


RESEARCH

Open Access



# Exploring sex differences in auditory saliency: the role of acoustic characteristics in bottom-up attention

Naoya Obama<sup>1</sup>, Yoshiki Sato<sup>2</sup>, Narihiro Kodama<sup>1</sup>, Yuhei Kodani<sup>1,3</sup>, Katsuya Nakamura<sup>1,4</sup>, Ayaka Yokozeki<sup>5</sup> and Shinsuke Nagami<sup>6\*</sup> 

## Abstract

**Background** Several cognitive functions are related to sex. However, the relationship between auditory attention and sex remains unclear. The present study aimed to explore sex differences in auditory saliency judgments, with a particular focus on bottom-up type auditory attention.

**Methods** Forty-five typical adults (mean age:  $21.5 \pm 0.64$  years) with no known hearing deficits, intelligence abnormalities, or attention deficits were enrolled in this study. They were tasked with annotating attention capturing sounds from five audio clips played in a soundproof room. Each stimulus contained ten salient sounds randomly placed within a 1-min natural soundscape. We conducted a generalized linear mixed model (GLMM) analysis using the number of responses to salient sounds as the dependent variable, sex as the between-subjects factor, duration, maximum loudness, and maximum spectrum of each sound as the within-subjects factor, and each sound event and participant as the variable effect.

**Results** No significant differences were found between male and female groups in age, hearing threshold, intellectual function, and attention function (all  $p > 0.05$ ). Analysis confirmed 77 distinct sound events, with individual response rates of 4.0–100%. In a GLMM analysis, the main effect of sex was not statistically significant ( $p = 0.458$ ). Duration and spectrum had a significant effect on response rate ( $p = 0.006$  and  $p < 0.001$ ). The effect of loudness was not statistically significant ( $p = 0.13$ ).

**Conclusions** The results suggest that male and female listeners do not differ significantly in their auditory saliency judgments based on the acoustic characteristics studied. This finding challenges the notion of inherent sex differences in bottom-up auditory attention and highlights the need for further research to explore other potential factors or conditions under which such differences might emerge.

**Keywords** Auditory attention, Bottom-up attention, Sex differences, Saliency judgments, Acoustic characteristics

\*Correspondence:  
Shinsuke Nagami  
shinsuke.nagami.0514@gmail.com  
Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

## Introduction

Attention is the function of allocating finite neural resources to appropriate stimuli in the external world, and auditory attention involves the ability to focus on specific sounds. Auditory attention can be divided into Top-down attention, which is consciously controlled, and Bottom-up attention, which is an automatic response to external stimuli [1, 2]. Top-down mechanisms: Attention is guided by our goals and expectations. For example, if you're listening for a specific name in a crowded room, you'll focus your attention on that sound [3, 4]. Bottom-up mechanisms: Attention is captured by salient stimuli in the environment. For example, the sound of a loud noise will grab your attention, even if you're not expecting it [5]. Top-down auditory attention helps us to focus on specific sounds, while bottom-up auditory attention helps us to stay aware of changes in the environment that may be important. Both types of attention are essential for our daily life.

As an academic background, attention research has progressed mainly in the visual field, but the exploration of auditory attention, especially bottom-up attention, has been a field of rapid interest in recent years [6]. Early attempts borrowed models from visual saliency, creating an "auditory saliency map" [7–9]. However, these models proved inadequate due to the unique characteristics of sound as a time-varying stimulus. By interpreting the time (T)-frequency (F) spectrogram as an auditory image, these models treated the T-F dimension as if it were a spatial X–Y axis, without fully considering time as a distinct dimension. Consequently, studies have explored various approaches, recognizing its dynamic nature: Experiments using auditory saliency stimuli while performing active tasks [10, 11], distractor paradigms introducing background noise [12, 13], auditory saliency judgment tasks where participants indicate salient moments in audio stimuli [5, 14, 15]. In particular, auditory saliency judgment tasks have been shown to account for bottom-up attentional effects, although they are not free from top-down attentional effects [16]. Research in this field emphasizes that loudness, pitch, duration, context, and timbre play a major role in the saliency of sounds that captured auditory attention [5–15]. For example, the sound of an explosion or a conversation between female is undeniably a prominent event. Moreover, context plays a crucial role in identifying subtle sounds as noteworthy events, such as the chirping of crickets in a quiet setting.

Previous research has largely overlooked sex differences when examining the relationship between the subjective saliency of sound and its acoustic characteristics. For instance, studies have demonstrated that infant cries elicit stronger reactions in females, whereas cries from

adult females tend to have a more pronounced effect on males [17]. Additionally, it has been observed that females exhibit greater responsiveness to aggressive vocalizations [18]. These findings suggest that auditory responses vary between sexes; however, the specific acoustic properties driving these differences remain poorly understood. This gap in our understanding provides the impetus for the current study, which seeks to rigorously explore how sex differences in auditory attention correlate with specific acoustic characteristics, thereby addressing a significant oversight in the field.

The purpose of this study was to gain psychoacoustic insight into sex differences in the dimensions of auditory saliency and their interactions. The approach taken here is that the listener listens to an Audio set and annotates any salient sound events. We defined "saliency" as easy to notice [19], and the sound events that yielded response rates above the median of the intersubject agreement were considered saliency sounds derived from bottom-up auditory attention [15]. In this exploratory study, we specifically aimed to examine the relationship between the acoustic characteristics of the salient sounds obtained and the response by sex. Based on this experimental model, we will test for the first time whether acoustic characteristics can explain the sex differences in bottom-up attention revealed in human behavioral experiments. The results of this study will be contrasted with psychoacoustic findings from previous behavioral experiments and will serve as a springboard for exploring sex differences in attentional function in auditory field.

