JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

Rapid ocular responses are modulated by bottom-up driven auditory salience

https://doi.org/10.1523/JNEUROSCI.0776-19.2019

Cite as: J. Neurosci 2019; 10.1523/JNEUROSCI.0776-19.2019

Received: 5 April 2019
Revised: 28 June 2019
Accepted: 12 July 2019

Rapid ocular responses are modulated by bottom-up driven auditory salience

Sijia Zhao[1], Nga Wai Yum[1], Lucas Benjamin[1], Elia Benhamou[2], Makoto Yoneya[3], Shigeto Furukawa[3], Fred Dick[4,5], Malcolm Slaney[6], Maria Chait[1]

[1]Ear Institute, University College London, London WC1X 8EE, UK

[2] Dementia Research Centre, Department of Neurodegenerative Disease, University College London, London WC1N 3AR, UK

[3] NTT Communication Science Laboratories, NTT Corporation, Atsugi, 243-0198 Japan

[4]Department of Psychological Sciences, Birkbeck College, London, WC1E 7HX

[5] Department of Experimental Psychology, University College London, WC1H 0DS

[6]Machine Hearing Research, Google, Mountain View, CA 94043

Short title: Microsaccades modulated by auditory salience

Corresponding Authors:
Maria Chait
m.chait@ucl.ac.uk
Ear Institute, University College London
332 Gray's Inn Road, London WC1X 8EE, UK

Sijia Zhao
sijia.zhao.10@ucl.ac.uk
Ear Institute, University College London
332 Gray's Inn Road, London WC1X 8EE, UK

Number of figures: 7
Number of extended data figures: 2
Abstract length: 194 words
Introduction length: 613 words
Discussion length: 1484 words

Acknowledgements: This work was supported by an EC Horizon 2020 grant and a BBSRC international partnering award to MC.

Keywords: Microsaccades; Attention; Pupil Dilation; Superior Colliculus; Crowd-computing; acoustic roughness.

Abstract

Despite the prevalent use of alerting sounds in alarms and human-machine interface systems and the long-hypothesized role of the auditory system as the brain's 'early warning system', we have only a rudimentary understanding of what determines auditory salience—the automatic attraction of attention by sound—and which brain mechanisms underlie this process. A major roadblock has been the lack of a robust, objective means of quantifying sound-driven attentional capture. Here we demonstrate that: (1) a reliable salience scale can be obtained from crowd-sourcing (N=911), (2) acoustic roughness appears to be a driving feature behind this scaling, consistent with previous reports which implicate roughness in the perceptual distinctiveness of sounds, and (3) crowdsourced auditory salience correlates with objective autonomic measures. Specifically, we show that a salience ranking obtained from online raters correlated robustly with the superior colliculus (SC)-mediated ocular freezing response - microsaccadic inhibition (MSI) - measured in naïve, passively listening human participants (of either sex). More salient sounds evoked earlier MSI, consistent with a faster orienting response. These results are in line with the hypothesis that MSI reflects a general re-orienting response which is evoked by potentially behaviorally important events irrespective of their modality.

Significance statement: Microsaccades are small, rapid, fixational eye movements, measurable with sensitive eye-tracking equipment. We reveal a novel, robust link between microsaccade dynamics and the subjective salience of brief sounds (salience rankings obtained from a large number of participants in an online experiment): Within 300 ms of sound onset, the eyes of naïve, passively listening participants demonstrate different microsaccade patterns as a function of the sound's crowdsourced salience. These results position the superior colliculus (hypothesized to underlie microsaccade generation) as an important brain area to investigate in the context of a putative multi-modal salience-hub. They also demonstrate an objective means for quantifying auditory salience.

Introduction

Our perception of our surroundings is governed by a process of competition for limited resources. This involves an interplay between task-focused and bottom-up-driven processes which automatically bias perception towards certain aspects of the world, to which our brain, through experience or evolution, has been primed to assign particular significance. Understanding the neural processes which underlie such involuntary attentional capture is a topic of intense investigation in systems neuroscience (Itti and Koch, 2001, 2000; Kaya and Elhilali, 2014; Kayser et al., 2005).

Research in vision has capitalized on the fact that attentional allocation can be 'objectively' decoded from ocular dynamics: Observers free-viewing complex visual scenes tend to demonstrate consistent fixation, saccade and microsaccade patterns that can be used to infer the attributes that attract bottom-up visual attention (Hafed et al., 2009; Krauzlis et al., 2018; Parkhurst et al., 2002; Peters et al., 2005; Veale et al., 2017; Yuval-Greenberg et al., 2014). The underlying network for (micro-)saccade generation is centered on the superior colliculus (SC; Hafed et al., 2009) with a contribution from the frontal eye fields (Peel et al., 2016), consistent with a well-established role for these regions in computing the visual salience map and controlling overt attention (Veale et al., 2017; White et al., 2017b,a).

The appearance of new events is also associated with two types of rapid orienting responses: **(1)** an 'ocular freezing' ('microsaccadic inhibition'; MSI) response—a rapid transient decrease in the incidence of microsaccades, hypothesized to arise through suppression of ongoing activity in the SC by new sensory inputs (Engbert and Kliegl, 2003; Hafed and Clark, 2002; Hafed and Ignashchenkova, 2013; Rolfs et al., 2008); and **(2)** A phasic pupil dilation response (PDR; Wang et al., 2017; Wang and Munoz, 2015). The PDR has been linked to potentially SC-mediated (Wang and Munoz, 2015) spiking activity in the Locus Coeruleus (Joshi et al, 2016) which constitutes the source of Norepinephrine

92 (Noradrenaline) to the central nervous system, and therefore controls global vigilance and
93 arousal.

94 Both MSI and PDR have been shown to systematically vary with visual salience
95 (Bonneh et al., 2014; Rolfs et al., 2008; Wang et al., 2017, 2014; Wang and Munoz, 2014)
96 and are theorized to reflect the operation of an interrupt process that halts ongoing activities
97 so as to accelerate an attentional shift towards a potentially survival-critical event.

98 Interestingly, sounds can also drive these ocular responses. Abrupt, or otherwise out-
99 of-context auditory events evoke pupil dilation, and cause MSI (Liao et al. 2014; 2016,
100 Wetzel et al. 2016; Wang et al 2014; Wang and Munoz, 2014, 2015; Rolfs et al 2005,.2008)
101 in line with a proposal that these responses reflect the operation of a modality-general
102 orienting mechanism. In fact, sounds cause faster responses than visual stimuli (Rolfs et al.,
103 2008; Wang et al., 2014), consistent with the 'early warning system' role of hearing (Murphy
104 et al, 2013). However, because only very simple stimuli have been used, the degree to which
105 sound-evoked ocular responses reflect acoustic properties beyond loudness (Huang and
106 Elhilali, 2017; Liao et al., 2016) remains unknown.

107 Here we sought to determine whether MSI and PDR are sensitive to auditory
108 salience. We used crowdsourcing to obtain a 'subjective' (e.g., ratings-based) salience
109 ranking of a set of brief environmental sounds. The obtained salience scale was also verified
110 with a small "in-lab" replication. Then in a lab setting, a group of naïve participants passively
111 listened to these sounds while their ocular dynamics and pupil dilation were recorded. We
112 demonstrate that MSI (but not PDR) is systematically modulated by auditory salience. This is
113 consistent with the hypothesis that MSI indexes a rapid, multi-modal orienting mechanism
114 which is sensitive to not just the onset, but also the specific perceptual distinctiveness of brief
115 sounds.

Methods

Stimuli

Eighteen environmental sounds drawn from Cummings et al., (2006) (a subset from Dick and Bussiere (2002) and Saygin et al., 2005), were used. All stimuli were 500ms long and RMS equated (see individual spectrograms in Figure 1A and Figure 1-1 in the Extended Data for sound files).

