



Citation: McMullin, M. A., Kumar, R., Higgins, N. C., Gygi, B., Elhilali, M., & Snyder, J. S. (2024). Preliminary Evidence for Global Properties in Human Listeners During Natural Auditory Scene Perception. *Open Mind: Discoveries in Cognitive Science*, 8, 333–365. https://doi.org/10.1162/opmi_a_00131

DOI:
https://doi.org/10.1162/opmi_a_00131


Received: 10 March 2023
Accepted: 10 February 2024

Competing Interests: The authors declare no conflict of interests.

Corresponding Author:
Margaret A. McMullin
mcmulm1@unlv.nevada.edu

Copyright: © 2024
Massachusetts Institute of Technology
Published under a Creative Commons
Attribution 4.0 International
(CC BY 4.0) license

Preliminary Evidence for Global Properties in Human Listeners During Natural Auditory Scene Perception

Margaret A. McMullin¹, Rohit Kumar², Nathan C. Higgins³ , Brian Gygi⁴, Mounya Elhilali², and Joel S. Snyder¹

¹Department of Psychology, University of Nevada, Las Vegas, Las Vegas, NV, USA

²Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA

³Department of Communication Sciences & Disorders, University of South Florida, Tampa, FL, USA

⁴East Bay Institute for Research and Education, Martinez, CA, USA

Keywords: auditory scene perception, acoustic analysis, natural scenes, computational modeling

ABSTRACT

Theories of auditory and visual scene analysis suggest the perception of scenes relies on the identification and segregation of objects within it, resembling a detail-oriented processing style. However, a more global process may occur while analyzing scenes, which has been evidenced in the visual domain. It is our understanding that a similar line of research has not been explored in the auditory domain; therefore, we evaluated the contributions of high-level global and low-level acoustic information to auditory scene perception. An additional aim was to increase the field's ecological validity by using and making available a new collection of high-quality auditory scenes. Participants rated scenes on 8 global properties (e.g., open vs. enclosed) and an acoustic analysis evaluated which low-level features predicted the ratings. We submitted the acoustic measures and average ratings of the global properties to separate exploratory factor analyses (EFAs). The EFA of the acoustic measures revealed a seven-factor structure explaining 57% of the variance in the data, while the EFA of the global property measures revealed a two-factor structure explaining 64% of the variance in the data. Regression analyses revealed each global property was predicted by at least one acoustic variable ($R^2 = 0.33\text{--}0.87$). These findings were extended using deep neural network models where we examined correlations between human ratings of global properties and deep embeddings of two computational models: an object-based model and a scene-based model. The results support that participants' ratings are more strongly explained by a global analysis of the scene setting, though the relationship between scene perception and auditory perception is multifaceted, with differing correlation patterns evident between the two models. Taken together, our results provide evidence for the ability to perceive auditory scenes from a global perspective. Some of the acoustic measures predicted ratings of global scene perception, suggesting representations of auditory objects may be transformed through many stages of processing in the ventral auditory stream, similar to what has been proposed in the ventral visual stream. These findings and the open availability of our scene collection will make future studies on perception, attention, and memory for natural auditory scenes possible.



INTRODUCTION

Every day, our auditory system undertakes the complex task of organizing various incoming sounds in a coherent manner, allowing us to not only decipher where sounds are coming from, but to also interpret what we are listening to. For example, when conversing with a friend at a noisy café, the auditory system maintains the exceptional ability to segregate the noisy background (e.g., music, espresso machines, other conversations) from a friend's voice and further group the sound components of their speech into an intelligible stream of words. The process of perceptually segregating and grouping numerous acoustic objects is known as 'Auditory Scene Analysis' (ASA; Bregman, 1990).

