture continuity by introducing a silent gap near the change significantly reduced this bias. However, when a louder excerpt was inserted instead of the gap (which would not be expected to break the continuity), it did not influence the judgements. This indicates that this statistical accumulation operates on sound streams.

In behavioural and brain imaging experiments, Skerritt-Davis and Elhilali (2018) showed that higher order statistics (covariance between successive inputs) about stochastic sound sequences are tracked by the brain. They used sequences of tones for which entropy increased or decreased midway through the sequence. The participants' task was to detect that change. The authors proposed a Bayesian model which can explain the experimental results, but only when higher order statistics are included. The model maintains multiple hypotheses about the state of the auditory scene and uses prediction error to weight them according to incoming evidence.

### 9.4.4. Salience in a Bayesian framework

If one sees salience as rarity or novelty (Tsuchida and Cottrell 2012), it can be naturally linked to violating expectations in Bayesian inference and to the prediction error directly.

Itti and Baldi (2009) proposed a definition of surprise which states that an input is

surprising if it changes our belief about the world (the prior, or the model), and so the surprise can be measured as the distance between the posterior and the prior. They show that surprise, based on Kullback-Leibler divergence between posterior and prior, was able to predict human gaze in natural recordings better than other metrics, including visual salience metric. They conclude that Bayesian surprise attracts attention. Huang and Elhilali (2017) also emphasise the importance of short- and long-term context for salience.

Some auditory salience models, especially more recent ones, draw inspiration from these theories of perception, and use input statistics or Bayesian surprise to predict novelty – for example, the Kalman filter-based model by Kaya and Elhilali (2014), which will be used in Chapters 10 and 11. The following section gives a short review of the Kalman filter.

## 9.5. Kalman filter

The Kalman filter is a tool that can be used for estimating the state of a linear system from noisy measurements and some existing knowledge about that system. It is a recursive algorithm, which at each iteration: 1) predicts (updates) an *a priori* estimate of the system state, and 2) corrects this prediction according to a measurement, obtaining an *a posteriori* estimate.

Take a process which can be described by a linear equation:

$$x_k = Ax_{k-1} + w_k \tag{9.2}$$

where $x_k$ is the state of the process at iteration (time step) $k$, and $w$ is the process noise, which is normally distributed with covariance matrix $Q$.

The matrix $A$, which describes how process states change in consecutive steps, as well

as the variance of the noise described by $Q$ are both known, but the actual state of the system is not directly available. However, some noisy measurements can be obtained from the process. They can be related to the actual state by matrix $H$ and measurement noise $v$ (again, with normal distribution and covariance $R$):

$$z_k = Hx_k + v_k$$

Knowing measurements $z$, as well as matrices $A$ and $H$, the Kalman filter will try to estimate the process state $x$. For iteration $k$, the **prediction** step of the filter calculates a priori predictions of the next state and error simply from what it knows about the system and the previous estimation:

$$\hat{x}_k^- = A\hat{x}_{k-1} \tag{9.3}$$

$$P_k^- = AP_{k-1}A^T + Q \tag{9.4}$$

Note that $\hat{x}_k^-$ is the *a priori* estimate of the process state $x_k$, while $\hat{x}_{k-1}$ is the *a posteriori* estimate of the state $x_{k-1}$. Similarly, $P_k^-$ is the a priori covariance of estimate errors, while $P_{k-1}$ is the a posteriori error covariance (calculated for iteration $k-1$).

Following the prediction, Kalman gain is calculated, which weights the a priori prediction versus measurement, based on the measurement error covariance $R$ and the estimate error covariance $P$. In the proposed model, the measurement error is kept constant (for each feature), so in practice, as the estimate error gets smaller, the estimate is weighted higher (treated as more reliable), and the measurement is weighted less. The Kalman gain is:

$$K_k = P_k^- H^T (HP_k^- H^T + R)^{-1} \tag{9.5}$$

Next, in the **correction** step, the a posteriori state estimate and error are calculated

from the noisy measurement $z_k$, a priori estimates, and Kalman gain:

$$\hat{x}_k = \hat{x}_k^- + K_k(z_k - H\hat{x}_k^-) \tag{9.6}$$

$$P_k = (I - K_k H)P_k^- \tag{9.7}$$

Figure 9.5 summarises the algorithm.

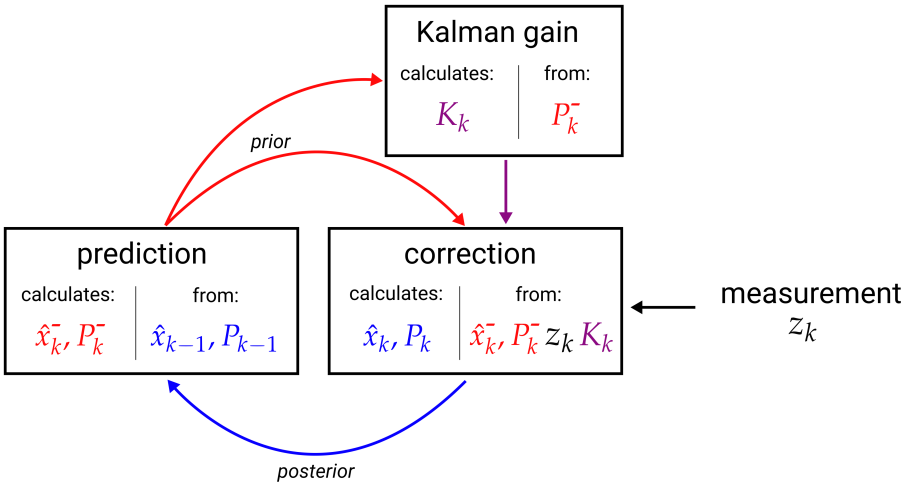



**Figure 9.5.:** Kalman filter steps.

Notice that the term $z_k - H\hat{x}_k$ in Eq. (9.6) represents the difference between actual measurement and its prediction (and it is weighted by Kalman gain). This term is called the model *innovation* and will represent surprise in the model.

## 9.6. Summary

Bayesian inference can serve as a model for auditory salience. The brain constructs models of the environment and predicts future inputs. If the actual input is different from the prediction, it is surprising – salient, and therefore attracts attention.

An example of a model based on this principle is the one by Kaya and Elhilali (2014), which uses Kalman filters to track regularities in acoustic features. The Kalman filter represents most of the ideas described in Section 9.4 – it predicts future input based on present input and the model, it updates the model according to the prediction error, and it weights the model versus data depending on variance of the input (however, it operates on discrete values rather than probabilities).

The following chapters will demonstrate how spatial information can be used in such a model and give an example practical application. In Chapter 10, a way of incorporating the spatial location of a sound into a salience model will be proposed and it will be shown that such a model can successfully predict experimental results. In Chapter 11, an example application of a salience model will be presented, which combined with a machine learning technique can enhance the detection of environmental sounds in audio recordings.

# 10

# Predicting spatial surprise

This chapter will show how a Kalman-based salience model can produce data similar to the pupil dilation results of the distraction experiment described in Chapter 6.

The model described here was based on the work of Kaya and Elhilali (2014), who model the Mismatch Negativity (MMN) brain response with predictive coding implemented as multiple Kalman filters, which track changes in feature streams. The choice of this model is based on the relevance of predictive coding in human perception, and how well it relates to surprise and novelty (see Section 9.4.2).

## 10.1. Features

A practical model of salience requires some form of a feature extraction block. However, here, this stage was omitted for a few reasons. Firstly, what is proposed here is not a full, practical model, but rather a general example which aims to demonstrate how a model based on principles described in the previous chapter can match experimental data. Secondly, even after a lot of careful consideration about feature extraction details and parameters, that step is likely to add some amount of

noise – coming from both algorithm error (e.g. in localisation algorithms) and the sub-optimal choice of features (as there are many possible ones to chose from). Instead, this model will work on optimal features with not too much noise, to focus on demonstrating the principles of the salience model.

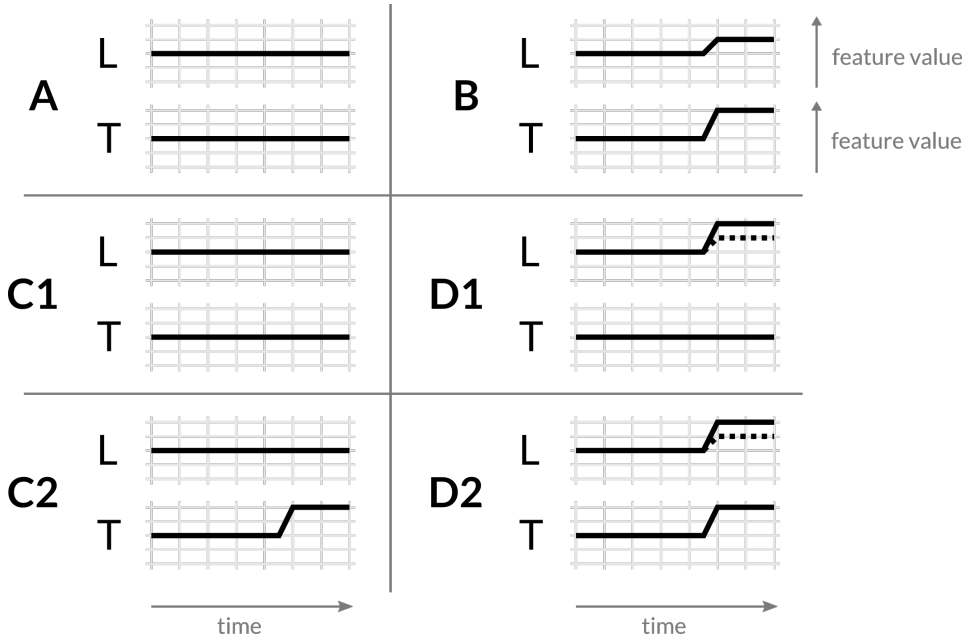



**Figure 10.1.:** Illustration of the location and type features (L – location, T – type). Value of L reflected the actual distance between context and distractor (1 or 2). In condition B, the value for L is quite arbitrary: a sound appears in a new location, but there is no location context as such.

Therefore, the variables used in the Distraction experiment will be used as features directly – namely, spatial location and type of sound. These were represented symbolically, in a way which shows the match and mismatch between context and distractor sounds over time. Figure 10.1 shows how each experimental condition was modelled. The *type* variable for the duration of the context was either 0 or 1, and for the distractor, either 1 (matching type) or 2 (non-matching). The *location* variable was always 0 for the duration of the context, and for the distractor, either 0, 0.5 or 1, depending on the spatial separation between the two. In the distractor-only condition

(B), the *location* variable for the distractor was chosen to be 0.5. The length of context and distractor sounds corresponded to the actual ones in each trial. Gaussian noise with standard deviation 0.05 was added to these to make them more realistic and allow for some algorithm errors, which would normally occur with noisy data.[1] The amount of noise added is arbitrary, but – as will be described below – the model has a parameter which represents the noise covariance of the input. By adjusting it to match the noise in the features, the model can be made to work with a range of different noise levels.

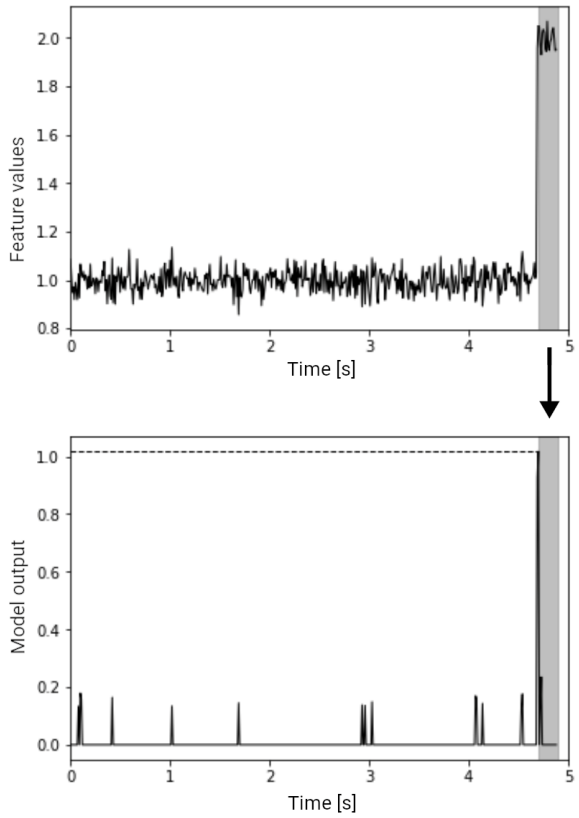



**Figure 10.2:** Example of how the salience score was calculated. Top: artificial feature (sound type) with added noise; the context builds an expectation of a sound of one type (value 1), but in the last 200 ms, the sound type is changed (value 2). Bottom: output of the salience model; the single-number score is the maximum over the duration of the distractor – here, the grey area.

[1] Note that the features are noisy, but if averaged over time, they are always correct (no feature extraction error).

## 10.2. Deviance detection

### 10.2.1. Initialisation

As in Kaya and Elhilali (2014), in the model described here multiple Kalman filters run in parallel on the same feature stream. To determine the number of filters at the start of an audio recording, clustering is performed on the first 500 ms of each feature stream and a Kalman filter is initialised for each of the resulting clusters.

While Kaya and Elhilali (2014) used k-means clustering, in the proposed model Gaussian Mixture Models were used instead, as they gave better results in estimating clusters from simulated feature streams. The advantage of Gaussian Mixture Models is that, apart from the cluster centres, they can also estimate a full covariance matrix for each cluster. This allows for non-spherical clusters and clusters of different shapes (see Figure 4.2). The number of clusters to be initialised was determined by minimising the Akaike information criterion (AIC), with the maximum of 4.

### 10.2.2. Kalman parameters

Each initialised Kalman filter tracks its separate regularity stream within a feature vector. The state of this "process" (as described in Section 9.5) is coded as a matrix containing a feature value and the difference between the last two consecutive feature values (equation 10.1). Taken together with the system matrix A shown below, this means that at any point in time, the feature vector is expected to continue changing in the same manner it has most recently changed.

$$X_n = \begin{bmatrix} x_n \\ x_n - x_{n-1} \end{bmatrix}, A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \tag{10.1}$$

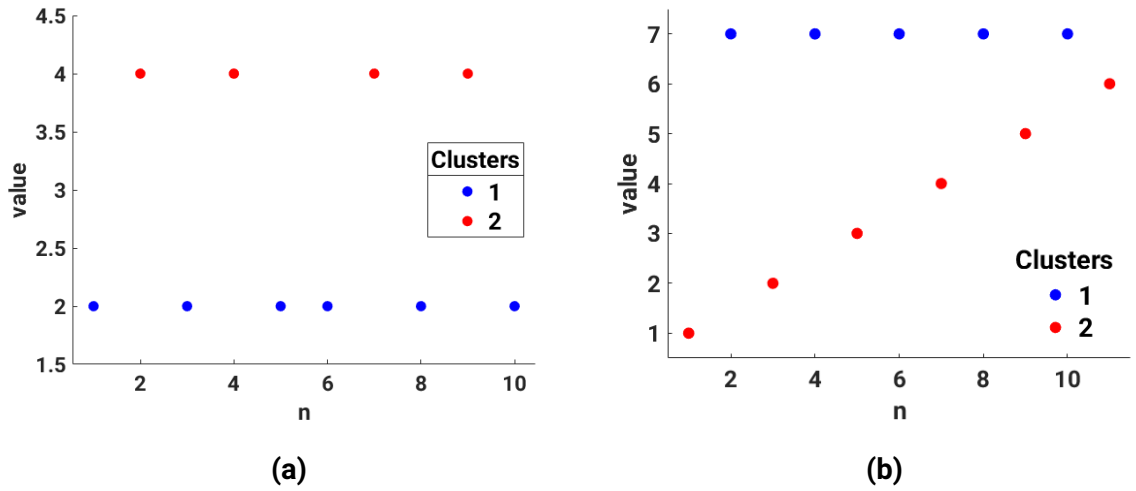



**Figure 10.3.:** Examples of two simulated feature streams, with colours indicating the result of GMM clustering. Each point represents one feature value at a time n, and colours show how they were assigned to one of two clusters. Notice that in panel b), the two clusters have different shapes – something that a k-means algorithm could not correctly deal with.

