JASA ARTICLE

Sound identity, salience, and perceived importance in complex auditory environments[a)]

Yu-Jeh Liu and Mounya Elhilali[b)]
Department of Electrical and Computer Engineering, Laboratory for Computational Audio Perception, Johns Hopkins University, Baltimore, Maryland 21218, USA

ABSTRACT:
Human listeners effortlessly identify salient sounds in their environments, yet the relationship between sound class identity, auditory salience, and perceived importance in complex auditory scenes remains poorly understood. In this study, we investigate these connections with scores derived from subject responses using a scoring mechanism, combined with auditory salience and pupillometry data. By leveraging both psychophysical experiments as well as a large-scale annotated dataset, our findings reveal biased responses and higher importance rankings for specific sound classes, such as alarm sounds and speech, and highlight a consistent perceptual ordering of sounds based on their identity. Salience judgments and pupillary responses further support this distinction, showing that the level of heightened arousal follows the same sound class order. The results underscore the influence of semantic mappings on both bottom-up and top-down sensory processing, suggesting that sound identity plays a crucial role in shaping perceptual judgment and neural responses. Despite dataset limitations, our findings offer insights into auditory scene analysis and provide a novel framework for understanding how auditory perception prioritizes sounds based on both their inherent properties and learned semantic associations. © *2025 Acoustical Society of America*.
https://doi.org/10.1121/10.0039710
(Received 4 March 2025; revised 7 October 2025; accepted 10 October 2025; published online 27 October 2025)
[Editor: Christian Lorenzi] Pages: 3489–3502

I. INTRODUCTION

Auditory scene analysis seeks to transform the cacophony of everyday sounds into a manageable set of perceptual tokens (auditory events) that can be treated as time bound "objects" by the nervous system (Bizley *et al.*, 2013; Griffiths and Warren, 2004). In this work, we reserve the term auditory event for any temporally bounded token whose acoustic attributes (e.g., pitch, timbre, loudness, temporal envelope) combine with higher level schemas that encode its semantic identity and real-world source. Consider an everyday manifestation of the classic phenomenon of the *cocktail-party problem* (Bregman, 1990), a bustling café during a lunch rush, a sharp, repetitive pattern of a mobile phone ringtone is differentiated by its distinct spectral signature and abrupt onsets, and it is simultaneously categorized as an "incoming call," allowing it to emerge from overlapping speech, clattering dishes, and background music. Canonical grouping cues, such as common onset, harmonicity, and spectral proximity, facilitate this segregation (Darwin, 1997; Oh *et al.*, 2022; Wagemans *et al.*, 2012); however, this process is not purely bottom-up. These bottom-up Gestalt regularities operate in concert with top-down category expectations and semantic knowledge so that perceptual segmentation of auditory events is finely tuned when acoustic evidence matches familiar schemas (e.g., telephone alert, mechanical hiss). This interaction between signal-driven structure and semantic prediction motivates the present investigation: do certain sound categories, by virtue of their ecological or cognitive relevance, obtain privileged perceptual status within complex auditory scenes?

Vision research provides a compelling precedent for such biases. Not only do low level cues segment visual scenes into objects, but systematic regularities also tilt attention toward particular objects and positions. Spatial and semantic structures can accelerate the learning of face–scene associations, facilitating rapid contextual predictions (Zhou and Geng, 2024). Positional regularities (e.g., the tendency for "object tops" or scene "bottoms") influence similarity judgments, with strongest sensitivity near the scene center where fixations cluster (Langley and McBeath, 2023; Odegaard *et al.*, 2015). Moreover, hierarchical preferences for faces and bodies emerge in the first months of life and influence later attentional development, pointing to an interplay of innate predispositions and experience (Bindemann *et al.*, 2010; Frank *et al.*, 2014). In addition, complementary findings show that canonical object configurations (e.g., a lamp above a table) are recognized more efficiently than improbable ones, underscoring the role of semantic priors in shaping perceptual gain. These converging data highlight that object recognition is influenced not only by segmentation rules, but also by enduring biases that assign special status to certain visual categories and their expected spatial relations. Crucially, quantitative work by Spain (Spain and Perona, 2011) demonstrates that observers rank the *perceptual importance* of scene elements

[a)]This paper is part of the special issue on Ecological Perspectives on Hearing.
[b)]Email: mounya@jhu.edu

by jointly weighing bottom-up salience and task-driven semantic relevance, providing direct evidence that certain categories systematically receive priority. Together, these studies indicate that object recognition is influenced not only by segmentation rules, but also by enduring biases that assign special status to particular visual categories and their expected spatial relations.

Parallel phenomena appear in audition, although they have been less thoroughly catalogued. Semantically congruent warning beeps, coupled with visual hazard icons, shorten reaction times and boost accuracy in audiovisual tasks, illustrating how meaning modulates auditory priority (Isherwood and McKeown, 2017). In contrast, semantic incongruency between a target sound and its background auditory scene leads to more accurate identification, which highlights the complex mechanisms underlying semantic association processing (Gygi and Shafiro, 2011). Electroencephalography and functional magnetic resonance imaging (MRI) reveal that neonates already display cortical tuning for melodic contour and tonal harmony, implying that musical biases are present before extensive cultural exposure (Perani *et al.*, 2010). In adults, neuroimaging shows partially segregated frontotemporal networks for speech and for music that extend well beyond primary auditory cortex and display distinct time course dynamics (Koelsch, 2005; Leaver and Rauschecker, 2010; Norman-Haignere *et al.*, 2015). Perceptually, speech tends to "pop out" in multi-talker mixtures, whereas identically loud environmental sounds do not; a hallmark of privileged processing. Yet, we still lack a systematic account of how sound *identity* (e.g., speech, music, animal vocalizations) modulates perceptual priority when multiple events compete for attention in real scenes. Addressing this gap is the aim of the present study, which asks whether listeners assign consistent priorities to sound categories and how these priorities map onto measurable salience and physiological indices.

To tackle this question, we invoke the broader construct of *salience*, the stimulus driven conspicuity that funnels attentional resources toward a perceptual locus. Salience is traditionally described as a fusion of an event's physical attributes with top-down biases rooted in meaning and context (Huang and Elhilali, 2020; Kaya and Elhilali, 2017). Comparable principles operate in vision, where edge and shape segregation bootstraps both low- and high-level object representations (Driver and Baylis, 1995; Hoffman and Singh, 1997). In audition, loudness, spectrotemporal contrast, and abrupt onsets confer a bottom-up pull, but semantic content (such as spoken language, familiar melodies, biologically relevant calls) can amplify or dampen the pull even when low level energy is matched (Broderick *et al.*, 2019; Kothinti and Elhilali, 2023). Visual experiments demonstrate that text or objects sharing semantic relationships attract gaze more strongly than equally salient but unrelated items (Hwang *et al.*, 2011; Wang and Pomplun, 2012; Wu *et al.*, 2014). By analogy, speech and music may recruit dedicated cortical circuitry that endows them with an attentional advantage over acoustically similar environmental sounds. Disentangling these contributions is instrumental for developing predictive models of attention that go beyond energy-based detectors and incorporate categorical knowledge. Essentially, salience is not synonymous with *importance*. A flashing neon sign may dominate visual attention while conveying little behavioral relevance when compared with a dull but informative traffic light in the driver's periphery (Wang *et al.*, 2010). Vision studies demonstrate that object ranking reflects a negotiated balance between low level salience and higher order semantics, modulated by task goals, scene contexts, and learned contingencies (de Haas *et al.*, 2019; Nuthmann *et al.*, 2020; Schomaker *et al.*, 2017; Wang *et al.*, 2018). Computational models that incorporate object–context interactions outperform purely salience driven models in predicting which items observers later recall or act upon (Tian *et al.*, 2022). Despite clear parallels between visual and auditory scene analysis, the joint influence of acoustic salience and semantic identity on perceived importance of auditory events remains largely unexplored. Clarifying this relationship will inform assistive listening technologies, automatic audio summarization, and neuro-ergonomic design of warning systems, all of which must decide *which* sounds merit priority.

The present work therefore examines how listeners parse dynamic acoustic scenes when salience, semantics, and judged importance intersect. Across two laboratory experiments, we continuously record frame-level overt responses, subjective salience estimates and high temporal resolution pupillometry while participants listen to naturalistic mixtures recorded from everyday environments. Each mixture contains temporally overlapping events drawn from diverse sources. Additionally, on a trial-level scale, we introduce a data-driven scoring framework that positions events along a perceptual continuum based on sound identity. From the collected measures, we test whether listeners' trial-level importance rankings, frame-level subjective salience judgments, and pupil arousal converge along the continuum. Hierarchical modeling allows us to disentangle shared variance due to low level acoustics from category-specific contributions, revealing that certain sounds retain an importance premium even when equivalent in loudness and temporal position to competing sounds. By integrating perceptual and physiological measures, our study aims to reveal how biases toward specific sound categories shape object identification and, ultimately, comprehension of complex auditory scenes.

II. METHODS

A. Audio stimuli

The audio stimuli used in this study are sourced from the Google AudioSet (Google Inc., Mountain View, CA) evaluation dataset and its subset, the detection and classification of acoustic scenes and events (DCASE) challenge Task 4a public set. This particular dataset is selected for a few reasons: (1) the audio clips encompass diverse sound classes within a single clip, (2) the audio clips are collected from different YouTube (YouTube Inc., San Bruno, CA) creators; hence, reflecting a wide range of recording setups, (3) strong labels, which are sound class annotations with

https://doi.org/10.1121/10.0039710

precise start and end time boundaries, are available throughout the duration of the audio clips.