## Materials and methods

### Participants

Between November 1, 2021, and February 28, 2022, 50 typical college students (25 males and 25 females) aged 20–23 years were recruited and provided written informed consent. Males ranged in age from 20 to 23 years (mean 21.4, SD 0.99), females ranged in age from 21 to 23 years (mean 21.48, SD 0.51). Inclusion criteria were those who did not normally use English and for whom the authors determined that the English of the experimental stimuli would not bias the measurement. Exclusion criteria included individuals with hearing deficits (> 20 dB hearing loss in 125–8000 Hz range), intellectual abnormalities (MMSE score < 24), attention deficits (> 1.5 SD below mean in CAT tasks), or those who had difficulty understanding the experimental procedures. The participants' hearing ability was assessed using standard pure tone audiometry to ensure their hearing ability was within the normal range. To assess subjects' intellectual functioning, we used the Mini-Mental State Examination [20]. Visual Cancellation and Auditory Detection testing tasks

from the Clinical Assessment for Attention [21] were used to assess visual and auditory attentional functioning. This study was approved by the Ethics Committee of the Kawasaki University of Medical Welfare, Okayama, Japan (approval no.: 21-012) and was conducted in accordance with the Declaration of Helsinki.

### Materials

We presented audio set to the participants, who then responded to the sound that captured their attention by simply pressing a button on their hands. The audio set was generated digitally using a personal computer (FMVU93C3BZ, FUJITSU, Tokyo, Japan), transformed via Bluetooth connection, and presented through a loudspeaker (HT-X8500, SONY, Tokyo, Japan). A computer program (PsychoPy version 21.2.3) controlled all participants' responses, instructions, and audio set. During the experiment, the participants sat in front of the loudspeakers at a distance of 90 cm. They were provided with a pencil and questionnaire form. The evaluator waited behind them and responded only when requested. A video camera (HC-W590M, Panasonic, Tokyo, Japan) was also placed behind the participants to record the experiment.

### Audio set

The internationally standardized Urban Sound Dataset (USD) [22] was used as the audio set. The USD consists of various speech sounds, including the sounds of daily life and nature. In this study, the audio set were created using the following procedure. First, 1-min sounds, which are everyday noises such as conversations and street sounds, were used as the Background Token (BT) of audio clip. Next, 10 salient sounds based on USD [22], such as alarm and drill sounds, were randomly placed in the BT as Foreground Token (FT). These FTs were classified based on the USD Sound Class Classification Method (<https://urbansounddataset.weebly.com/taxonomy.html>) as "Nature," "Human," "Mechanical," or "Music" based on the USD Sound Classification Method. One FT was defined as the time from the start of one sound to the end of FT. For example, the sound of "dog bark" was extracted from the start of a dog bark until the end of the bark. Accordingly, 10 FTs were randomly placed within 1 min of BT. In total, we created five audio clip using Wondershare Filmora9, version 9.5.2.11. BT sounds consist of everyday environmental sounds. FT sounds include everyday sounds except when labeled "human," which represent atypical sounds (e.g., English expressions of surprise or delight). Table 1 shows these five audio clips.

**Table 1** List of audio clips

Audio clip	Token	Taxonomy	
		Sound classes	Sound events
Clip 1	Background		Night ambience_Backyard
		Foreground	
		Nature	Dog bark (1)
		Mechanical	Car horn (1)
		Human	Footsteps (1)
		Human	Dog bark (2)
		Mechanical	Car riding (1)
		Mechanical	Mechanical (1)
		Nature	Dog howl
		Nature	Footsteps (2)
	Mechanical	Mechanical (2)	
	Nature	Animal tweet	
Clip 2	Background		Saturday ambience
		Foreground	
		Human	Child voice
		Human	Child shouting (1)
		Human	Child shouting (2)
		Human	Child singing
		Human	Child shouting (3)
		Human	Child speech
		Human	Child shouting (4)
		Human	Child shouting (5)
	Human	Child shouting (6)	
	Human	Child shouting (7)	
Clip 3	Background		Street party
		Foreground	
		Mechanical	Alarm (1)
		Mechanical	Mechanical (3)
		Mechanical	Mechanical noise (1)
		Mechanical	Car horn (2)
		Mechanical	Alarm (2)
		Mechanical	Alarm (3)
		Mechanical	Police siren
		Music	Music live
	Mechanical	Train horn	
	Nature	Bird tweet	
Clip 4	Background		Helicopter over neighborhood
		Foreground	
		Mechanical	Mechanical noise (2)
		Mechanical	Car brakes
		Mechanical	Drilling (1)
		Mechanical	Car horn (3)
		Mechanical	Car riding (2)
		Mechanical	Bus pneumatics
		Mechanical	Drilling (2)
		Mechanical	Mechanical (4)
	Mechanical	Mechanical (5)	
	Mechanical	Drilling (3)	

**Table 1** (continued)

Audio clip	Token	Taxonomy	
		Sound classes	Sound events
Clip 5	Background		High-technology generator
	Foreground	Mechanical	Alarm (4)
		Mechanical	Air conditioner
		Mechanical	Jackhammer (1)
		Mechanical	Siren (1)
		Mechanical	Jackhammer (2)
		Mechanical	Siren (2)
		Mechanical	Mechanical noise (3)
		Mechanical	Mechanical noise (4)
		Mechanical	Jackhammer (3)
Mechanical	Mechanical (6)		

Token = Indicate whether each sound event is a background sound or a foreground sound placed in a background sound; Taxonomy = Sound class and name of each sound event based on the classification of USD [22]

### Procedure

The audio set consisted of five different audio clips, and each stimulus was presented once. These audio clips were played randomly, with a 1-min interval. During the 1-min interval between two audio clips, the participants answered a questionnaire about the previous audio clip. They completed the practice task after receiving the instruction; after, they understood how to answer the questionnaire before starting the main task. The experiment was conducted by the first and second authors.

To avoid the effect of top-down auditory attention resulting from an instruction, the assigned authors provided the following instructions to the participants [14]: “You will listen to an audio clip containing various sounds. While listening, please press the button immediately if any sound captures your attention without you intentionally focusing on it.”

### Behavioral paradigm

In this study, participants’ reactions were recorded using the PsychoPy software, which captured all sounds that prompted a response. Participants also identified their reaction sites by watching the video recording after the experiment to reconfirm their reaction sites. The study’s approach to categorizing responses was guided by an auditory saliency model referenced in the literature [14]. According to this model, responses were divided based on the level of inter-participant agreement. Sounds that were deemed more salient tended to draw a uniform

response from many participants, indicating a predominance of bottom-up auditory attention—this refers to the instinctive reaction to sound stimuli. Conversely, less salient sounds captured the attention of fewer participants, suggesting a greater influence of top-down auditory attention, which involves more cognitive processing of sounds. Thus, we categorized a response with broad intersubject agreement as dominated by bottom-up auditory attention. Specifically, “salient sounds to which bottom-up attention contributes” were defined as those that received a median level of agreement or higher among participants [15].