The matrix relating $X_n$ to the "measured" signal value $z_n$ is: $H_n = \begin{bmatrix} 1 & 0 \end{bmatrix}$, which means each incoming feature value is a direct representation of the first element ($x_n$) of the process state matrix.

The measurement and system noise covariance matrices are as follows:

$$Q = \begin{bmatrix} \sigma_w^2 & 0 \\ 0 & \sigma_b^2 \end{bmatrix}, R = \sigma_v^2 \tag{10.2}$$

where the variances are empirically chosen for each feature.

As in Kaya and Elhilali (2014), following the clustering stage, the state and covariance matrices are initialised based on the last two feature values. For each cluster, a filter initialised at step $n$ has the following starting estimates:

$$\hat{X}_n^- = \begin{bmatrix} z_{n-1} + (z_{n-1} - z_{n-2}) \\ z_{n-1} - z_{n-2} \end{bmatrix} \tag{10.3}$$

$$\hat{P}_n = \begin{bmatrix} 5\sigma_v^2 + 2\sigma_w^2 + \sigma_b^2 & \sigma_w^2 + 3\sigma_v^2 + \sigma_b^2 \\ \sigma_w^2 + 3\sigma_v^2 + \sigma_b^2 & 2\sigma_v^2 + \sigma_w^2 + 2\sigma_b^2 \end{bmatrix} \tag{10.4}$$

If a new feature value is not correctly predicted by any of the currently running filters, a new filter is initialised with the same initial estimates as above. The decision whether to start a new filter is based on the following threshold:

$$|z_n - H\hat{X}_n^-| \leq 2\sqrt{P_{[1]} + \sigma_v^2} \tag{10.5}$$

The left side of Eq. (10.5) is the model innovation, while the right side equals two standard deviations of the innovation (the innovation covariance matrix being $S = HPH^T + R = P + R$). $P_{[1]}$ indicates the first element of the matrix $P$. If a filter is accurate, the innovation error will be low (and the estimate error will be low, as measurement noise is constant), and smaller changes will initialise new filters. Note that this is slightly different from Kaya and Elhilali (2014), who specify the left side of Eq. (10.5) as: $|z_n - HA\hat{X}_n|$, which in effect measures the difference between the feature value $z_n$ and the estimated measurement value at the *following* step, $n + 1$.

Once this threshold is exceeded, a new filter is initialised and an MMN spike is produced, with amplitude equal to the innovation – the bigger the difference between the prediction and the actual value, the bigger the surprise. If a filter has not correctly predicted any feature values for 100 ms, it is closed.

## 10.3. Results

The artificial features – location and type – are fed into the Kalman-based salience model, as described previously in this chapter. The outcome scores from both

variables are then summed together. As a result, for each trial, a salience score is obtained over the time of the trial. Because the pupil dilation responses were measured as a response to the distractor sound, the single number output from the model is the maximum of the salience score for the duration of the distractor (see Figure 10.2).

Figure 10.5 shows the modelled and actual measurement data for each condition. Note that much of the variability in the experimental data was due to participants' variation in pupil dilation range, however these individual differences were not modelled here. Figure 10.6 shows medians of the pupil dilation responses plotted versus the median calculated salience scores. The modelled values correlate very well both with mean PDRs (Pearson correlation coefficient $= 0.95$) and peak PDRs (Pearson correlation coefficient $= 0.98$).

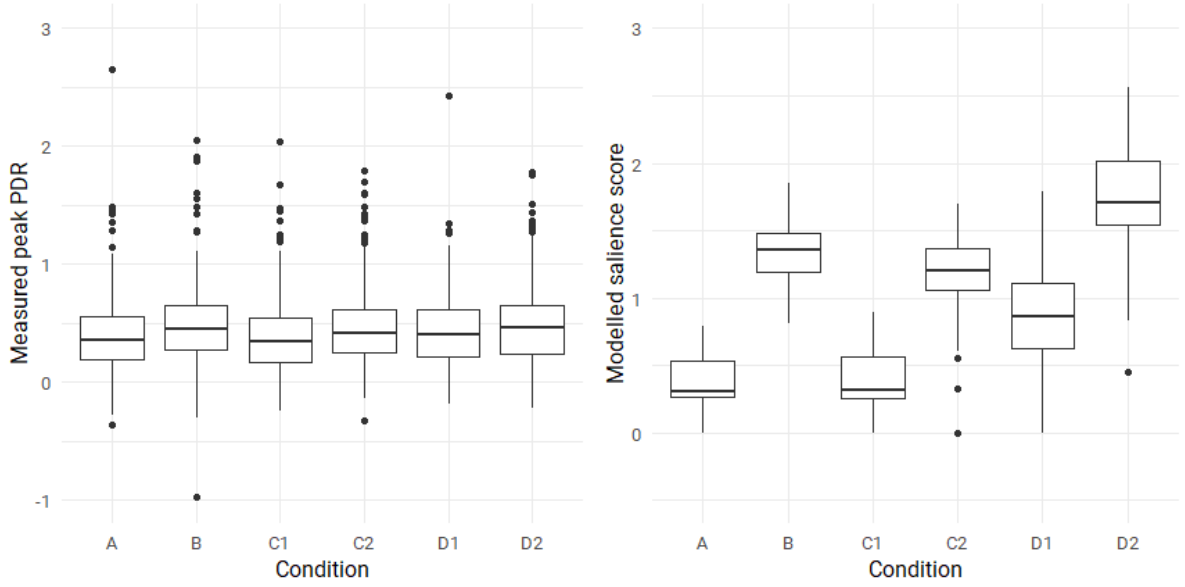



**Figure 10.5.:** Left panel: measured peak pupil dilation responses (in mm), right panel: modelled salience scores. Boxplots show the median, 25th and 27th percentile of the responses and modelled scores for all experimental trials.

To confirm that the model outputs can predict the experimental data, a simple linear

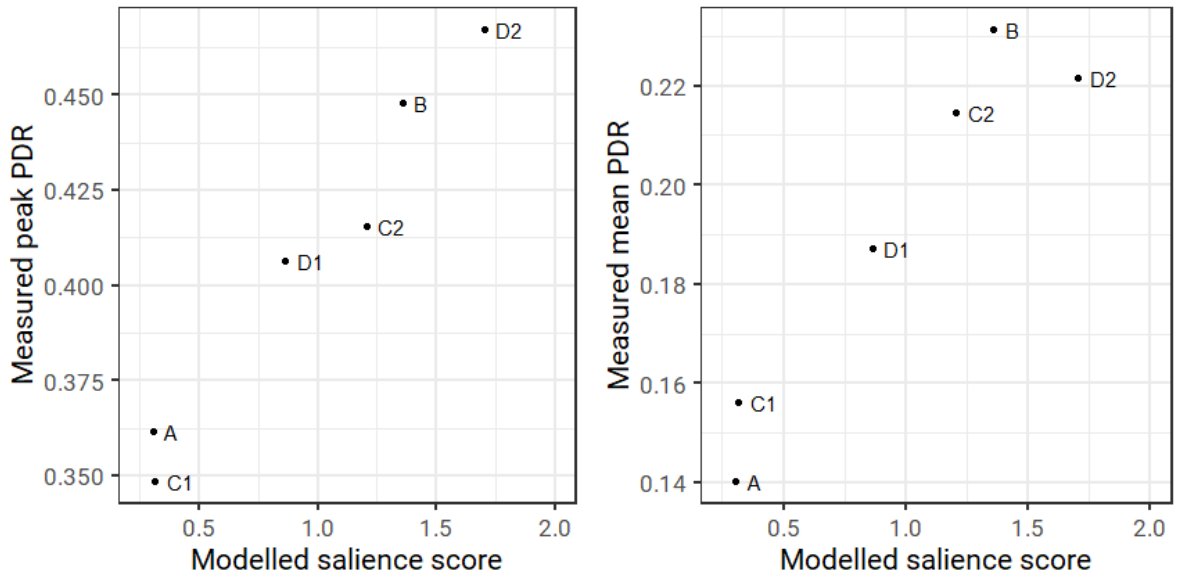



**Figure 10.6.:** Correlation of the median modelled salience scores with median peak (right panel) and average (left panel) pupil dilation responses, in millimetres.

regression was used with the modelled salience score for each experimental trial as a fixed effect. The effect of the salience score was statistically significant both for outcome variable peak pupil dilation ($Est. = 0.069, p < 0.0001$) and mean pupil dilation response ($Est. = 0.069, p = 0.0001$).

## 10.4. Summary

A simple way to extend an auditory salience model to include spatial information was presented in this chapter. Spatial position of sounds was used as a feature which was tracked by a deviance detection model. This allowed for detection of unexpected changes in spatial position and predicted pupil dilation responses obtained in an experiment described in Chapter 6 well. The results shown in this chapter will be further discussed in Chapter 13.

# 11

# Example application – AED

To further test the applicability of the Kalman-based salience model as described in Section 10.2, it was used in an Acoustic Event Detection task (AED). The goal of AED is to automatically find sound events and estimate their position in time in an audio signal. The salience model was combined with a Non-negative Matrix Factorization (NMF) model (Sobieraj, Rencker and Plumbley 2018). Non-negative Matrix Factorization is a machine learning method based on dictionary learning. The resulting method works for AED on weakly labelled data, that is, data for which we do not have exact information of when the interesting sound occurs, but just a tag of which sounds are present in a given audio excerpt.

The structure of the model is shown in Figure 11.1. The sound signal is analysed in frames in parallel by the salience and NMF models. In principle, the salience model should detect onsets of any interesting events, regardless of whether they are the target in the task. Therefore, its output is combined with the NMF output, which can differentiate between target and non-target events.

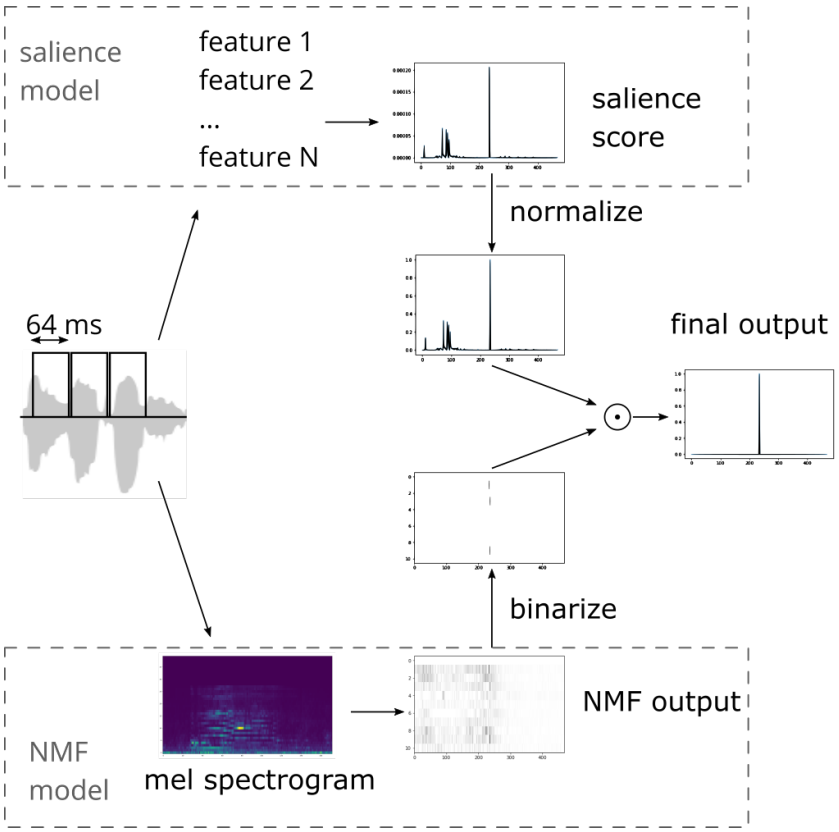



**Figure 11.1.:** Structure of the model combining NMF with auditory salience.

## 11.1. Kalman-based salience model

The Kalman-based salience model was the same as described in Chapter 10, except for an added feature extraction stage. Six features were extracted using the pyAudioAnalysis library (Giannakopoulos 2015), with a 64 ms window: energy, energy entropy, spectral centroid, spectral rolloff, spectral entropy and zero-crossing rate. Each feature extracted from the signal is tracked by one or multiple Kalman filters simultaneously.

This process produces vectors of salience spikes $s_i$ (one from each feature). The resulting salience score for frame $n$ is obtained by applying feature-specific and

between-feature weights and summing the resulting vectors, as follows:

$$s(n) = \sum_{i \in [1,N]} s_i(n) \left( w_i + \sum_{j \in [1,N], j \neq i} w_{ij} \max_{k \in [-1,1]} s_j(n+k) \right). \tag{11.1}$$

The weights $w_{ii}$ and $w_{ij}$ used in Eq. 11.1 were trained with a constrained logistic regression, where the binary output variable was the presence of an event in an audio file, predictor values were the mean $s_i$ for each feature, and the weights were constrained to be positive. Recordings of 30 seconds are used for training.

The salience score $s(n)$ per frame is computed for each test sample, forming a vector $s$, which is then normalised to its maximum. Finally, in the last step, $s$ is multiplied by the output of the NMF to create the final result, indicating a possible event onset.

The combined model was evaluated on rare event detection using only weakly labeled data from the audio recordings of the TUT Rare Sound Events 2017, which were provided for Task 2 of the DCASE2017 challenge (Mesaros et al. 2017). The dataset consists of around 100 isolated sound examples for three target classes: *gunshot*, *baby crying* and *glass breaking*, together with background audio which is part of the TUT Acoustic Scenes 2016 dataset (Mesaros, Heittola and Virtanen 2016).

## 11.2. Results and discussion

Table 11.1 presents the results of the evaluation on the test set. For comparison, alongside the proposed method, the results for each of the NMF and salience models separately are shown. In the salience model, an event was detected for every frame in which the salience score exceeded 50% of the maximum salience score in that test recording.

The F-score (which takes into account both precision and recall of the model) of the

| Event type | **Proposed** | | NMF | | Salience | |
|---|---|---|---|---|---|---|
| | ER | F1 | ER | F1 | ER | F1 |
| Gunshot | **0.76** | **65%** | 0.80 | 64 % | 1.45 | 36% |
| Glass breaking | **1.07** | 46% | 1.23 | 41% | 1.12 | **54%** |
| Baby crying | **1.04** | 36% | 1.07 | **37%** | 1.71 | 32% |

**Table 11.1.:** Evaluation results on detection of events of different types. Error Rate (ER) and F-score (F1) are reported for the proposed method, NMF only and Salience model only. Lowest ER and highest accuracy for each target sound are shown in bold.

combined method ranges between 36% and 65% depending on the sound class, and the error rate is between 1.04 and 0.76. The total performance of the combined model aside, it is interesting to look at ways in which the salience model brings a benefit or outperforms the NMF model.

First, the difference that the salience model makes is not the same for all sound classes. The results show that adding the auditory salience model to the NMF detector improves its performance for gunshot and glass breaking events, but not the baby crying event, for which it decreases the error rate, but does not improve the F-score, suggesting a low hit rate.