FIGURE 1 ABOUT HERE

Subjective salience via Crowdsourcing

Experimental Design and Statistical Analysis

The experiment was conducted on the Amazon Mechanical Turk (MTurk) platform - an online marketplace that allows remote participants ('workers' per MTurk nomenclature) to perform tasks via a web-interface. Raters were selected from a large pool of prequalified 'workers' in the US who were judged to be reliable over the course of many previous (non-salience related) experiments. Each session is delivered through a 'human intelligence task' (HIT) page, which contains instructions and the stimuli for that session (see Figure 1-2 in the Extended Data section for an example of a HIT page used in the present experiment). Since we were interested in the extent to which a relatively free listening environment can result in meaningful data, we did not impose constraints on sound delivery (e.g. Woods et al., 2017) or level (though it was suggested that the participants listen over headphones).

Participants made relative salience judgments on sounds presented in pairs. In total, a pool of 7038 pairs (153 possible pairs x 2 orders to control for order effects x 23 repetitions) were generated. The pairs were then arranged in 207 different HITs by randomly selecting subsets of 34 different pairs from the above pool. Each HIT also included six randomly interspersed 'catch trials', where the two sounds were identical. Each HIT, therefore,

contained 40 sound pairs along with task instructions. Each pair had its own ‘Play’ button, which when pressed started the presentation of the corresponding sound pair, with a 500 ms silent gap between the two sounds. Participants were asked which sound was ‘more salient or noticeable. Which sound would you think is more distracting or catches your attention?’ Participants could only listen to each pair once before responding by selecting one of the ‘first’, ‘second’ or ‘identical’ buttons to progress to the next sound pair. Participants were instructed to choose the ‘identical’ button only if the sounds were physically identical (catch trials). Failure to respond appropriately to the catch trials (or choosing the ‘identical’ response for the non-catch trials) indicated lack of engagement with the experiment and resulted in the data from that session being excluded from analysis (~10% exclusion rate, see below). Participants were offered financial compensation roughly equal to the minimum US wage and prorated for the 5-minute experiment time. To encourage participant engagement, we paid a small bonus when participants correctly responded to identical sounds and subtracted a small amount for each miss. Each HIT was run by 5 unique workers for an overall number of 1035 sessions. The time limit for task completion was set to 60 minutes, though we expected the experiment to last an average of 3 minutes. Figure 2A plots the actual duration distribution. Most sessions were completed within 3 minutes.

Each participant was free to complete up to a maximum of 9 different HITs. A distribution of HITs per worker is in Figure 2B. Most (71.4%) completed one HIT only, whilst 52 workers (12.4%) completed the maximum number of HITs. We did not find any relationship between participants' number of HITs and performance on catch trials. From the total of 1035 sessions completed, 57 included a single missed catch trial, 11 included two missed catch trials and 11 included more than two missed catch trials. Fifty-eight sessions contained false positives. Overall 124 sessions were excluded. The remaining sessions were comprised of 384 unique workers, of which 270 completed only one HIT and the rest completed multiple HITs.

Salience ranking was computed by pooling across all HITs and counting the proportion of pairs on which each sound was judged as more salient. Variability was estimated using bootstrap re-sampling (1000 iterations), where on each iteration, one session for each of the 207 unique HITs was randomly selected for the ranking analysis. The error bars in Figure 1B are one standard deviation from this analysis. The same ranking was obtained after removing sessions with durations exceeding the 90th percentile (14.09 min, N=820 remaining) or the 75th percentile (5.98 min, N=683 remaining).

FIGURE 2 ABOUT HERE

Acoustic analysis

The salience data were analyzed to examine possible correlations with several key acoustic features previously implicated in perceptual salience.

An **overall loudness** measure was produced by a model in which the acoustic signal was filtered with a bank of bandpass filters of width 1 ERB (Moore and Glasberg, 1983) and centre frequencies spaced 1/2 ERB from 30 Hz to 16 kHz. The instantaneous power of each filter output was smoothed with a 20 ms window and elevated to power 0.3 to approximate specific loudness (Hartmann, 1996). Outputs were then averaged across channels to produce a single value. This model was preceded by a combination of high-pass and low-pass filters to approximate effects of outer and middle ear filtering (Killion, 1978).

Several key **measures of salience derived from the model of Kayser et al. (2005)** were examined. Paralleling work in the visual modality (Itti & Koch, 2001), this model produces an auditory saliency map in the form of a frequency x time representation, indicating the spectro-temporal loci that are hypothesized to be particularly perceptually salient. The representation is computed by independently extracting several key auditory features (loudness, spectral and temporal contrast), which are normalized to create a feature-independent scale and then linearly combined together to create the overall map. For the present analysis, we extracted several parameters from the salience map computed for

each sound-token: the maximum value within the saliency map (this is the parameter used in the experiments reported in Kayser et al, 2005), the mean saliency value across the entire map, and max/mean gradient across the frequency and/or time dimensions.

Roughness was calculated from the modulation power spectrum, computed using the approach described in Elliott and Theunissen (2009; see also Arnal et al., 2015). Roughness is associated with energy in the high end (>30Hz) of the amplitude modulation spectrum, though it also depends on modulation depth and other spectral factors (Pressnitzer and McAdams, 1999). As is typical of natural wide-band sounds, in our sound set we found a strong correlation (Spearman r=0.808, p<0.0001) between power at high modulation rates (30-100 Hz) and those below 30 Hz (0-30 Hz). We also noted that salience (MTurk derived; see results) significantly correlated with power at modulations between 30-100 Hz (Spearman, r=0.585 p=0.01) but not with the low frequency (0-30 Hz) modulations (Spearman r=0.222 p=0.376). Controlling for low frequency modulations as a covariate (partial correlation), yielded a substantially stronger correlation (Spearman r=0.707 p=0.002), suggesting that the high modulation rates, independently of overall modulation power, contributed to salience. Therefore, to specifically isolate the contribution of high modulation rates, and control for overall power across the modulation spectrum, 'roughness' was quantified as the ratio between power at modulations between 30-100 Hz and power between 0-100 Hz (i.e. across the full range). See also Figure 3, for a comparison of how roughness in the present set relates to the range of roughness (calculated in the same way) obtained from a diverse set of environmental sounds.

FIGURE 3 ABOUT HERE

Subjective salience via in-lab replication

Participants

To verify that the online salience ranking data also hold when tested in a more controlled environment, eighteen paid participants (15 females, average age 23.8, range 18–

31) took part in an in-lab replication study; all reported normal hearing and no history of neurological disorders. Experimental procedures were approved by the research ethics committee of University College London and written informed consent was obtained from each participant.

Experimental Design and Statistical Analysis

We again used a pairwise task with identical presentation parameters as in the MTurk experiment, but that every participant was presented with the full set of all possible pairs (78 pairs x 2 possible orders x 2 repetitions) for a total of 312 pairs of sounds. These were presented in a random order in six consecutive blocks (~8 min). Each block also contained eight randomly interspersed catch trials of identical sounds. Participants were allowed a short rest between blocks.

The stimuli were delivered to the participants' ears by Sennheiser HD558 headphones (Sennheiser, Germany) via a UA-33 sound card (Roland Corporation) at a comfortable listening level, self-adjusted by each participant. Stimulus presentation and response recording were controlled with the Psychtoolbox package (Psychophysics Toolbox Version 3; Brainard, 1997) with MATLAB (The MathWorks, Inc.). Participants were tested in a darkened and acoustically shielded room (IAC triple-walled sound-attenuating booth). The session lasted for 1 hour, starting with the same instructions given in the MTurk experiment. Participants were instructed to fixate their gaze on a white cross at the centre of the computer screen while listening to the stimuli, and to respond by pressing one of 3 keyboard buttons to indicate 'sound A more salient'/ 'sound B more salient'/ 'identical sounds'. The participant's response initiated the following trial with a random inter-trial interval of 1.5 to 2 s. Blocks featuring incorrect responses – whether a miss or a false alarm – to any of the eight catch trials indicated lack of engagement and the whole block was discarded from the analysis. In this instance, all participants performed perfectly with a 100% hit rate and 0% false alarm rate resulting in no exclusions.

Eye tracking

Participants

This experiment was performed by a total of 30 paid participants (28 females; aged 18-29, average 23.33), with 15 participants initially (14 females; aged 21~28, average 23.53), subsequently supplemented by an additional group of 15 participants so as to have a better measure of variability across the population. No participants were excluded from this experiment. All reported normal hearing and no history of neurological disorders. All participants were naïve to the aims of the experiment and none had participated in the 'in-lab' ranking experiment described above. Experimental procedures were approved by the research ethics committee of University College London and written informed consent was obtained from each participant.