Historically, theories of both auditory and visual scene analysis have suggested that our perception of a scene relies on the identification and segregation of multiple objects within it, resembling a detail-oriented processing style (Biederman, 1987; Bregman, 1990). However, it is possible that a more global process may also occur when observers evaluate auditory scenes. In the visual modality, there is evidence for global properties that enable visual scenes to be rapidly recognized, even without recognition of individual objects comprising the scene (Greene & Oliva, 2009a, 2009b; Ross & Oliva, 2010). The significance of the scene-centered approach proposed by Greene and Oliva (2009a) is that the representation exists at the level of the entire scene and not just individual objects. Instead of just building a visual representation using local geometric information about individual objects, the visual system also uses global properties that provide information about the scene's structure, function, and overall layout to guide perception. The global properties identified by Oliva and colleagues fall into three categories: structural properties (openness, expansion, mean depth), constancy properties (temperature, transience), and functional properties (concealment, navigability). Greene and Oliva (2009b) asked participants to view a series of visual scenes and indicate whether each scene was consistent with a basic-level category (e.g., identifying a scene as a mountain or waterfall) or a global category (e.g., identifying a scene as an open environment or hot place). The authors controlled the presentation rate of each image to maintain participant classification accuracy at 75%. The results indicated participants required less viewing time to perform the global categorization task at the same level of accuracy as the basic-level task. Other studies have demonstrated the importance of global image features (Greene & Oliva, 2009a; Oliva & Torralba, 2006; Ross & Oliva, 2010) and collectively suggest that these global properties could act as automatic heuristics to analyze natural visual scenes.

Recent work has also highlighted the importance of individual objects in scene categorization. Wiesmann and Vö (2022) asked participants to rapidly categorize visual scenes with various levels of spatial resolutions (ranging from low to high) into a superordinate category (indoor or outdoor) and also a basic-level category (e.g., kitchen, living room, beach, forest, etc.). Participants were able to perform this task well above chance, but performance was higher for scenes with higher spatial resolutions. Furthermore, participants were even more accurate at identifying individual objects in scenes with higher resolutions, suggesting rapid scene categorization cannot be explained by global properties alone, but likely also relies on information provided by individual objects. This finding was extended by Wiesmann and Vö (2023), who showed that participants were able to successfully identify scenes that were reduced to a single object (e.g., TV, car, bed, etc.) at speeds less than 50 msec. These findings demonstrate the complexity of visual scene processing and consider the roles played by the setting of scenes as well as the individual objects within them. Our primary research goal was to evaluate the contribution of high-level semantic knowledge and low-level acoustic information during auditory scene perception. As previously mentioned, many studies have demonstrated the usefulness of global information in rapid scene categorization, but more recent

endeavors suggest individual objects also play a crucial role in scene categorization. In the auditory domain, the role of high-level semantic information and low-level acoustic features has been explored in the change deafness literature. Characterized as a perceptual error, change deafness is the inability to detect changes in auditory scenes (Snyder et al., 2012). This error is useful in our study of ASA because it informs us of our auditory system's limitations. To study change deafness, participants are typically presented with a simultaneous array of sounds (e.g., dog barking, piano, phone ringing, bell). After a short interruption (usually white noise), the scene is presented again. Finally, participants must indicate whether the second scene was the same as or different than the first scene. In one such study, Gregg and Samuel (2009) presented participants with two types of different trials (i.e., the scene changed in some way from the first to second presentation). These trials exhibited either a within-category change (e.g., a small dog bark changing to a large dog bark) or a between-category change (e.g., a small dog bark changing to a bell chime). Results demonstrated that participants' ability to detect a within-category change was significantly worse than their ability to detect a between-category change. As the low-level acoustic features were controlled for, the authors hypothesized that semantic information is useful when constructing representations of auditory scenes. They further observed that listeners used both high-level semantic information and low-level acoustic information when constructing auditory representations of sounds, but the low-level acoustic information was not used to the same extent as semantic knowledge of sounds.

Additional research has also evaluated the influence of acoustic features of sounds on listeners' ability to identify and discriminate recognizable objects, both in isolation and when presented concurrently with complex auditory scenes. Leech et al. (2009) addressed the possible existence of semantic knowledge-driven expectancies about auditory scenes. In their study, participants were presented with multiple target sounds embedded into an auditory scene. The target sounds were either congruent (e.g., the target being a goat and the background being a farm) or incongruent (e.g., the target being a goat and the background being a casino) with the auditory scene in which the sounds were embedded. An acoustic analysis of all target sounds and background auditory scenes was conducted to evaluate whether acoustic similarity or dissimilarity between the targets and backgrounds may have influenced target identification. Participants more accurately identified target sounds that were contextually incongruent with the background scenes and the acoustic variables that significantly influenced this effect were correlogram-based pitch measures and peak autocorrelation statistics. However, since acoustic similarity was not an exclusive predictor of target congruency or incongruency with the background scene, the findings from this study suggest that high-level semantic factors may significantly influence listeners' ability to detect and identify meaningful sounds within complex auditory scenes.

