TESTING SPATIAL ASPECTS OF AUDITORY SALIENCE

Zuzanna Podwinska, Bruno M Fazenda, William J Davies

University of Salford
School of Computing, Science & Engineering
43 Crescent, Salford, M5 4WT, UK
z.podwinska@edu.salford.ac.uk

ABSTRACT

Auditory salience describes the extent to which sounds attract the listener's attention. So far, there have not been any published studies testing if the location of sound relative to the listener influences its salience. In fact, not many experiments in general test auditory attention in a fully spatialised setting, with sounds in front and behind the listener. We modified two experimental methods from the literature so that they can be used to test spatial salience - one based on oddball detection and artificially created sounds, the other based on self-reported attention tracking in a more ecologically valid scenario. Each of these methods has its advantages and each presents different challenges. However, they both seem to indicate that high frequency sounds arriving from the back are slightly less salient. We believe this result could likely be explained by loudness differences.

1. INTRODUCTION

1.1. Motivation

Certain sounds in the environment involuntarily attract attention. This happens outside of the listener's control, and depends on the properties of the sound itself. It is also task-independent: even if the listener is consciously paying attention to a radio programme or a piece of music, her attention will be drawn to a new *salient* sound in the environment. Salience can be defined as the property of sound which makes it stand out among other sounds [1].

Although there have been studies on salience of acoustic features such as loudness, brightness or tempo [2, 9, 10], no studies so far have shown how salience might be related to the location of the sound. To date, spatial attention studies have focused on target-distractor separation and relied on focused top-down attention (e.g. [3, 4]). But do different locations of sound around the listener have inherently different salience, regardless of what the person is focused on? It is not unreasonable to suspect that it might be the case. For example, one could argue that there would be an evolutionary advantage to humans being more alert to sounds arriving from behind them, where vision provides little useful information. The difficulty in studying this question lies mainly in determining where a person's attention was directed. Unlike in vision, where eye-tracking is often used, humans do not have auditory organs which would indicate the direction of the attentional 'spotlight'.

In this work, we propose two methods of testing spatial auditory salience, which are extensions of previously published salience experiments.

1.2. Measuring auditory salience

There is not one widely agreed upon way of behaviourally measuring auditory salience. Perhaps the most straightforward way of testing whether a sound is salient is asking human subjects directly. For example, in an annotation task [5], participants were asked 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' [6, 7, 8]. 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 (especially since there is no single, universally accepted definition of auditory salience). Some researchers [9] avoid this issue by asking participants to indicate where their *attention* is, and to do so in real time. This is somewhat analogous to gaze tracking in visual attention, but a less direct representation of the phenomenon, as it also involves conscious tracking of one's attention.

Another way of testing salience is through sound detection - for example, of sound in noise [6]. This is more objective but seems more removed from the notion of salience. It assumes that more salient sounds will be easier to detect, which might not be strictly true. Another task involved detection of a salient event in a scene [10], which still might confound salience and energetic masking issues. A different paradigm is based on oddball detection - detecting 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 indicators of stimulus salience (e.g. [2, 11]).

Finally, some experiments use task interference paradigms, where participants are asked to perform a task while unrelated distracting stimuli are played to them. Sound salience is assumed to be directly related to the amount of distraction caused, so changes in response time and error rate are an indication of stimulus salience.

In the following section, we present two methods which are spatial extensions of two of the published salience measuring paradigms discussed above [2, 9].

2. EXPERIMENT 1

2.1. Methods

The first experiment is based on the Segregation of Asynchronous Patterns (SOAP) paradigm [12]. It is based on the idea that two perceived auditory streams compete for attentional resources, and as a result one of them becomes *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 become the foreground. The main assumption here is that it will be easier to detect changes in the foreground (more salient) stream.

In the original SOAP experiment, two sound patterns were presented dichotically through headphones. Both patterns consisted of short birdsong excerpts separated by constant interstimulus 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 auditory object. The participants' task was to detect a change in ISI in one of the streams. No instructions were given about which stream should be attended to. According to the SOAP framework, listeners should be statistically more likely to attend to, and detect changes in, the more salient stream.