A total of 50 audio clips serve as our audio stimuli. In a first identification experiment, the stimuli are extracted directly from the dataset and are each 10 s long. In a second salience experiment, the 10 s audio clips are extended to their original samples from the initial YouTube sources and set to 30 s duration. The stimuli have varying original sampling rates ranging from 44.1 to 192 kHz. All stimuli are volume equalized using the Fast Forward Moving Picture Experts Group (FFmpeg) EBU R128 loudness normalization tool and then resampled at 44.1 kHz.

Each recording in the stimulus set includes at least one of the 12 sound types of interest: (1) speech, (2) music, (3) cat, (4) dog, (5) wild animals, (6) dishes, (7) frying, (8) alarm bell, (9) electric shaver, (10) blender, (11) wind, and (12) running water—following the initial annotation set in AudioSet. These 12 sound types are specifically chosen to span across the top level sound types indicated by the AudioSet ontology, which are the six aggregate sound classes: (1) speech, (2) music, (3) animal, (4) domestic sounds, (5) alarm, and (6) natural sounds. The analysis was conducted based on the six sound types. However, some clips may contain additional sound classes beyond the ones mentioned above. A more detailed description of the stimuli is included in the Appendix.

B. Experimental procedure

For this study, two experiments are carried out following a study protocol approved by the Johns Hopkins Institutional Review Board (IRB). See Ethics Approval for details.

1. Event identification experiment

- *Experimental setup*: The first experiment is conducted in a sound booth with soundproofing insulation under supervision of an experimenter. Subjects are seated in the booth, and stimuli are presented with an ASUS Xonar Essence STX sound card (ASUS Inc., Taipei, Taiwan) over a pair of Sennheiser HD595 headphones (Sennheiser electronic SE & Co. KG, Wedemark, Germany).
- *Experimental paradigm*: The experiment consists of 50 trials and has a total duration of around 50 min. Each trial presents the subject with a 10 s stereo audio stimulus while they fixate on a crosshair displayed on a computer screen. After each stimulus, the subject is presented with text boxes containing all possible sound classes through a graphical interface. Subjects are instructed to identify and rank sound classes that stand out the most to them in a trial. The ranking process is done by subjects arranging text boxes from top to bottom using a drag-and-drop menu. In addition to subject responses collected from the ranking interface, pupillometry is tracked and recorded by an EyeLink 1000 eyetracking camera (SR Research, Ltd., Oakville, ON, Canada) at a sampling rate of 2000 Hz throughout the entire trial.
- *Participants*: A total of 17 human subjects (nine male, seven female, 1 non-binary/unspecified), average age of 27.4 years (standard deviation 3.5 years), are recruited for the task.

2. Auditory salience experiment

- *Experimental setup*: The second experiment is an online study run on the Amazon Mechanical Turk (MTurk) platform (Amazon Inc., Seattle, WA). The presentation of the experiment is implemented using the jsPsych library (de Leeuw, 2015). Subjects are instructed to use headphones only. An auditory test is employed at the start of the experiment. Testing sound clips are played on one side each time, and subjects need to indicate which side a sound is played. It requires a 100% accuracy to continue to the experiment and serves to ensure the participants are using headphones. Amazon Web Services (AWS) (Amazon Inc., Seattle, WA) is used to host the experiment, and the execution of the interface is enabled by the psiTurk framework (Gureckis *et al.*, 2016).
- *Experimental paradigm*: The experiment employs dichotic listening following the procedure developed in Kothinti *et al.* (2021). Each subject is presented with 15 stimulus pairs drawn at random from the pool of 50 stimuli, without repetition. In each trial, the computer screen is divided by two vertical lines, segregating it into three distinct sections: left, middle, and right. Subjects are instructed to listen to both scenes simultaneously and indicate their attentional focus by continuously moving the cursor. A training video is played before trials to clarify the instructions. If the scene played on the right captures the subject's attention, subjects move the cursor to the right. When subjects deem scenes in both ears to be attention-grabbing or when there is no focus, they keep the cursor in the middle of the screen.
- *Participants*: A total of 570 subjects (419 male, 145 female, six non-binary/unspecified) initially participated in the study (average age 32.7 years, standard deviation 7.9 years). After a quality control analysis procedure (elaborated below), 192 subjects are retained for further analysis. A retention rate of $\sim 33\%$ is typical for this online paradigm, given the high response noise as previously established in Kothinti *et al.* (2021).

C. Subject data analysis

All analyses are conducted across all participants in both experiments. Repeating the analyses for gender groups male and female shows no differences. There are not enough participants in the other gender groups (non-binary/unspecified) for a conclusive statistical analysis.

1. Importance scores

For the event identification experiment, subjects identify sound classes in the order they find particularly salient after each trial is complete. This process results in a ranked choice matrix for each trial. Notably, the subjects' choices reflect a complex interaction among factors, such as bottom-up salience, semantic biases, top-down recognition difficulty, etc., that leads to the trial-level ranking decisions. Adopting a procedure proposed by Spain and Perona (2011), an importance score is generated by treating the ranking

process as an urn problem without replacement. The process of ball selection from the urn corresponds to the selection of a certain sound class from the drag-and-drop interface. The importance score of a given sound class in a given trial is therefore defined as the probability of the ball being drawn first.

Following the urn model, the probability of sound class i from K total classes is defined as $P(S_i)$ with the constraint

$$\begin{cases} \sum_{i=1}^{K} P(S_i) = 1, \\ P(S_i) < 1 \quad \forall i \in \{1, \cdots, K\}. \end{cases} \tag{1}$$

Given that the importance scores are essentially probabilities $P(S_i)$, solving for the importance scores of sound classes in a single trial is equivalent to solving a maximum likelihood problem defined by the likelihood function

$$\mathcal{L} = \prod_{n=1}^{N} \prod_{m=1}^{M_n} P(S^{\{m,n\}} | S^{\{m-1,n\}}, S^{\{m-2,n\}}, \ldots, S^{\{1,n\}}), \tag{2}$$

where N is the number of subjects, M_n is the number of ranked objects chosen by subject n, and $S^{\{m,n\}}$ is the sound class ranked m by subject n and takes the value $S^{\{m,n\}} = S_i$ for some i.

Here, because we are drawing without replacement, the probability $P(S^{\{m,n\}} | S^{\{m-1,n\}}, \ldots, S^{\{1,n\}})$ is equal to

$$= \begin{cases} P(S^{\{m,n\}} | S^{\{m-1,n\}}, \cdots, S^{\{1,n\}}) \\ 0 \qquad \text{if } S^{\{m,n\}} \in \{S^{\{1,n\}}, \ldots, S^{\{m-1,n\}}\}, \\ \dfrac{P(S^{\{m,n\}})}{1 - \sum_{i=1}^{m-1} P(S^{\{i,n\}})} \qquad \text{otherwise.} \end{cases} \tag{3}$$

Finally, the importance score for each sound class is defined as the maximum likelihood estimator (MLE) of the likelihood function defined in Eqs. (2) and (3) using subjects' responses, under constraints set in Eq. (1). Given that the defined likelihood function is not convex, Basin-hopping in conjunction with sequential least squares programming (SLSQP) in the Python scipy optimize module are used to find the global minimum for the MLE estimation. Different analyses of the optimization procedure using the loss function and Monte Carlo methods confirm that stable global maxima of the importance score are achieved. Based on empirical testing, optimization results after 8000 iterations of basin hopping are taken as the final importance scores used for further analysis. An example of computed importance scores for one trial derived from subjects' ranking responses is shown in Fig. 1, bottom right panel.

2. Labeling scores from a public dataset

In order to extend the importance score analysis to a wider set of acoustic stimuli and to verify the validity of the derived importance scores, we analyze identification results from a public dataset. The multi-annotator estimated strong labels (MAESTRO) dataset consists of complex auditory scene recordings capturing different acoustic scenes and is annotated to estimate a soft label that reflects the divergence in judgments between multiple annotators. The dataset contains real-life recordings of everyday scenes that are annotated by two expert annotators and synthetic scenes generated with randomly placed auditory events annotated through MTurk. The procedure for deriving soft labels from raw annotations is described in Martin-Morato and Mesaros (2023). For the current study, we analyze 250 min of MAESTRO scenes that contain sound events determined to match the six general classes identified in the event identification experiment, using the mapping in Table I.