### Acoustical setting

The experiment was performed in a soundproof room with a background noise level of  $\leq 48$  dB, and the minimum volume of audio set was set to a signal-to-noise ratio of +10 dB or higher [23]. All participants confirmed that the default volume was set at a comfortable listening level during the practice task. All sound effects on the loudspeaker audio output were turned off. The laboratory’s background noise level and maximum loudness of the audio clip were measured using a sound level meter (NL-27 K, RION, Tokyo, Japan) with a C-weighted maximum sound pressure level ( $L_{max}$ ). The maximum sound pressure level ( $L_{max}$ ) during a sound event is defined as the maximum loudness of that sound event. Then, all audio clips were then processed using a Fast Fourier Transform with a Hanning window to extract frequencies from the maximum spectrum. The extraction of the maximum spectrum of each sound event and the sound waveform and spectrogram of each Audio clip were displayed in MATLAB version 23.2.0.2409890 (R2023b). In the sound waveform, the amplitude of each audio clip was displayed along the time axis. In the spectrogram, the normalized frequency of each audio clip was displayed along the time axis.

### Verification of participant engagement with audio clip

After listening to each audio clip, participants completed a questionnaire assessing the accuracy of the content. Audio clips for which the questionnaire responses were incorrect were excluded from further analysis. For example, after listening to an audio clip containing human speech, the participants were asked the following question: “What sound did this audio clip contain?” They were then encouraged to select one of the following five sounds: “insect sound,” “ringtones,” “human speech,” “thunder,” and “wave sound.” The response “human speech” was the correct answer, and data regarding other choices (false answers) were excluded. In addition, to accurately identify the sound event to which the participants responded, we identified the response points by

checking the experimental video with the participants after the experiment. This approach helped ensure that only data from participants who accurately engaged with the audio clips were included in the analysis, thereby improving the reliability of the results [24].

**Sample size calculation**

In the linear multiple regression analysis, the sample size was calculated a priori using G\*Power, version 3.1.9.2. The effect size was 0.26, the power (1-β) was 0.8, and the significance level was set at 0.05, based on previous studies using this research model [15]. We estimated that 47 participants were required to detect the effects of duration, maximum loudness, and maximum spectrum on the response rate of the audio clip. The effect size was based on previous research [15], which reported moderate effect sizes for associations between acoustic features and salience judgments.

**Statistical analysis**

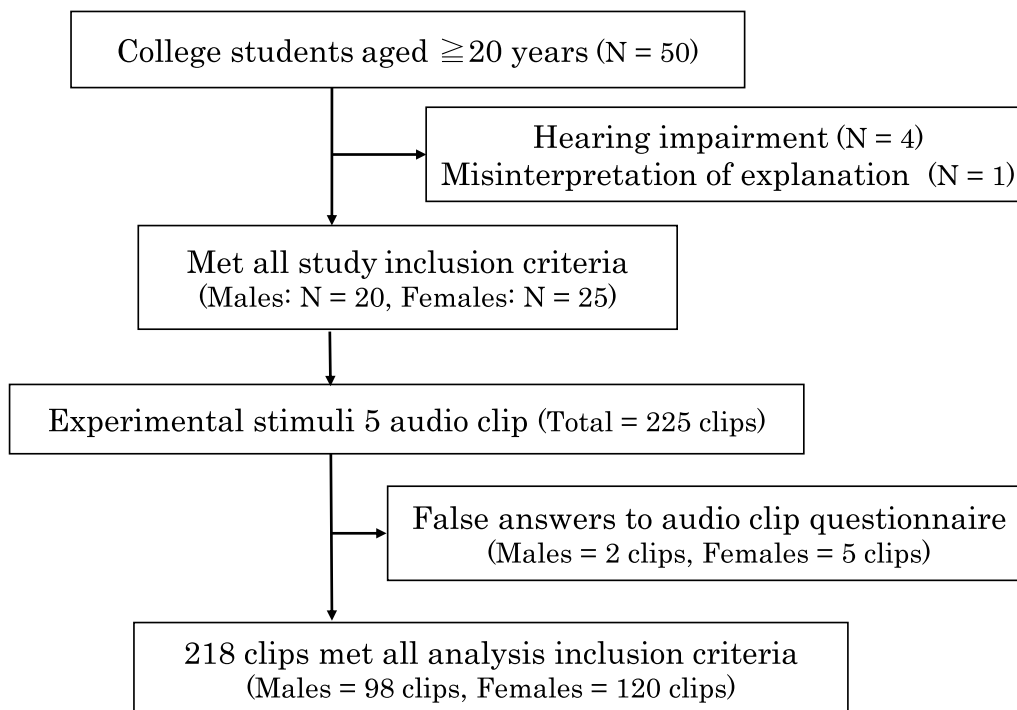
We conducted independent sample t-tests to compare demographic characteristics between sexes. We conducted a generalized linear mixed model (GLMM) analysis to examine the relationship between sex and acoustic characteristics. The GLMM was chosen because of the binary (0 or 1) nature of the response data, the unbalanced data structure of repeated measures, and the need

to account for variate effects (individual differences and differences in sound events). The model treated sex as a between-subjects factor and acoustic characteristics (duration, loudness, and spectrum) as a within-subjects factor. All acoustic characteristics were Z-standardized. We included sex and acoustic characteristics as fixed effects and participants and sound events as variable effects. The statistical analyses were carried out with EZR (Version 1.54). A p-value of <0.05 was considered statistically significant.