In order for the SOAP framework to account for spatial effects, we modified it so that sound patterns arrive at the listener from 2 out of 6 locations around them, rather than just left and right. The participant was seated in an acoustically treated listening room, surrounded by loudspeakers as in Figure 1. In [12], participants were asked to choose between the left or right stream. However, in this experiment we wanted to avoid requiring participants to localise sounds, as we were not interested in their localisation ability as such. Therefore, we decided to use two distinctively different stimulus types: short noise bursts, either high- or low-pass filtered at 2 kHz. Each pattern contained only one type of stimulus, and participants were asked to detect a shortened ISI and indicate whether it occurred in the high or low frequency pattern. The sounds were designed so that there was no overlapping spectral content, to ensure that it was easy to segregate and follow one of the streams without too much interference from the other. To ensure asynchrony, one of the two patterns always included shorter stimuli than the other (200 versus 150 ms). This resulted in one pattern sounding faster than the other (a property which is referred to here as tempo). Independent variables were then: sound location (1 to 6, as shown on figure 1), frequency (high and low), and tempo (fast and slow). Each participant was exposed to all conditions.

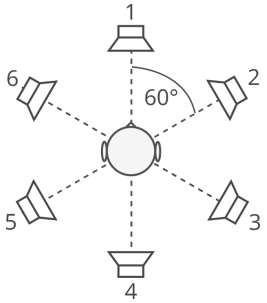


Figure 1: Loudspeaker set-up in the listening room.

Before the main experiment, participants completed a short training session and a baseline test, where only one pattern was present at a time. 19 volunteers took part in the experiment, all with self-reported normal hearing, average age 30.4, 4 female, 18 right-handed.

2.2. Results

Time elapsed from the end of the shortened ISI to the button press was recorded as response time (RT). Only correct responses were taken into account. The data was analysed with a Generalised Linear Mixed Model (lme4 package in R [13]) with an inverse Gaussian distribution and an identity link function, to account for a non-normal distribution of response times. Fixed effects were location, frequency, and tempo, and random effects were participant and background sound location. A model including frequency-tempo and frequency-location interactions was used as it gave the best fit (based on the Akaike information criterion).

Table 1: GLMM results on response time data. Significant predictors are in bold.

Fixed effects	Coeff.	SE	Z	p-value
(Intercept)	**0.843**	**0.031**	**27.13**	**< 0.0001**
Location 2	0.019	0.026	0.74	0.458
Location 3	-0.007	0.025	-0.29	0.773
Location 4	-0.038	0.024	-1.56	0.119
Location 5	-0.026	0.024	-1.09	0.276
Location 6	-0.029	0.024	-1.18	0.238
Frequency (high)	-0.005	0.026	-0.21	0.835
Tempo (fast)	0.003	0.014	0.21	0.831
Frequency:Tempo	**-0.070**	**0.020**	**-3.47**	**0.0005**
Location2:Frequency	-0.032	0.034	-0.92	0.356
Location3:Frequency	0.013	0.034	0.38	0.707
Location4:Frequency	**0.138**	**0.035**	**3.89**	**< 0.0001**
Location5:Frequency	0.061	0.034	1.79	0.074
Location6:Frequency	0.059	0.034	1.73	0.084
Random effects	Standard deviation			
Participant	0.103			
Background location	0.016			

The results, shown in Table 1, indicate that there are significant interactions: frequency-tempo and frequency-location. A post-hoc analysis of contrasts shows that, for low frequency stimuli, there are no significant differences between locations. However, for high frequency stimuli, there are significant differences between front and back locations ($p = 0.0018$), back and right-front ($p = 0.0002$), and back and right-back ($p = 0.005$). Figure 2 shows estimated mean response times and confidence intervals for the two interactions.

2.3. Discussion

There was no difference between participants' responses to different locations and tempo when the stimuli were low frequency noise. However, for high frequency stimuli, responses were on average 67ms faster for fast compared to slow patterns. Additionally, for high frequency stimuli, responses were significantly slower (about 100 ms) if target sound was behind the listener, than if it was in front of them.