The reason for this difference in performance for different event classes may be that the first two – gunshot and glass breaking – usually have sudden onsets, while the last one – baby crying – can start rather slowly. The salience model is designed to detect sudden changes in features, but will adapt to changes that are too slow. While this property makes it useful in some types of backgrounds (see below), it also means that it might not be suited for events which develop slowly, or might need a larger frame window for them.

Looking closer at individual sound recordings, there were a number of cases where

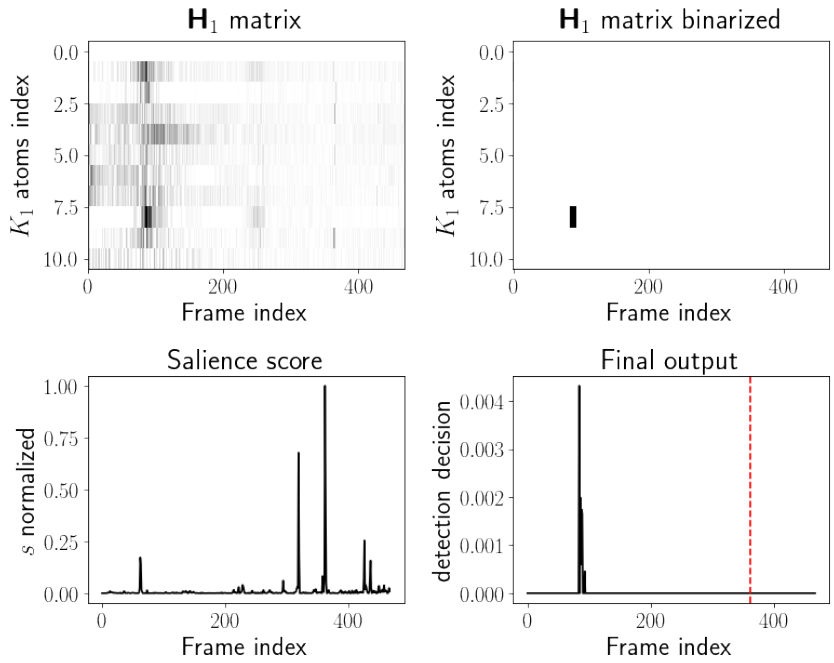



**Figure 11.2.:** Results for a gunshot event over a residential area background, with a loud car passing in the first half of the file. Top row: $\mathbf{H}_1$ matrix from the NMF model. Bottom left: salience model output $s$. Bottom right: final output of the model, from which an event is detected for any value larger than 0. Red dashed line shows the position of the target event. Even though it was correctly recognised by the salience model, the combined models do not detect it.

the salience model was able to detect an event when the NMF was not. This is also evident from the fact that the salience model outperforms both the NMF and the proposed method for the glass breaking event. One situation where the salience model presents an advantage is when the background noise significantly but slowly increases in level – e.g. a train passing (see Figure 11.2). Because a Kalman filter-based model is not sensitive to sudden feature changes, it is able to adapt to this background, and only flag a detection when changes in feature values correspond to new, "surprising" events. It also seems to perform well in loud cafeteria-type backgrounds (see Figure 11.3).

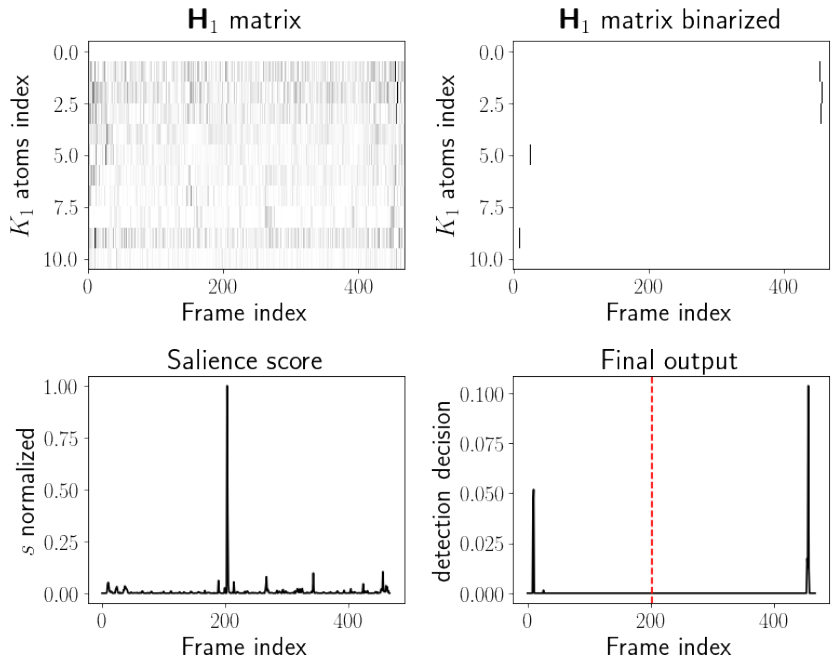



**Figure 11.3.:** Results for a gunshot event over a cafe/restaurant background. Top row: $\mathbf{H}_1$ matrix from the NMF model, before and after binarization. Bottom left: salience model output $s$, after normalization. Bottom right: final output of the model, from which an event is detected for any value larger than 0. Red dashed line shows the position of the target event, which was correctly recognised by the salience model, but not the NMF model.

## 11.3. Summary

This chapter described an example of a practical application of an auditory salience model based on deviance detection. The model was combined with a method based on Non-Negative Matrix Factorization and used to detect events in acoustic signals. It improved performance of the algorithm on sound events with sudden onsets and noisy backgrounds.

# 12

# Summary

Part II discussed modelling of spatial auditory salience.

Models based on prediction and deviance detection in particular correspond well with known mechanisms of salience. Chapter 9 reviewed these mechanisms and different approaches to modelling auditory salience. None of the models take spatial information of the sound into account in their salience calculations.

Chapter 10 showed that adding information about spatial location of sound to a model based on Kalman filters can predict pupil dilation responses to breaking spatial expectations.

Then, in Chapter 11, an application of a deviance-detection model was shown, where it was used to improve acoustic event detection performance of a Non-Negative Matrix Factorization algorithm.

# Part III.

# General discussion

# 13

# Discussion

## 13.1. Auditory salience and spatial location

Part I of this thesis describes perceptual experiments designed to test the effect of the location of a sound on auditory salience. As there is no standard testing method for auditory salience, four different methods were used, each of which addressed the question from a slightly different perspective. The results suggest that the spatial position of sound, alone, does not directly affect its salience. In this section, the evidence for this claim will be discussed in the context of low-level and high level salience, the effects of loudness, perceptual load, and expectations.

### Low- and high-level salience

The experiment in Chapter 3 used an oddball detection method, based on that used by Tordini et al. (2013), but modified to include six spatial locations around the listener. The stimuli used in this experiment were band-pass filtered noise. The results showed no difference between participants' responses to different locations when the stimuli were low frequency noise. However, for high frequency stimuli, responses were

significantly slower (about 100 ms) for target sounds behind the listener, compared to sounds in front or on the right. Participants were also about twice as likely to be incorrect in detecting the oddball for rear rounds. These results suggest that spatial salience might be related to the spectral content of the sound. For high frequency noise, sounds in the back appear to be less salient than those in the front, whereas no such effect exists for low frequency noise.

One of the shortcomings of the oddball experiment was that the stimuli were simple, synthetic sounds. Although this allowed for straightforward manipulation of the sound, it could be argued that the perception and responses to those stimuli does not accurately represent everyday listening situations.

Chapter 4 tested auditory salience in a self-reporting experiment in a more ecologically valid environment, where top-down attention also played a role. Participants listened to recordings of real-life sound events and reported their attentional focus in real-time. As expected, participants paid attention to louder sounds more often, which is in agreement with other studies on salience of loudness (Kaya and Elhilali 2014; Huang and Elhilali 2017). The results also suggest an interaction between brightness and location of sound – there is a small decline in salience of sounds arriving from behind the listener, but only for high brightness sounds.

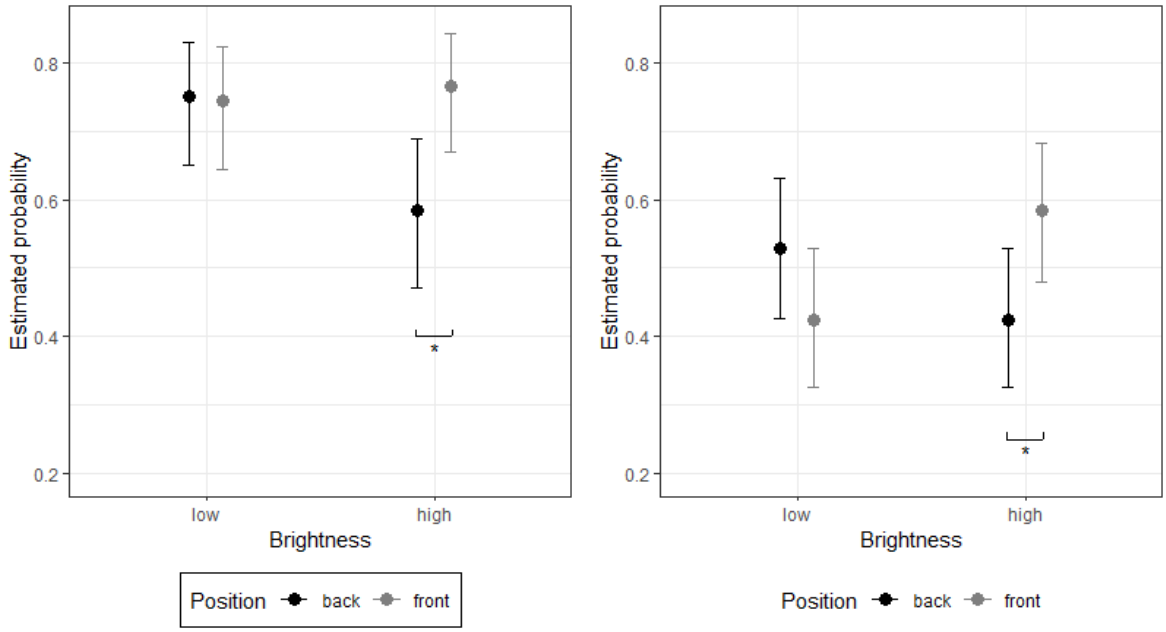
Because natural sounds were used as targets, other factors not taken into account in the design could influence the results, especially participant-specific subjective effects, such as personal experience or an emotional reaction to a sound. For example, one participant noted that they thought they paid more attention to sneezing and coughing sounds, as they instinctively wanted to avoid a source of viruses and bacteria. There is also evidence to suggest that emotional environmental sounds can influence spatial attentional orienting by causing attentional avoidance (orienting *away*) from taboo or emotionally negative sounds presented to the left side (Bertels

et al. 2013).

### Loudness

Both the oddball detection and the self-reporting experiments show a similar interaction between sound location and its spectral content – specifically, only sounds with high brightness (including high-pass noise) are less salient when they are behind the listener, than when they are in front. Fig. 13.1 shows a comparison of these results: the data points are probabilities of a correct response (oddball experiment) and attended event (self-reporting experiment), but it is assumed here that they both represent how likely people are to attend to sounds which are in a particular location around them. Note that both the high and low-pass noise stimuli had spectral centroid values outside of the range of the natural sounds used here (spectral centroid of the low noise: 812 Hz, high noise: 13691 Hz, natural sounds: 1000-5500 Hz). Still, they show a similar pattern.

These results could likely be explained by loudness differences. Pinna-shadowing can affect the loudness of high frequency or broadband sounds arriving from behind the listener, and this decrease in loudness can make these sounds less salient. Sivonen and Ellermeier (2006) measured loudness for different locations around the listener (only on the left, however, as they assumed symmetry), shown in Figure 13.3. Their results show lower sensitivity from the back for 5 kHz sounds, and almost no difference for 400 Hz and 1000 Hz sounds (third-octave noise bands), consistent with the results of the salience experiments. However, they also found higher sensitivity to sounds arriving from the side, which was not seen here.




**(a)** Oddball detection experiment **(b)** Self-reporting experiment

**Figure 13.1.:** Estimated probabilities that a sound was attended to, for front and back location vs frequency or brightness from the two first experiments, with 95% confidence intervals.

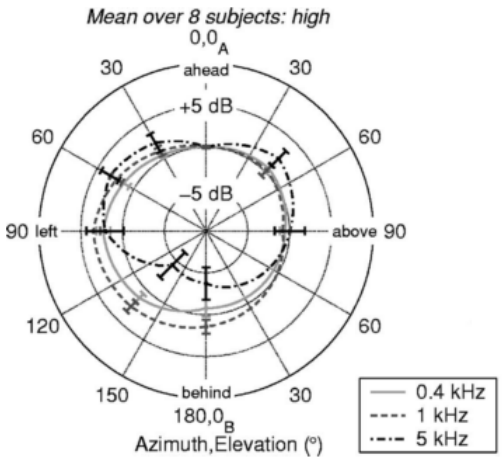



**Figure 13.3:** Directional loudness sensitivities at 65 dB SPL. Reproduced from Sivonen and Ellermeier (2006).

### Perceptual load

The experiment in Chapter 5 tested auditory salience in a dual-task experiment, in which perceptual load was manipulated by adding sound sources to the scene.

Salience was determined from participants' performance on the secondary task, which was detection of a known stimulus.

Firstly, the results show a main effect of front/back position on detection of the secondary target ($p = 0.00001$). The targets behind the listener were 1.4 times more likely to be detected than the ones in front, and this effect was independent of perceptual load. In addition, in low perceptual load conditions, participants were more likely to detect targets located to their right side, than to the left ($p = 0.02$, odds ratio at set size 1 = 2, at set size 2 = 1.7). This effect, however, disappeared at higher perceptual load levels. The expectation before performing the experiment was that, because at high perceptual load levels only the most salient sounds are noticed, increasing the load could perhaps reveal some auditory salience differences which are otherwise missed. However, the interaction that was found is the opposite – increasing perceptual load diminished an apparent spatial effect. It could point towards a right-ear advantage which is only available with free perceptual capacity. Usually however, a right-ear advantage is associated with speech stimuli (Marsh, Pilgrim and Sörqvist 2013) and there have even been arguments for a *left-ear* advantage for non-speech sounds (Hadlington, Bridges and Darby 2004). These left/right advantages have often been attributed to brain pathways and faster processing of certain type of stimuli in one of the hemispheres and there is evidence that speech and melody are processed in different hemispheres (Albouy et al. 2020). However, it has also been argued that the right ear advantage is accounted for by attention (Hiscock and Kinsbourne 2011). These results differ from results of experiments described in Chapters 3 and 4 – this will be discussed later in this chapter.

There is also a question of whether the strategy participants used to perform the tasks could change the interpretation of the results. The assumption was that they would focus mostly on the primary task, and only detect the secondary task when they had free perceptual capacity, or when the target was sufficiently salient. However, there is

a possibility that, instead, they first decided which primary target was present, and then switched their attention fully to the secondary task. If this was indeed the case, increasing the number of sound sources might not really correspond to a decrease in free perceptual capacity. It would still likely cause increased task difficulty, simply because after spending more time on the primary task, there was less time left to find the secondary target. An attempt to prevent participants from adopting this strategy was to instruct them to respond as quickly as possible, and to stop the sound clip as soon as they responded. The previously published non-spatial version of this experiment (Remington and Fairnie 2017) avoided this problem by using very short stimuli (100 ms), but this was sacrificed to make the experiment closer more ecologically valid. At least three different ways of explaining dual-task interference have been proposed: capacity limitations, bottleneck (task-switching) and processing "cross-talk" between tasks (Pashler 1994).