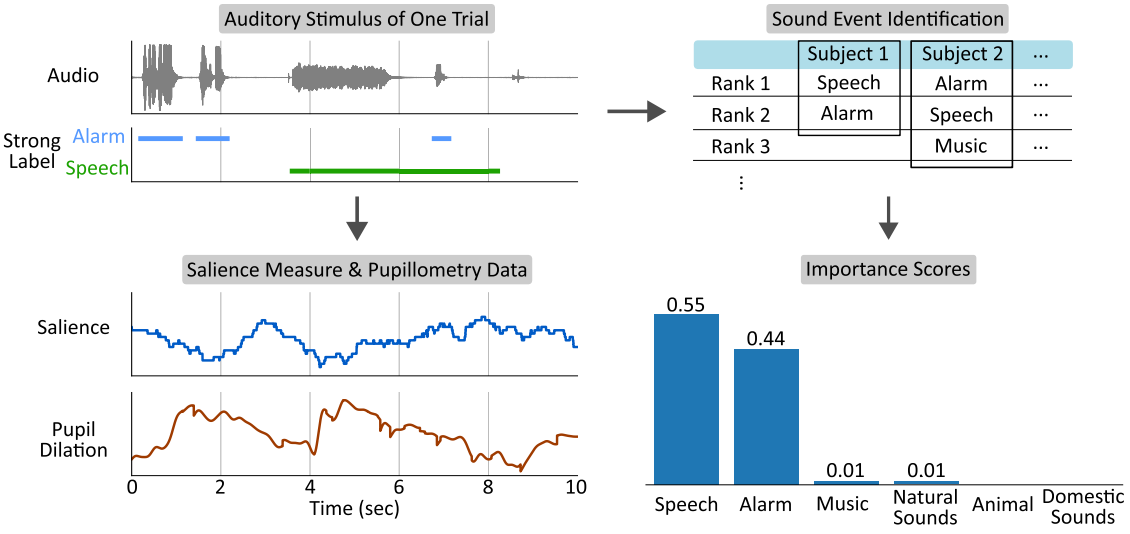

FIG. 1. An overview of experimental paradigms and data. Top left depicts a typical audio stimulus for one trial along with timestamps of sound events (strong labels) provided in the DCASE dataset. Bottom left depicts subjects' salience judgment obtained from an online experiment and pupil dilation measured concurrently with the sound identification experiment. Top right represents typical class identification and ranking reported by different subjects. Bottom right illustrates output of the importance score optimization analysis based on subjects' ranked responses.

https://doi.org/10.1121/10.0039710

TABLE I. Sound class mapping from the MAESTRO dataset to this study.

MAESTRO sound class	Class grouping
Siren	Alarm
People talking	Speech
Announcement	Speech
Bird singing	Animal
Dog barking	Animal
Street music	Music
Cutlery and dishes	Domestic sounds
Coffee machine	Domestic sounds
Door opens/closes	Domestic sounds
Furniture dragging	Domestic sounds
Wind blowing	Natural sounds

3. Analysis of identification scores

Both importance and labeling scores provide a distribution of judgments for each sound class over a [0, 1] support, reflecting its perceived importance and priority. In order to compare class-specific distributions for each of these measures, the probability density function for each class is estimated using kernel density estimation (KDE) with a Gaussian kernel to estimate the distribution density with 100 bins over the support. Next, the Wasserstein distance (Earth mover's distance) using the squared Euclidean distance as the cost function is used to compare pairs of class-specific distributions in order to capture differences in distribution shape. This metric is chosen because it penalizes large differences between distributions and captures subtle changes in the shape of the distributions. Pairwise comparisons between the distributions of the six classes are then combined using dendrograms to analyze the hierarchical relationships between them (Frades and Matthiesen, 2010; Wierzchoń and Kłopotek, 2018). The dendrogram analysis is constrained using an optimal leaf ordering (OLO) with a condition that maximizes the sum of similarity between every leaf and all other leaves in the adjacent cluster. This procedure allows us to identify groupings of sound classes and assess similarities in these groupings in order to pinpoint whether some sound classes tend to generate perceptual responses that are more similar to each other. More importantly, the procedure enables us to compare grouping across the two measures explored in this study: importance scores derived from the event identification experiment and labeling scores obtained from the MAESTRO dataset.

To verify the statistical significance and stability of the hierarchical grouping, we employ bootstrapping as a resampling technique. A total of 75% of the data from each class-specific distribution are systematically resampled, and hierarchical clustering is reapplied to generate a dendrogram using OLO. This procedure is repeated for 1000 iterations. Kendall's tau is then computed between the original and each of the bootstrap-generated dendrograms in order to evaluate the consistency in the ordering of clusters across different bootstrap runs. A bias correction of Kendall's tau is computed to account for bias introduced due to the small sample size. Confidence intervals of corrected tau values are then evaluated.

4. Inter-subject agreement from the identification experiment

To evaluate the inter-subject agreement of reported classes in the identification experiment, subject responses for each trial are arranged into binary matrices for each class (1 when a class has been selected by a subject in a trial; 0 otherwise). A Hamming distance is used as a measure of similarity across pairs of binary sequences comparing all possible pairs of subjects. The distance values are normalized (ranging from 0 to 1) by the length of the response vector and reflect the ratio of disagreement. This analysis yields a class-specific distribution of distances whose mean is compared to two baseline measures. First, a random binary matrix is generated from a fair Bernoulli distribution coin toss representing random selection of sound classes. Pairs of distance are computed 136 (unique pairs between 17 subjects) times to form a class-agnostic baseline of random responses. Second, for each class, a randomly shuffled binary matrix is generated, and pairs of distances are computed 136 times. Naturally for a normalized measure, classes with fewer occurrences (e.g., alarms) result in different baselines compared to frequently occurring classes.

To further assess the inter-subject agreement trends across classes, a bootstrapping procedure is used to examine higher-order moments of distance distributions. For each class and each bootstrapping round, half trials are selected at random, and the average inter-subject distribution across subjects is evaluated. Each bootstrapping distribution yields a variance, skewness, and kurtosis measures that are compiled across 500 iterations. Quantitatively similar results are obtained with different percentage of trials selected.

5. Button press response from the identification experiment

In addition to class identity, response times are also analyzed using the timestamp when a sound class text-box is clicked. By averaging the number of button clicks per class across trials, it is evident that subjects generally click only once per trial when a sound class is deemed present (speech = 1.11 clicks / trial, music = 1.07 clicks / trial, animal = 1.14 clicks / trial, domestic sounds = 1.10 clicks / trial, alarm = 1.13 clicks / trial, natural sounds = 1.17 clicks / trial]. Therefore, only the first instance of clicking for a sound class within a trial is taken into account for further analysis to eliminate noise from accidental clicking. Due to potential variability from hardware and user factors, button press timing provides a coarse rather than precise temporal measure. Still, it is analyzed in the current study for statistical significance that reflects correlations with independently computed importance scores.

6. Salience measure from the auditory salience experiment

Similar to the importance scores, subjects' responses obtained from the auditory salience experiments capture the

intrinsic dynamics between bottom-up and top-down factors. However, these responses operate on a more local, frame-level scale, in contrast to the more global, trial-level perspective of the importance scores.

First, they are analyzed for quality control to flag outlier subjects and trials based on extreme behaviors following the steps proposed by Kothinti *et al*. (2021). Subjects with outlier behavior based on switching rate between left and right side and erratic cursor behavior are excluded from further analysis. Next, dichotic responses to left/right-ear attentional engagement are recorded as binary sequences. A value of 1 represents the subject focusing on a given scene, while 0 represents attention to the opposite scene or neither scene. By averaging the binary sequences across all subjects, a temporal salience measure is generated for the particular scene.

Salience judgments, defined as the temporal salience measure averaged across subjects, near the onset of each sound event (based on AudioSet's strong labels) are analyzed over 1.5 s post onset. Since the identification experiment only flags presence or absence of a sound class for each trial, only one windowed response, with the greatest absolute slope value, is chosen per sound class per trial to account for the maximal response induced by a specific sound class. The time-aligned salience curves are then derived for each sound class and a linear fit is used to derive linear slope values from 100 rounds of 30-sample bootstrapping curves.

For comparison, a reference curve is established. Sample windowed responses lasting 1.5 s are collected preceding event onsets, specifically where no other sound class onsets occur 3 s before these onsets. This ensures noninterference from preceding sound events.

The relationship between the trial-level importance scores and the frame-level salience measure is also investigated. Three values are extracted for Pearson's correlation test against the importance scores across all sound classes. These three values are: response maximum, response mean, and response change. Same as the sound event alignment just described, a 1.5 s windowed response with the greatest absolute slope value is chosen per sound class per trial to extract the values. The response maximum is defined as the single maximum value inside the windowed response. The response mean is computed by averaging response values inside the windowed response. Third, the response change is defined as the absolute slope value to capture the provocation rate (either direction). Over the 50 trials, there are 75 pairs of values for each response measure against the importance scores. A Pearson's correlation analysis is conducted, and the significance is further verified with 100 rounds of 37-sample bootstrapping.

D. Pupillometry data analysis from the identification experiment

Following the procedure proposed by Kret and Sjak-Shie (2019), pupillometry data collected during the event identification experiment is preprocessed through four steps: First, intervals with discontinuities longer than 75 ms, likely linked to blinking, are identified. Samples near the discontinuities (±50 ms) are removed. Second, only gaps measuring less than 250 ms are filled using linear interpolation. Third, the samples are detrended and z-score normalized. Last, the curves are smoothed using a fifth-order Butterworth low-pass filter (with 4 Hz cutoff).

Similar to the analysis of subjective salience judgments, pupillary responses near the onset of each sound event are analyzed over a 1.5 s post event onset. Time-aligned pupillometry curves are derived for each sound class and a logistic function fit is employed to capture the spontaneous nature of pupil dilations. Growth rate values from the logistic fit are computed from 100 runs of bootstrapping curves. Given that pupillometry data results in missing samples for some subjects, bootstrapping is conducted at each time point, gathering half of the available samples. A reference curve is also computed away from the onset of any sound class following the same procedure as salience judgments.

E. Acoustic attribute analysis

To investigate the influence of acoustic properties on importance judgments, distributions of 16 acoustic attributes are obtained for each sound class. These attributes are extracted after mapping the acoustic waveform into a time-frequency auditory spectrograms at a frame rate of 125 Hz with a biomimetic model (Wang and Shamma, 1994). They form a comprehensive collection that covers common temporal and spectral properties of an audio signal and have been frequently used in studies of auditory salience and perception of complex acoustic scenes (Huang and Elhilali, 2017; Kothinti and Elhilali, 2023; McMullin *et al*., 2024).