Experimental Design and Statistical Analysis

Sixteen sounds out of the original set were used in this experiment. Sound #3 and sound #9 (see Figure 1) were excluded due to experiment length constraints.

The effective onset (the time point to which the eye tracking analysis is time locked) was adjusted for each sound-token to a time where level exceeded a fixed threshold. The threshold was defined as the 20^{th} percentile of the distribution of power (per time sample; over the initial 50 ms) pooled across sound-tokens. Further controls for onset energy, based on correlating loudness at onset with the various eye tracking measures, are described in the Results section.

Participants listened passively to the sounds, which were presented in random order, with a randomized inter-trial interval between 6 and 7 seconds. In total, 320 trials (16 sound-tokens x 20 repetitions of each) were presented. Stimuli were diotically delivered to participants' ears using Sennheiser HD558 headphones (Sennheiser, Germany) via a Creative Sound Blaster X-Fi sound card (Creative Technology, Ltd.) at a comfortable listening level, self-adjusted by each participant. Stimulus presentation and response recording were controlled with the Psychtoolbox package (Psychophysics Toolbox Version 3;

Brainard, 1997) on MATLAB (The MathWorks, Inc. R2018a). Participants sat with their head fixed on a chinrest in front of a monitor (viewing distance 65cm) in a dimly lit and acoustically shielded room (IAC triple-walled sound-attenuating booth). They were instructed to continuously fixate at a black cross presented at the centre of the screen against a grey background and to passively listen to the sounds (no task was performed). A 24-inch monitor (BENQ XL2420T) with 1920x1080 pixel resolution and 60 Hz refresh rate presented the fixation cross and feedback. The visual display remained the same throughout. To avoid pupillary light reflex effects, display and ambient room luminance were kept constant throughout the experiment. To reduce fatigue, the experiment was divided into nine 4-minute blocks, each separated by a 4-minute rest period.

Pupil measurement

An infrared eye-tracking camera (Eyelink 1000 Desktop Mount, SR Research Ltd.) positioned just below the monitor continuously tracked gaze position and recorded pupil diameter, focusing binocularly with a sampling rate of 1000 Hz. The standard five-point calibration procedure for the Eyelink system was conducted prior to each experimental block. Participants were instructed to blink naturally. They were also encouraged to rest their eyes briefly during inter-trial intervals. Prior to each trial, the eye-tracker automatically checked that the participants' eyes were open and fixated appropriately; trials would not start unless this was confirmed.

Analysis of eye blinks

Eye blinks are commonly observed as an involuntary response to abrupt sounds, part of the brainstem-mediated startle reflex (Blumenthal and Goode, 1991; Davis, 1984; Knudson and Melcher, 2016). The elicitation of blinks has been shown to be sensitive to a range of stimulus manipulations (Blumenthal, 2015, 1988) and it was, therefore, important to relate eyeblink incidence to the measures of salience used here.

Since the blink reflex occurs rapidly after stimulus presentation (Blumenthal, 1988), we analyzed data from the first 500 ms after sound onset. For each subject and sound token,

298 eyeblink incidence was computed by tallying the number of trials that contained a blink
299 (defined as full or partial eye closure). Though the incidence of blinks was low overall
300 (<10%), it varied substantially across participants. For the correlation analyses reported
301 below (Figure 4), the average rate, across participants, was computed for each sound
302 condition.

303 **Analysis of pupil diameter data**

304 To measure the sound-evoked pupil dilation responses (Figure 5), the pupil diameter
305 data of each trial were epoched from 0.5s before to 3s after sound onset. Intervals where the
306 eye tracker detected full or partial eye closure (manifested as loss of the pupil signal) were
307 automatically treated as missing data and recovered with shape-preserving piecewise cubic
308 interpolation; epochs with more than 50% missing data were excluded from analysis. On
309 average, less than two trials per participant were rejected.

310 To compare results across blocks, conditions, and participants, the epoched data
311 within each block were z-score normalized. A baseline correction was then applied by
312 subtracting the median pupil size over the pre-onset period; subsequently, data were
313 smoothed with a 150 ms Hanning window.

314 **Microsaccade analysis**

315 Microsaccade detection was based on the algorithm proposed by Engbert and Kliegl
316 (2003). In short, microsaccades were extracted from the continuous horizontal eye-
317 movement data based on the following criteria: (a) a velocity threshold of $\lambda = 6$ times the
318 median-based standard deviation within each block; (b) above-threshold velocity lasting for
319 longer than 5ms but less than 100ms; (c) the events are binocular (detected in both eyes)
320 with onset disparity less than 10ms; and (d) the interval between successive microsaccades
321 is longer than 50 ms.

322 Extracted micro-saccade events were represented as unit pulses (Dirac delta). Two
323 complementary analysis approaches were employed. The first involved tallying MS events,

collapsed across subjects and trials (for more details, see the Results section). The second approach entailed analysing MS rate time series: For each sound-token, in each participant, the event time series were summed and normalized by the number of trials and the sampling rate. Then, a causal smoothing kernel $\omega(\tau) = \alpha^2 \times \tau \times e^{-\alpha\tau}$ was applied with a decay parameter of $\alpha = \frac{1}{50}$ ms (Dayan and Abbott, 2001; Rolfs et al., 2008; Widmann et al., 2014), paralleling a similar technique for computing neural firing rates from neuronal spike trains (Dayan and Abbott, 2001; see also Joshi et al., 2016; Rolfs et al., 2008). The obtained time series was then baseline corrected over the pre-onset interval. Due to the low baseline incidence of microsaccades per participant (roughly 2 events per second) and the small number of presentations per sound token (n=20; required to prevent perceptual adaptation) a within-subject analysis was not possible. Mean microsaccade rate time series, obtained by averaging across participants for each sound token, are used for the analyses reported here. Robustness is verified using bootstrap resampling (see results).

Correlation analysis

To control for outlier effects, all reported bivariate and partial correlations were performed using the conservative Spearman's rank correlation method (two-tailed). The one exception to this was the direct comparison of MSI- and PDR- based correlations, where we computed Pearson correlations between crowdsourced salience and the various eye tracking measures discussed in the results (MSI latency, PDR peak amplitude, PDR derivative peak amplitude; see defined below). Differences in Pearson correlation coefficients were tested using the procedures for testing statistical differences between correlations using the implementation in the R package cocor (Diedenhofen and Musch, 2015).

Results

(1) Crowdsourced salience ranking yielded a meaningful and stable salience scale.

Eighteen environmental sounds (see Figure 1A for spectrograms and Figure 1-1 in the Extended data section for sound files) originally used by Cummings et al. (2006), were selected for this study. The stimulus set represents the variety of sounds that may be encountered in an urban acoustic environment, including animal sounds, human non-speech sounds (kiss, sneeze), musical instrumental sounds, impact sounds (golf ball, tennis, impact, coins), and an assortment of mechanical sounds (car alarm, alarm clock, camera shutter, pneumatic drill, lawn mower etc.). All stimuli were 500 ms long and RMS equated.

We obtained salience rankings of this sound set from 911 online participants via the Amazon Mechanical Turk (MTurk) platform (see Methods). Sounds were presented in pairs, and participants were required to report which one was 'more salient or attention-grabbing'.

Over all responses, a small but robust order-of-presentation bias for selecting the second sound in a pair was observed ($t = -9.240$, $p<0.001$; mean probability to choose the 1^{st} sound = 0.47, mean probability to choose the 2^{nd} sound = 0.53). However, because the order of presentation was counterbalanced across pairs, this bias did not affect the rating results.

To derive a relative measure of salience for each sound, we counted the proportion of pairs (collapsed across all data from all participants) on which each sound was judged as more salient, producing a measure of relative salience ranging between 0 and 1 (Figure 1B). It is striking that a clear scale of subjective salience can be captured across these 18 brief, arbitrarily selected, sounds. Variability was estimated using bootstrap re-sampling (1000 iterations), where on each iteration, salience was computed over a subset of the data (see Methods). The error bars in Figure 1B are one standard deviation from this analysis.