### Expectations

Chapter 6 explores spatial auditory salience in the context of surprise and expectations. Participants performed an auditory task in 6 different experimental conditions: the baseline condition with no distracting sounds, a single distractor condition, and 4 conditions with different combinations of a context sound followed by a distractor (with matching or non-matching location, and matching or non-matching sound type). The results show that introducing various distraction conditions affected both response times and pupil dilation responses (PDRs), compared to the no-distraction baseline. However, each of these two response metrics was affected differently by the experimental conditions.

Pupil dilation data show clear, statistically significant responses to all the conditions in which a new sound was introduced to the environment, regardless of context type.

The effect is very similar in size to one reported in previous studies: 0.08 mm (Marois, Marsh and Vachon 2019). Changing the location of a sound between context and distractor, when the type of sound was the same, had a smaller, but significant effect (around 0.04 mm). This can be interpreted as a response to breaking the listener's expectations about what sound sources are in the environment and, to a smaller degree, about their spatial position.

Notably, the condition which breaks both expectations does not seem to evoke a larger response than the condition which breaks only one (type). The average pupil size baseline for all trials in the experiment was 4.5 mm, which was just over half of the maximum of pupil dilations (Mathôt 2018), so it is unlikely that the total range of possible dilations was exceeded. On the other hand, the mean pupil dilation response in these two conditions was 0.47 mm, which might be close to the upper limit of task-evoked PDRs, which are typically in the order of 0.1 to 0.5 mm (Winn et al. 2018).[1] However, if this pattern in results is not simply an artefact of the pupil dilation range, it could mean that effects of breaking multiple expectations are not necessarily additive. In this particular situation, one could argue that a new sound in a new location is simply not more surprising than a new sound in an expected location, because there was nothing in this scenario that would lead participants to believe that sounds will always only come from one location (in fact, in the real world, this is almost never the case). The implication of this would be that spatial auditory expectations are mostly built with relation to specific sound streams (*the bird is likely to continue singing on my right*), rather than with relation to locations in space (*it is unlikely that any sounds will come from my left*). Changing location of a sound is therefore only relevant (salient) if it breaks continuity of an otherwise consistent stream.

[1] This response is much larger than the 0.08 mm expectation-breaking effect reported earlier, because it includes any potential responses to the task – for example, processing the target, deciding on the response, deciding on which button should be pressed etc.

Crucially, when no expectations are violated (a distractor matches context in both location and type), the pupil dilation response is no different than in the baseline condition, where only the target sound was present. In other words, a distracting sound which is a continuation of a previously heard, predictable stream did not elicit a pupil dilation response, even though it was presented at the same time as the target. This confirms that the significant PDRs recorded for other conditions did not simply reflect processing of additional auditory information.

Response times show a less clear picture. All three conditions in which there was a mismatch between context and distractor were significantly different from the no-distraction condition, but an analysis of contrasts shows no significant difference between them. The condition with context matching the distractor in both location and type was not significantly different from the no-distraction condition. Somewhat surprisingly, neither is the distractor-only condition, where there was no context sound.

Clearly the behavioural responses have a different pattern from the pupil dilation responses. In trying to understand where these differences originate, it might be worth looking more closely at responses to a condition for which the difference between response times and pupil dilation was perhaps most interesting: the distractor-only condition. The initial prediction was that a single sound preceded by silence will be very distracting, as it clearly breaks expectations about the environment. This prediction, however, was only confirmed in PDR, but not behavioural data.

Let us compare this distractor-only condition with the condition in which neither location nor type match between context and distractor. It can be argued that they both break spatial expectations in a similar way, with a sound appearing in a new spatial location, and they both introduce a new sound to the environment. The main difference between them is the presence or absence of the context sound. It is possible that it is this presence of the context sound that makes sounds more distracting and

increases response times, but has no effect on pupil dilation responses. A hint towards the greater importance of the context for response times than for pupil dilation is also the larger and more consistent effect of context length on the former. One reason for this could be that breaking expectations from a sound (context) causes larger behavioural distraction than breaking expectations from silence (no context). In fact, a distinction has been made in the literature between *initial orienting response* – to the first stimulus in a sequence – and *change orienting response* – when an aspect of an existing stimulus sequence changes (Näätänen and Gaillard 1983). However, larger PDRs have previously been found to the change OR, than the initial OR (Steiner and Barry 2011), which was not confirmed in this experiment. A different, simple way of explaining this is that the mere presence of a context sound distracts participants and slows them down, perhaps by involuntarily drawing their attention to itself – even though participants were informed about the location of the incoming target word. This could then cause a response delay by having to switch attention back from the context sound to the target, which may not be reflected in the pupil dilation responses.

Finally, no effect of the absolute spatial position of the distractors on neither behavioural responses or pupil dilation responses was found. This result is consistent with the results of experiments described in Chapter 3 and Chapter 4.

### Summary

Three of the experiments discussed above suggest that location of sound around the listener does not modulate auditory salience. However, the experiment in Chapter 5 indicated that sounds behind the listener might be more salient than those in front. What are possible explanations for the discrepancy in the experimental results between the dual-task and other experiments?

The cause likely lies with the different methods used. This difference could indicate

that something about the detection task in the dual-task experiment from Chapter 5 makes it fundamentally different from the experiments in Chapters 3, 4 and 6, and it measures a different perceptual or cognitive process. Therefore, these spatial effects might arise not due to salience, but to a mechanism only present in the dual-task case.

When the definitions of salience described in Chapter 2 are considered, the three experiments use methods that are closer to the "attentional" definition of salience, while the dual-task experiment might be closer to the 'detection' definition. However, the distinction is not absolutely precise and there could be arguments for explaining some of the experiments in terms of both attentional orienting and detection. It is also possible that the reason is not due to the detection task itself, but to the dual-task paradigm, which was unique to the experiment in Chapter 5. Even though the distraction experiment (Chapter 6) was similar, as participants' main focus was also on something unrelated to the stimuli, they were still only performing one task, as the stimuli were not relevant.

It is worth noting that the effect sizes for these spatial effects were rather small, and it is possible that they arose due to chance. To be sure, the results would need to replicated in a spatial dual-task experiment. Taken together, the evidence from the four experiments does not support the idea that spatial position modulates auditory salience.

Although the literature on the salience of sounds in the rear is lacking, an argument could be made for a benefit to noticing sounds that are behind the person, outside of the field of view. Being alerted to these sounds, allows the person to turn their head and use vision to investigate. It has indeed been shown that orienting attention to a sound enhances visual processing in that location (Spence 2010). On the other hand, one could argue that the result suggesting that no particular spatial location is more salient than any other is not at all surprising, especially in the view of the 'Bayesian brain' theory (described in Chapter 9). If one views salience as the violation of

expectations, sounds are only salient in a context, when they do not agree with the listener's predictions. Therefore, a sound in a particular location can only be salient if, according to the listener's mental model, it is not likely to appear there. The results of the distraction experiment reported in Chapter 6 are indeed in agreement with this view of the brain and more specifically of auditory processing. In this experiment, no particular location on its own caused a larger or smaller reaction than other locations. On the other hand, there was a small but significant pupil dilation response to a sound unexpectedly changing its spatial position, suggesting that violations of spatial expectations are salient. Still, this response was smaller than that to a new sound – i.e. changing expectations about which sound is present in the scene.

## 13.2. Measuring auditory salience

An experiment which compared three different measurement methods confirmed that they are likely to respond differently to effects of low- and high-level salience. Additionally, pupil dilation responses have been measured to spatially unexpected sounds in a distraction experiment, which supports the effectiveness of physiological methods for measuring auditory salience.

### Behavioural methods

The problem of determining the most appropriate methods of measuring auditory salience remains open. In this thesis, four different paradigms were used to investigate the issue. Each method was different and had its own advantages and potential issues. Table 13.1 gives an overview of the differences and similarities between the methods.

In each experiment, participants performed a different task. The goal of the oddball detection task in Chapter 3 was to detect a shorter inter-stimulus interval in one of

| | Chapter 3 | Chapter 4 | Chapter 5 | Chapter 6 |
|---|---|---|---|---|
| Task | oddball detection | self-reporting | detection (secondary) | classification |
| Stimuli relevance | relevant | relevant | relevant | not relevant |
| Salient sounds will... | ...be attended to more often | ...be attended to more often | ...be easier to detect | ...cause automatic attentional orienting |
| Collected metrics | RT (speed), hit rate | hit rate | hit rate, pupil dilation | RT, pupil dilation |
| Localisation | not required | required and can be an issue | not required | not required |
| Stimuli | noise | natural sounds | natural sounds | natural sounds |
| Energetic masking | unlikely | possible | possible (for high load levels) | unlikely |
| Top-down effects | minimal | significant | minimal | minimal |
| PCA loadings | | | | |
| Dim. 1 | 0.06/0.01 | 0.99 | – | 0.26 |
| Dim. 2 | 0.87/0.88 [speed/hit rate] | 0.03 | – | 0.72 |

**Table 13.1.:** Comparison of experimental methods developed and used in this thesis. *Task* is the task on which performance is evaluated. PCA results are from the comparison experiment in Chapter 7.

two competing streams. Because of the streams' asynchrony, it was very difficult – if not impossible – to follow both simultaneously, so the inter-stimulus difference was only evident once a stream was attended to. Therefore, the participant would statistically detect more oddballs, and detect them quicker, in the stream which was more often attended to, and therefore – more salient. In the self-reporting experiment in Chapter 4, participants were free to point towards a sound which attracted their attention in real time. The stimuli's position in time relative to the background sounds

was randomised. The idea is, again, that more salient sounds will attract attention more often. These two experiments fit with the "attentional" definition of salience discussed in Chapter 2. The experiment in Chapter 5, where perceptual load was modulated, was a dual-task experiment, in which salient sounds were assumed to be more often detected in the secondary task. It could be thought of more as based on a "detection" definition of salience. Finally, the method in the distraction experiment in Chapter 6 differs from the previous ones in that the stimuli being tested were not actually relevant to the participants' task. Instead, they pulled their attention away from the task. The more salient sounds will be more "successful" at causing this attentional orienting and result in longer response times. This is, again, close to the "attentional" definition of salience.

All of the experiments were based on existing mono or stereo-based methods, and extended to include more spatial positions. They utilised various positions around the listener, including behind them. Arguably, this type of a paradigm allows for a more ecologically valid listening situation than many of the other salience or attention experiments, which tend to be performed over headphones or a loudspeaker in front of the listener (as discussed in Chapter 2). After all, in the real world, sounds do arrive at the listener from all possible locations, and auditory phenomena such as spatial release from masking are only possible when sounds are spatially separated.

Another benefit is that multiple sounds can be presented simultaneously and – as long as they are spatially separated – spatial release from masking helps reduce energetic masking effects. In the distraction experiment (Chapter 6) spatial separation between two sounds presented simultaneously was always at least 90°, in the oddball detection experiment (Chapter 3) – 60°, and in the dual-task experiment the minimum distance was 30°. However, in the self-reporting experiment (Chapter 4), it was possible for two simultaneous sounds to be spatially overlapping.

Although spatialisation of these experimental methods has its clear benefits, it is

worth noting that it also comes with certain problems and limitations. For example, in classic dichotic experiments, the standard way of presenting competing sound streams is through headphones. This allows for complete separation of the two channels which facilitates easy stream segregation. In addition, when an experiment with dichotic presentation requires a participant to respond by choosing one of the streams – for example, an oddball task like Tordini et al. (2013), or a free-listening experiment like Huang and Elhilali (2017) – the usual way is to indicate the left or right ear, usually by pressing a button on the left or right side. This straightforward ear-button mapping makes the response intuitive, and with streams being completely separated in the left and right headphone, the task should not pose a challenge. However, with the streams in more positions, some more difficult to localise than the extreme left and right (Blauert 1997), an analogous way of pointing to a selected stream is not as straightforward, and can significantly increase average response times. The self-reporting experiment described in Chapter 4 confirmed that localisation errors can be a significant issue, even when all the sounds were concentrated around front/back and left/right axes. This makes interpretation of responses in such an experiment difficult.

The experiment in Chapter 3, which used an oddball detection method, attempted to avoid the localisation problem altogether. In it, participants were asked to differentiate between two types of stimuli, and instead of the *location* of the stream, indicate which *stimulus* stream was built of (e.g. high frequency or low frequency noise). This is less intuitive than indicating left/right – also because it requires remembering which button to press for which stimulus – and requires prior training. In order not to make the task more difficult (which could introduce more errors), the oddball detection experiment described in Chapter 3 used stimuli which were very simple – noise bursts. This, however, reduced environmental validity. A way of making the task a bit more intuitive could be to use speech (words or just vowels), but

this also limits the types of characteristics of sound which can be tested. In general, this approach only lets the researcher test limited types of stimuli, and usually requires prior training. Therefore, from the point of view of localisation errors, a distraction experiment like in Chapter 6 or a dual-task detection (Chapter 5) might be better, as they do not require the participants to explicitly localise any sounds.

Apart from presenting stimuli around the listener, another aspect adding to the ecological validity was that all experiments, except for the oddball detection (Chapter 3), used recordings of real-world sounds as stimuli. This is perhaps especially an improvement in the distraction experiment (Chapter 6), as many of the distraction methods in the literature use the standard sound/deviant sound paradigm, where the standard is a simple sound such as a tone repeated regularly. The experiment in Chapter 6 introduced a paradigm in which the standard sound instead develops continuously, in a natural way, over time.

Although all of these methods are based on previously published salience and attention research, the comparison analysis described in Chapter 7 shows that not all of the methods measure the same phenomenon. The non-spatial equivalents of methods used in Chapters 3, 4 and 6 were compared by running them each with the same stimuli. A Principle Components Analysis of the results shows at least two different dimensions of salience, each correlated with different methods. Chapter 7 gave a detailed discussion of the PCA results. In general, the two dimensions seem to describe low-level and high-level aspects of salience. Attention has been shown to operate on multiple levels, from the most basic, local changes in signal characteristics, to more global deviations in patterns (Chennu et al. 2013). In addition, on this higher level sounds can be salient because of their meaning (Moray 1959) or emotional content (Thierry and Roberts 2007).

Of the experiments in this thesis, the self-reporting one in Chapter 4 is most likely to be affected by top-down attention and higher-level characteristics of sound, such as

meaning, or emotional connotations. This also means that it will capture more variation in which sounds different people find salient. The other three experiments either kept the participant's top-down attention focused on something unrelated to the stimuli (Chapters 4 and 6), or used only two types of simple stimulus, and made sure that the type of stimulus was balanced across all other conditions (in case a participant had a preference for one over the other). The experiment in Chapter 6 can most likely measure the lowest-level attentional orienting out of the four methods, because in it, the stimuli are irrelevant to the task, and there is no benefit of attending to them at all. Participants might even actively try to block them out, so they attract attention in a truly automatic way. Brain responses to "oddball" (deviant) stimuli have been recorded even in sleep (Atienza, L. Cantero and Gómez 1997), so it is known to be a very low-level process.

These results can not determine if any of the methods is more "correct" for measuring salience. Instead, what they underline is that salience, defined as the ability to attract attention, is complex, and different measurement methods will capture its different aspects. This is something that should therefore always be taken into account when interpreting results of auditory salience experiments. As the comparison study was limited in scope, with the number of data points relatively small and certainly unrepresentative of all possible methods, a more comprehensive analysis of available auditory salience measurement methods would be beneficial for the field.

### Pupil dilation responses

In addition to behavioural metrics, such as response times or task accuracy, pupil dilation responses (PDR) have been collected for two of the experiments (Chapters 5 and 6). Interestingly, the PDRs did not agree entirely with behavioural data.

Chapter 5 found an effect of spatial position of the target sound in behavioural data

(detection accuracy), but not for pupil dilation responses. Perhaps even more importantly, while behavioural task performance clearly decreased with increasing perceptual load, the pupil dilation responses show a different pattern. While task accuracy decreased in a close to linear manner with each sound source added to the scene, PDRs were similar for set sizes 1 and 2, significantly larger for set size 6, and the largest for set size 4. Therefore, pupil dilation failed to reflect the increased perceptual load in a gradual manner.