- Loudness (LD) is the average energy in envelopes computed on 28 bark frequency bands (Zwicker *et al*., 1991).
- Raw energy (RE) is the energy values computed from the spectrogram bands.
- Energy rates (ER) is the energy from spectrotemporal modulation decomposition.
- Pitch (P) is the pitch value determined based on template matching as demonstrated in the optimum processor method (Goldstein, 1973).
- Harmonicity (H) is the measure of the strength of voicing and represents the match between spectral slices and their matched pitch templates.
- Brightness (BR) is the spectral centroid at each temporal slice.
- Bandwidth (BW) is the spectral spread around spectral centroid, computed as the average of the absolute difference between the spectral centroid and frequencies with magnitude spectrum as the weights.
- Irregularity (IR) is a measure of jaggedness in the spectrum, computed as a sum of squares of the spectral magnitude difference between consecutive spectral bins and divided by the sum of squares of all spectral magnitudes.
- Flatness (FL) represents the flatness in the spectrum and is calculated as the geometric mean of spectral

https://doi.org/10.1121/10.0039710

magnitudes divided by the arithmetic mean of spectral magnitudes.
- Average slow temporal modulations, or low rates (LR) is the average energy in frequencies $\leq$20 Hz.
- Average fast temporal modulations, or high rates (HR) is the average energy in frequencies $\geq$20 Hz and measures roughness.
- Rate centroid (RC) is computed as the centroid of temporal modulations for frequencies $\leq$32 Hz.
- Rate maximum (RM) is the maximum energy value across all rates.
- Absolute rate centroid (ARC) is the rate centroid calculated with the magnitude of rate in the weighted average.
- Scale centroid (SC) is the scale centroid computed from spectral modulations.
- Scale maximum (SM) is the maximum energy value across all scales.

The analysis is repeated for both stimuli used for the event identification experiment as well as the MAESTRO dataset. Audio waveforms from both datasets are analyzed for segments that align with a strong label for each of the classes. Mean values over time for each acoustic attribute and each sound class are computed, and then compared to the importance scores and identification scores for the experimental and MAESTRO datasets, respectively. Both within and across group correlations between importance and acoustic attributes are examined using Kendall's tau. A total of 500 rounds of permutation are utilized as null hypothesis to validate statistical significance.

III. RESULTS

A. Identification scores reveal a consistent perceptual ranking of sound classes

The ranked choice of sound classes in the identification experiment results in importance scores that reflect the weight or priority that listeners give to each sound class (see Methods). Importance scores vary between 0 and 1 reflecting low to high ranking for a given sound class across scenes and subjects. By visually examining the importance distributions across classes, Fig. 2(A) (left) reveals that some classes, such as alarm, speech, and to some extent animal, are fairly uniform with heavy tails on both ends and a pronounced density near 1 suggesting that likelihood of identification of these events is generally consistent and prioritized among listeners. Other classes, particularly domestic and natural sounds, tend to be unimodal and skewed towards zero, suggesting less priority by listeners to report or identify these classes. In contrast, music appears to be fairly bimodal likely owing to the perception of music as both an important presence in a scene as well as a background element. The same analysis is replicated in the large-scale dataset MAESTRO based on cross-subject soft label identifications (total 250 min of audio compared to 8.3 min in the identification experiment). Figure 2(A) (right) shows generally similar trends as identification-based importance scores; although the alarm, speech, and animal classes are not very uniform but have heavier tails on both ends. The domestic and natural sound classes lack density near 1.

To quantitatively explore similarities of labeling distributions across sound classes, hierarchical clustering is implemented using dendrograms with pairwise Wasserstein distances [Fig. 2(B), left]. The trends of identification scores reveal that alarm and speech have closely related importance distributions, different from animal and music, which themselves are clustered separately from domestic sounds and natural sounds. The clustering procedure uses OLO that minimizes total distance between adjacent leaves along the dendrogram. Hence, it reveals a *gradual* arrangement of distribution shapes among the six sound classes. A bootstrapping procedure over 1000 iterations confirms that the progression of response types among sound classes is statistically significant (bias-corrected Kendall's tau, $p = 0$). When the importance scores are separated based on gender, no statistical significance is detected. The same analysis is repeated for the MAESTRO dataset and results in identical dendrogram clustering [Fig. 2(B), right]. The MAESTRO groupings are also statistically significant (bias-corrected Kendall's τ, $p = 0.0085$).

B. Inter-subject agreement supports the same perceptual sound class ordering

While the importance scores examine ranked choices across listeners in the identification experiment, inter-subject agreement probes consistency in reported scores. Analysis of subject responses between pairs of subjects reveals high agreement among subjects for all classes [Fig. 2(C), red boxes). The inter-subject agreement (measured as a Hamming distance) is significantly lower than a baseline model of random responses from a fair coin toss process (light gray box) [unpaired t-test with alarm (t-stat $= 72.7$, $p \approx 10^{-179}$), speech (t-stat $= 55.9$, $p \approx 10^{-150}$), animal (t-stat $= 63.1$, $p \approx 10^{-163}$), music (t-stat $= 67.1$, $p \approx 10^{-170}$), domestic sounds (t-stat $= 62.8$, $p \approx 10^{-163}$), and natural sounds (t-stat $= 43.4$, $p \approx 10^{-124}$)]. Furthermore, a class-specific random floor (dark gray boxes) also shows that subjects' responses are statistically lower than this baseline [unpaired t-test with alarm (t-stat $= 34.7$, $p \approx 10^{-101}$), speech (t-stat $= 56.8$, $p \approx 10^{-152}$), animal (t-stat $= 62.8$, $p \approx 10^{-163}$), music (t-stat $= 58.8$, $p \approx 10^{-156}$), domestic sounds (t-stat $= 35.9$, $p \approx 10^{-105}$), and natural sounds (t-stat $= 31.7$, $p \approx 10^{-93}$)]. The class-specific random baseline varies as it reflects the differences in number of occurrences across classes.

A one-way analysis of variance (ANOVA) on inter-subject agreement shows a statistical significant difference in distribution means across the six classes ($p \approx 10^{-30}$). Looking closely at the shape of the agreement distributions themselves reveals consistent trends across classes, captured by higher moments of the distributions. Second, third, and fourth moments corresponding to the variance, skewness, and kurtosis of the agreement of subject responses across trials are