(2) The crowdsourced salience scale is strongly correlated with in-lab salience judgements

An in-lab replication was conducted in order to validate the salience scale obtained from the MTurk experiment. The paradigm was essentially identical, but the experiments were performed in a lab setting, and under a controlled listening environment. The main differences were that (1) to reduce test time, the sound set was reduced to 13 sounds (selected to capture the salience range of the full set and indicated in orange bars, Figure 1B); (2) all participants listened to all sound pairs during an hour-long session.

Because the 'in-lab' experiment was designed to mirror the MTurk experiment, the planned analysis involved collapsing across trials and subjects in the same way as described above. The in-lab ranking showed a strong correlation with the online ranking ($r = 0.857$, $p<0.0001$; Figure 1C).

The small number of trials per sound-pair (n=4), which was necessary to reduce perceptual adaptation (and to fit into time constraints), produces rather noisy subject-level data. Regardless, we attempted to assess the association between MTurk and in-lab rating on an individual level using a repeated measures correlation analysis (rmcorr package in R; Bakdash and Marusich, 2017). The results confirmed that subject level correlation with the MTurk data was significant (Pearson $r=0.428$, $p<0.0001$; Spearman $r=0.455$, $p<0.0001$), though, as expected, accounting for noise at the individual subject level resulted in a decreased explained variance. Overall, the individual level analysis supports the conclusion that the group level data can be taken as representative of single subjects.

(3) Crowdsourced salience correlates with acoustic roughness.

Though the present set of sounds is too small to systematically pinpoint the sound features that contribute to auditory salience, we sought to understand whether the obtained 'subjective' salience scale correlates with several key acoustic features, previously hypothesized as contributing to salience:

We found that, despite the fact that **loudness** is known to be a prominent contributor to perceptual salience (Huang and Elhilali, 2017; Kayser et al., 2005; Liao et al., 2016), the crowdsourced salience scale in the present set did not significantly correlate with loudness ($r=0.428$, $p=0.078$; see Methods for details about the loudness measure). This may be partly because the level of these sounds was RMS-normalised thus removing some of the larger differences in loudness between sounds.

Next, we tested the relationship between crowdsourced salience and **measures of salience derived from the model of Kayser et al. (2005)**. Several relevant parameters were examined (see Methods). Only correlations with the gradient along the frequency dimension were significant (Spearman's $r=0.525$ $p=0.027$ for the maximum gradient and $r=0.488$, $p=0.049$ for the mean gradient; for the rest of the comparisons $p \geqslant 0.155$), indicating that perceptual salience may be associated with salience maps in which salient regions are sparsely spread across the spectrum.

Motivated by previous work (Arnal et al., 2015; Huang and Elhilali, 2017; Sato et al., 2007), we also investigated the correlation between perceptual salience and **roughness** –a perceptual quality that is associated with energy in the high end (>30Hz) of the amplitude modulation spectrum (e.g. Arnal et al., 2015; see methods)**.** Here, the correlation between crowdsourced salience and roughness yielded a significant effect ($r=0.709$, $p=0.001$; Figure 1D), consistent with accumulating evidence that roughness is a major contributor to salience.

(4) Crowdsourced salience correlates with objective measures from ocular dynamics

Next, we asked whether acoustic salience automatically (i.e. without a remit from a task) modulates ocular orienting responses. A subset of 16 out of the 18 original sounds (two sounds, 3 and 9, were excluded due to experimental time constraints) were presented to naïve, centrally-fixating subjects who listened passively to the sounds, without performing any task, while their gaze position and pupil diameter were continuously tracked. Sounds

were presented in random order, and with a random inter-sound interval between 6 and 7 seconds. Overall, each sound was presented 20 times across the experimental session. This small number of repetitions was chosen so as to minimize potential effects of perceptual adaptation to the stimuli. The analysis is therefore based on group-level correlations. Re-sampling based analyses were conducted to derive an estimate of the distribution of correlation strengths in the population.

We analyzed two types of rapid orienting responses: the 'ocular freezing' (MSI) response (Engbert and Kliegl, 2003; Hafed and Clark, 2002; Hafed and Ignashchenkova, 2013; Rolfs et al., 2008) and the pupil dilation response (PDR; Wang et al., 2017; Wang and Munoz, 2015). We also analyzed the incidence of eye blinks and their possible relationship to perceptual salience. Eye blinks are a component of the brainstem-mediated startle reflex (Davis, 1984), hypothesized to reflect an automatic defensive response to abrupt or threatening stimuli. The startle eye blink response is commonly elicited by loud, rapidly rising sounds (Blumenthal and Goode, 1991; Knudson and Melcher, 2016) but has been shown to be sensitive to a range of stimulus manipulations (Blumenthal, 2015, 1988).

The incidence of eye blinks was not correlated with crowdsourced salience

The incidence of eye blinks was low overall (<10%) and did not significantly correlate with any of the measures reported here.

FIGURE 4 ABOUT HERE

Measures of pupillary dilation were not correlated with crowdsourced salience

The temporal evolution of the normalised pupil diameter (the pupil dilation response, PDR) is presented in Figure 5. The pupil starts to dilate around 0.5 s after sound onset, and peaks at approximately 1.12 seconds (ranging from 1.02 to 1.33 s). We did not observe significant correlation between crowdsourced salience rating and any key parameters

associated with PDR dynamics (see Figure 5C for statistics), including the PDR peak amplitude and the peak of the PDR derivative (maximum rate of change of the PDR).

FIGURE 5 ABOUT HERE

Crowdsourced salience is correlated with microsaccadic inhibition (MSI)

The microsaccade results are shown in Figure 6. In line with previous demonstrations (e.g. Rolfs et al., 2008), we observed an abrupt inhibition of microsaccadic activity following sound presentation. The drop in microsaccade rate begins at ~0.3 s after onset and reaches a minimum at 0.45 s.

We conducted two different analyses in order to determine the extent to which MSI differs across sounds. The first approach is based on pooling MS data across trials and subjects and counting MS events. We defined a window spanning a 500 ms interval from 200-700 ms after sound onset. This window encompasses the interval before the beginning of MSI and after it has settled (See Figure 6A; the extent of the interval is also shown in Figure 6F). We then tallied the MS events for each sound token. This measure correlated with crowdsourced salience such that more salient sounds were associated with fewer MS events (= a lower **MS rate**) within the window ($r=-0.627$, $p=0.009$; Figure 6B). As can be seen in Figure 6B the number of MS events is small overall and the differences between conditions are narrow, reflecting the low incidence of micro-saccades. The robustness of the observed correlation was confirmed with bootstrap resampling (see Figure 6C; 5000 iterations; balanced) where on each iteration we selected 30 participants with replacement to compute the tally. This analysis (Figure 6C) confirmed a negatively skewed distribution of r values centred around -0.5 (median $r=-0.458$), with 98.72% of r values smaller than 0 ($p=0.013$) and a left skewed distribution of associated p values (Figure 6D). This effect was maintained for windows spanning up to one second from sound onset.

The MS rate demonstrated a significant correlation with roughness (r=-0.607 p=0.013) but not with loudness (r-0.353 p=0.18). The correlation between MS rate and crowdsourced salience was no longer significant when controlling for roughness as a covariate (partial correlation r(13)=-0.350 p=0.201), suggesting that dependence on roughness is a major contributor to the correlation between MSI and crowdsourced salience.

Next, we aimed to understand more precisely how the dynamics of microsaccadic inhibition vary with salience by quantifying the **MSI latency** for each sound. This was accomplished by computing a MS rate time series for each token (Figure 6E; see methods). MSI latency was then determined by computing a grand-mean microsaccade rate time series (averaged across sound tokens; see Figure 6E), identifying its mid-slope amplitude (horizontal dashed line in Figure 6F), and obtaining the time at which the microsaccade rate time series associated with each sound token intersected with this value. This latency, hereafter referred to as the '*microsaccadic inhibition latency*' (MSI latency), correlated with the crowdsourced salience rating (r=-0.627, p=0.009; Figure 6G), such that increasing salience was associated with earlier MSI.