It is possible that this reflects a property of pupil dilation. As PDRs for set size 6 were significantly smaller than for set size 4, it is unlikely to be a simple ceiling effect on dilations. There is evidence that pupil dilation responses might decrease for very difficult conditions, such as high SNRs in a speech-in-noise test (Wendt et al. 2018), possibly demonstrating some level of disengagement from a task which becomes too challenging. This might be the effect observed in these data. It is also worth noting that it is not straightforward to interpret pupillometry for this experiment because of the inconsistent timing of the trials. Not only would the sound clip stop when the participant responded to the first question, making each trial a different length, but also there is no way of knowing the moment in which a participant notices the secondary target. It is possible that more subtle effects got lost in the noise.

On the other hand, in the distraction experiment (Chapter 6) the measured pupil dilation revealed participants' automatic responses to unexpected stimuli, which were not apparent in their behavioural performance. This indicates that physiological metrics – such as pupil dilation – might reveal salience effects not present in behavioural responses. As they are automatic, they will not be susceptible to participant errors and have the potential to give more objective measurements. Also, because any behavioural responses include effects of cognitive and motor processes associated with making decisions about a response and physically responding (e.g. pressing a button), they are likely to be more noisy. Other researchers have also found

differences between behavioural responses and pupil dilation responses. Marois and Vachon (2018), for example, recorded significant PDR to deviant sounds, but not an effect in performance in a reading comprehension task, and suggest that "the PDR may be a more sensitive attention-capture index than behavioural measures".

Pupil dilation has been previously measured to various auditory stimuli, including unpredictable or infrequent events (Zekveld, Koelewijn and Kramer 2018). However, although distraction to location deviants has been demonstrated in behavioural task performance and brain responses (Roeber, Widmann and Schröger 2003; Corral and Escera 2008), it has not been previously shown in pupil dilation. The experiment presented in Chapter 6 demonstrates that this effect can also be detected with pupil dilation responses.

## 13.3. Implications for modelling

The problem of not having consistent auditory salience ground truth still exists and will be an issue for attempts to develop and assess reliable auditory salience models (Kaya and Elhilali 2017) until some form of standard testing is established in the field. Modelling attempts based on ground truth collected with different experimental methods might well give very different results.

### Violation of expectations

The results described in this thesis, in particular in Chapter 6, support the idea of modelling salience as a deviation from prediction. Such a model will usually consist of two crucial parts: prediction of the incoming input, and comparison between the prediction and the actual input. Often, this will be accompanied by building some form of an internal model of the environment and updating that model as new input

emerges. The models will also often have a way of accounting for the degree of confidence in the prediction (or how accurate it is).

In Chapter 10, an example of such a model was described, based on multiple Kalman filters, and it was shown that it can predict experimental data from Chapter 6 with a correlation coefficient above 0.9. In Chapter 11, an application of the Kalman-filter-based model was presented, where it is combined with a Non-negative Matrix Factorization (NMF) algorithm for acoustic event detection. Adding this salience module improved the performance of the NMF for two sound classes: the more impulsive "gunshot" and "glass breaking", but not for the more gradually evolving "baby crying". It also outperformed NMF in recordings with background noise which was at a high level, but changing slowly. In general, the deviance detection-based salience module was able to adapt to slowly changing environments, and only flag events when sound features changed suddenly.

Other models based on this principle have been proposed, particularly more recently, for example ones which calculate Bayesian surprise – the difference between prior and posterior probabilities of the input (Schauerte et al. 2011), or the log-surprise as the difference between predicted input distributions at two consecutive steps (Rodríguez-Hidalgo, Peláez-Moreno and Gallardo-Antolín 2018). Tsuchida and Cottrell (2012) used a similar approach by calculating signal statistics and comparing them to each new frame (see also Kaya and Elhilali 2017 for a discussion). Bayesian surprise models have also been used for acoustic event detection (Schauerte and Stiefelhagen 2013; Rodríguez-Hidalgo, Peláez-Moreno and Gallardo-Antolín 2018).

However, this is not always the approach developers of salience models take. The classic saliency map of Kayser et al. (2005) – following the visual equivalent of Itti, Koch and Niebur (1998) – used a center-surround differentiation method, which finds areas on the spectrogram that locally differ from their surroundings, by subtracting spectrograms calculated with coarser time scales from the finer ones. This allows it to

find discontinuities in features (temporal and frequency contrast). However, although this approach finds discontinuities in features, it only works at the most local level, and does not have the prediction component in it. And, as Kaya and Elhilali (2017) point out, these models ignore the fact that time is directional, and there is temporal build up from the past to the future.

Some models first calculate acoustic features such as loudness (Kim et al. 2014) or pitch (Kaya and Elhilali 2012), while other determine salience straight from spectrograms (Kayser et al. 2005) or results of modelled peripheral auditory processing (Tsuchida and Cottrell 2012). Although almost all models agree on using some sort of an energy or loudness representation, other features are less clear. There are studies in the field which focus on trying to identify the most important salience features (Kaya and Elhilali 2014; Tordini, Bregman and Cooperstock 2016). However, considering the Bayesian principles, the question of which features exactly make sounds salient might not be as relevant as having a general deviance detection mechanism, which operates on multiple features. From what is known about the way the brain processes auditory information, it is likely that the salience mechanism utilises whatever information is available to it, or the most useful features. Therefore, if there is not enough variation in the preferred feature, it will find deviations in other features which can trigger the salience mechanism.

Furthermore, some inconsistencies in experimental results could potentially be explained in the Bayesian paradigm. Take brightness, for example – there have been studies which argue that brighter sounds are more salient that darker ones (Huang and Elhilali 2017), and there have been ones claiming the opposite (Tordini, Bregman and Cooperstock 2016). Perhaps these differences could be explained by different expectations, or context, against which the stimuli were set. For example, the study by Tordini, Bregman and Cooperstock (2016) used bird song recordings – could there be a general expectation for bird sounds to sound bright? This kind of an expectation

could make darker bird songs more surprising.

What many of these prediction-based models still lack, is a way to detect deviations in more complex or abstract patterns or rules. For example, Schröger et al. (2007) found mismatch negativity (MMN) brain responses to violations of an abstract rule ("second tone in a pair has a frequency 26% higher than the first") even when participants were ignoring the sounds. The type of model used in this thesis would not really be able to deal with this problem. Some have addressed the issue of different timescales of attention – for example, the multiple Kalman filters used by (Kaya and Elhilali 2014) are meant to work on different timescales, but they are not – by design – able to detect repeating patterns or complex rules.

Finally, the Bayesian framework allows for convenient inclusion of individual differences in how people perceive – and react to – salient sounds. Some people may have a lower salience threshold than others, making them more distractable. Looking at it through a Bayesian perspective: some people may put more emphasis on their internal models (or priors) and need very strong data for them to be affected. Others may adjust their models more easily when they receive external data, which will make them pay more attention to it. This "sensitivity" aspect is key in personalisation of models but poses a significant challenge when universal modelling of experimental data is attempted.

### Spatial models

Even though the absolute spatial position of sound is not likely to impact its salience directly, it is still worth developing models which take it into account. For example, if a salience model is to be included in an object-oriented broadcasting system, it would be beneficial for the model to have access to each sound object's location information. As mentioned earlier in this chapter, there are properties of sound which change with

location and are known to influence salience – perhaps most notably – loudness. If a model is not using a recording from a head and torso simulator (or in-ear microphones), it might miss the subtle but important differences in loudness between sounds in different positions. Tracking sound positions and taking them into account in both salience and loudness calculations may therefore be beneficial.

Additionally, as shown in the literature (Chan, Merrifield and Spence 2005; Roeber, Widmann and Schröger 2003) and confirmed by the experiment in Chapter 6, deviations in sound location can be salient, which means that location should be included in a salience model's deviation detection algorithm. In Chapter 10, spatial location information was added to a Kalman filter-based model by adding it as a tracked signal feature. It was shown to be able to predict the results of the experimental data from Chapter 6. There is actually some evidence that the brain automatically and continuously tracks spatial location of sounds – Deouell et al. (2006) demonstrated the mismatch negativity (MMN) brain response in reaction to changes in spatial location which was proportional to the degree of the spatial deviation.

The Kalman filter model described in Chapter 10 was able to detect salient, unexpected changes of spatial location of sound. Specifically, the modelled salience score for each experimental condition correlated well with measured pupil dilation responses. Although binaural signals have been used in salience models to aid grouping (Wrigley and Brown 2004), the model developed in this thesis is the first model which derived salience directly from sound location information.

It is however a simple, conceptual-level model. It did not include a feature extraction module and the values of the *type* and *location* features used were based on an educated guess, and could be argued to be rather arbitrary. The final stage of combining salience scores from both features was also simplified. A more comprehensive model would perhaps require a form of training to determine the extent to which deviations in location influence salience, compared to deviations in

other features. The results of experiments in this thesis suggest that the *location* feature was less important than the *type* feature. This *type* feature might however be explained by a number of different characteristics – not only changes in the spectral content of the sound, or even its temporal envelope, but also higher-level brain processes categorizing a sound as a new sound source.

## 13.4. Summary

The results of the experiments presented in this thesis indicate that spatial position of sound does not directly influence its salience. This means that, for example, the specific positioning of auditory alarm signals around the listener might not be important. However, a sound unexpectedly changing its spatial location is likely to attract attention. In addition, as loudness does tend to modulate salience, it is important to keep in mind how it might change with spatial location.

Therefore, keeping information about spatial location of sounds in the environment can be useful in auditory salience models. For example, it can be used in a deviance-detection-based model as a feature, in which any unexpected changes are identified as salient. The work presented in this thesis emphasises the importance of spatial auditory models, and advocates for auditory salience models based on deviance detection and prediction.

Finally, obtaining reliable ground truth for model training and assessment is crucial, but not straightforward for auditory salience models. The lack of standard measurement methods, or available testing datasets, means that models based on different experimental methods might in fact predict different aspects of auditory salience. The comparison of experimental methods described in this thesis identified two relevant dimensions of salience – one correlated with the more automatic and

low-level experimental methods, and the other with higher-level ones.

# 14

# Conclusions

This thesis addressed the problem of measuring and modelling spatial auditory salience. Different methods of defining and assessing auditory salience were discussed in Chapter 2. There is no consensus in the literature about which methods are best to use, or how exactly auditory salience is defined, but it is often described as the ability of sounds to attract attention. The remaining chapters of Part I used four different experimental methods to investigate how spatial position of a sound influences auditory salience.

**Experimental data suggests that the absolute spatial location of a sound alone does not modulate its salience.** The experiments described in Chapters 3, 4 and 6, each based on a different method, found no effects of spatial location of sound on auditory salience. However, unexpected changes in spatial position of an auditory stream did evoke pupil dilation responses in a distraction experiment. This confirms previous reports from the literature that breaking expectations about spatial position of a sound causes an attentional response.

**Experimental methods used to measure auditory salience vary on at least two different dimensions of salience, indicating that they measure different aspects of**

**salience.** The experiment in Chapter 7 compared methods based on oddball detection in competing streams, real-time self-reporting, and distraction, and compared them with published salience scores. The analysis found that the methods occupy a two-dimensional space of salience, where one dimension can be interpreted as more low-level, and the other high-level, with potential top-down influences. Additionally, it was shown that **pupil dilation responses can be measured to sounds unexpectedly changing their spatial position.** In a distraction experiment described in Chapter 6, pupil dilation responses revealed reactions to broken expectations which were not apparent in behavioural task performance data.

Part II explored how the spatial location of sound can be used to improve auditory salience models. Chapter 9 discussed methods of modelling salience, with emphasis on approaches based on prediction and expectations.

It was shown in Chapter 10 how **spatial location information can be incorporated into a deviance detection-based model and successfully predict pupil dilation responses to broken spatial expectations.** Sound location changing in time was used as a feature in a model based on Kalman filters, which detects unexpected deviations as salient.

In light of the findings described above, it is important for future research to more carefully define what is meant by "salience". When designing and performing auditory salience experiments, one should consider which attentional level is being measured by the chosen experimental method. Where practical, pupil dilation should be considered as a measurement method, since the work described in this thesis provides one more data point in support of it being sensitive to auditory salience.

The definition is also important to clarify for auditory salience models, and which type of salience is of interest depending on the model's intended application. Additionally, the spatial position of sounds and how it changes over time should be

considered in these models. Taking into account the spatial position can also be beneficial because salience might be influenced by features of sound which do vary with spatial position, such as loudness.

## 14.1. Further work

In future work, it would be interesting take a step further and study the relationship between salience and the continuous movement of sound sources. While it has been shown that sounds moving towards the listener tend to be more salient than those moving away (e.g. Baumgartner et al. 2017), the effects of other types of movement are less clear. Different aspects of movement such as direction, speed, starting and end position could be considered. It would be particularly interesting to see if the results confirm predictions from a Kalman filter-based model, which would be likely to mark fast movements as salient and ignore slow movement, but would not necessarily differentiate between directions or starting points.

Additionally, there is certainly more work that could be done on modelling. A model could be developed with spatial position as one of the features, and crucially, with a module which integrates it with other features in a manner which reflects ground truth data. This would require designing experiments which can reveal the relative importance of spatial position compared to other features. How the deviance detection mechanism utilizes different features could also be investigated in more general terms, for example, whether the general mechanism is more important than the particular features of sound which it utilizes.

Finally, in this thesis, it was assumed salience is independent of localisation, as it was defined and treated as a property of sound. However, it could also be argued that any potential assignment of salience to a particular spatial location happens not outside of

the head but *inside* – that it is the brain that assigns importance to one position over another. To this end, it would be interesting to see if salience depends not on *location* but rather *localisation* of a sound. This could be done, for example, by designing experiments in which the same sound stimuli are both localised and assessed on their salience.

# References

Aitchison, Laurence and Máté Lengyel (2017). 'With or without you: predictive coding and Bayesian inference in the brain'. In: Current opinion in neurobiology 46, pp. 219–227.

Akaike, H. (1974). 'A new look at the statistical model identification'. In: IEEE Transactions on Automatic Control 19.6, pp. 716–723.

Al Noori, Ahmed, Philip Duncan and Francis Li (June 2017). 'Training "On the Fly" to Improve the Performance of Speaker Recognition in Noisy Environments'. In: Audio Engineering Society Conference: 2017 AES International Conference on Audio Forensics. Paper 2-2.

Alain, Claude, Stephen R. Arnott and Benjamin J Dyson (2014). 'Varieties of auditory attention.' In: Oxford library of psychology. The Oxford handbook of cognitive neuroscience, Vol. 1. Core topics. Ed. by Kevin N. Ochsner, Stephen Kosslyn, Claude Alain, Stephen R. Arnott and Benjamin J. Dyson. Oxford University Press, pp. 215–236.

Albouy, Philippe, Lucas Benjamin, Benjamin Morillon and Robert J Zatorre (2020). 'Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody'. In: Science 367.6481, pp. 1043–1047.

Arnal, Luc H., Adeen Flinker, Andreas Kleinschmidt, Anne-Lise Giraud and David Poeppel (2015). 'Human Screams Occupy a Privileged Niche in the Communication Soundscape'. In: Current Biology 25.15, pp. 2051–2056.

Atienza, Mercedes, José L. Cantero and Carlos M Gómez (1997). 'The mismatch negativity component reveals the sensory memory during REM sleep in humans'. In: Neuroscience Letters 237.1, pp. 21–24.

Avan, Paul and Pierre Bonfils (1992). 'Analysis of possible interactions of an attentional task with cochlear micromechanics'. In: Hearing Research 57.2, pp. 269–275.