**Results**

**Participant characteristics**

Fifty typical adults (25 males and 25 females) were recruited for the study. According to predefined exclusion criteria, five males were excluded: four due to below-average hearing and one due to a misinterpretation of the instructions. The incorrect interpretation of the instructions specifically involved participants responding to a sound that captured their attention and continuing to press the button even after the sound had stopped, although they were only supposed to press it once. In this study, the analysis included data from 45 individuals (20 males and 25 females), after excluding 5 individuals who met the exclusion criteria (see Fig. 1). Statistical analysis was used to compare demographic characteristics between sexes. No significant



**Fig. 1** Participant enrollment and audio set

differences were found between male and female groups in age ( $r=0.06$ ,  $p=0.7$ ), MMSE scores ( $r=0.1$ ,  $p=0.489$ ), hearing thresholds (right ear:  $r=0.12$ ,  $p=0.43$ ; left ear:  $r=0.05$ ,  $p=0.728$ ), or attention function tests (all  $p>0.05$ ). Table 2 shows the demographic characteristics of the subjects.

Participant enrollment and audio clips flowchart with the inclusion and exclusion criteria used for establishing the audio clip datasets and the participant datasets; n, number of participants in the dataset.

### Participant responses to audio clips

In this experiment, each of the 45 participants was presented with 5 audio clips, amounting to a total of 225 audio clips. Out of these, 7 clips were excluded from the analysis due to incorrect questionnaire responses, leaving 218 clips (98 from male and 120 from female) for further analysis (see Fig. 1). The sound waveforms

and spectrograms for each audio clip are displayed in Fig. 2.

### Analysis of participant responses to audio stimuli

Participants' responses to audio clips were plotted on the time scale of sound waveforms as shown in Fig. 2. A PsychoPy analysis of the responses to the 5 audio clips (a total of 218 clips) identified 77 sound events. The participants' response rates for these events ranged from 4.0 to 100%. The median response rate was calculated to be 33.0%. In this study, salient sounds—those to which bottom-up attention primarily contributes—were defined as those with response rates of 33.0% or greater. The analysis revealed 47 salient sounds. These are listed by sex in Table 3.

To evaluate the participants' ability to distinguish between salient and non-salient sounds, we calculated key metrics including the hit rate, false alarm rate, and d-prime. In this study, the hit rate represents the proportion of responses to sound events with a response rate

**Table 2** Participant characteristics

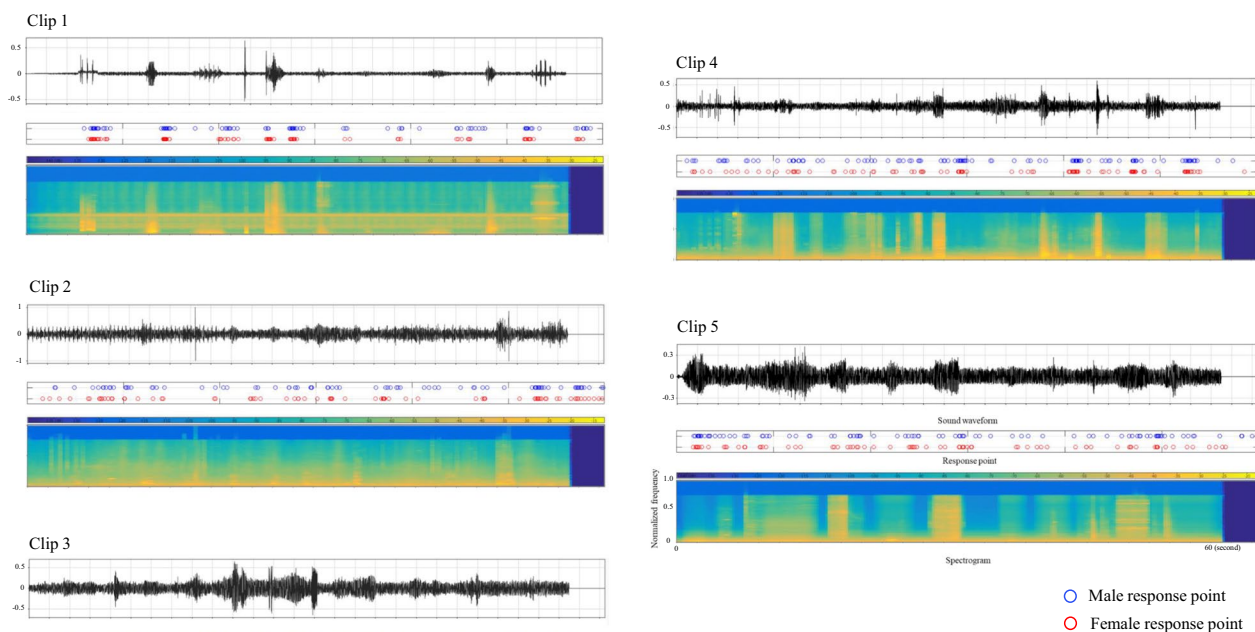
Variable	Mean (SD)	SE	CV	95% CI	r	p value
Age <sup>1)</sup>						
Male	21.4 (0.99)	0.22	0.04	20.93–21.86	0.06	0.7
Female	21.48 (0.5)	0.10	0.02	21.26–21.69		
MMSE <sup>1)</sup>						
Male	29.70 (0.92)	0.20	0.03	29.26–30.13	0.1	0.489
Female	29.84 (0.55)	0.11	0.01	29.61–30.06		
Right ear (dBHL) <sup>2)</sup>						
Male	5.62 (4.12)	0.92	0.73	3.69–7.55	0.12	0.43
Female	4.75 (3.08)	0.61	0.64	3.47–6.02		
Left ear (dBHL) <sup>2)</sup>						
Male	5.25 (3.23)	0.72	0.61	3.73–6.76	0.05	0.728
Female	4.92 (3.07)	0.61	0.62	3.65–6.18		
Visual cancellation task correct answer (%) <sup>1)</sup>						
Male	99.73 (0.83)	0.18	0.008	99.34–100.11	0.01	0.927
Female	99.78 (0.53)	0.10	0.005	99.56–100		
Accuracy (%)						
Male	100 (0)	0	0	100–100	–	–
Female	100 (0)	0	0	100–100		
Auditory detection task correct answer (%) <sup>1)</sup>						
Male	99.6 (0.82)	0.18	0.008	99.21–99.98	–	1
Female	99.56 (1.00)	0.20	0.01	99.15–99.97		
Accuracy (%) <sup>1)</sup>						
Male	99.7 (0.97)	0.21	0.009	99.24–100.15	0.24	0.109
Female	100 (0)	0	0	100–100		