The results show an interaction between tempo and frequency: slow patterns were more salient if they were low frequency, and fast patterns were more salient if they were high frequency. Interestingly, the study this experiment was based on [2] also found

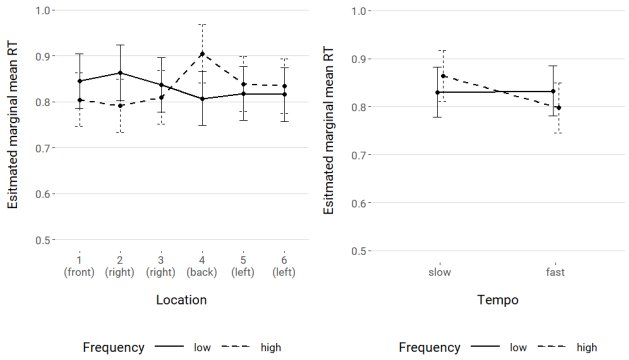


Figure 2: Estimated marginal means with 95% confidence intervals show interactions between spectrum of the noise bursts and and location (left panel) and tempo (right panel).

an interaction between these variables, but in the opposite direction: "the sounds with higher salience [...] are those with faster tempo and lower spectral centroid". They also concluded that in general, sounds with a lower spectral centroid were more salient, which was not found here. This last result is also in contrast to some other studies, such as [9], which found a significant increase in brightness in salient events. Our experiment did not find a significant effect of spectral centroid on salience.

3. EXPERIMENT 2

3.1. Methods

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

The goal of the second experiment was to test spatial salience in a more ecologically valid scenario. The experimental procedure was inspired by [9], 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, a mouse and a visual interface were used.

A similar procedure was used here, but with stimuli arriving from different locations 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 [9], 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 [9] 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). Position in time of distractors was randomised but the same for all participants. 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:

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

3.1.1. Target sounds

Targets were short clips from recordings of real-world sounds (from [14], [15] and ([16]), on average 3 seconds long. Time spacing between consecutive stimuli varied randomly from 2 to 4 s. The stimuli belonged to three different semantic categories, which were determined based on the soundscape taxonomy established in a sorting experiment by [17]. The categories were: Nature (subcategory: Animals), People (subcategory: Voices), Manmade (subcategory: Industrial).

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

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

where $Y(n)$ is the amplitude of the nth 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 sounding unnatural, sound spectra were not manipulated. Instead, events were chosen so that their spectral centroid falls within one of two groups: 1000-2500 Hz or 4000-5500 Hz.

Short-term loudness of sound was calculated using the Dynamic Loudness Model [18] available through the PsySound3 toolbox in Matlab [19]. As an indication of 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 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.

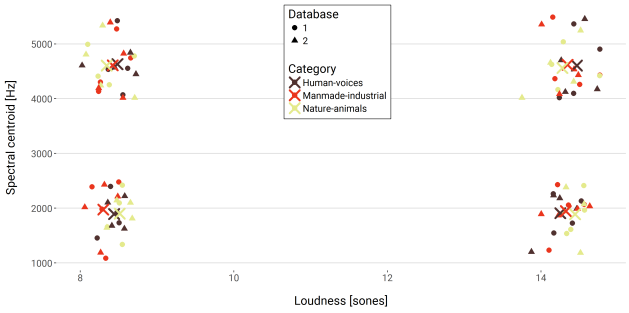
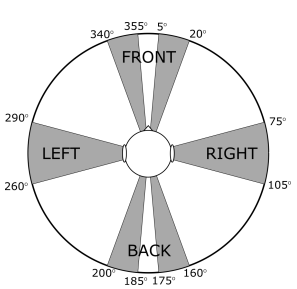


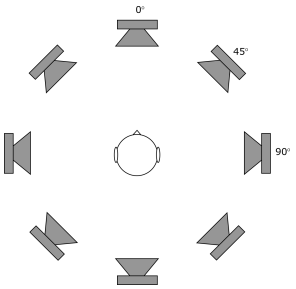
Figure 3: Stimuli used in the experiment. Database corresponds to two stimuli groups, used with different backgrounds. Colours indicate one of the three semantic categories. Recordings were chosen to fall within the two spectral centroid levels, and then their loudness was manipulated, while keeping the pairs of brightness groups as similar as possible.

Each sound was assigned to either one of the two levels in a way that minimised mean and variance differences between brightness levels. Figure 3 shows all targets on the loudness-brightness spectrum.

Targets were placed at one of four 30° areas (cones in Figure 4a) 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 in order to minimise front-back confusion effects (see Figure 4a). 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.