Bates, Douglas, Martin Mächler, Ben Bolker and Steve Walker (2015). 'Fitting Linear Mixed-Effects Models Using lme4'. In: Journal of Statistical Software 67.1, pp. 1–48.

Baumgartner, Robert, Darrin K. Reed, Brigitta Tóth, Virginia Best, Piotr Majdak, H. Steven Colburn and Barbara Shinn-Cunningham (Aug. 2017). 'Asymmetries in behavioral and neural responses to spectral cues demonstrate the generality of auditory looming bias'. In: Proceedings of the National Academy of Sciences 114.36, pp. 9743–9748.

BBC Sound Effects Library (2018). URL: `https://www.sound-ideas.com/Product/152/BBC-Sound-Effects-Library-Original-Series` (visited on 17/07/2018).

Bertels, Julie, Régine Kolinsky, Déborah Coucke and José Morais (2013). 'When a bang makes you run away: Spatial avoidance of threatening environmental sounds'. In: Neuroscience Letters 535, pp. 78–83.

Best, Virginia, Frederick J. Gallun, Antje Ihlefeld and Barbara G. Shinn-Cunningham (2006). 'The influence of spatial separation on divided listening'. In: The Journal of the Acoustical Society of America 120.3, pp. 1506–1516.

Best, Virginia, Erol J. Ozmeral, Norbert Kopčo and Barbara G. Shinn-Cunningham (2008). 'Object continuity enhances selective auditory attention'. In: Proceedings of the National Academy of Sciences 105.35, pp. 13174–13178.

Blauert, Jens (1997). Spatial hearing: the psychophysics of human sound localization. MIT press.

Boes, M., D. Oldoni, B. De Coensel and D. Botteldooren (2012). 'Attention-driven auditory stream segregation using a SOM coupled with an excitatory-inhibitory ANN'. In: The 2012 International Joint Conference on Neural Networks (IJCNN), pp. 1–8.

Boes, Michiel, Karlo Filipan, Bert De Coensel and Dick Botteldooren (Jan. 2018). 'Machine Listening for Park Soundscape Quality Assessment'. In: Acta Acustica united with Acustica 104.1, pp. 121–130.

Bones, Oliver Bones, Trevor John Cox and William Jonathan Davies (2018). 'Sound categories: category formation and evidence-based taxonomies'. In: Frontiers in Psychology 9, p. 1277.

Botteldooren, Dick and Bert De Coensel (2009). 'Informational masking and attention focussing on environmental sound'. eng. In: NAG/DAGA, Proceedings. Ed. by Marinus M Boone. Rotterdam, the Netherlands: NAG; DEGA, pp. 399–402.

Breebaart, Jeroen, Steven van de Par and Armin Kohlrausch (2001). 'Binaural processing model based on contralateral inhibition. III. Dependence on temporal parameters'. In: The Journal of the Acoustical Society of America 110.2, pp. 1105–1117.

Bregman, Albert S (1990). Auditory scene analysis: The perceptual organization of sound. MIT Press.

Bressler, Scott, Salwa Masud, Hari Bharadwaj and Barbara Shinn-Cunningham (2014). 'Bottom-up influences of voice continuity in focusing selective auditory attention'. In: Psychological research 78.3, pp. 349–360.

Broadbent, D. E. (1958). Perception and Communication.

Bruce, N. D. B. and J. K. Tsotsos (May 2005). 'An attentional framework for stereo vision'. In: The 2nd Canadian Conference on Computer and Robot Vision (CRV'05), pp. 88–95.

Cabrera, Densil, Sam Ferguson, Farhan Rizwi and Emery Schubert (2008). 'PsySound3: a program for the analysis of sound recordings'. In: The Journal of the Acoustical Society of America 123.5, pp. 3247–3247.

Carlin, Michael and Mounya Elhilali (2015). 'Modeling attention-driven plasticity in auditory cortical receptive fields'. In: Frontiers in Computational Neuroscience 9, p. 106.

Carlyon, Robert P., Rhodri Cusack, Jessica M. Foxton and Ian H. Robertson (2001). 'Effects of attention and unilateral neglect on auditory stream segregation'. In: J. Exp. Psychol.: Human Percept. Perform. 27.1, pp. 115–127.

Chalupper, Josef and Hugo Fastl (2002). 'Dynamic loudness model (DLM) for normal and hearing-impaired listeners'. In: Acta Acustica united with Acustica 88.3, pp. 378–386.

Chan, Jason S., Katherine Merrifield and Charles Spence (2005). 'Auditory Spatial Attention Assessed in a Flanker Interference Task'. In: Acta Acustica united with Acustica 91.3, pp. 554–563.

Chennu, Srivas, Valdas Noreika, David Gueorguiev, Alejandro Blenkmann, Silvia Kochen, Agustín Ibánez, Adrian M Owen and Tristan A Bekinschtein (2013). 'Expectation and attention in hierarchical auditory prediction'. In: Journal of Neuroscience 33.27, pp. 11194–11205.

Corral, Maria-Jose and Carles Escera (Oct. 2008). 'Effects of sound location on visual task performance and electrophysiological measures of distraction'. In: Neuroreport 19.15, pp. 1535–1539.

Cusack, Rhodri, John Decks, Genevieve Aikman and Robert P. Carlyon (2004). 'Effects of Location, Frequency Region, and Time Course of Selective Attention on Auditory Scene Analysis.' In: Journal of Experimental Psychology: Human Perception and Performance 30.4, pp. 643–656.

De Coensel, Bert and Dick Botteldooren (2008). 'Modeling auditory attention focusing in multisource environments'. In: The Journal of the Acoustical Society of America 123.5, pp. 3255–3260.

– (2010). 'A model of saliency-based auditory attention to environmental sound'. In: 20th International Congress on Acoustics, ICA 2010 August, pp. 1–8.

Deouell, Leon Y., Diana Deutsch, Donatella Scabini, Nachum Soroker and Robert T Knight (2007). 'No disillusions in auditory extinction: perceiving a melody comprised of unperceived notes.' In: Frontiers in Human Neuroscience 1.March, p. 15.

Deouell, Leon Y., Ariel Parnes, Natasha Pickard and Robert T. Knight (2006). 'Spatial location is accurately tracked by human auditory sensory memory: evidence from the mismatch negativity'. In: European Journal of Neuroscience 24.5, pp. 1488–1494.

Deutsch, Diana (1975). 'Two-channel listening to musical scales'. In: The Journal of the Acoustical Society of America 57.5, pp. 1156–1160.

Dietz, Mathias, Stephan D. Ewert and Volker Hohmann (2011). 'Auditory model based direction estimation of concurrent speakers from binaural signals'. In: Speech Communication 53.5. Perceptual and Statistical Audition, pp. 592–605.

Duangudom, V. and D. V. Anderson (Sept. 2007). 'Using auditory saliency to understand complex auditory scenes'. In: 2007 15th European Signal Processing Conference, pp. 1206–1210.

– (2013). 'Identifying salient sounds using dual-task experiments'. In: 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 1–4.

Duncan, J (Dec. 1984). 'Selective attention and the organization of visual information'. In: Journal of experimental psychology. General 113.4, pp. 501–517.

Duncan, John (2006). 'EPS Mid-Career Award 2004: brain mechanisms of attention'. In: The Quarterly Journal of Experimental Psychology 59.1, pp. 2–27.

Elhilali, Mounya (2013). 'Bayesian inference in auditory scenes'. In: Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE. IEEE, pp. 2792–2795.

Eltiti, S, D Wallace and E Fox (2005). 'Selective target processing: perceptual load or distractor salience?' In: Perception and psychophysics 67.5, pp. 876–885.

Filipan, Karlo, Annelies Bockstael, Bert De Coensel, Hrvoje Domitrovic and Dick Botteldooren (2016a). 'Auditory attention modeling within SONORUS ESR 10'. In: INTER-NOISE and NOISE-CON Congress and Conference Proceedings. Vol. 253. 1, pp. 7149–7154.

Filipan, Karlo, Annelies Bockstael, Bert De Coensel and Marc Schönwiesner (2016b). 'A novel auditory saliency prediction model based on spectrotemporal modulations'. In: Proceedings of the 22nd International Congress on Acoustics.

Font, Frederic, Gerard Roma and Xavier Serra (21/10/2013 2013). 'Freesound Technical Demo'. In: ACM International Conference on Multimedia (MM'13). ACM. Barcelona, Spain: ACM, pp. 411–412.

Friston, Karl (2009). 'The free-energy principle: a rough guide to the brain?' In: Trends in cognitive sciences 13.7, pp. 293–301.

– (2018). 'Does predictive coding have a future?' In: Nature Neuroscience.

Frith, Christopher D. and Heidelinde A. Allen (1983). 'The skin conductance orienting response as an index of attention'. In: Biological Psychology 17, pp. 27–39.

Fritz, Jonathan B, Mounya Elhilali, Stephen V David and Shihab A Shamma (2007). 'Auditory attention—focusing the searchlight on sound'. In: Current Opinion in Neurobiology 17.4. Sensory systems, pp. 437–455.

Genesis (2009). Loudness Toolbox. Retrieved from `http://genesis-acoustics.com/en/loudness_online-32.html`. Accessed: 2018-04-14.

Giannakopoulos, Theodoros (Dec. 2015). 'pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis'. In: PLOS ONE 10.12, pp. 1–17.

Giard, Marie-Hélène, Lionel Collet, Patrick Bouchet and Jacques Pernier (1994). 'Auditory selective attention in the human cochlea'. In: Brain Research 633.1, pp. 353–356.

Glasberg, Brian R and Brian CJ Moore (2002). 'A model of loudness applicable to time-varying sounds'. In: Journal of the Audio Engineering Society 50.5, pp. 331–342.

Hadlington, Lee J., Andrew M. Bridges and C. Philip Beaman (2006). 'A left-ear disadvantage for the presentation of irrelevant sound: Manipulations of task requirements and changing state'. In: Brain and Cognition 61.2, pp. 159–171.

Hadlington, Lee, Andrew M Bridges and Richard J Darby (2004). 'Auditory location in the irrelevant sound effect: The effects of presenting auditory stimuli to either the left ear, right ear or both ears'. In: Brain and Cognition 55.3, pp. 545–557.

Handel, S. (1988). 'Space is to time as vision is to audition: seductive but misleading'. In: Journal of Experimental Psychology: Human Perception and Performance 14.2, pp. 315–317.

Heilbron, Micha and Maria Chait (2017). 'Great expectations: Is there evidence for predictive coding in auditory cortex?' In: Neuroscience.

Hiscock, Merrill and Marcel Kinsbourne (2011). 'Attention and the right-ear advantage: What is the connection?' In: Brain and Cognition 76.2. Dichotic Listening Anniversary Special Issue, pp. 263–275.

Huang, Nicholas and Mounya Elhilali (2017). 'Auditory salience using natural soundscapes'. In: The Journal of the Acoustical Society of America 141.3, pp. 2163–2176.

Hughes, Robert W. (2014). 'Auditory distraction: A duplex-mechanism account'. In: PsyCh Journal 3.1, pp. 30–41.

Itti, L., C. Koch and E. Niebur (1998). 'A model of saliency-based visual attention for rapid scene analysis'. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 20.11, pp. 1254–1259.

Itti, Laurent and Pierre Baldi (2009). 'Bayesian surprise attracts human attention'. In: Vision research 49.10, pp. 1295–1306.

Itti, Laurent and Christof Koch (2001). 'Computational modelling of visual attention.' In: Nature Reviews Neuroscience 2.3, pp. 194–203.

James, William (1890). The Principles of Psychology. Dover Publications.

Johnson, Palmer Oliver and Jerzy Neyman (1936). 'Tests of certain linear hypotheses and their application to some educational problems.' In: Statistical research memoirs.

Kakouros, Sofoklis, Okko Rasanen and Unto K. Laine (May 2013). 'Attention based temporal filtering of sensory signals for data redundancy reduction'. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE.

Kalinli, O. and S. Narayanan (2009). 'Prominence Detection Using Auditory Attention Cues and Task-Dependent High Level Information'. In: IEEE Transactions on Audio, Speech, and Language Processing 17.5, pp. 1009–1024.

Kalinli, Ozlem and Shrikanth S Narayanan (2007). 'A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech'. In: INTERSPEECH-2007, pp. 1941–1944.

Kassner, Moritz, William Patera and Andreas Bulling (2014). 'Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction'. In: Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. UbiComp '14 Adjunct. Seattle, Washington: ACM, pp. 1151–1160.

Kauramäki, Jaakko, Liro P. Jääskeläinen and Mikko Sams (2007). 'Selective attention increases both gain and feature selectivity of the human auditory cortex'. In: PLOS ONE 2.9.

Kaya, E. M. and M. Elhilali (2012). 'A temporal saliency map for modeling auditory attention'. In: 2012 46th Annual Conference on Information Sciences and Systems (CISS), pp. 1–6.

Kaya, Emine Merve and Mounya Elhilali (Mar. 2013). 'A model of auditory deviance detection'. In: 2013 47th Annual Conference on Information Sciences and Systems (CISS). IEEE.

– (2014). 'Investigating bottom-up auditory attention'. In: Frontiers in Human Neuroscience 8, p. 327.

– (2017). 'Modelling auditory attention'. In: Philosophical Transactions of the Royal Society B: Biological Sciences 372.1714.

Kaya, Emine Merve, Nicolas Huang and Mounya Elhilali (Aug. 2020). 'Pitch, Timbre and Intensity Interdependently Modulate Neural Responses to Salient Sounds'. In: Neuroscience 440, pp. 1–14.

Kayser, Christoph, Christopher I. Petkov, Michael Lippert and Nikos K. Logothetis (2005). 'Mechanisms for Allocating Auditory Attention: An Auditory Saliency Map'. In: Current Biology 15.21, pp. 1943–1947.

Kidd, Gerald, Tanya L. Arbogast, Christine R. Mason and Frederick J. Gallun (2005). 'The advantage of knowing where to listen'. In: The Journal of the Acoustical Society of America 118.6, pp. 3804–3815.

Kim, Kyungtae, Kai-Hsiang Lin, Dirk B. Walther, Mark A. Hasegawa-Johnson and Tomas S. Huang (2014). 'Automatic detection of auditory salience with optimized linear filters derived from human annotation'. In: Pattern Recognition Letters 38, pp. 78–85.

Knill, David C and Alexandre Pouget (2004). 'The Bayesian brain: the role of uncertainty in neural coding and computation'. In: TRENDS in Neurosciences 27.12, pp. 712–719.

Koch, Christof and S. Ullman (1985). 'Shifts in selective visual attention: Towards the underlying neural circuitry'. In: Human Neurobiology 4.4.

Koelewijn, Thomas, Hilde de Kluiver, Barbara G. Shinn-Cunningham, Adriana A. Zekveld and Sophia E. Kramer (2015). 'The pupil response reveals

increased listening effort when it is difficult to focus attention'. In: Hearing Research 323, pp. 81–90.

Kret, Mariska E and Elio E Sjak-Shie (2019). 'Preprocessing pupil size data: Guidelines and code'. In: Behavior Research Methods 51.3, pp. 1336–1342.

Kubovy, Michael (2017). 'Concurrent-pitch segregation and the theory of indispensable attributes'. In: Perceptual organization. Routledge, pp. 55–98.

Lartillot, Olivier and Petri Toiviainen (2007). 'A Matlab toolbox for musical feature extraction from audio'. In: International Conference on Digital Audio Effects (DAFx-07). Bordeaux, pp. 237–244.

Lavie, Nilli (June 1995). 'Perceptual load as a necessary condition for selective attention'. In: Journal of Experimental Psychology. Human Perception and Performance 21.3, pp. 451–468.