Statistical analysis; r = Effect size; Statistical analysis

SD Standard Deviation, SE Standard error, CV Coefficient of Variation, CI Credible Interval, dBHL Decibels Hearing Level

<sup>1)</sup> Mann–Whitney U-test

<sup>2)</sup> t-test



**Fig. 2** Acoustical feature waveforms of all audio clips in the order of number

of 33% or higher, while the false alarm rate indicates the proportion of responses to sound events with a response rate of 33% or lower. The d-prime value, derived from the z-scores of the hit rate and false alarm rate, provides a measure of the participants’ discriminability between these sound categories. As shown in Table 4, the analysis yielded a hit rate of 0.61 and a false alarm rate of 0.38. The corresponding d-prime value was calculated to be 0.56, indicating moderate discriminability. These findings suggest that while participants were generally able to distinguish between salient and non-salient sounds, there remains some degree of overlap in their responses.

From the top row, the sound waveforms, response points by sex, and spectrogram are displayed for each audio clip. At the bottom of each sound waveform, the reaction points of the subjects are shown on a time scale. Males are indicated by blue plots and females by red plots.

**Sex differences analysis of response rates to acoustic features**

This analysis examined the relationship between response rates and acoustic features across all 77 sound events. The GLMM results showed that the main effect of sex was not statistically significant ( $\beta = -0.04$ ,  $SE = 0.06$ ,  $r = -0.11$ ,  $p = 0.458$ ), suggesting no significant difference in overall response rates between sexes. Regarding

acoustic characteristics, duration had a significant positive effect on the response ( $\beta = 0.06$ ,  $SE = 0.02$ ,  $r = 0.28$ ,  $p = 0.006$ ). Spectrum also showed a significant negative effect ( $\beta = -0.08$ ,  $SE = 0.02$ ,  $r = -0.34$ ,  $p < 0.001$ ). However, loudness did not have a significant effect ( $\beta = -0.03$ ,  $SE = 0.02$ ,  $r = -0.16$ ,  $p = 0.13$ ).

Furthermore, we examined the interactions between sex and each acoustic characteristic. The interaction between sex and duration was not statistically significant ( $\beta = -0.02$ ,  $SE = 0.01$ ,  $r = -0.02$ ,  $p = 0.103$ ). However, the interaction between sex and loudness was significant ( $\beta = 0.04$ ,  $SE = 0.01$ ,  $r = 0.05$ ,  $p = 0.001$ ), indicating that loudness may influence response rates differently depending on sex. The interaction between sex and spectrum was not significant ( $\beta = 0.009$ ,  $SE = 0.01$ ,  $r = 0.01$ ,  $p = 0.489$ ). The findings from these analyses are summarized in Table 5.

**Comparative analysis of acoustic features in Bottom-up vs. Top-down auditory saliency**

The analysis revealed notable differences between bottom-up and top-down auditory attention across various acoustic characteristics. For sound duration, bottom-up sounds had a mean duration of 2.00 s ( $SD = 1.33$ ), while top-down sounds averaged 1.45 s ( $SD = 1.09$ ). Although this difference approached statistical significance ( $r = 0.19$ ,  $p = 0.091$ ), it was not statistically significant.

**Table 3** List of salient sounds by sex

Audio clip	Token	Taxonomy		Duration [Second]	Loudness [dBSPL]	Spectrum [Hz]	Response rate [%]	
		Sound classes	Sound events				Male	Female
Clip 1	Foreground	Nature	Dog bark (1)	2.03	64.9	1370	83	75
		Mechanical	Car horn (1)	1.12	69.1	1000	89	67
		Human	Footsteps (1)	4.9	64.9	73	83	58
		Human	Dog bark (2)	0.14	71.2	511	39	54
		Mechanical	Car riding (1)	2.08	68.3	64	78	66
		Nature	Footsteps (2)	3.53	62.1	263	61	42
		Mechanical	Mechanical (2)	2.09	64.2	1603	67	50
		Nature	Animal tweet	3	60.7	5282	56	38
Clip 2	Background	Noise	Car noise	1.59	69.3	9105	–	45
	Foreground	Human	Child shouting (1)	0.57	64.7	2962	39	39
		Human	Child shouting (2)	1.51	67.5	1233	33	30
		Human	Child shouting (3)	1.09	70	1053	33	26
		Human	Child shouting (4)	2.13	68.2	1090	–	39
	Background	Noise	Car noise	3.48	77.9	9591	37	–
	Foreground	Human	Child shouting (5)	1.39	69.4	1284	33	57
		Human	Child shouting (6)	2.43	63.6	1169	17	35
Human		Child shouting (7)	2.01	70.1	1071	89	78	
Clip 3	Background	Noise	Car noise	4.45	67.9	8421	84	64
	Foreground	Mechanical	Alarm (1)	1.02	61.8	883	63	36
		Mechanical	Mechanical (3)	1.1	63.7	121	42	40
		Mechanical	Mechanical noise (1)	1.09	61.5	309	37	–
	Background	Noise	Car noise	2	71.5	2177	35	–
	Foreground	Mechanical	Car horn (2)	2.51	69.6	999	90	84
		Mechanical	Alarm (2)	1.54	69.2	610	58	52
		Mechanical	Alarm (3)	0.58	68.7	3064	60	56
Mechanical		Police siren	5.02	71.4	1140	74	64	
Music		Music live	4.5	66.7	393	42	48	
Clip 4	Foreground	Mechanical	Train horn	1.5	69.1	438	69	48
		Mechanical	Drilling (1)	2.45	61.5	30	69	48
	Background	Mechanical	Car riding (2)	3.02	60.3	130	42	36
		Noise	Noise	0.18	67.4	6492	40	–
	Foreground	Mechanical	Drilling (2)	1.21	66.1	122	90	68
		Noise	Noise	4.5	68.3	810	–	44
	Foreground	Mechanical	Mechanical (4)	2.02	63	1617	100	76
		Mechanical	Mechanical (5)	1.41	65.3	1797	79	80
Noise		Dog bark	0.47	65.5	5700	35	–	
Clip 5	Background	Mechanical	Drilling (3)	2.46	66.5	130	79	72
	Background	Noise	Noise	0.37	69.4	2898	35	30
	Foreground	Mechanical	Alarm (4)	0.11	58.8	151	47	–
		Mechanical	Air conditioner	1.04	60.9	57	53	39
		Mechanical	Jackhammer (1)	3	61.7	100	58	48
		Mechanical	Siren (1)	1.16	61.4	1169	52	48
		Mechanical	Jackhammer (2)	3.59	60.8	97	68	57
		Mechanical	Siren (2)	2.03	59.3	487	42	–
Mechanical		Mechanical noise (3)	0.55	63.5	483	47	35	
Mechanical	Mechanical noise (4)	0.45	56.2	487	58	–		
	Mechanical	Jackhammer (3)	3.59	58.1	5097	79	52	