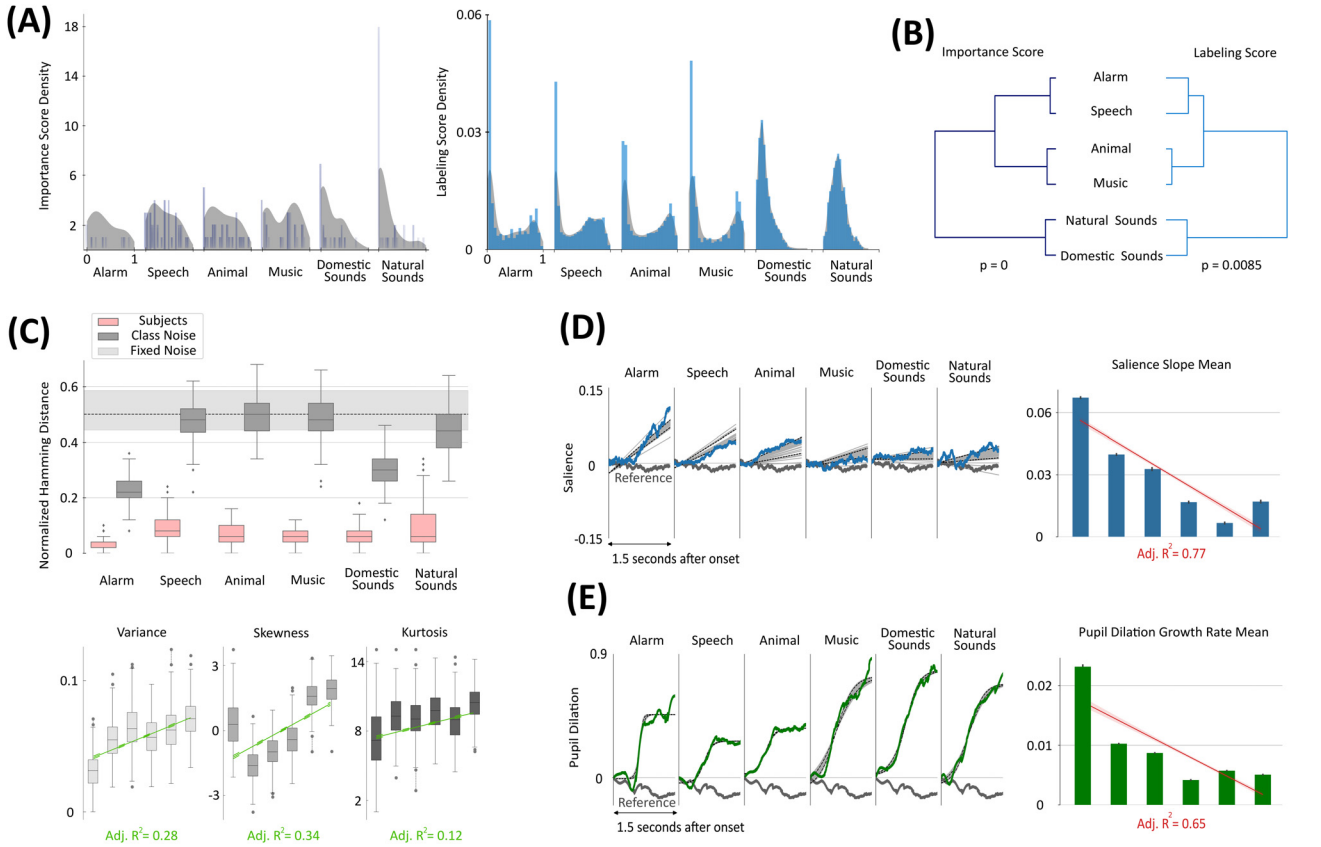

FIG. 2. (A) Left: Importance score distributions and KDE fits grouped by sound class. Right: Labeling score distributions and KDE fits grouped by sound class. (B) Left: Hierarchical clustering dendrogram based on the importance scores. Right: Hierarchical clustering dendrogram based on the labeling scores. (C) Top: Normalized Hamming distance distribution comparing subjects' responses with class-specific and fixed noise floor sequences. Bottom: Bootstrapped results of the variance, skewness, and kurtosis of response consistency, arranged by sound class following the same order as on the top. (D) Left: Salience mean curves with bootstrapping (Black dashed lines represent edges of the 95% confidence interval.). Right: Bootstrapped slope mean values arranged by sound class following the same order as on the left. (E) Left: Pupil dilation mean curves with bootstrapping (black dashed lines represent edges of the 95% confidence interval.) Right: Bootstrapped growth rate mean values from bootstrapping arranged by sound class following the same order as on the left.

obtained using a bootstrapping procedure [Fig. 2(C), bottom]. Examining the second moment shows a gradual increase in variance suggesting a smaller spread of response variability across subjects for sounds, such as alarm, speech, and wider spread for natural or domestic sounds (linear regression fit, adjusted $R^2 = 0.28, F = 1180, p \approx 10^{-217}$). The third moment reveals higher positive skewness for natural or domestic sounds, indicating lower across-subject agreement relative to alarm or speech sounds (adjusted $R^2 = 0.34$, $F = 1580, p \approx 10^{-217}$). The fourth moment again shows the same linear trend, confirming a heavier tail, higher percentage of outlier subject responses, for natural or domestic sounds relative to alarm or speech sounds (adjusted $R^2 = 0.12$, $F = 400, p \approx 10^{-83}$).

C. Changes in salience judgments follow the identical perceptual ordering of sound classes

Subjects' salience judgments reflect the degree to which sound events stand out and attract attention within a scene. Figure 2(D) shows a significant increase in salience relative to a baseline salience (defined away from onsets of any identifiable sound events) near onset of sound events (defined by the strong labels provided in the AudioSet database). As shown by bootstrapping, the salience slope is systematically positive and varies across sound classes [alarm = 0.0673 (0.0610, 0.0709); speech = 0.0399 (0.0375, 0.0457); animal- = 0.0328 (0.0160, 0.0372); music = 0.0168 (0.0134, 0.0246); domestic sounds = 0.0067 (0.0004, 0.0111); natural sounds = 0.0171 (0.0069, 0.0217), where each number represents the mean slope value (5% percentile, 95% percentile)]. For sound classes, such as alarm and speech sounds with higher positive slope values, it indicates that attention is drawn more rapidly because of them. As comparison, the baseline salience away from event onsets has a flat slope of −0.0071 (−0.0223, 0.0088). Comparing the slopes across the six sound classes reveals a linearly decreasing trend following the same perceptual ordering of sound classes as mentioned in Secs. III A and III B sections (linear regression fit, adjusted $R^2 = 0.77, F = 1953, p \approx 10^{-190}$). Gender analysis with the same ordering of sound classes reveals statistical significance for both males and females (male, adjusted $R^2 = 0.29, F = 252, p \approx 10^{-47}$; female, adjusted $R^2 = 0.31, F = 266, p \approx 10^{-50}$).

https://doi.org/10.1121/10.0039710

D. Pupillometry curve growth rate at sound class onsets displays similar sound-class–specific behavior

Sound-class–dependent behavior is also observed when pupillometry curves are time aligned at the sound event onsets. The pupillometry mean curves and logistic function curve fits are shown in Fig. 2(E) and reveal a steeper increase in growth rate for certain sounds, such as the alarm and speech [alarm = 0.0232 (0.0194, 0.0279); speech- = 0.0103 (0.0098, 0.0107); animal = 0.0087 (0.0082, 0.0092); music = 0.0042 (0.0036, 0.0048); domestic sounds = 0.0057 (0.0053, 0.0061); natural sounds = 0.0051 (0.0046, 0.0056), where each number represents the growth rate mean (5% percentile, 95% percentile)]. Similar to behavior salience, this indicates that alarm and speech sounds cause more rapid pupil dilation. In comparison, the baseline pupillometry curve away from event onsets is nearly flat and does not approximate or fit well with a logistic function. The right panel shows the growth rate mean obtained through bootstrapping (linear regression fit, adjusted $R^2 = 0.65, F = 1123, p \approx 10^{-139}$). Similar to salience judgments, pupillometry provides biomarker evidence of a graded response to different sound classes. However, males or females pupillometry data alone do not show significance for the particular ordering of sound classes (male, adjusted $R^2 = 0.004, F = 3.20, p = 0.07$; female, adjusted $R^2 = -0.001, F = 0.15, p = 0.70$). By inspection, multiple outlier growth rate values lead to the test results, which is likely caused by the missing pupillometry data and limited number of subjects.

E. Perceptual importance is strongly correlated with reaction time

During the ranking process of the event identification trials, subjects' button response is also biased by the identities of the specific sound classes. A one-way ANOVA test shows that there is significant difference ($p = 0.006$) among the reaction time mean of the six sound classes. Specifically, from the longest to the shortest button press time mean, we have domestic sounds {4.62 s [standard deviation (std) = 2.03]}, followed by natural sounds [4.58 (std = 1.89) sec], alarm [3.98 (std = 2.22) sec], speech [3.65 (std = 1.33) sec], animal [3.46 (std = 1.31) sec], and music [3.18 (std = 1.09) sec]. Further analysis with a *post hoc* Tukey's honestly significant difference (HSD) test confirms that only two pairs of the sound classes possess significantly different reaction time mean: music/domestic sounds ($p = 0.047$) and music/natural sounds ($p = 0.014$). The rest of the sound classes have statistically similar reaction time mean among themselves.

Furthermore, a correlation analysis comparing button response time and importance scores shows a negative correlation across sound classes ($r = -0.59$, $p \approx 10^{-14}$). The correlation remains significant when considering male and female subjects separately (male, $r = -0.23$, $p = 0.007$; female, $r = -0.40$, $p \approx 10^{-7}$). The observation implies that it takes the subjects a longer time to press a button when a sound event is deemed less important in a given trial. This is illustrated in Fig. 3(A) (left). A separate verification with tail samples removed (importance score < 0.05) again confirms a significant correlation ($r = -0.62$, $p \approx 10^{-12}$). In addition, examining the correlation value per sound class shows that all sound classes exhibit similar significant correlation statistics. [alarm ($r = -0.73$, $p = 0.03$), speech ($r = -0.68$, $p \approx 10^{-6}$), animal ($r = -0.68$, $p \approx 10^{-4}$), music ($r = -0.66$, $p \approx 10^{-4}$), domestic sounds ($r = -0.52$, $p = 0.05$), and natural sounds ($r = -0.55$, $p = 0.003$)].

The relationship between response time and importance score does not address whether there is a relationship between the sound class selected first and its subsequent highest ranking. Alternatively, it is possible that sound classes are identified later but still assigned a high ranking. Analyzing the sound class first selected and the one ranked highest by each participant in each trial reveals that, on average, 89% of the cases, subjects opt for a sound class first if they rank it on top of the others. Similar high ratios are observed across all sound classes (alarm = 92%, speech = 91%, animal = 87%, music = 92%, domestic sounds = 84%, and natural sounds = 88%).

F. Duration correlates positively with how important subjects perceive a sound class

One of the open questions is whether event duration or position influences importance scores reported by listeners. Using strong labels derived from the original DCASE dataset, we explore the correlation between the importance score and these properties. Overall, sound duration reveals a statistically significant correlation ($r = 0.37$, $p = 0.0004$). The relationship is verified with tail samples removed (sound duration > 9 s) ($r = 0.45$, $p = 0.0001$). The correlation remains significant when considering genders separately (male, $r = 0.28$, $p = 0.007$; female, $r = 0.42$, $p \approx 10^{-5}$). In contrast, the sound position does not show any correlation with importance scores [Figs. 3(B) and 3(C)]. This is assessed by considering the position of the first occurrence of a sound class ($r = -0.07$, $p = 0.50$) or the last occurrence ($r = 0.16$, $p = 0.14$). Gender analysis again shows agreeing results both considering the first occurrence (male, $r = -0.01$, $p = 0.93$; female, $r = -0.17$, $p = 0.11$) and the last occurrence (male, $r = 0.19$, $p = 0.07$; female, $r = -0.06$, $p = 0.61$).