The correlation between MSI latency and crowdsourced salience was significantly different from the PDR correlations with crowdsourced salience reported above (MSI vs PDR: z = 2.6369, p = 0.0042; MSI vs PDR derivative: z = 3.0640, p = 0.0011; see Methods). It was further confirmed that MSI latency did not significantly correlate with blink rates within the first 500 ms of sound onset (Figure 4C).

FIGURE 6 ABOUT HERE

Additional analyses to confirm effect robustness (Figure 7) used bootstrap resampling to estimate the stability of the correlation between MSI latency and crowdsourced salience across the subject pool. This involved computing a distribution of p and r values for sub-group sizes of 30 and 15 subjects (with replacement). We iteratively (5000 iterations) selected n samples (n=15 or 30) from the full pool of N=30. For each subset, we computed

the correlation between MSI latency and crowdsourced salience. The distribution of associated correlation coefficients demonstrated a moderate correlation (median r=-0.5063 for N=15, r=-0.4615 for N=30) between MSI latency and crowdsourced salience. The distributions of p values are significantly left-skewed (Fisher's Method; $p<0.0001$; further details in the figure)—indicative of a true effect.

FIGURE 7 ABOUT HERE

To determine what acoustic information might have driven the observed microsaccade effect, we correlated the MSI latency with the measures obtained from the Kayser et al. (2005) model (see Methods). This analysis revealed no significant correlations ($p \geqslant 0.203$ for all).

We also correlated MSI latency for each sound with roughness and loudness estimates computed between 0-300 ms (window sizes of 50, 100, 150, 200, 250 and 300 ms) - e.g., over the interval between sound onset and the average onset time of ocular inhibition. For loudness, none of the correlations reached significance ($p>0.152$), This suggests that though the sounds used had clearly differing distributions of power at onset, this did not contribute primarily to the correlation with micro-saccadic inhibition. The correlation between MSI latency and crowdsourced salience was maintained even when controlling for loudness at onset (0-50 ms window from onset; partial correlation; $r(13)=-0.666$ $p=0.007$; same holds for longer intervals).

In contrast to the lack of a stable link between loudness and MSI latency, a significant correlation with roughness was present from 250 ms onwards ($p \leqslant 0.028$, $r \geqslant -0.547$), confirming the previous observations of a strong link between roughness and MSI rate. The correlation between MSI latency and salience was no longer significant when controlling for

roughness as a covariate (partial correlation r(13)=-0.455 p=0.088), suggesting that dependence on roughness is a major contributor to the correlation between MSI and crowdsourced salience.

Discussion

The main aim of this work was to understand whether/how ocular orienting responses in naïve listeners are modulated by acoustic salience. We showed that a crowdsourced 'subjective' (i.e. rating based) salience ranking of brief, non-spatial, environmental sounds robustly correlated with the ocular freezing response measured in naïve, passively listening participants. Sounds ranked by a large group of online participants as more salient evoked earlier microsaccadic inhibition (MSI), consistent with a faster orienting response (Figure 6). These results establish that information about auditory salience is conveyed to the superior colliculus (SC), the primary generator of micro saccades, within ~300 ms after sound onset. That sounds systematically modulated microsaccade activity demonstrates that the mechanisms which drive microsaccadic inhibition are sensitive to a broad range of salient events beyond the visual modality.

Crowdsourced Salience

We demonstrated that a robust measure of perceptual salience can be obtained from a web-based mass-participant experimental platform. Online experimenting is gaining popularity within cognitive science (see Stewart et al., 2017 for a review), including in the auditory modality (Woods et al., 2017; Woods and McDermott, 2018). However, there are various potential drawbacks to this approach relating to lack of control over the participants' listening devices and environment. These may be especially severe for perceptual judgement experiments which demand a high level of engagement from participants. However, the limitations are offset by important unique advantages, including the opportunity

of obtaining a large amount of data in a short period of time, and running brief 'one-shot' experiments which are critical for avoiding perceptual adaptation. Furthermore, in the context of salience, the variability of the sound environment may in fact provide 'real world' validity to the obtained scale. Here we established that despite the various concerns outlined above, capitalizing on big numbers makes it possible to acquire a stable, informative, salience scale with relatively minimal control of the listeners and their environment. Indeed, the salience scale obtained online correlated robustly with in-lab ranking measures as well as with certain acoustic features previously established as contributing to perceptual salience.

Specifically, we found a strong correlation with 'roughness' - the perceptual attribute that is associated with 'raspy', 'buzzing' or 'harsh' sounds. This correlation arose 'organically' in the sense that the sounds in the present study were not selected to vary across this or other acoustic dimensions. The link between roughness and salience is in line with previous reports (e.g. Arnal et al., 2015; Huang and Elhilali, 2017; Sato et al., 2007) which established a clear role for this feature in determining the perceptual prominence of sounds. Most recently this was demonstrated in the context of the distinctiveness of screams (Arnal et al., 2015; though Arnal et al. used the term 'fearful' as opposed to 'salient' in their experiments).

Kayser et al. (2005) have proposed a model for auditory salience, inspired in its architecture by the well-established model for visual salience (Itti and Koch, 2001). We found limited correlation between the parameters derived from that model and the present crowdsourced scale. This is possibly because the Kayser model is better suited to capturing 'pop-out'-like saliency, associated with attentional capture by an object which stands out from its background. Instead here we focused on brief sounds reflecting single acoustic sources.

It is important to stress that the present sound set is too small for an extensive exploration of the features that might drive perceptual salience. Roughness likely stood out here because of the primacy of that feature and because our sounds spanned a large enough roughness range (Figure 3). The robustness of the crowdsourced judgements

suggests that a similar crowd-sourcing approach but with a larger, and perhaps more controlled, set of sounds may reveal other relevant sound features. In particular, recent advances in sound synthesis technologies make it possible to systematically and independently vary acoustic features towards a controlled investigation of acoustic salience.

Acoustic salience did not modulate pupil responses

The pupil dilation response (PDR) indexes activity within the LC-norepinephrine system (Aston-Jones and Cohen, 2005; Joshi et al., 2016), which is proposed to play a key role in controlling global vigilance and arousal (Sara, 2009; Sara and Bouret, 2012). In previous work that reported an association between the PDR and sound salience, the dominant driving feature for the correlation was loudness (Huang and Elhilali, 2017; Liao et al., 2016). In contrast, differences along this dimension were minimized in the present stimuli to allow us to focus on subtler, but potentially behaviourally important, contributors to perceptual salience. Our failure to observe a modulation of the PDR by salience suggests that, at least in the context of auditory inputs, pupil dilation may reflect a non-specific arousal response, evoked by stimuli which cross a certain salience threshold. This account is consistent with the relatively late timing of the PDR (peaking about 1 sec after sound onset) thereby potentially reflecting a later stage of processing than that captured by microsaccades (see below).

Microsaccadic inhibition (MSI) is a correlate of acoustic salience

We revealed a robust correlation between MSI latency and crowdsourced salience: Sounds judged by online raters as more salient were associated with a more rapid (Figure 6G) and extensive (as reflected by decreased incidence; Figure 6B) inhibition of microsaccadic activity. The effect arose early – from roughly 350 ms after sound onset, pointing to fast underlying circuitry. Correlation analyses indicated that the bulk of this effect is driven by a correlation with roughness, suggesting that this information is computed sufficiently early to affect the body's automatic re-orienting response.

606 The brain mechanisms which respond to acoustic roughness are poorly understood.
607 Response signatures have been observed in both auditory cortical and subcortical areas
608 (e.g., Schnupp et al., 2015), including the amygdala—a key brain centre for fear/risk
609 processing (Adolphs et al., 1995; Arnal et al., 2015; Bach et al., 2008; Ciocchi et al., 2010;
610 Nader et al., 2000) . Arnal et al (2015) reported that the amygdala, but not auditory cortex,
611 exhibited specific sensitivity to temporal modulations within the roughness range. This was
612 interpreted as suggesting that rough sounds activate neural systems associated with the
613 processing of danger. The present findings, demonstrating an association between
614 salience/roughness and rapid orienting responses, are consistent with this conclusion.