*Duration* Duration of each sound in seconds, *Loudness* Maximum loudness of sound, *Spectrum* Maximum spectrum of sound, *Response rate* Percentage agreement of responses to sound events by sex, *dBSPL* Sound pressure level, *Hz* Frequency



**Table 4** Metrics for participants' sound event discrimination ability

Metric	Value
Hit	47
-Hit rate	0.61
- z-score (Hit rate)	0.28
False alarm	30
-False alarm rate	0.38
- z-score (False alarm rate)	- 0.28
d-prime	0.56

Hit = Number of responses to sound events with a response rate of 33% or above; False alarm = Number of responses to sound events with a response rate below 33%; H = Hit rate; F = False alarm rate

**Table 5** GLMM analysis results

Fixed effects	Estimate (β)	SE	r	p-value
(intercept)	0.38	0.05	0.68	< 0.001
Sex	- 0.04	0.06	- 0.11	0.458
Duration (Z)	0.06	0.02	0.28	0.006
Loudness (Z)	- 0.03	0.02	- 0.16	0.13
Spectrum (Z)	- 0.08	0.02	- 0.34	< 0.001
Sex*Duration (Z)	- 0.02	0.01	- 0.02	0.103
Sex*Loudness (Z)	0.04	0.01	0.05	0.001
Sex* Spectrum (Z)	0.009	0.01	0.01	0.489

ZZ standardization change number

Regarding loudness, the mean loudness for bottom-up sounds was 65.55 dB (SD=4.36), compared to 66.71 dB (SD=5.29) for top-down sounds, with this difference not reaching statistical significance ( $r=0.11$ ,  $p=0.316$ ). The spectral analysis indicated a significant difference between the two attention types; bottom-up sounds had a lower mean spectral frequency (1811.34 Hz, SD= 2457.25) compared to top-down sounds (4164.5 Hz, SD= 4231.33), with this difference being statistically

significant ( $r=0.32$ ,  $p=0.005$ ). These results suggest that spectral characteristics, in particular, play a crucial role in differentiating between bottom-up and top-down auditory attention. The findings from these analyses are summarized in Table 6.

**Sex differences analysis of response rates to various sound classes**

We conducted a Welch's t-test to analyze response rates for the different sound classes. For Human, 43.77% of males and 46.22% of females ( $r=0.16$ ,  $p=0.504$ ), for Nature 66.66% of males and 51.66% of females ( $r=0.43$ ,  $p=0.296$ ), and for Noise 40.37% of males and 36.25% of females ( $r=0.19$ ,  $p=0.459$ ), with no statistically significant differences among them. On the other hand, for Mechanical, 65% of males and 51.23% of females ( $r=0.39$ ,  $p=0.005$ ) found statistically significant differences. For Music, the analysis was inadequate. Overall, significant sex differences were found primarily in responses to Mechanical, with no significant differences found for the other sound classes. The findings from these analyses are summarized in Table 7.

**Discussion**

The purpose of this exploratory study was to investigate potential sex differences in bottom-up auditory attention. Our findings suggest that acoustic characteristics, particularly duration and spectrum, have a significant effect on the response rate, although no significant differences by sex were found.

Previous studies have extensively explored auditory saliency without adequately addressed sex differences, primarily focusing on characteristics of sound events (such as loudness, frequency, timbre, and duration) and elements of the overall soundscape (such as context, timing, and probability theory) [5–15]. In this study, we show that the effect of acoustic characteristics is more important than sex differences in determining auditory

**Table 6** Acoustic feature analysis: duration, loudness, and spectrum differences in auditory attention systems

Variable	Mean (SD)	SE	CV	95% CI	r	p value
Duration <sup>1)</sup>						
Bottom up	2.00 (1.33)	0.19	0.66	1.6–2.39	0.19	0.091
Top down	1.45 (1.09)	0.19	0.75	1.04–1.86		
Loudness <sup>2)</sup>						
Bottom up	65.55 (4.36)	0.63	0.06	64.27–66.83	0.11	0.316
Top down	66.71 (5.29)	0.96	0.07	64.74–68.69		
Spectrum <sup>1)</sup>						
Bottom up	1811.34 (2457.25)	358.42	1.35	1089.86–2532.81	0.32	0.005
Top down	4164.5 (4231.33)	772.53	1.01	2584.49–5744.5		

Statistical analysis; 1) Mann–Whitney U-test, 2) t-test

**Table 7** Sound classes response rate statistical analysis

Variable	Mean (SD)	SE	CV	95% CI	r	p value
Human <sup>1)</sup>						
Male	43.77 (24.85)	8.28	0.56	24.66–62.88	0.16	0.504
Female	46.22 (16.68)	5.56	0.36	33.39–59.04		
Mechanical <sup>2)</sup>						
Male	65 (17.25)	3.38	0.26	58.03–71.96	0.39	0.005
Female	51.23 (17.01)	3.33	0.33	44.35–58.1		
Music						
Male	42	–	–	–	–	–
Female	48	–	–	–		
Nature <sup>2)</sup>						
Male	66.66 (14.36)	8.29	0.21	30.98–102.34	0.43	0.296
Female	51.66 (20.3)	11.72	0.39	1.22–102.1		
Noise <sup>1)</sup>						
Male	40.37 (18.15)	6.41	0.44	25.19–55.55	0.19	0.459
Female	36.25 (14.25)	5.03	0.39	24.33–48.16		