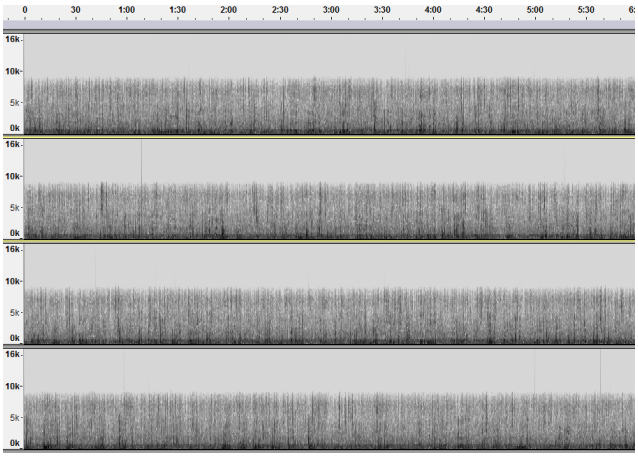
Figure 4: Target locations and experimental setup.

3.1.2. Scenes

These targets were used in two different sound scenes, each about 5 minutes long, each with different background sound and distracting events. Targets were divided into 2 balanced groups (this is represented by different shapes in Figure 3) 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 (from [20]). 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, and no one reported knowing any of the languages well enough to understand any of the sentences. The speech was originally recorded at 16000 Hz sampling frequency. Spectrogram of the *speech* background is shown in figure 5.

Figure 5: Spectrogram of the *speech* background. Each row represents one channel. For clarity, only 4 of the 8 channels are shown.



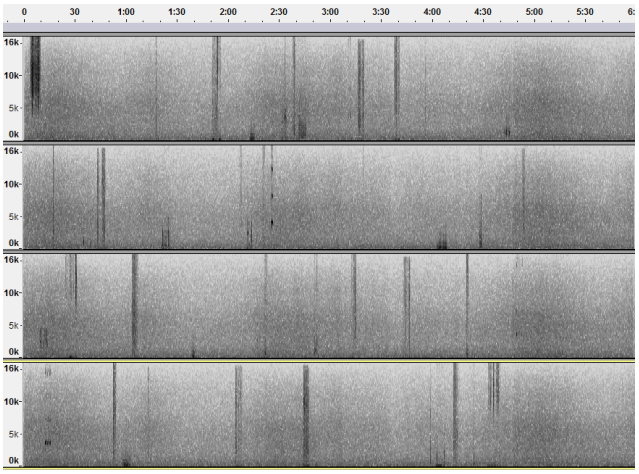
The second scene (*nature*) had a steady wind sound as background, and distracting sound events from the semantic category Nature, but different subcategories than the targets: 48 were sounds of insects, 32 of leaves and branches, and 16 of water, all positioned evenly across all 8 channels. These distractors were distributed over the background in a similar manner as target events, with one or two distractors present at any given time, and 2-4 s gaps in-between. Some distractors overlapped with targets, but because of the randomisation of target positions and timings, this overlap was different for each participant. Average background loudness was 4.3 sones, and average distractor loudness: 11 sones. Spectral centroid of distractors ranged between 780 Hz and 13600 Hz. Figure 6 shows a spectrogram of this background.

3.1.3. Reproduction system and participants

The target stimuli were reproduced over a 2nd order ambisonic system, using the Higher Order Ambisonic Library Matlab toolbox [21]. The reproduction system was 8 loudspeakers placed on an octagon, at ear-level (see Figure 4b). Background was not ambisonic but rather an 8-channel signal sent directly to the loudspeakers. All sounds were reproduced with a 44100 Hz sampling frequency.

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

Figure 6: Spectrogram of the *nature* background. Each row represents one channel. For clarity, only 4 of the 8 channels are shown.



3.2. Results

3.2.1. Data preprocessing

Figure 7 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 correct responses, unnecessarily reducing participants' performance.

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 7 shows the chosen centre area and response quadrants.

3.2.2. 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 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.