615 Microsaccades are increasingly understood to index an active attentional sampling
616 mechanism which is mediated by the SC (Hafed et al., 2015; Krauzlis et al., 2018; Rolfs,
617 2009; Rucci and Poletti, 2015; Wang and Munoz, 2015). Accumulating work suggests that
618 MS occurrence is not automatic but rather modulated by the general state of the participant
619 and by the availability of computational capacity, such that microsaccade incidence is
620 reduced under high load (Dalmaso et al., 2017; Gao et al., 2015; Widmann et al., 2014;
621 Yablonski et al., 2017). MSI is an extreme case for such an effect of attentional capture on
622 ocular dynamics, interpreted as reflecting an interruption of ongoing attentional sampling so
623 as to prioritize the processing of a potentially important sensory event. The dominant account
624 for MSI is that sensory input to the SC causes an interruption to ongoing activity by disturbing
625 the balance of inhibition and excitation (Rolfs et al., 2008). Previously reported effects of
626 visual salience on MSI (Bonneh et al., 2015; Rolfs et al., 2008; Wang et al., 2017) were
627 therefore interpreted as indicating that visual salience may be coded at the level of the SC
628 (see also Mizzi and Michael, 2014; Veale et al., 2017; White et al., 2017a, 2017b). We
629 showed that the perceptual salience of sounds also modulates this response, consistent with
630 a well-established role for the SC as a multi-sensory hub (Meredith et al., 1987; Meredith and
631 Stein, 1986; Wallace et al., 1998; Wang et al., 2017). Importantly, this effect was observed

during diotic presentation—sounds did not differ spatially and were perceived centrally, within the head.

The present results thus suggest that an investigation of SC responses to sound may provide important clues to understanding auditory salience. There is evidence for projections from the auditory cortex to the SC (Meredith and Clemo, 1989; Zingg et al., 2017) which might mediate the effects observed here, or they may arise via a subcortical pathway with the IC (e.g. Xiong et al., 2015), or the amygdala, as an intermediary.

Finally, the present experiments focused on the salience of brief sounds presented in silence. However, the ongoing context within which sounds are presented is known to play a critical role in determining their perceptual distinctiveness (Leech et al., 2007; Krishnan et al., 2012; Kaya and Elhilali, 2014; Sohoglu and Chait, 2016; Southwell and Chait, 2018). In the future, the paradigm established here can be easily expanded to more complex figure-ground situations or to tracking salience within realistic sound mixtures. A further question relates to understanding whether ocular dynamics reflect perceptual salience primarily linked to basic, evolutionary-driven sound features such as roughness or whether they can also be modulated by arbitrary sounds endowed with salience via association or reinforcement (e.g. ones' mobile ring tone).

References

Adolphs, R., Tranel, D., Damasio, H., Damasio, A.R., 1995. Fear and the human amygdala. J. Neurosci. 15, 5879–5891. https://doi.org/10.1523/JNEUROSCI.15-09-05879.1995

Arnal, L.H., Flinker, A., Kleinschmidt, A., Giraud, A.-L., Poeppel, D., 2015. Human Screams Occupy a Privileged Niche in the Communication Soundscape. Curr. Biol. 25, 2051–2056. https://doi.org/10.1016/j.cub.2015.06.043

Aston-Jones, G., Cohen, J.D., 2005. An Integrative Theory of Locus Coeruleus-Norepinephrine Function: Adaptive Gain and Optimal Performance. Annu. Rev. Neurosci. 28, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Bach, D.R., Schächinger, H., Neuhoff, J.G., Esposito, F., Salle, F.D., Lehmann, C., Herdener, M., Scheffler, K., Seifritz, E., 2008. Rising Sound Intensity: An Intrinsic Warning Cue Activating the Amygdala. Cereb. Cortex 18, 145–150. https://doi.org/10.1093/cercor/bhm040

Bakdash, J.Z., Marusich, L.R., 2017. Repeated Measures Correlation. Front. Psychol. 8. https://doi.org/10.3389/fpsyg.2017.00456

Blumenthal, T.D., 2015. Presidential Address 2014: The more-or-less interrupting effects of the startle response. Psychophysiology 52, 1417–1431. https://doi.org/10.1111/psyp.12506

Blumenthal, T.D., 1988. The Startle Response to Acoustic Stimuli Near Startle Threshold: Effects of Stimulus Rise and Fall Time, Duration, and Intensity. Psychophysiology 25, 607–611. https://doi.org/10.1111/j.1469-8986.1988.tb01897.x

Blumenthal, T.D., Goode, C.T., 1991. The Startle Eyeblink Response to Low Intensity Acoustic Stimuli. Psychophysiology 28, 296–306. https://doi.org/10.1111/j.1469-8986.1991.tb02198.x

Bonneh, Y., Fried, M., Arieli, A., Polat, U., 2014. Microsaccades and drift are similarly modulated by stimulus contrast and anticipation. J. Vis. 14, 767–767. https://doi.org/10.1167/14.10.767

Bonneh, Y.S., Adini, Y., Polat, U., 2015. Contrast sensitivity revealed by microsaccades. J. Vis. 15, 11. https://doi.org/10.1167/15.9.11

Brainard, D.H., 1997. The Psychophysics Toolbox. Spat. Vis. 10, 433–436.

Ciocchi, S., Herry, C., Grenier, F., Wolff, S.B.E., Letzkus, J.J., Vlachos, I., Ehrlich, I., Sprengel, R., Deisseroth, K., Stadler, M.B., Müller, C., Lüthi, A., 2010. Encoding of conditioned fear in central amygdala inhibitory circuits. Nature 468, 277–282. https://doi.org/10.1038/nature09559

Cummings, A., Čeponienė, R., Koyama, A., Saygin, A.P., Townsend, J., Dick, F., 2006. Auditory semantic networks for words and natural sounds. Brain Res. 1115, 92–107. https://doi.org/10.1016/j.brainres.2006.07.050

Dalmaso, M., Castelli, L., Scatturin, P., Galfano, G., 2017. Working memory load modulates microsaccadic rate. J. Vis. 17, 6–6. https://doi.org/10.1167/17.3.6

Davis, M., 1984. The Mammalian Startle Response, in: Eaton, R.C. (Ed.), Neural Mechanisms of Startle Behavior. Springer US, Boston, MA, pp. 287–351. https://doi.org/10.1007/978-1-4899-2286-1_10

Dayan, P., Abbott, L.F., 2001. Theoretical neuroscience. Cambridge, MA: MIT Press.

Dick, F., Bussiere, J., 2002. The effects of linguistic mediation on the identification of environmental sounds 14, 9.

Diedenhofen, B., Musch, J., 2015. cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. PLOS ONE 10, e0121945. https://doi.org/10.1371/journal.pone.0121945

Elliott, T.M., Theunissen, F.E., 2009. The Modulation Transfer Function for Speech Intelligibility. PLOS Comput. Biol. 5, e1000302. https://doi.org/10.1371/journal.pcbi.1000302

Engbert, R., Kliegl, R., 2003. Microsaccades uncover the orientation of covert attention. Vision Res. 43, 1035–1045. https://doi.org/10.1016/S0042-6989(03)00084-1

Gao, X., Yan, H., Sun, H., 2015. Modulation of microsaccade rate by task difficulty revealed through between- and within-trial comparisons. J. Vis. 15, 3–3. https://doi.org/10.1167/15.3.3

Hafed, Z.M., Chen, C.-Y., Tian, X., 2015. Vision, Perception, and Attention through the Lens of Microsaccades: Mechanisms and Implications. Front. Syst. Neurosci. 9. https://doi.org/10.3389/fnsys.2015.00167

Hafed, Z.M., Clark, J.J., 2002. Microsaccades as an overt measure of covert attention shifts. Vision Res. 42, 2533–2545. https://doi.org/10.1016/S0042-6989(02)00263-8

Hafed, Z.M., Goffart, L., Krauzlis, R.J., 2009. A Neural Mechanism for Microsaccade Generation in the Primate Superior Colliculus. Science 323, 940–943. https://doi.org/10.1126/science.1166112

Hafed, Z.M., Ignashchenkova, A., 2013. On the Dissociation between Microsaccade Rate and Direction after Peripheral Cues: Microsaccadic Inhibition Revisited. J. Neurosci. 33, 16220–16235. https://doi.org/10.1523/JNEUROSCI.2240-13.2013

Hartmann, E., 1996. Outline for a theory on the nature and functions of dreaming. Dreaming 6, 147–170. https://doi.org/10.1037/h0094452

Huang, N., Elhilali, M., 2017. Auditory salience using natural soundscapes. J. Acoust. Soc. Am. 141, 2163–2176. https://doi.org/10.1121/1.4979055

Itti, L., Koch, C., 2001. Feature combination strategies for saliency-based visual attention systems. J. Electron. Imaging 10, 161–169. https://doi.org/10.1117/1.1333677

Itti, L., Koch, C., 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Res. 40, 1489–1506.