Lê, Sébastien, Julie Josse and François Husson (2008). 'FactoMineR: A Package for Multivariate Analysis'. In: Journal of Statistical Software 25.1, pp. 1–18.

Lee, Jae and Charles Spence (July 2015). 'Audiovisual crossmodal cuing effects in front and rear space'. In: Frontiers in Psychology 6.

Lenth, Russell (2019). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.4.1.

Liao, Hsin-I, Shunsuke Kidani, Makoto Yoneya, Makio Kashino and Shigeto Furukawa (2015). 'Correspondences among pupillary dilation response, subjective salience of sounds, and loudness'. In: Psychonomic Bulletin & Review, pp. 412–425.

Liao, Hsin-I, Makoto Yoneya, Shunsuke Kidani, Makio Kashino and Shigeto Furukawa (2016). 'Human Pupillary Dilation Response to Deviant Auditory Stimuli: Effects of Stimulus Properties and Voluntary Attention'. In: Frontiers in Neuroscience 10, p. 43.

Lindemann, W. (1986). 'Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals'. In: The Journal of the Acoustical Society of America 80.6, pp. 1608–1622.

Lo, Steson and Sally Andrews (2015). 'To transform or not to transform: using generalized linear mixed models to analyse reaction time data'. In: Frontiers in Psychology 6, p. 1171.

Macken, William J., Sébastien Tremblay, Robert J. Houghton, Alastair P. Nicholls and Dylan M. Jones (2003). 'Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory.' In: Journal of Experimental Psychology: Human Perception and Performance 29.1, pp. 43–51.

Maison, Stéphane, Christophe Micheyl and Lionel Collet (2001). 'Influence of focused auditory attention on cochlear activity in humans'. In: Psychophysiology 38.1, pp. 35–40.

Marois, Alexandre, Katherine Labonté, Mark Parent and François Vachon (2018). 'Eyes have ears: Indexing the orienting response to sound using pupillometry'. In: International Journal of Psychophysiology 123, pp. 152–162.

Marois, Alexandre, John E. Marsh and François Vachon (2019). 'Is auditory distraction by changing-state and deviant sounds underpinned by the same mechanism? Evidence from pupillometry'. In: Biological Psychology 141, pp. 64–74.

Marois, Alexandre and François Vachon (2018). 'Can pupillometry index auditory attentional capture in contexts of active visual processing?' In: Journal of Cognitive Psychology 30.4, pp. 484–502.

Marsh, John, Lea Pilgrim and Patrik Sörqvist (2013). 'Hemispheric specialization in selective attention and short-term memory: a fine-coarse model of left- and right-ear disadvantages'. In: Frontiers in Psychology 4, p. 976.

Masutomi, Keiko, Nicolas Barascud, Makio Kashino, Josh H McDermott and Maria Chait (2015). 'Sound Segregation via Embedded Repetition Is Robust to

Inattention.' In: Journal of experimental psychology. Human perception and performance 42.3, pp. 386–400.

Mathôt, Sebastiaan (2018). 'Pupillometry: Psychology, physiology, and function'. In: Journal of Cognition 1.1.

McDermott, Josh H., Michael Schemitsch and Eero P. Simoncelli (2013). 'Summary statistics in auditory perception'. In: Nature Neuroscience 16.4, pp. 493–498.

McWalter, Richard and Josh H McDermott (2018). 'Adaptive and selective time averaging of auditory scenes'. In: Current Biology 28.9, pp. 1405–1418.

Mesaros, A., T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj and T. Virtanen (2017). 'DCASE 2017 Challenge Setup: Tasks, Datasets and Baseline System'. In: Proc. of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), pp. 85–92.

Mesaros, A., T. Heittola and T. Virtanen (2016). 'TUT database for acoustic scene classification and sound event detection'. In: Proc. of the 24th European Signal Processing Conference 2016 (EUSIPCO 2016).

Michie, P T, E L Le Page, N Solowij, M Haller and L Terry (1996). 'Evoked otoacoustic emissions and auditory selective attention.' In: Hearing research 98.1-2, pp. 54–67.

Moray, Neville (1959). 'Attention in dichotic listening: Affective cues and the influence of instructions'. In: Quarterly Journal of Experimental Psychology 11.1, pp. 56–60.

Näätänen, R. and A.W.K. Gaillard (1983). 'The Orienting Reflex and the N2 Deflection of the Event-Related Potential (ERP)'. In: Tutorials in Event Related Potential Research: Endogenous Components. Ed. by Anthony W.K. Gaillard and Walter Ritter. Vol. 10. Advances in Psychology. North-Holland, pp. 119–141.

Neuhoff, John G. (2003). 'Pitch variation is unnecessary (and sometimes insufficient) for the formation of auditory objects'. In: Cognition 87.3, pp. 219–224.

Nöstl, Anatole, John E. Marsh and Patrik Sörqvist (Nov. 2012). 'Expectations Modulate the Magnitude of Attentional Capture by Auditory Events'. In: PLOS ONE 7.11, pp. 1–7.

NSL Auditory-Cortical Matlab Toolbox (2008). URL: http://nsl.isr.umd.edu/downloads.html.

Parmentier, Fabrice B.R., Jane V. Elsley, Pilar Andrés and Francisco Barceló (2011). 'Why are auditory novels distracting? Contrasting the roles of novelty, violation of expectation and stimulus change'. In: Cognition 119.3, pp. 374–380.

Pashler, Harold (1994). 'Dual-task interference in simple tasks: Data and theory.' In: Psychological Bulletin 116.2, pp. 220–244.

Peck, Rachael B., Michael D. Hall, Jeremy R. Gaston and Kelly Dickerson (2018). 'Evidence of Change Deafness with Continuously Moving Targets'. In: Auditory Perception & Cognition 1.1-2, pp. 66–96.

Podwinska, Z., I. Sobieraj, B. M. Fazenda, W. J. Davies and M. D. Plumbley (May 2019). 'Acoustic event detection from weakly labeled data using auditory salience'. In: ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 41–45.

Politis, Archontis (2016). Microphone array processing for parametric spatial audio techniques.

R Core Team (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Rao, Rajesh PN and Dana H Ballard (1999). 'Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects'. In: Nature neuroscience 2.1, p. 79.

Remington, Anna and Jake Fairnie (2017). 'A sound advantage: Increased auditory capacity in autism'. In: Cognition 166, pp. 459–465.

Rennies, Jan, Jesko L Verhey and Hugo Fastl (2010). 'Comparison of loudness models for time-varying sounds'. In: Acta Acustica united with Acustica 96.2, pp. 383–396.

Rodríguez-Hidalgo, A., C. Peláez-Moreno and A. Gallardo-Antolín (2017). 'Towards multimodal saliency detection: An enhancement of audio-visual correlation

estimation'. In: 2017 IEEE 16th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC), pp. 438–443.

Rodríguez-Hidalgo, Antonio, Carmen Peláez-Moreno and Ascensión Gallardo-Antolín (2018). 'Echoic log-surprise: A multi-scale scheme for acoustic saliency detection'. In: Expert Systems with Applications 114, pp. 255–266.

Roeber, Urte, Andreas Widmann and Erich Schröger (2003). 'Auditory distraction by duration and location deviants: a behavioral and event-related potential study'. In: Cognitive Brain Research 17.2, pp. 347–357.

Santangelo, Valerio, Marta Olivetti Belardinelli and Charles Spence (Feb. 2007). 'The suppression of reflexive visual and auditory orienting when attention is otherwise engaged'. In: Journal of experimental psychology. Human perception and performance 33.1, pp. 137–148.

Schauerte, Boris, Benjamin Kühn, Kristian Kroschel and Rainer Stiefelhagen (2011). 'Multimodal saliency-based attention for object-based scene analysis'. In: Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on. IEEE, pp. 1173–1179.

Schauerte, Boris and Rainer Stiefelhagen (2013). '"Wow!" Bayesian surprise for salient acoustic event detection'. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, pp. 6402–6406.

Schröger, Erich, Alexandra Bendixen, Nelson J. Trujillo-Barreto and Urte Roeber (Nov. 2007). 'Processing of Abstract Rule Violations in Audition'. In: PLoS ONE 2.11. Ed. by Sheng He, e1131.

Schröger, Erich, M.-H Giard and Ch Wolff (2000). 'Auditory distraction: event-related potential and behavioral indices'. In: Clinical Neurophysiology 111.8, pp. 1450–1460.

Schröger, Erich and Christian Wolff (1998). 'Behavioral and electrophysiological effects of task-irrelevant sound change: a new distraction paradigm'. In: Cognitive Brain Research 7.1, pp. 71–87.

Shamma, Shihab (2001). 'On the role of space and time in auditory processing'. In: Trends in Cognitive Sciences 5.8, pp. 340–348.

Shamma, Shihab A., Mounya Elhilali and Christophe Micheyl (2011). 'Temporal coherence and attention in auditory scene analysis'. In: Trends in Neurosciences 34.3, pp. 114–123.

Shinn-Cunningham, Barbara G. (2008). 'Object-based auditory and visual attention'. In: Trends in Cognitive Sciences 12.5, pp. 182–186.

Shinn-Cunningham, Barbara and Virginia Best (2015). 'Auditory selective attention'. In: The handbook of attention 99.

Sivonen, Ville Pekka and Wolfgang Ellermeier (2006). 'Directional loudness in an anechoic sound field, head-related transfer functions, and binaural summation'. In: The Journal of the Acoustical Society of America 119.5, pp. 2965–2980.

Skerritt-Davis, Benjamin and Mounya Elhilali (2018). 'Detecting change in stochastic sound sequences'. In: PLoS Computational Biology 14.5, pp. 1–24.

Slaney, M., T. Agus, S. Liu, M. Kaya and M. Elhilali (2012). 'A model of attention-driven scene analysis'. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 145–148.

Sobieraj, I., L. Rencker and M. D. Plumbley (2018). 'Orthogonality-Regularized Masked NMF for Learning on Weakly Labeled Audio Data'. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2436–2440.

Southwell, Rosy, Anna Baumann, Cécile Gal, Nicolas Barascud, Karl Friston and Maria Chait (2017). 'Is predictability salient? A study of attentional capture by auditory patterns'. In: Philosophical Transactions of the Royal Society B: Biological Sciences 372.1714, p. 20160105.

Spence, Charles (Mar. 2010). 'Crossmodal spatial attention'. In: Annals of the New York Academy of Sciences 1191.1, pp. 182–200.

Spence, Charles J. and Jon Driver (1994). 'Covert spatial orienting in audition: Exogenous and endogenous mechanisms.' In: Journal of Experimental Psychology: Human Perception and Performance 20.3, pp. 555–574.

Spence, Charles, Jae Lee and Nathan Van der Stoep (2020). 'Responding to sounds from unseen locations: crossmodal attentional orienting in response to sounds presented from the rear'. In: European Journal of Neuroscience 51.5, pp. 1137–1150.

Steiner, Genevieve Z. and Robert J. Barry (2011). 'Pupillary responses and event-related potentials as indices of the orienting reflex'. In: Psychophysiology 48.12, pp. 1648–1655.

Stekelenburg, J.J. and A. Van Boxtel (2002). 'Pericranial muscular, respiratory, and heart rate components of the orienting response'. In: Psychophysiology 39.6, pp. 707–722.

Sussman, Elyse, István Winkler, Minna Huotilainen, Walter Ritter and Risto Näätänen (May 2002). 'Top-down effects can modify the initially stimulus-driven auditory organization'. In: Cognitive Brain Research 13.3, pp. 393–405.

Tang, Y and TJ Cox (Sept. 2018). 'Improving intelligibility prediction under informational masking using an auditory saliency model'. In: International Conference on Digital Audio Effects, pp. 113–119.

Thierry, Guillaume and Mark V Roberts (2007). 'Event-related potential study of attention capture by affective sounds'. In: Neuroreport 18.3, pp. 245–248.

Timpe-Syverson, GK and TN Decker (1999). 'Attention effects on distortion-product otoacoustic emissions with contralateral speech stimuli'. In: Journal of the American Academy of Audiology 10.7, pp. 371–378.

Tordini, Francesco, Albert S. Bregman and Jeremy R. Cooperstock (2015). 'The loud bird doesn't (always) get the worm: Why computational salience also needs brightness and tempo'. In: 21st International Conference on Auditory Display (ICAD2015), July 6-10, 2015, Graz, Styria, Austria, pp. 236–243.

Tordini, Francesco, Albert S. Bregman and Jeremy R. Cooperstock (Sept. 2016). 'Prioritizing foreground selection of natural chirp sounds by tempo and spectral centroid'. In: Journal on Multimodal User Interfaces 10.3, pp. 221–234.

Tordini, Francesco, Albert S. Bregman, Jeremy R. Cooperstock, Anupryia Ankolekar and Thomas Sandholm (2013). 'Toward An Improved Model Of Auditory Saliency'. In: Proceedings of the 19th International Conference on Auditory Display (ICAD2013). International Community for Auditory Display, pp. 189–196.

Tsiami, A., A. Katsamanis, P. Maragos and A. Vatakis (2016). 'Towards a behaviorally-validated computational audiovisual saliency model'. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2847–2851.

Tsuchida, Tomoki and Garrison Cottrell (2012). 'Auditory saliency using natural statistics'. In: Proceedings of the Annual Meeting of the Cognitive Science Society. Vol. 34.

Vachon, François, Robert W Hughes and Dylan M Jones (Jan. 2012). 'Broken expectations: violation of expectancies, not novelty, captures auditory attention'. In: Journal of experimental psychology. Learning, memory, and cognition 38.1, pp. 164–177.

Vachon, François, John E Marsh and Katherine Labonté (Nov. 2019). 'The automaticity of semantic processing revisited: Auditory distraction by a categorical deviation'. In: Journal of experimental psychology. General.

Vuilleumier, Patrik (Dec. 2005). 'How brains beware: neural mechanisms of emotional attention'. In: Trends in Cognitive Sciences 9.12, pp. 585–594.

Wacongne, Catherine, Etienne Labyt, Virginie van Wassenhove, Tristan Bekinschtein, Lionel Naccache and Stanislas Dehaene (2011). 'Evidence for a hierarchy of predictions and prediction errors in human cortex'. In: Proceedings of the National Academy of Sciences 108.51, pp. 20754–20759.

Walsh, Kyle P, Edward G Pasanen and Dennis McFadden (2015). 'Changes in otoacoustic emissions during selective auditory and visual attention'. In: The Journal of the Acoustical Society of America 137.5, pp. 2737–2757.

Wang, J., M. P. Da Silva, P. Le Callet and V. Ricordel (June 2013). 'Computational Model of Stereoscopic 3D Visual Saliency'. In: IEEE Transactions on Image Processing 22.6, pp. 2151–2165.

Wendt, Dorothea, Thomas Koelewijn, Patrycja Książek, Sophia E. Kramer and Thomas Lunner (2018). 'Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test'. In: Hearing Research 369. Aging & Speech Communication 2017, pp. 67–78.

Whiteley, Louise and Maneesh Sahani (2012). 'Attention in a Bayesian framework'. In: Frontiers in human neuroscience 6, p. 100.

Winn, Matthew B., Dorothea Wendt, Thomas Koelewijn and Stefanie E. Kuchinsky (2018). 'Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started'. In: Trends in Hearing 22, pp. 1–32.

Wolmetz, Michael and Mounya Elhilali (2016). 'Attentional and contextual priors in sound perception'. In: PloS ONE 11.2, pp. 1–17.

Wood, N and N Cowan (Jan. 1995). 'The cocktail party phenomenon revisited: how frequent are attention shifts to one's name in an irrelevant auditory channel?' In: Journal of experimental psychology. Learning, memory, and cognition 21.1, pp. 255–260.

Woodcock, James S, Bruno M Fazenda, Trevor J Cox and William J Davies (2019). 'Pupil dilation reveals changes in listening effort due to energetic and informational masking'. In: Proceedings of the 23rd International Congress on Acoustics, pp. 6193–6198.