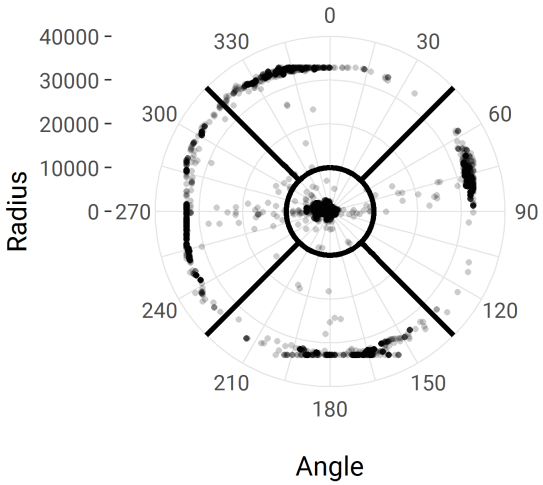


Figure 7: Raw joystick movement data for one of the baseline experiment participants. Dots are joystick positions sampled at regular time intervals. The darker the region, the more data points there are. Solid black lines show how the space was divided into quadrants and the centre area.

A GLMM model confirmed that none of the other factors (loudness, brightness, category) had an effect on localisation accuracy, nor were there any significant interactions between them.

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

3.2.3. 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 was pooled together, forming a new variable in the analysis - experiment type. By looking at interactions between 'Experiment' and other variables, we can see if adding distracting sounds - in other words, introducing attentional effects - had an effect on any of these variables.

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 2. Wald tests indicate 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, and loudness should not affect localisation. However, there is also a possibility that some of this effect is due to different levels of energetic masking.

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 seem 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 8. 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.

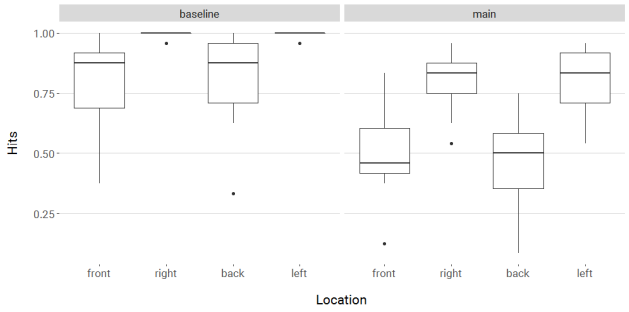
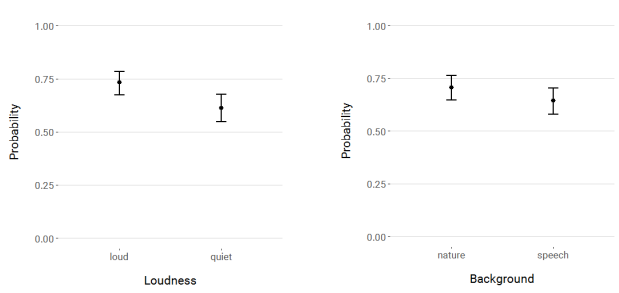


Figure 8: Responses to sounds in different positions for the baseline localisation experiment (left panel) and the main experiment (right panel). Boxplots show hit scores calculated for a particular condition and for each participant.



(a) Effect of loudness (b) Effect of background type

Figure 9: Probability of attending to target sounds in the main experiment, based on model in Table 3. Error bars show 95% confidence intervals.

To see if there were any interactions between independent variables, we analysed the main experiment data separately from the baseline data. A GLMM model with the best fit based on AIC included one interaction: location/brightness (see Table 3). The model indicates that brightness significantly changes 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 10.

The model also confirms a significant main effect of loudness, and suggests that there is a significant effect of background type,

Table 2: Coefficient estimates of the interactions in the fitted model, their standard errors, Z statistics and p-values. Note that we are mostly interested in how the main experiment interacted with other variables, not in the main effects. Predictors in bold are statistically significant.

Fixed effects	Coeff.	SE	Z	p-value
(Intercept)	**1.54**	**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	**-2.18**	**0.31**	**-7.07**	**< 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	Standard deviation			
Participant	0.51			

with higher probability of attending to targets in the *nature* background. This is not surprising, as compared to *speech*, the *nature* background was less busy. Figure 9 shows both of these effects.