Joshi, S., Li, Y., Kalwani, R.M., Gold, J.I., 2016. Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. Neuron 89, 221–234. https://doi.org/10.1016/j.neuron.2015.11.028

Kaya, E.M., Elhilali, M., 2014. Investigating bottom-up auditory attention. Front. Hum. Neurosci. 8. https://doi.org/10.3389/fnhum.2014.00327

Kayser, C., Petkov, C.I., Lippert, M., Logothetis, N.K., 2005. Mechanisms for Allocating Auditory Attention: An Auditory Saliency Map. Curr. Biol. 15, 1943–1947. https://doi.org/10.1016/j.cub.2005.09.040

Killion, M.C., 1978. Revised estimate of minimum audible pressure: Where is the ''missing 6 dB''? J. Acoust. Soc. Am. 63, 1501–1508. https://doi.org/10.1121/1.381844

Knudson, I.M., Melcher, J.R., 2016. Elevated Acoustic Startle Responses in Humans: Relationship to Reduced Loudness Discomfort Level, but not Self-Report of Hyperacusis. J. Assoc. Res. Otolaryngol. 17, 223–235. https://doi.org/10.1007/s10162-016-0555-y

Krauzlis, R.J., Bogadhi, A.R., Herman, J.P., Bollimunta, A., 2018. Selective attention without a neocortex. Cortex, The Unconscious Guidance of Attention 102, 161–175. https://doi.org/10.1016/j.cortex.2017.08.026

Liao, H.-I., Kidani, S., Yoneya, M., Kashino, M., Furukawa, S., 2016. Correspondences among pupillary dilation response, subjective salience of sounds, and loudness. Psychon. Bull. Rev. 23, 412–425. https://doi.org/10.3758/s13423-015-0898-0

Meredith, M.A., Clemo, H.R., 1989. Auditory cortical projection from the anterior ectosylvian sulcus (Field AES) to the superior colliculus in the cat: An anatomical and electrophysiological study. J. Comp. Neurol. 289, 687–707. https://doi.org/10.1002/cne.902890412

Meredith, M.A., Nemitz, J.W., Stein, B.E., 1987. Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. J. Neurosci. 7, 3215–3229. https://doi.org/10.1523/JNEUROSCI.07-10-03215.1987

Meredith, M.A., Stein, B.E., 1986. Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. J. Neurophysiol. 56, 640–662. https://doi.org/10.1152/jn.1986.56.3.640

Mizzi, R., Michael, G.A., 2014. The role of the collicular pathway in the salience-based progression of visual attention. Behav. Brain Res. 270, 330–338. https://doi.org/10.1016/j.bbr.2014.05.043

Moore, B.C., Glasberg, B.R., 1983. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J. Acoust. Soc. Am. 74, 750–753.

Nader, K., Schafe, G.E., Doux, J.E.L., 2000. Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. Nature 406, 722. https://doi.org/10.1038/35021052

Parkhurst, D., Law, K., Niebur, E., 2002. Modeling the role of salience in the allocation of overt visual attention. Vision Res. 42, 107–123. https://doi.org/10.1016/S0042-6989(01)00250-4

751 Peel, T.R., Hafed, Z.M., Dash, S., Lomber, S.G., Corneil, B.D., 2016. A Causal Role for the Cortical
752 Frontal Eye Fields in Microsaccade Deployment. PLOS Biol. 14, e1002531.
753 https://doi.org/10.1371/journal.pbio.1002531

754 Peters, R.J., Iyer, A., Itti, L., Koch, C., 2005. Components of bottom-up gaze allocation in natural
755 images. Vision Res. 45, 2397–2416. https://doi.org/10.1016/j.visres.2005.03.019

756 Pressnitzer, D., McAdams, S., 1999. Two phase effects in roughness perception. J. Acoust. Soc. Am.
757 105, 2773–2782. https://doi.org/10.1121/1.426894

758 Rolfs, M., 2009. Microsaccades: Small steps on a long way. Vision Res. 49, 2415–2441.
759 https://doi.org/10.1016/j.visres.2009.08.010

760 Rolfs, M., Kliegl, R., Engbert, R., 2008. Toward a model of microsaccade generation: The case of
761 microsaccadic inhibition. J. Vis. 8, 5–5. https://doi.org/10.1167/8.11.5

762 Rucci, M., Poletti, M., 2015. Control and Functions of Fixational Eye Movements. Annu. Rev. Vis. Sci.
763 1, 499–518. https://doi.org/10.1146/annurev-vision-082114-035742

764 Sara, S.J., 2009. The locus coeruleus and noradrenergic modulation of cognition. Nat. Rev. Neurosci.
765 10, 211–223. https://doi.org/10.1038/nrn2573

766 Sara, S.J., Bouret, S., 2012. Orienting and Reorienting: The Locus Coeruleus Mediates Cognition
767 through Arousal. Neuron 76, 130–141. https://doi.org/10.1016/j.neuron.2012.09.011

768 Sato, S., You, J., Jeon, J.Y., 2007. Sound quality characteristics of refrigerator noise in real living
769 environments with relation to psychoacoustical and autocorrelation function parameters. J.
770 Acoust. Soc. Am. 122, 314–325. https://doi.org/10.1121/1.2739440

771 Saygin, A.P., Dick, F., Bates, E., 2005. An on-line task for contrasting auditory processing in the verbal
772 and nonverbal domains and norms for younger and older adults. Behav. Res. Methods 37,
773 99–110.

774 Schnupp, J.W.H., Garcia-Lazaro, J.A., Lesica, N.A., 2015. Periodotopy in the gerbil inferior colliculus:
775 local clustering rather than a gradient map. Front. Neural Circuits 9.
776 https://doi.org/10.3389/fncir.2015.00037

777 Sohoglu, E., Chait, M., 2016. Detecting and representing predictable structure during auditory scene
778 analysis. eLife 5, e19113. https://doi.org/10.7554/eLife.19113

779 Southwell, R., Chait, M., 2018. Enhanced deviant responses in patterned relative to random sound
780 sequences. Cortex 109, 92–103. https://doi.org/10.1016/j.cortex.2018.08.032

781 Stewart, N., Chandler, J., Paolacci, G., 2017. Crowdsourcing Samples in Cognitive Science. Trends
782 Cogn. Sci. https://doi.org/10.1016/j.tics.2017.06.007

783 Veale, R., Hafed, Z.M., Yoshida, M., 2017. How is visual salience computed in the brain? Insights from
784 behaviour, neurobiology and modelling. Phil Trans R Soc B 372, 20160113.
785 https://doi.org/10.1098/rstb.2016.0113

786 Wallace, M.T., Meredith, M.A., Stein, B.E., 1998. Multisensory Integration in the Superior Colliculus of
787 the Alert Cat. J. Neurophysiol. 80, 1006–1010. https://doi.org/10.1152/jn.1998.80.2.1006

788 Wang, C.-A., Blohm, G., Huang, J., Boehnke, S.E., Munoz, D.P., 2017. Multisensory integration in
789 orienting behavior: Pupil size, microsaccades, and saccades. Biol. Psychol. 129, 36–44.
790 https://doi.org/10.1016/j.biopsycho.2017.07.024