G. Discrepancy between strong labels and human subject labels exists only for few trials

While a portion of this study relies on the belief that the strong label annotations available from the public AudioSet database are reliable and accurate, validation is needed. The subject responses and corresponding importance scores are compared against the strong labels in order to quantify their inconsistency—defined as the discrepancy score. In this context, a trial's discrepancy score is computed as the cumulative sum of the importance scores for sound classes

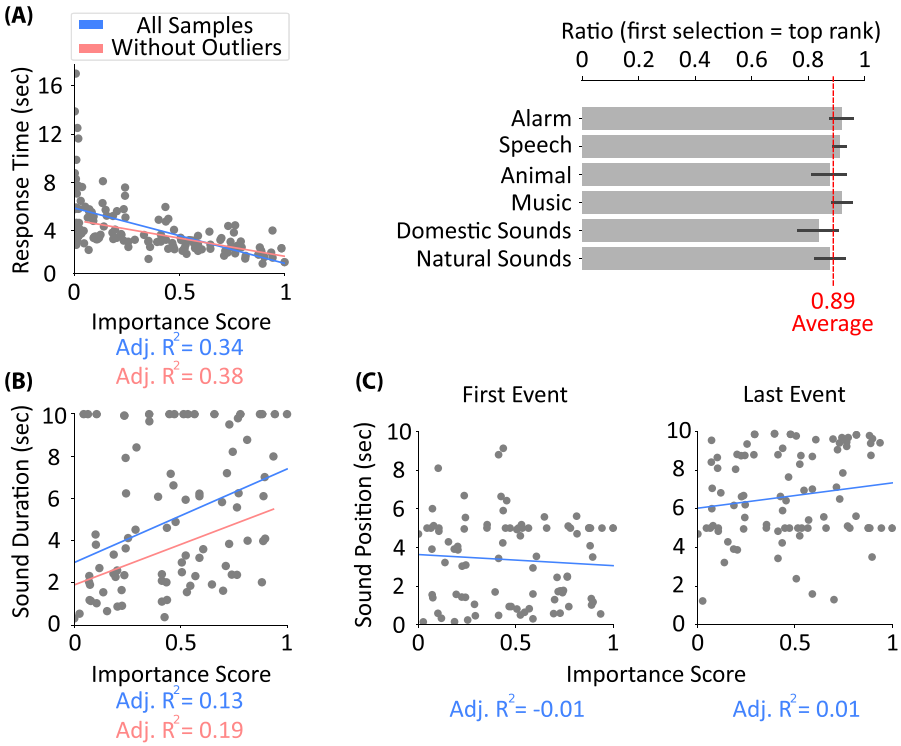

FIG. 3. (A) Left: Button response time plotted against the importance scores across all trials and sound classes. Right: The ratio of the number of times a sound class is ranked top against the number of times a sound class is selected first. (B) Sound durations plotted against the importance scores across all trials and sound classes. (C) Left: Sound positions of the first sound events plotted against the importance scores across all trials and sound classes. Right: Sound positions of the last sound events plotted against the importance scores across all trials and sound classes.

reported by more than half of the subjects (importance score $> 10^{-5}$) but are omitted from the strong labels. As depicted in Fig. 4, among the 50 trials, only five surpass a discrepancy score of 0.5. Among these five trials, the music sound class emerges as the primary contributor to the observed disparities.

Moreover, upon examining the ratio of events reported by subjects but are absent in strong labels, natural sounds (38.5%), music (33.3%), and animal (15.0%) sound classes are the only ones among the six sound classes that show non-zero percentages of discrepancy occurrences. Given that discrepancy is limited to only few classes and few trials ($\sim$24%), we reason that strong labels are reliable for any follow-up analyses.

H. Acoustic profiles of individual sound classes do not reflect statistical bias that drives perceptual importance

While the analysis reported here suggests the existence of perceptual discrimination towards sound classes and their prominence, it leaves open the possibility that effects are driven by acoustic properties of sound events in addition/or in lieu of their semantic labels. To explore both possibilities, a within and across-group correlation is explored. This analysis is repeated for both the experimental stimuli and the MAESTRO dataset where correlation analyses between acoustic profiles and perceptual importance/soft labels are examined. Within and across-class analysis with samples spanning of all sound events reveal no significant correlation between perceptual importance and any of the acoustic attribute (Kendall's tau with permutation test, $p > 0.05$).

IV. DISCUSSION

In this article, we explore how human auditory perception is shaped by sound event identity in complex, naturalistic acoustic scenes. The study introduces an importance scoring framework that quantifies listeners' subjective judgments about which sound events are most perceptually prominent. These scores are analyzed alongside auditory salience estimates, pupillometry data, and cross-validation using the MAESTRO dataset. Together, these measures reveal systematic biases in how listeners prioritize sound categories, providing new insights into the cognitive and perceptual structure of auditory scene analysis.

Clear patterns emerge in how sound classes are treated perceptually. Alarms and speech sounds tend to receive higher importance scores, with distributions that are heavy-tailed toward high ranking [Fig. 2(A)]. In contrast, domestic and natural sounds are more often ranked low in importance,

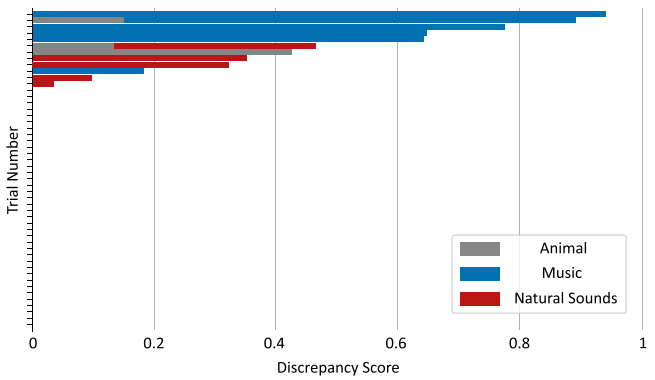

FIG. 4. Discrepancy score breakdown by trial and sound class.

https://doi.org/10.1121/10.0039710

with unimodal distributions skewed toward zero. Music and animal sounds show intermediate profiles indicating context-dependent variability. These trends are echoed in the MAESTRO dataset, which uses a different annotation approach but yields consistent category-level differences. Hierarchical clustering of these distributions reveals the same optimal grouping across both datasets [Fig. 2(B)].

An interesting observation arises for the music class, which yields a notable bimodal distribution, likely reflecting its unique functional ambiguity within everyday auditory environments. Unlike alarm or speech sounds, which are typically associated with clear behavioral affordances (e.g., action, comprehension, or response), music serves a wide range of purposes from focal entertainment to passive ambiance; and listeners adjust their perceptual engagement accordingly. In many environments, music is deliberately relegated to the background to facilitate other tasks—for example, in retail stores, waiting rooms, or elevators. In such contexts, listeners implicitly treat music as a non-informative, non-urgent auditory stream that helps shape mood or reduce perceived silence, but not one that requires immediate cognitive resources. This is supported by studies showing that background music is often semantically "deprioritized" and easily filtered out unless it is disruptive or unusually salient (Kämpfe *et al.*, 2011). Music can be effectively ignored under dual-task conditions or when listeners focus on more goal-relevant stimuli (Dalton and Fraenkel, 2012). This capacity to fade into the perceptual background may be partly due to its predictable structure, which allows the auditory system to allocate minimal attention for tracking it (Kidd *et al.*, 2005). On the other hand, music can function as a central feature of a scene—especially when it carries intentional communicative, social, or emotional content. For example, in a concert hall, film soundtrack, or ritual setting, music is not only salient but expected to be meaningful and affectively rich. Neuroscientific studies show that music under these conditions engages specialized networks including auditory, motor, and limbic regions, and may even synchronize across listeners at the neural level (Zatorre *et al.*, 2007; Koelsch, 2014). Furthermore, musical expectancy and familiarity can modulate how prominently music is perceived in complex scenes (Tillmann, 2012). Thus, the bimodal nature of importance scores in our data reflects this functional duality of music sounds.

Effects of privileged treatment, or biased responses, toward some sound classes are also evident in the consistency of listener judgments. Alarm and speech sounds elicit highly consistent responses across participants, with low variance and narrower distributions. In contrast, domestic and natural sounds show more variable and less predictable response patterns, with distributions that are skewed and heavy-tailed. Music and animal sounds again fall between these extremes, consistent with their flexible contextual roles in auditory scenes. These findings suggest that sound identity plays a central role in structuring perceptual judgments.

Crucially, these patterns cannot be explained solely by recognizability. While recognizing a sound is likely a prerequisite for judging its importance, our findings go beyond this baseline. First, our analysis focuses not only on whether a sound is reported, but on the *order* in which it is reported at the end of each trial. If recognition alone governed reporting behavior, one might expect variability in order or recency-based effects. Instead, we observe consistent response rankings across participants, with specific categories, such as speech and alarm sounds, systematically prioritized. This suggests that listeners do more than simply recognize events; they rank them according to perceived relevance. Second, we validate these findings using an independent dataset (MAESTRO) featuring a different stimulus set and real-time annotation methodology, which yields convergent category-level effects. Third, participant responses align closely with strong labels from the AudioSet corpus, indicating that the majority of sound events were correctly identified. These results argue against a recognizability-based account and support the view that importance judgments reflect structured perceptual biases linked to event identity and salience, rather than mere identification success.

Pupillometry data provide physiological support for this interpretation. Sound categories, such as speech and alarm sounds, in contrast to natural and domestic sounds, elicit earlier and larger pupil dilations; a known marker of increased cognitive effort and attentional engagement. Prior work has shown that pupil size tracks listening effort, task difficulty, and emotional shifts (Bradley *et al.*, 2008; Kahneman and Beatty, 1966; Liao *et al.*, 2016; Porter *et al.*, 2007). These responses, while involuntary, align well with the category-level patterns observed in importance and salience ratings, suggesting that certain sound types engage greater mental resources even in passive or non-instructed contexts.