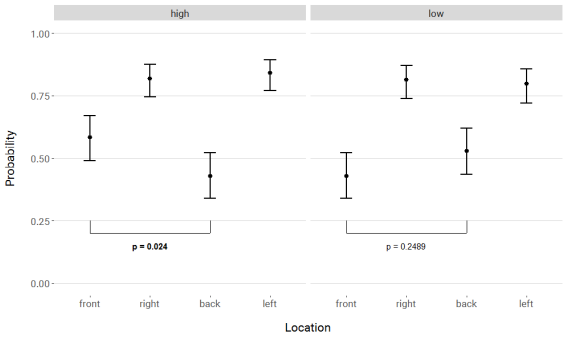


Figure 10: Probability of attending to sounds in different positions in the main experiment, split by brightness of the sound. Error bars show 95% confidence intervals. Based on model in Table 3.

3.3. Discussion

As expected, participants paid attention to louder sounds more often, which is in agreement with other studies on salience of loudness [10, 9]. 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.

There are significant differences between sound categories. This 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

Table 3: Results of the GLMM model fitted with main experiment data. Significant predictors in bold.

Fixed effects	Coeff.	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	Standard deviation			
Participant	0.43			

other properties of the sounds (e.g. impulsiveness) which vary between the categories. A more thorough analysis of the acoustic properties of sounds in different categories could be useful.

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 emotional reaction to a sound. With enough data points, these effects should average out, leaving the effects of the target sounds themselves. These effects will be the focus of a further study.

4. GENERAL DISCUSSION

4.1. Comparison of methods

Each of the two experiments used a different method to study the effect of sound location on auditory salience. There are a few important differences between them. Firstly, the tasks used in the two experiments were very different. It is reasonable to assume that tracking one's attention - Experiment 2 - is a more complex task, more prone to errors than the oddball detection task used in Experiment 1.

Secondly, unlike Experiment 1, Experiment 2 used a method which relied to some extent on sound localisation, which is not always perfect, and might add additional errors. This introduced the need for a way to separate localisation errors from attentional effects.

Thirdly, although no instructions about what to listen to were given in Experiment 2, it allowed for possible effects of top-down attention and personal preference for a specific sound or sound category. This makes Experiment 2 more sensitive to subjectivity and processes beyond bottom-up attention.

Finally, the experiments used very different sounds as stimuli. The advantage of Experiment 2 was its use of real-world sounds and a more ecologically valid listening environment.

4.2. Comparison of results

Despite the differences in the two methods, both experiments were designed to test auditory salience in a spatial setting. Both experiments seem to show a small decline in saliency of sounds arriving from behind the listener, but only for higher frequency sounds. This effect could potentially be explained by loudness differences

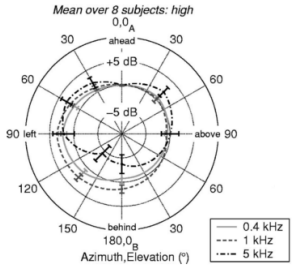


Figure 11: Directional loudness sensitivities at 65 dB SPL, reproduced from [22]

caused by pinna shadowing. [22] measured loudness for different locations around the listener (only on the left, however, as they assumed symmetry). Their results, shown in Figure 11, suggest 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 two experiments. No other effects of spatial salience were found.

Neither of the experiments showed a clear main effect of spectral content of the sound on salience, in contrast to the original studies the experiments were based on. However, it is worth pointing out that the two original studies provide contradicting results. While [2] report that lower sound patters were more salient, in [9], an increase in brightness causes an increase in salience. Both studies use the spectral centroid as a representation of brightness, however [2] only found the significant effect when the spectral centroid was calculated on sounds previously weighted with equal-loudness contours. It might be that the relationship between spectral content of sound and its salience is more complex and needs further research.

5. CONCLUSIONS AND FUTURE WORK

We have designed and tested two methods for testing spatial aspects of auditory salience. Having these methods not only lets us study the effect of location of sound on salience, but also allows conducting salience experiments in a more ecologically valid listening situation. Method used in Experiment 1 gives results which are easier to interpret, however it is difficult to use more natural sounds as stimuli. On the other hand, the method used in Experiment 2 is more prone to errors and effects of top-down attention, but allows a more natural listening environment, with real-life sounds.

The results suggest that high frequency sounds arriving from behind the listener are less salient, but the effect is not large and could probably be explained by loudness differences. If this indeed is the case, it confirms the usefulness of sound for interfaces, where an auditory alert can be placed anywhere around the person and still effectively attract their attention.