Data for Music class is based on a single observation, so statistics are not robust. Statistical analysis; 1) Mann–Whitney U-test, 2) t-test

saliency. The negative spectral effects suggest that high-frequency sounds are associated with lower response rates. While it is generally known that the human auditory system is most sensitive to the mid-frequency range, this study did not specifically address this aspect. Previous research has shown that low-frequency sounds can cause discomfort and stress in humans, inhibiting cognitive activity [25, 26]. Although there is no direct evidence to date that low-frequency sounds specifically attract human attention, this study may have highlighted the potential for low-frequency sounds to draw attention due to their association with discomfort and other negative effects. The positive effect of duration suggests that longer sounds are perceived more prominently. This supports previous research findings that sound duration plays an important role in capturing auditory attention [1, 6, 27]. Interestingly, the effect of loudness was not significant. This is in contrast to previous studies [6, 19, 27] that found loudness to be a major determinant of auditory saliency. This discrepancy may be due to our experimental design or stimulus characteristics and requires further study. Because previous study differs from previous studies in that it focuses on the interaction of bottom-up and top-down attention in natural soundscapes. Previous simple studies have only examined competition for attentional resources in more constrained situations. This study, however, elucidates the mechanisms of attentional control in a more complex auditory environment. Similar to the present study, studies using natural soundscapes have shown that the relationship to background sounds plays a role in attracting bottom-up attention, especially because background sounds are powerful

attention attractors [1]. In other words, the relationship between auditory saliency and acoustic features that has been demonstrated in simple auditory environments does not hold true for experiments on attentional control in complex auditory environments such as the present study. While our results provide initial evidence for these sex-based differences, they are preliminary and should be further explored through more extensive studies to confirm these trends and fully understand their implications.

In the paired t-test conducted earlier, the response rates to mechanical sounds differed significantly between males and females. This result was based on the specific data set analyzed and may have been influenced by particular sound classes and characteristics. The higher response rate of males to mechanical sounds might indicate greater sensitivity and interest in these stimuli, potentially related to evolutionary, cultural, and social factors. However, the GLMM results showed no significant overall sex differences, indicating that while differences may occur in specific cases (as in our subset), these differences are not generalizable to all sound types and conditions. This underscores the importance of considering context and specific acoustic characteristics when assessing response rates.

Our assessment of auditory saliency was grounded in behavioral measures derived from participants' subjective evaluations. Some studies have measured auditory saliency by pupil dilation responses [19, 28], but this method does not necessarily correlate with auditory saliency. Moreover, manual annotation method may not adequately capture the active, continuous scene scanning required to measure auditory saliency effectively [15]. In

our study, auditory salience was evaluated using participants' manual annotations of sound events in response to 1-min audio clips. Although manual annotation is commonly used [29, 30], it can be influenced by top-down attention. This limitation was addressed by adjusting for consistency in responses across multiple participants [5, 14, 15]. Our study refines previous approaches by better distinguishing between bottom-up and top-down auditory attention based on participants' responses. The experimental paradigm minimizes the influence of top-down processes, allowing us to explore the cognitive and subjective aspects of auditory attention through active tasks. However, incorporating passive paradigms could provide valuable insights into automatic bottom-up attention mechanisms. Future research should combine active and passive paradigms to compare subjective experiences with automatic responses, offering a more comprehensive understanding of auditory salience. For example, using passive measures like mismatch negativity could help capture unconscious responses to auditory changes, further clarifying the interplay between bottom-up and top-down attention mechanisms. These differences highlight the need for further research in the auditory domain to fully understand these processes.

This study does have several limitations. Firstly, the acoustic analysis of the audio clips was confined to duration, maximum loudness, and maximum spectrum. Therefore, the results are applicable only to these specific acoustic features and their impact on auditory salience judgments concerning sex differences. Secondly, the audio clips used might be biased and may not adequately represent other types of auditory stimuli, such as human voices or natural sounds, potentially skewing the results. Lastly, the use of the GLMM has allowed us to uncover complex relationships between sex and acoustic characteristics, but at the same time may complicate the interpretation of the results. In addition, it does not account for interactions between acoustic characteristics, and future models that include these interactions should be considered. Future studies would benefit from testing the generalizability of our findings by including a wider range of age groups and subjects from different cultural backgrounds. Exploring the neural mechanisms involved in processing acoustic characteristics using techniques such as functional brain imaging would further deepen our understanding.

## Conclusions

Our results suggest that while there may be nuanced differences in response rates between sexes in specific contexts or datasets, the overarching trend does not support a substantial sex difference. Instead, acoustic characteristics like duration and spectrum play a more critical role

in shaping responses. These findings may contribute to a better understanding of the mechanisms of auditory attention and to the development of sound environment design and auditory interfaces.

## Acknowledgements

This research received no specific grants from any funding agency in the public, commercial, or not-for-profit sectors.

## Author contributions

NO contributed to the protocol design and writing of the manuscript. YS contributed to the data collection and review of the manuscript. NK contributed to the acoustic analysis of the data and review of the manuscript. KY, KN and AY contributed to the review of the manuscript. SN contributed to the data analysis and writing of the manuscript.

## Funding

Not applicable.

## Availability of data and materials

The main data are available in the main text or the Supplementary Information. All other data are available from the authors upon reasonable request. <https://datadryad.org/stash/share/fl-ZCUc1tieLcmXvEltGIIEmjKX7WD6hsPKNZppGgEs>.

## Declarations

### Ethics approval and consent to participate

This study was approved by the Ethics Committee of the Kawasaki University of Medical Welfare, Okayama, Japan (approval no.: 21–012) and was conducted in accordance with the Declaration of Helsinki.

### Consent to publication

Informed consent was obtained from all participants, and consent for publication was secured through a signed written form.

### Competing interests

The authors declare no competing interests.