Wrigley, S. N. and G. J. Brown (2004). 'A computational model of auditory selective attention'. In: IEEE Transactions on Neural Networks 15.5, pp. 1151–1163.

Xeno-canto (n.d.). [Accessed: 17/07/2018]. URL: https://www.xeno-canto.org.

Zatorre, Robert J., Todd A. Mondor and Alan C. Evans (1999). 'Auditory Attention to Space and Frequency Activates Similar Cerebral Systems'. In: NeuroImage 10.5, pp. 544–554.

Zekveld, Adriana A., Thomas Koelewijn and Sophia E. Kramer (2018). 'The Pupil Dilation Response to Auditory Stimuli: Current State of Knowledge'. In: Trends in Hearing 22, pp. 1–25.

Zhao, Sijia, Nga Wai Yum, Lucas Benjamin, Elia Benhamou, Makoto Yoneya, Shigeto Furukawa, Fred Dick, Malcolm Slaney and Maria Chait (2019). 'Rapid Ocular Responses Are Modulated by Bottom-up-Driven Auditory Salience'. In: Journal of Neuroscience 39.39, pp. 7703–7714.

# A

# Sound recordings used in the free-listening experiment

**Table A.1.:** Sound clips used in database 1. In some cases, more than one clip has been taken from the same audio file.

| No | FileSource | FileID | Category | Event | L [sone] | SC [Hz] | Len. [s] |
|---|---|---|---|---|---|---|---|
| 1 | Freesound | 151213 | people/voices | cough | 17.6 | 4071 | 0.92 |
| 2 | Freesound | 211197 | people/voices | cough | 22.3 | 4019 | 3.20 |
| 3 | Freesound | 251489 | people/voices | cough | 22.9 | 4904 | 0.72 |
| 4 | Freesound | 109759 | people/voices | laughter | 22.6 | 4096 | 5.00 |
| 5 | Freesound | 59460 | people/voices | laughter | 17.9 | 4554 | 2.18 |
| 6 | Freesound | 156844 | people/voices | sneeze | 17.3 | 4534 | 1.65 |
| 7 | Freesound | 369297 | people/voices | sneeze | 17.4 | 5424 | 0.46 |
| 8 | Freesound | 156843 | people/voices | sneeze | 22.5 | 5364 | 0.35 |
| 9 | Freesound | 187104 | people/voices | voice | 22.6 | 1546 | 0.30 |
| 10 | Freesound | 353925 | people/voices | voice | 17.2 | 1456 | 4.00 |
| 11 | Freesound | 85292 | people/voices | cough | 22.7 | 2132 | 1.36 |
| 12 | Freesound | 79769 | people/voices | laughter | 17.3 | 2398 | 4.87 |
| 13 | Freesound | 79775 | people/voices | laughter | 17.9 | 1731 | 2.38 |
| 14 | Freesound | 270301 | people/voices | scream | 22.3 | 2264 | 1.14 |
| 15 | Freesound | 54505 | people/voices | sneeze | 22.5 | 1725 | 2.02 |

**Table A.1.:** Sound clips used in database 1. In some cases, more than one clip has been taken from the same audio file.

| No | FileSource | FileID | Category | Event | L [sone] | SC [Hz] | Len. [s] |
|---|---|---|---|---|---|---|---|
| 16 | Freesound | 34783 | people/voices | yawn | 17.8 | 1922 | 2.65 |
| 17 | Freesound | 27880 | manmade/industrial | bike horn | 22.7 | 4259 | 0.42 |
| 18 | Freesound | 239030 | manmade/industrial | car horn | 22.9 | 4430 | 0.32 |
| 19 | Freesound | 54086 | manmade/industrial | car horn | 17.7 | 4744 | 2.58 |
| 20 | Freesound | 106486 | manmade/industrial | car engine | 17.9 | 4133 | 5.00 |
| 21 | Freesound | 195451 | manmade/industrial | car engine | 22.3 | 4364 | 5.00 |
| 22 | Freesound | 96519 | manmade/industrial | car engine | 17.3 | 4302 | 4.70 |
| 23 | Freesound | 196139 | manmade/industrial | car horn | 22.3 | 2056 | 0.62 |
| 24 | Freesound | 175855 | manmade/industrial | car horn | 22.7 | 2071 | 2.42 |
| 25 | Freesound | 22882 | manmade/industrial | can engine | 17.4 | 1983 | 5.00 |
| 26 | Freesound | 50454 | manmade/industrial | car engine | 17.9 | 2479 | 2.73 |
| 27 | Freesound | 50661 | manmade/industrial | car engine | 17.3 | 2391 | 5.00 |
| 28 | Freesound | 106015 | manmade/industrial | car engine | 17.6 | 1085 | 4.42 |
| 29 | Freesound | 186938 | manmade/industrial | car engine | 22.4 | 2430 | 2.85 |
| 30 | Freesound | 240671 | manmade/industrial | car engine | 22.6 | 1233 | 5.01 |
| 41 | Freesound | 374 | manmade/industrial | door lock | 22.2 | 5490 | 4.95 |
| 42 | BBC SFX | CD5-17 | manmade/industrial | car horn | 17.4 | 5273 | 1.84 |
| 31 | Freesound | 38560 | nature/animals | birds | 22.1 | 4627 | 2.17 |
| 32 | Freesound | 72547 | nature/animals | birds | 22.3 | 5038 | 5.00 |
| 33 | Freesound | 159609 | nature/animals | birds | 17.7 | 4411 | 4.44 |
| 34 | Freesound | 196251 | nature/animals | cat | 17.2 | 4254 | 3.35 |
| 35 | Freesound | 146964 | nature/animals | cat | 22.8 | 1965 | 5.01 |
| 36 | Freesound | 110389 | nature/animals | dog | 22.4 | 1611 | 0.34 |
| 37 | Freesound | 30344 | nature/animals | dog | 17.7 | 1336 | 4.69 |
| 38 | Freesound | 157695 | nature/animals | dog | 17.3 | 1659 | 4.75 |
| 39 | Freesound | 180256 | nature/animals | dog | 22.3 | 1536 | 4.57 |
| 40 | Freesound | 192236 | nature/animals | dog | 18.0 | 2097 | 4.43 |
| 43 | BBC SFX | CD6-31 | nature/animals | robin | 22.7 | 4166 | 1.36 |
| 44 | BBC SFX | CD6-31 | nature/animals | robin | 17.3 | 4991 | 2.01 |
| 45 | BBC SFX | CD6-40 | nature/animals | cockatoo | 17.8 | 4779 | 1.93 |

**Table A.1.:** Sound clips used in database 1. In some cases, more than one clip has been taken from the same audio file.

| No | FileSource | FileID | Category | Event | L [sone] | SC [Hz] | Len. [s] |
|---|---|---|---|---|---|---|---|
| 46 | BBC SFX | CD6-06 | nature/animals | cat | 22.9 | 4419 | 2.83 |
| 47 | xeno-canto | XC155713 | nature/animals | crow | 22.7 | 2414 | 2.68 |
| 48 | xeno-canto | XC155713 | nature/animals | crow | 17.5 | 2420 | 1.74 |

**Table A.2.:** Sound clips used in database 2

| No | FileSource | FileID | Category | Event | L [sone] | SC [Hz] | Len. [s] |
|---|---|---|---|---|---|---|---|
| 49 | Freesound | 221518 | people/voices | sneeze | 25.5 | 4174 | 3.34 |
| 50 | Freesound | 108017 | people/voices | cough | 25.2 | 1988 | 5.02 |
| 51 | Freesound | 119450 | people/voices | laughter | 14.8 | 2103 | 1.13 |
| 52 | Freesound | 254869 | people/voices | crying | 15.3 | 2221 | 7.52 |
| 53 | Freesound | 411924 | people/voices | whistle | 24.9 | 1203 | 0.85 |
| 54 | Freesound | 53663 | people/voices | cough | 24.8 | 4123 | 3.30 |
| 55 | Freesound | 328892 | people/voices | sneeze | 15.4 | 4840 | 2.89 |
| 56 | Freesound | 382906 | people/voices | laughter | 14.9 | 1680 | 2.39 |
| 57 | Freesound | 270301 | people/voices | scream | 24.6 | 2182 | 1.44 |
| 58 | Freesound | 119102 | people/voices | sneeze | 15.0 | 4569 | 0.63 |
| 59 | Freesound | 34783 | people/voices | yawn | 15.2 | 1625 | 4.32 |
| 60 | Freesound | 132295 | people/voices | whistle | 14.5 | 4607 | 3.34 |
| 61 | Freesound | 118104 | people/voices | sneeze | 24.7 | 4701 | 2.59 |
| 62 | Freesound | 194533 | people/voices | sneeze | 25.1 | 5455 | 2.81 |
| 63 | Freesound | 156844 | people/voices | sneeze | 15.4 | 4448 | 4.92 |
| 64 | Freesound | 411638 | people/voices | whistle | 25.3 | 2235 | 0.65 |
| 65 | Freesound | 174840 | manmade/industrial | car horn | 25.5 | 4432 | 1.86 |
| 66 | Freesound | 18527 | manmade/industrial | car engine | 15.0 | 1190 | 5.00 |
| 67 | Freesound | 331542 | manmade/industrial | car horn | 15.1 | 4015 | 3.14 |
| 68 | Freesound | 243783 | manmade/industrial | car engine | 15.0 | 2430 | 5.00 |
| 69 | Freesound | 148398 | manmade/industrial | car engine | 14.7 | 2018 | 3.99 |
| 70 | Freesound | 119455 | manmade/industrial | car engine | 24.9 | 4085 | 3.67 |
| 71 | Freesound | 175846 | manmade/industrial | car horn | 24.6 | 1893 | 2.78 |

| No | FileSource | FileID | Category | Event | L [sone] | SC [Hz] | Len. [s] |
|---|---|---|---|---|---|---|---|
| 72 | Freesound | 243773 | manmade/industrial | car engine | 15.3 | 2214 | 5.00 |
| 73 | Freesound | 50455 | manmade/industrial | car engine | 25.3 | 1989 | 2.92 |
| 74 | Freesound | 38682 | manmade/industrial | car engine | 14.8 | 4193 | 4.94 |
| 75 | Freesound | 125520 | manmade/industrial | car horn | 25.2 | 2036 | 4.89 |
| 76 | Freesound | 174840 | manmade/industrial | car horn | 15.3 | 4824 | 2.02 |
| 77 | Freesound | 83465 | manmade/industrial | car horn | 24.6 | 5356 | 0.50 |
| 78 | Freesound | 351421 | manmade/industrial | car engine | 24.9 | 1867 | 3.96 |
| 79 | BBC SFX | CD5-02 | manmade/industrial | car driving | 15.1 | 5393 | 2.96 |
| 80 | BBC SFX | CD5-05 | manmade/industrial | motorcycle | 25.2 | 4537 | 4.86 |
| 81 | Freesound | 96950 | nature/animals | crow | 15.2 | 2100 | 4.35 |
| 82 | Freesound | 100038 | nature/animals | birds | 24.7 | 4657 | 3.35 |
| 83 | Freesound | 214759 | nature/animals | cat | 25.2 | 2076 | 3.11 |
| 84 | Freesound | 214759 | nature/animals | cat | 14.9 | 2143 | 4.92 |
| 85 | Freesound | 242414 | nature/animals | dog | 24.8 | 1940 | 1.08 |
| 86 | Freesound | 191687 | nature/animals | dog | 14.9 | 1643 | 4.50 |
| 87 | Freesound | 138344 | nature/animals | crow | 15.4 | 4015 | 5.01 |
| 88 | Freesound | 130034 | nature/animals | cat | 25.3 | 5245 | 3.00 |
| 89 | Freesound | 257839 | nature/animals | birds | 24.9 | 4017 | 5.00 |
| 90 | Freesound | 207124 | nature/animals | dog | 25.2 | 1184 | 5.00 |
| 91 | Freesound | 212454 | nature/animals | dog | 15.4 | 1811 | 4.10 |
| 92 | BBC SFX | CD6-07 | nature/animals | dog | 24.7 | 2383 | 2.50 |
| 93 | BBC SFX | CD6-31 | nature/animals | robin | 14.8 | 4807 | 1.45 |
| 94 | BBC SFX | CD6-09 | nature/animals | dog | 25.3 | 4311 | 1.46 |
| 95 | BBC SFX | CD6-42 | nature/animals | parakeet | 14.9 | 5336 | 2.44 |
| 96 | xeno-canto | 402795 | nature/animals | bird | 15.0 | 4243 | 2.23 |

# B

# Sound recordings used in the distraction experiment

The following recordings were used to create stimuli for the experiment in Chapter 6. Most stimuli were short excerpts of these recordings. In some cases, more than one clip has been taken from the same audio file.

| File source | File ID | Sound type |
|---|---|---|
| Birdsong recordings (context and distractors) | | |
| Xeno-canto | 62259 | Canada goose |
| Xeno-canto | 130583 | Willow tit |
| Xeno-canto | 135492 | Black-headed grosbeak |
| Xeno-canto | 183650 | Blyth's reed warbler |
| Xeno-canto | 199077 | Common redshank |
| Xeno-canto | 285296 | Common whitethroat |
| Xeno-canto | 330250 | Willow warbler |
| Xeno-canto | 362008 | Boreal owl |
| Xeno-canto | 371426 | Great tit |
| Xeno-canto | 379910 | Black-naped monarch |

| File source | File ID | Sound type |
|---|---|---|
| Xeno-canto | 380355 | Great tit |
| Xeno-canto | 400921 | Common blackbird |
| Xeno-canto | 402795 | Great tit |
| Xeno-canto | 402994 | Great tit |
| Xeno-canto | 433102 | Pavonine cuckoo |
| Xeno-canto | 433343 | Pavonine cuckoo |
| Xeno-canto | 443004 | Rattling cisticola |
| Xeno-canto | 451797 | Planalto tapaculo |
| Xeno-canto | 463248 | Sardinian warbler |
| BBC SFX Library / CD6-Animals & Birds | 31 | Robin |
| FreeSound | 72547 | birdsong |
| FreeSound | 242490 | birdsong |
| Other recordings (distractors) | | |
| BBC SFX Library / CD6-Animals & Birds | 03 | Cat |
| BBC SFX Library / CD6-Animals & Birds | 07 | Dog |
| BBC SFX Library / CD6-Animals & Birds | 08 | Dog |
| BBC SFX Library / CD6-Animals & Birds | 09 | Dogs barking |
| BBC SFX Library / CD6-Animals & Birds | 17 | Hen |
| BBC SFX Library / CD6-Animals & Birds | 30 | Donkey |
| BBC Sound Effects | 07022498 | Crash: teapot broken |
| BBC Sound Effects | 07058028 | Chopping tree |
| BBC Sound Effects | 07058171 | Canned drink opened |
| BBC Sound Effects | 07063116 | Walking |
| BBC Sound Effects | 07065075 | Bottle put onto shelf |
| BBC Sound Effects | 07070149 | Clock |
| BBC Sound Effects | 07074124 | Chains rattling |
| BBC Sound Effects | 07074131 | Clock cartoon |
| BBC Sound Effects | 07074135 | Cork pop |

| File source | File ID | Sound type |
|---|---|---|
| FreeSound | 103995 | Knock |
| FreeSound | 115920 | Clapping |
| FreeSound | 151212 | Cough |
| FreeSound | 197435 | Clapping |
| FreeSound | 100032 | Dog |
| FreeSound | 114587 | Dog |
| FreeSound | 9032 | Dog |
| FreeSound | 12654 | Water drop |
| FreeSound | 50623 | Water drop |
| FreeSound | 156026 | Frog |
| FreeSound | 15689 | Frog |