Because of the possible effects of subjectivity and top-down attention in Experiment 2, more participants will be invited to participate in it in the future. With more data points, we will be more confident that the errors caused by subjective effects average out, leaving only the effect of the sound itself. Additionally, it might be interesting to confirm this result in a distraction-type experiment, as well as to more carefully account for differences in loudness of sounds in different locations.

6. REFERENCES

[1] F. Tordini, A. S. Bregman, and J. R. Cooperstock, "Prioritizing foreground selection of natural chirp sounds by tempo and spectral centroid," *Journal on Multimodal User Interfaces*, vol. 10, no. 3, pp. 221–234, Sep 2016.

[2] ——, "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*, 2015, pp. 236–243. [Online]. Available: https://smartech.gatech.edu/handle/1853/54145

[3] C. J. Spence and J. Driver, "Covert Spatial Orienting in Audition: Exogenous and Endogenous Mechanisms," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 20, no. 3, pp. 555–574, 1994.

[4] V. Best, F. J. Gallun, A. Ihlefeld, and B. G. Shinn-Cunningham, "The influence of spatial separation on divided listening," *The Journal of the Acoustical Society of America*, vol. 120, no. 3, pp. 1506–1516, 2006.

[5] K. Kim, K. H. Lin, D. B. Walther, M. A. Hasegawa-Johnson, and T. S. Huang, "Automatic detection of auditory salience with optimized linear filters derived from human annotation," *Pattern Recognition Letters*, vol. 38, no. 1, pp. 78–85, 2014.

[6] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: An auditory saliency map," *Current Biology*, vol. 15, no. 21, pp. 1943–1947, 2005.

[7] V. Duangudom and D. V. Anderson, "Using auditory saliency to understand complex auditory scenes," *European Signal Processing Conference*, no. Eusipco, pp. 1206–1210, 2007.

[8] T. Tsuchida and G. W. Cottrell, "Auditory saliency using natural statistics," *Proc. Annual Meeting of the Cognitive Science*, pp. 1048–1053, 2012.

[9] N. Huang and M. Elhilali, "Auditory salience using natural soundscapes," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 2163–2176, 2017.

[10] E. M. Kaya and M. Elhilali, "Investigating bottom-up auditory attention," *Frontiers in human neuroscience*, vol. 8, no. May, p. 327, 2014.

[11] R. Southwell, A. Baumann, C. Gal, N. Barascud, K. J. Friston, and M. Chait, "Is predictability salient? A study of attentional capture by auditory patterns," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, no. 1714, p. 20160105, feb 2017.

[12] F. Tordini, A. S. Bregman, and J. R. Cooperstock, "Toward an improved model of auditory saliency," in *ICAD*, 2013, pp. 189–196.

[13] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[14] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *ACM International Conference on Multimedia (MM'13)*, ACM. Barcelona, Spain: ACM, 21/10/2013 2013, pp. 411–412.

[15] Bbc sound effects library. [Online]. Available: https://www.sound-ideas.com/Product/152/BBC-Sound-Effects-Library-Original-Series

[16] Xeno-canto, https://www.xeno-canto.org, [Accessed: 17/07/2018]. [Online]. Available: https://www.xeno-canto.org

[17] O. B. Bones, T. J. Cox, and W. J. Davies, "Sound categories: category formation and evidence-based taxonomies," *Frontiers in Psychology*, vol. 9, p. 1277, 2018.

[18] J. Chalupper and H. Fastl, "Dynamic loudness model (dlm) for normal and hearing-impaired listeners," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 378–386, 2002.

[19] D. Cabrera, S. Ferguson, and E. Schubert, "'psysound3': Software for acoustical and psychoacoustical analysis of sound recordings." Georgia Institute of Technology, 2007. [Online]. Available: http://www.psysound.org

[20] A. Al Noori, P. Duncan, and F. Li, "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*, Jun 2017. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=18744

[21] A. Politis, "Microphone array processing for parametric spatial audio techniques," 2016. [Online]. Available: https://uk.mathworks.com/matlabcentral/fileexchange/54833-higher-order-ambisonics-hoa-library

[22] V. P. Sivonen and W. Ellermeier, "Directional loudness in an anechoic sound field, head-related transfer functions, and binaural summation," *The Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 2965–2980, 2006.