Additional findings link importance to behavioral dynamics. Reaction times correlate with importance ratings: highly ranked sounds tend to be reported earlier. This mirrors sensory prioritization in other modalities, such as rapid visual saccades toward faces (Crouzet *et al.*, 2010) or fast olfactory responses to biologically relevant odors (Boesveldt *et al.*, 2010). Our data also show that sound duration plays a role: longer events, regardless of when they occur, are more likely to be judged important. This aligns with prior research suggesting that temporally persistent sounds are more easily segregated from the background and integrated into scene perception (Bregman, 1990; Seifritz *et al.*, 2002).

Taken together, these findings suggest that listeners implicitly rank sound events based on both their acoustic prominence and their semantic or communicative relevance. This pattern points toward an emergent perceptual structure that may help explain how complex scenes are parsed and understood.

We propose that this structure can be conceptualized as a perceptual continuum—one that reflects the degree to which a sound category conveys communicative or behaviorally relevant information. At one end of this continuum are sounds, such as alarms and speech, which are consistently prioritized, elicit stronger physiological responses,

and are reported with high consistency. At the other end are domestic and natural sounds, which are less often selected as prominent and show more variable responses. Music and animal sounds occupy an intermediate space. This gradient of perceived importance suggests an emergent division between perceptually foregrounded and backgrounded sounds—not imposed by the task, but revealed through consistent listener behavior and cross-modal markers.

This foreground/background distinction becomes especially compelling when considering salience and pupillometry data. Sound categories that are consistently treated as foreground are also those that attract more attention and

TABLE II. Detailed description of the auditory stimuli.

Stimulus ID	Sound classes present (strong labels)	Text description
1	Speech, animal, natural sounds, others	A male and a female speaking while sailing
2	Speech, animal, others	Animal quacking with a male speaking
3	Speech, animal	Dog barking with a male speaking
4	Domestic sounds	Electric shaver with background music
5	Natural sounds, others	Boat sailing
6	Speech, animal	A female speaking with cow mooing
7	Domestic sounds, natural sounds	A person washing dishes
8	Music, others	Dancing music in Spanish
9	Speech, alarm	Male speaking and ringing a bell
10	Speech, domestic sounds	Male speaking and using a blender
11	Speech, music	Male speaking with background music
12	Speech, alarm	Male speaking with a phone ringing
13	Speech, music, others	Sports game broadcast
14	Animal, music	Dog whining with background music
15	Music, others	Male beatboxing
16	Alarm	Distant alarm sound
17	Alarm, animal	Dog whining to an alarm sound
18	Animal, speech	Female speaking and a cat chirping
19	Animal, natural sounds	Water flowing with distant dog barking
20	Animal	Distant dog barking
21	Animal, speech	Male speaking with dog barking
22	Speech, domestic sounds	Male speaking and using a blender
23	Speech, domestic sounds	Female speaking with pots clashing
24	Speech, domestic sounds	Female speaking and using a blender
25	Speech, animal	Female speaking with a cat meowing and a background music
26	Natural sounds	Water flowing with a background music
27	Speech, music, others	Music ending and a male speaking
28	Natural sounds	Water flowing with birds chirping
29	Alarm	School bell ringing
30	Alarm	Distant alarm sound
31	Speech, domestic sounds	Electric shaver and a male speaking
32	Speech, alarm	Male speaking and a doorbell ringing
33	Speech, animal	Female speaking and a cat meowing
34	Animal	Dog barking
35	Speech, music, animal, others	A movie trailer with dog barking, male narration, and background music
36	Speech, music, others	Television (TV) show with male speaking and female singing
37	None	Music playing with unknown animal chirping
38	Speech, domestic sounds	Multiple people speaking with dishes clashing
39	Speech, Domestic Sounds	Female speaking with pots, dishes clashing
40	Speech, Domestic Sounds	Male speaking with dishes clashing
41	Music, others	Female singing and guitar playing
42	Speech, animal	Male speaking with birds cawing
43	Animal, natural sounds	Dog barking with water dripping
44	Music	Instrument playing with background music
45	Animal, music, others	Bird chirping with background beats
46	Speech, natural sounds	Water flowing and female speaking
47	Speech, animal, music	Birds chirping with male speaking and background music
48	Speech, domestic sounds	Food frying with male speaking and background music
49	Speech, animal	Female speaking and dog barking
50	Speech, animal, others	Female speaking with deer bleating

https://doi.org/10.1121/10.0039710

cognitive resources. This finding aligns with prior studies showing that attentional capture in audition depends on a sound's novelty, emotional content, and learned associations (Parmentier *et al.*, 2010; Bonmassar *et al.*, 2020). Our results extend this literature by demonstrating that identity-driven foregrounding occurs even in rich, naturalistic scenes and across multiple measurement modalities.

Despite these findings, key caveats and limitations must be acknowledged. First, the sample size of the dataset curated for this study is limited. A total of 500 s of experimental data and 1500 s of crowd-sourced salience data are collected, complemented by 14 992 s from MAESTRO. Scene durations are brief (10 s), constrained by AudioSet labeling. This limits our ability to assess long-term dynamics, such as habituation and event-boundary effects. The use of six root categories from the AudioSet ontology (Gemmeke *et al.*, 2017) provides a useful structure, but may mask contextual nuances—for example, speech might be expected in a café but intrusive in a concert hall. Our current framework captures broad trends, while future work could explore more context-sensitive and fine-grained classifications.

Nonetheless, the convergence of behavioral, physiological, and computational evidence presented here supports a strong conclusion: sound identity plays a decisive role in how auditory scenes are parsed and prioritized. The proposed importance score offers a novel tool for quantifying perceptual prominence in contextually rich environments. Combined with salience models and pupillometry, it reveals a structured continuum of perceptual weighting that helps explain how listeners navigate complex auditory scenes. These findings highlight the interplay of top-down expectations and bottom-up acoustic attributes in real-world listening and open new avenues for modeling attention and importance in multi-source auditory environments.

ACKNOWLEDGEMENTS

This work was supported in part by Office of Naval Research Grant Nos. N00014-23-1–2050 and N00014-23-1–2086.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Ethics Approval

The experiments in this study follow a study protocol approved by the Johns Hopkins Institutional Review Board (IRB). The IRB number is HIRB00009248. Informed consent was obtained from all participants.

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy restrictions.

APPENDIX

Among the fifty, 10 s stimuli used in the event identification experiment, 14 of them contain sounds other than the six sound classes [total duration = 108 s, duration mean (std) = 7.7 (2.8) s]. The variability among the six sound classes is as listed: alarm [contained in seven stimuli, total duration = 43 s, duration mean (std) = 6.2 (3.2) s], speech [contained in 30 stimuli, total duration = 111 s, duration mean (std) = 3.7 (2.4) s], animal [contained in 20 stimuli, total duration = 72 s, duration mean (std) = 3.6 (2.8) s], music [contained in 12 stimuli, total duration = 93 s, duration mean (std) = 7.8 (3.3) s], domestic sounds [contained in 11 stimuli, total duration = 51 s, duration mean (std) = 4.6 (3.1) s], natural sounds [contained in eight stimuli, total duration = 68 s, duration mean (std) = 8.5 (2.9) s].

Table II is a detailed description of the 10 s auditory stimuli used in the event identification experiment.

Bindemann, M., Scheepers, C., Ferguson, H. J., and Burton, A. M. (**2010**). "Face, body, and center of gravity mediate person detection in natural scenes," J. Exp. Psychol.: Hum. Percept. Perform. **36**(6), 1477–1485.
Bizley, J. K., Walker, K. M. M., Nodal, F. R., King, A. J., and Schnupp, J. W. H. (**2013**). "Auditory cortex represents both pitch judgments and the corresponding acoustic cues," Curr. Biol. **23**(7), 620–625.
Boesveldt, S., Frasnelli, J., Gordon, A. R., and Lundström, J. N. (**2010**). "The fish is bad: Negative food odors elicit faster and more accurate reactions than other odors," Biol. Psychol. **84**(2), 313–317.
Bonmassar, C., Widmann, A., and Wetzel, N. (**2020**). "The impact of novelty and emotion on attention-related neuronal and pupil responses in children," Dev. Cogn. Neurosci. **42**, 100766.
Bradley, M. M., Miccoli, L., Escrig, M. A., and Lang, P. J. (**2008**). "The pupil as a measure of emotional arousal and autonomic activation," Psychophysiology **45**(4), 602–607.
Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA), pp. 1–773.
Broderick, M. P., Anderson, A. J., and Lalor, E. C. (**2019**). "Semantic context enhances the early auditory encoding of natural speech," J. Neurosci. **39**(38), 7564–7575.
Crouzet, S. M., Kirchner, H., and Thorpe, S. J. (**2010**). "Fast saccades toward faces: Face detection in just 100 ms," J. Vision **10**(4), 1–17.
Dalton, P., and Fraenkel, N. (**2012**). "Gorillas we have missed: Sustained inattentional deafness for dynamic events," Cognition **124**(3), 367–372.
Darwin, C. J. (**1997**). "Auditory grouping," Trends Cogn. Sci. **1**(9), 327–333.
de Haas, B., Iakovidis, A. L., Schwarzkopf, D. S., and Gegenfurtner, K. R. (**2019**). "Individual differences in visual salience vary along semantic dimensions," Proc. Natl. Acad. Sci. U.S.A. **116**, 11687–11692.
de Leeuw, J. (**2015**). "jsPsych: A JavaScript library for creating behavioral experiments in a Web browser," Behav. Res. **47**(1), 1–12.
Driver, J., and Baylis, G. C. (**1995**). "One-sided edge assignment in vision: 2. Part decomposition, shape description, and attention to objects," Curr. Dir. Psychol. Sci. **4**(6), 201–206.
Frades, I., and Matthiesen, R. (**2010**). "Overview on techniques in cluster analysis," Meth. Mol Biol. **593**, 81–107.
Frank, M. C., Amso, D., and Johnson, S. P. (**2014**). "Visual search and attention to faces during early infancy," J. Exp. Child Psychol. **118**(1), 13–26.
Gemmeke, J. F., Ellis, D. P. W., Freedman, F., Jansen, A., Lawrence, W., Moore, C., Plakal, M., Ritter, M., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (**2017**). "Audio Set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, pp. 776–780.
Goldstein, J. L. (**1973**). "An optimum processor theory for the central formation of the pitch of complex tones," J. Acoust. Soc. Am. **54**, 1496–1516.