### Author details

<sup>1</sup>Department of Speech and Hearing Sciences, Faculty of Rehabilitation, Kawasaki University of Medical Welfare, Kurashiki, Okayama, Japan. <sup>2</sup>Department of Rehabilitation, Kurashiki Central Hospital, Kurashiki, Okayama, Japan. <sup>3</sup>Graduate School of Health and Welfare Sciences, Okayama Prefectural University, Soja, Okayama, Japan. <sup>4</sup>Graduate School of Comprehensive Scientific Research, Prefectural University of Hiroshima, Shobara, Hiroshima, Japan. <sup>5</sup>Department of Neurosurgery, TAKAMATSU Red Cross Hospital, Takamatsu, Kagawa, Japan. <sup>6</sup>Department of Communication Disorders, School of Rehabilitation Sciences, Health Sciences University of Hokkaido, 1757, Ishikari-gun, Kanazawa, Tobetsu-cho, Hokkaido 061-0293, Japan.

Received: 13 May 2024 Accepted: 16 October 2024

Published online: 24 October 2024

## References

- Huang N, Elhilali M. Push-pull competition between bottom-up and top-down auditory attention to natural soundscapes. *eLife*. 2020;9:e52984.
- Awh E, Belopolsky AV, Theeuwes J. Top-down versus bottom-up attentional control: a failed theoretical dichotomy. *Trends Cogn Sci*. 2012;16:437–43.
- Cherry EC. Some experiments on the recognition of speech, with one and with two ears. *J Acoust Soc Am*. 1953;25:975–9.
- Baluch F, Itti L. Mechanisms of top-down attention. *Trends Neurosci*. 2011;34:210–24.
- Kothinti SR, Huang N, Elhilali M. Auditory salience using natural scenes: an online study. *J Acoust Soc Am*. 2021;150:2952.

6. Kaya EM, Elhilali M. Modelling auditory attention. *Philos Trans R Soc Lond B Biol Sci*. 2017. <https://doi.org/10.1098/rstb.2016.0101>.
7. Kayser C, Petkov CI, Lippert M, Logothetis NK. Mechanisms for allocating auditory attention: an auditory saliency map. *Curr Biol*. 2005;15:1943–7.
8. Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans Pattern Anal Mach Intell*. 1998;20:1254–9.
9. Chi T, Ru P, Shamma SA. Multiresolution spectrotemporal analysis of complex sounds. *J Acoust Soc Am*. 2005;118:887–906.
10. Southwell R, Baumann A, Gal C, Barascud N, Friston KJ, Chait M. Is predictability salient? A study of attentional capture by auditory patterns. *Philos Trans R Soc Lond B Biol Sci*. 2017. <https://doi.org/10.1098/rstb.2016.0105>.
11. Kaya EM, Huang N, Elhilali M. Pitch, timbre and intensity interdependently modulate neural responses to salient sounds. *Neuroscience*. 2020;440:1–14.
12. Petsas T, Harrison J, Kashino M, Furukawa S, Chait M. The effect of distraction on change detection in crowded acoustic scenes. *Hear Res*. 2016;341:179–89.
13. Vachon F, Labonté K, Marsh JE. Attentional capture by deviant sounds: a noncontingent form of auditory distraction? *J Exp Psychol Learn Mem Cogn*. 2017;43:622–34.
14. Kim K, Lin K, Walther DB, Hasegawa-Johnson MA, Huang TS. Automatic detection of auditory salience with optimized linear filters derived from human annotation. *Pattern Recognit Lett*. 2014;38:78–85.
15. Huang N, Elhilali M. Auditory salience using natural soundscapes. *J Acoust Soc Am*. 2017;141:2163.
16. Borji A, Itti L. State-of-the-art in visual attention modeling. *IEEE Trans Pattern Anal Mach Intell*. 2013;35:185–207.
17. Rigo P, De Pisapia N, Bornstein MH, Putnick DL, Serra M, Esposito G, et al. Brain processes in women and men in response to emotive sounds. *Soc Neurosci*. 2017;12:150–62.
18. Burra N, Kerzel D, Munoz D, Grandjean D, Ceravolo L. Early spatial attention deployment toward and away from aggressive voices. *Soc Cogn Affect Neurosci*. 2019;4(14):73–80.
19. Liao HI, Kidani S, Yoneya M, Kashino M, Furukawa S. Correspondences among pupillary dilation response, subjective salience of sounds, and loudness. *Psychon Bull Rev*. 2016;23:412–25.
20. Folstein MF, Folstein SE, McHugh PR. 'Mini-mental state'. A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*. 1975;12:189–98.
21. Kato M. The development and standardization of clinical assessment for attention (CAT) and clinical assessment for spontaneity (CAS). *Higher Brain Funct Res*. 2006;26:310.
22. Salamon J, Jacoby C, Bello JP. A dataset and taxonomy for urban sound research. <https://urbansounddataset.weebly.com/urbansound.html>. 2014.
23. Šrámková H, Granqvist S, Herbst CT, Švec JG. The softest sound levels of the human voice in normal subjects. *J Acoust Soc Am*. 2015;137:407–18.
24. Gilman TL, Shaheen R, Nylocks KM, Halachoff D, Chapman J, Flynn JJ, et al. A film set for the elicitation of emotion in research: a comprehensive catalog derived from four decades of investigation. *Behav Res Methods*. 2017;49:2061–82.
25. Araújo AJ, Neto PF, Torres SL, Remoaldo P. Low-frequency noise and its main effects on human health—a review of the literature between 2016 and 2019. *Appl Sci*. 2020;10:5205.
26. Javadi A, Pourabdian S, Forouharmajid F. The effect of low frequency noises exposure on the precision of human at the mathematical tasks. *Int J Prev Med*. 2022;23:13–33.
27. Kochanski G, Grabe E, Coleman J, Rosner B. Loudness predicts prominence: fundamental frequency lends little. *J Acoust Soc Am*. 2005;118:1038–54.
28. Wang CA, Boehnke SE, Itti L, Munoz DP. Transient pupil response is modulated by contrast-based saliency. *J Neurosci*. 2014;34:408–17.
29. Russell BC, Torralba A, Murphy KP, Freeman WT. Labelme: a database and web-based tool for image annotation. *Int J Comput Vis*. 2008;77:157–73.
30. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. *IEEE Conf Comput Vis Pattern Recognit*. 2009. <https://doi.org/10.1109/CVPR.2009.5206848>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.