791 Wang, C.-A., Boehnke, S.E., Itti, L., Munoz, D.P., 2014. Transient Pupil Response Is Modulated by
792 Contrast-Based Saliency. J. Neurosci. 34, 408–417.
793 https://doi.org/10.1523/JNEUROSCI.3550-13.2014

794 Wang, C.-A., Munoz, D.P., 2015. A circuit for pupil orienting responses: implications for cognitive
795 modulation of pupil size. Curr. Opin. Neurobiol., Motor circuits and action 33, 134–140.
796 https://doi.org/10.1016/j.conb.2015.03.018

797 Wang, C.-A., Munoz, D.P., 2014. Modulation of stimulus contrast on the human pupil orienting
798 response. Eur. J. Neurosci. 40, 2822–2832. https://doi.org/10.1111/ejn.12641

White, B.J., Berg, D.J., Kan, J.Y., Marino, R.A., Itti, L., Munoz, D.P., 2017a. Superior colliculus neurons encode a visual saliency map during free viewing of natural dynamic video. Nat. Commun. 8, 14263. https://doi.org/10.1038/ncomms14263

White, B.J., Kan, J.Y., Levy, R., Itti, L., Munoz, D.P., 2017b. Superior colliculus encodes visual saliency before the primary visual cortex. Proc. Natl. Acad. Sci. 114, 9451–9456. https://doi.org/10.1073/pnas.1701003114

Widmann, A., Engbert, R., Schröger, E., 2014. Microsaccadic responses indicate fast categorization of sounds: a novel approach to study auditory cognition. J. Neurosci. Off. J. Soc. Neurosci. 34, 11152–11158. https://doi.org/10.1523/JNEUROSCI.1568-14.2014

Woods, K.J.P., McDermott, J.H., 2018. Schema learning for the cocktail party problem. Proc. Natl. Acad. Sci. 115, E3313–E3322. https://doi.org/10.1073/pnas.1801614115

Woods, K.J.P., Siegel, M.H., Traer, J., McDermott, J.H., 2017. Headphone screening to facilitate web-based auditory experiments. Atten. Percept. Psychophys. 79, 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Xiong, X.R., Liang, F., Zingg, B., Ji, X., Ibrahim, L.A., Tao, H.W., Zhang, L.I., 2015. Auditory cortex controls sound-driven innate defense behaviour through corticofugal projections to inferior colliculus. Nat. Commun. 6, 7224. https://doi.org/10.1038/ncomms8224

Yablonski, M., Polat, U., Bonneh, Y.S., Ben-Shachar, M., 2017. Microsaccades are sensitive to word structure: A novel approach to study language processing. Sci. Rep. 7, 3999. https://doi.org/10.1038/s41598-017-04391-4

Yuval-Greenberg, S., Merriam, E.P., Heeger, D.J., 2014. Spontaneous Microsaccades Reflect Shifts in Covert Attention. J. Neurosci. 34, 13693–13700. https://doi.org/10.1523/JNEUROSCI.0582-14.2014

Zingg, B., Chou, X., Zhang, Z., Mesik, L., Liang, F., Tao, H.W., Zhang, L.I., 2017. AAV-Mediated Anterograde Transsynaptic Tagging: Mapping Corticocollicular Input-Defined Neural Pathways for Defense Behaviors. Neuron 93, 33–47. https://doi.org/10.1016/j.neuron.2016.11.045

Figure Legends

Figure 1: Crowdsourced ‘subjective’ salience rating for brief environmental sounds. (A) Spectrograms for all 18 sounds are displayed in order of ranking in (B). See Figure 1-1 in the Extended Data section for sound files. (B) Crowdsourced rating collected from MTurk (N=911). The sounds used in the in-lab replication are indicated by orange-coloured bars. Error bars are one standard deviation from bootstrap resampling. See Figure 1-2 in the Extended Data section for details of the instructions to online participants and MTurk page layout. (C) Crowdsourced salience rating is strongly correlated with the in-lab salience ranking. The dashed line indicates identical ranks (D) Crowdsourced salience rating is strongly correlated with acoustic ‘roughness’. All correlations are conducted using the Spearman rank method.

Figure 2: MTurk task data. (A) Distribution of time spent on each HIT (‘human intelligence task’). (B) Distribution of number of HITs completed per worker.

Figure 3: To estimate the range of 'roughness' values we might expect to encounter in the environment, we quantified roughness (see methods) for a large set of diverse natural sounds (N=274; sound duration = 500ms to match that in the present experiment) from set previously used in (Dick et al, 2002; Saygin et al, 2005). This information is presented in histogram form (grey bars). Roughness values for the sounds used in the present study are indicated by black diamonds. We also include roughness calculated for the scream sounds from Arnal et al, (2015; red diamonds). This analysis confirms that the set of sounds we used spans the range of roughness obtained from a diverse set of environmental sounds. The sounds at the top of the roughness range in our set, overlap with the roughness range defined by the scream sounds from Arnal et al (2015).

Figure 4: The incidence of eye blinks (= proportion of trials containing any blinks in the initial 500ms after sound onset per condition across subjects). The blink rate was not correlated with (A) the crowdsourced salience rating, (B) loudness or (C) microsaccadic inhibition latency. All correlations are conducted using the Spearman rank method.

Figure 5: Measures of pupillary dilation are not correlated with crowdsourced salience. (A) The PDR obtained from the full group (N=30). The solid lines represent the average normalized pupil diameter as a function of time relative to the onset of the sound. The line color indicates the MTurk salience ranking; more salient sounds are labelled in increasingly warmer colors. The solid black line is the grand-average across all conditions. The dashed line marks the peak average PDR. (B) The PDR derivative. The bottom panel shows the correlations (n.s.) between crowdsourced salience and peak PDR amplitude (left) and maximum of PDR derivative (right). None of the effects were significant. A similar analysis based on sound-token specific peaks (as opposed to based on the grand average) also did not yield significant effects. All correlations are conducted using the Spearman rank method.

Figure 6: Microsaccadic inhibition (MSI) is correlated with crowdsourced salience. (A) Raster plot of microsaccade events (pooled across all participants) as a function of time relative to sound onset. The Y axis represents single trials; each dot indicates the onset of a microsaccade. Trials are grouped by sound-token and arranged according to the MTurk-derived salience scale (increasingly hot colors indicate rising salience). The region of microsaccadic inhibition, between 0.2 and 0.7 second post-sound onset, is highlighted with a black rectangle. (B) Over this time interval, the MS rate (number of MS events per second) is correlated significantly with the crowdsourced salience rating. The result of bootstrap resampling is shown as (C) the distribution of correlation coefficients and (D) the distribution of associated p values. The vertical red dashed line indicates p=0.05. (E) Average microsaccade rate time-series for each sound (F) focusing on the MSI region. The solid black curve is the grand-average MS rate across all sound tokens. MSI commences at approximately 0.3-seconds post sound onset (open circle) and peaks around 0.45 seconds (solid black circle). The horizontal dashed line indicates the mid-slope of the grand average (amplitude = -0.04 a.u., time = 0.37 s). Black crosses mark the time at which the response to each sound intersects with this line, as a measure of MSI latency. (G) shows the correlation between these values and the crowdsourced salience rating. All correlations are conducted

using the Spearman rank method. Note identical correlation values in G and B are a chance occurrence (the two analyses are independent).

Figure 7: Estimation of the stability of the correlation between MSI latency and crowdsourced salience. (Left) Distribution of the Spearman correlation coefficients derived from re-sampling analyses with sub-group sizes of 15 or 30 participants. In both cases the distribution peaks around r=-0.5. (Right) Distribution of the p values associated with each N. The red vertical dashed line indicates p=0.05. A uniform distribution is expected under the null. The left skewed pattern observed here indicates a true effect. Skewness was formally confirmed by a X^2 test on p values <0.05 (n=30: $X^2(1066)=1405.42$, $p<0.0001$; n=15: $X^2(748)=903.2$, $p<0.0001$).

Extended Data

Figure 1-1: Sound files for the stimuli used in this study.

Figure 1-2: An example of a HIT ('human intelligence task') page used in the crowd-sourcing experiment.