Griffiths, T. D., and Warren, J. D. (**2004**). "What is an auditory object?," Nat. Rev. Neurosci. **5**(11), 887–892.
Gureckis, T., Martin, J., McDonnell, J., Rich, A., Markant, D., Coenen, A., Halpern, D., Hamrick, J. B., and Chan, P. (**2016**). "psiTurk: An open-source framework for conducting replicable behavioral experiments online," Behav. Res. **48**(3), 829–842.
Gygi, B., and Shafiro, V. (**2011**). "The incongruency advantage for environmental sounds presented in natural auditory scenes," J. Exp. Psychol. Hum. Percept. Perf. **37**(2), 551–565.
Hoffman, D. D., and Singh, M. (**1997**). "Salience of visual parts," Cognition **63**(1), 29–78.
Huang, N., and Elhilali, M. (**2017**). "Auditory salience using natural soundscapes," J. Acoust. Soc. Am. **141**(3), 2163–2176.
Huang, N., and Elhilali, M. (**2020**). "Push-pull competition between bottom-up and top-down auditory attention to natural soundscapes," eLife **9**, e52984.
Hwang, A. D., Wang, H.-C., and Pomplun, M. (**2011**). "Semantic guidance of eye movements in real-world scenes," Vision Res. **51**(10), 1192–1205.
Isherwood, S. J., and McKeown, D. (**2017**). "Semantic congruency of auditory warnings," Ergonomics **60**(7), 1014–1043.
Kahneman, D., and Beatty, J. (**1966**). "Pupil diameter and load on memory," Science **154**(3756), 1583–1585.
Kämpfe, J., Sedlmeier, P., and Renkewitz, F. (**2011**). "The impact of background music on adult listeners: A meta-analysis," Psychol. Music **39**(4), 424–448.
Kaya, E. M., and Elhilali, M. (**2017**). "Modelling auditory attention," Philos. Trans. R. Soc. London, Ser. B **372**(1714), 20160101.
Kidd, G., Jr., Arbogast, T. L., Mason, C. R., and Gallun, F. J. (**2005**). "The advantage of knowing where to listen," J. Acoust. Soc. Am. **118**(6), 3804–3815.
Koelsch, S. (**2005**). "Neural substrates of processing syntax and semantics in music," Curr. Opin. Neurobiol. **15**(2), 207–212.
Koelsch, S. (**2014**). "Brain correlates of music-evoked emotions," Nat. Rev. Neurosci. **15**(3), 170–180.
Kothinti, S. R., and Elhilali, M. (**2023**). "Are acoustics enough? Semantic effects on auditory salience in natural scenes," Front. Psychol. **14**, 1276237.
Kothinti, S. R., Huang, N., and Elhilali, M. (**2021**). "Auditory salience using natural scenes: An online study," J. Acoust. Soc. Am. **150**(4), 2952–2966.
Kret, M. E., and Sjak-Shie, E. E. (**2019**). "Preprocessing pupil size data: Guidelines and code," Behav. Res. **51**(3), 1336–1342.
Langley, M. D., and McBeath, M. K. (**2023**). "Vertical attention bias for tops of objects and bottoms of scenes," J. Exp. Psychol.: Hum. Percept. Perform. **49**(10), 1281–1295.
Leaver, A. M., and Rauschecker, J. P. (**2010**). "Cortical representation of natural complex sounds: Effects of acoustic features and auditory object category," J. Neurosci. **30**(22), 7604–7612.
Liao, H. I., Yoneya, M., Kidani, S., Kashino, M., and Furukawa, S. (**2016**). "Human pupillary dilation response to deviant auditory stimuli: Effects of stimulus properties and voluntary attention," Front. Neurosci. **10**(43), 43.
Martin-Morato, I., and Mesaros, A. (**2023**). "Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation," IEEE/ACM Trans. Audio. Speech Lang. Process. **31**, 902–914.
McMullin, M. A., Kumar, R., Higgins, N. C., Gygi, B., Elhilali, M., and Snyder, J. S. (**2024**). "Preliminary evidence for global properties in human listeners during natural auditory scene perception," Open Mind **8**, 333–365.
Norman-Haignere, S., Kanwisher, N. G., and McDermott, J. H. (**2015**). "Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition," Neuron **88**(6), 1281–1296.
Nuthmann, A., Schütz, I., and Einhäuser, W. (**2020**). "Salience-based object prioritization during active viewing of naturalistic scenes in young and older adults," Sci. Rep. **10**(1), 22057.
Odegaard, B., Wozny, D. R., and Shams, L. (**2015**). "Biases in visual, auditory, and audiovisual perception of space," PLoS Comput. Biol. **11**(12), e1004649.
Oh, Y., Zuwala, J. C., Salvagno, C. M., and Tilbrook, G. A. (**2022**). "The impact of pitch and timbre cues on auditory grouping and stream segregation," Front. Neurosci. **15**, 725093.
Parmentier, F. B., Elsley, J. V., and Ljungberg, J. K. (**2010**). "Behavioral distraction by auditory novelty is not only about novelty: The role of the distracter's informational value," Cognition **115**(3), 504–511.
Perani, D., Saccuman, M. C., Scifo, P., Spada, D., Andreolli, G., Rovelli, R., Baldoli, C., and Koelsch, S. (**2010**). "Functional specializations for music processing in the human newborn brain," Proc. Natl. Acad. Sci. U. S.A. **107**(10), 4758–4763.
Porter, G., Troscianko, T., and Gilchrist, I. D. (**2007**). "Effort during visual search and counting: Insights from pupillometry," Q. J. Exp. Psychol. **60**(2), 211–229.
Schomaker, J., Walper, D., Wittmann, B. C., and Einhäuser, W. (**2017**). "Attention in natural scenes: Affective-motivational factors guide gaze independently of visual salience," Vision Res. **133**, 161–175.
Seifritz, E., Neuhoff, J. G., Bilecen, D., Scheffler, K., Mustovic, H., Schächinger, H., Elefante, R., and Di Salle, F. (**2002**). "Neural processing of auditory looming in the human brain," Curr. Biol. **12**(24), 2147–2151.
Spain, M., and Perona, P. (**2011**). "Measuring and predicting object importance," Int. J. Comput. Vis. **91**(1), 59–76.
Tian, X., Xu, K., Yang, X., Du, L., Yin, B., and Lau, R. W. (**2022**). "Bi-directional object-context prioritization learning for saliency ranking," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, pp. 5872–5881.
Tillmann, B. (**2012**). "Music and language perception: Expectations, structural integration, and cognitive sequencing," Top. Cogn. Sci. **4**(4), 568–584.
Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., and von der Heydt, R. (**2012**). "A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization," Psychol. Bull. **138**(6), 1172–1217.
Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (**2018**). "GLUE: A multi-task benchmark and analysis platform for natural language understanding," arXiv:1804.07461.
Wang, H. C., and Pomplun, M. (**2012**). "The attraction of visual attention to texts in real-world scenes," J. Vision **12**(6), 26.
Wang, J., Chandler, D. M., and le Callet, P. (**2010**). "Quantifying the relationship between visual salience and visual importance," Proc. SPIE **7527**, 75270K.
Wang, K., and Shamma, S. A. (**1994**). "Self-normalization and noise-robustness in early auditory representations," IEEE Trans. Speech Audio Process. **2**, 421–435.
Wierzchoń, S. T., and Kłopotek, M. A. (**2018**). *Modern Algorithms of Cluster Analysis*, Studies in Big Data (Springer, Cham, Switzerland), Vol. 34.
Wu, C. C., Wick, F. A., and Pomplun, M. (**2014**). "Guidance of visual attention by semantic information in real-world scenes," Front. Psychol. **5**, 69374.
Zatorre, R. J., Chen, J. L., and Penhune, V. B. (**2007**). "When the brain plays music: Auditory–motor interactions in music perception and production," Nat. Rev. Neurosci. **8**(7), 547–558.
Zhou, Z., and Geng, J. J. (**2024**). "Learned associations serve as target proxies during difficult but not easy visual search," Cognition **242**, 105648.
Zwicker, E., Fastl, H., Widmann, U., Kurakata, K., Kuwano, S., and Namba, S. (**1991**). "Program for calculating loudness according to DIN 45631 (ISO 532B)," J. Acoust. Soc. Jpn. (E) **12**(1), 39–42.