The tasks just described are useful for the study of ASA. However, they typically use a mixture of simultaneously presented sounds from different recordings (Gregg & Samuel, 2008, 2009; Leech et al., 2009). This presents a fundamental limitation to studies of this type: the stimuli are somewhat artificial in nature, especially since some of the sound combinations may not typically occur in the real-world. An example of a study using more naturalistic sounds is by McDermott and Simoncelli (2011), which examined sound texture perception using a computational model of the human auditory system. Sound textures, which are the result of numerous similar acoustic events occurring in succession (e.g., rainstorm, galloping horses), were processed using an auditory model based on the tuning properties of neurons from the cochlea to the thalamus. To better understand how sound textures are represented in the brain, the authors then synthesized the sound textures based on the output of their

model (i.e., the statistics of real-world sounds). They hypothesized that if the novel synthesized sound textures were statistically matched with those of the real-world sounds, then the brain should be able to achieve texture recognition due to the synthesized signal sounding like a version of the originally presented sound texture. In a series of behavioral experiments, they found that synthetic sound textures were recognizable to participants but eliminating some of the statistics in the model reduced performance. Additionally, the authors modified the model so that it was less representative of the mammalian auditory system, which resulted in reduced recognizability of the synthetic sound textures. Some of the synthesized sounds (e.g., wind chimes, tapping rhythm, and a person speaking English) were not recognizable, though. These findings suggest that sound texture perception arises from the recognition of simple statistics in early auditory representations, which are potentially computed in neural populations downstream from the peripheral auditory system. Ultimately, the results of this study are important to our understanding of how the auditory system analyzes naturalistic sounds and could inform us of how the auditory system processes more complex stimuli, like naturalistic auditory scenes. By using naturalistic auditory scene stimuli, we hope to further increase the range of ecologically valid stimuli and abilities under study in the field of auditory perception.

Our secondary research goal was to record and use real-world auditory scenes. One major limitation in the current body of literature is the consistent use of pure tones, noise bursts, or artificially contrived auditory scenes as stimuli to study ASA. While using such stimuli has revealed much about the fundamental mechanisms of ASA and auditory perceptual awareness, the findings resulting from this work have limited power in educating us about natural auditory scene processing. In the field of visual scene perception, the use of naturalistic stimuli is highly evident (Greene & Oliva, 2009a, 2009b, 2010; Hansen et al., 2018; Harel et al., 2016; Ross & Oliva, 2010). This is perhaps the case because there are numerous large databases of natural visual scenes openly available for public use (Geisler & Perry, 2011; Xiao et al., 2010). While there are some databases that include high quality clips of individual sound objects (e.g., a single dog bark; Gygi & Shafiro, 2010), no database of high quality, real-world auditory scenes currently exists to our knowledge. To address this issue, we recorded a relatively large volume of audio/visual scenes, which are available for other researchers to use (see [Methods](#) for the database link).

By embracing the complexity of natural auditory scenes, a third goal of this study was to explore the relationship between complex auditory scenes and global perception through the lens of deep neural network models. In many ways, these artificial models parallel biological processing of sensory information, particularly in their ability to provide rich mappings from low-level features to more complex decompositions of sensory inputs (Richards et al., 2019; Saxe et al., 2021). Hierarchical feature analysis and abstraction is a hallmark of both deep neural networks and sensory systems, particularly in the ventral stream. This approach has been embraced in vision studies which showed that deep neural networks trained on identifying objects reveal progressions of hierarchical processing from simple features like edges and textures to more complex object categories mimicking gradients of processing in the visual what stream (Güçlü & van Gerven, 2015). Similar insights about processing gradients in the auditory stream have been gleaned using deep learning methods that not only shed light on cortical processing hierarchies, but also specialized processing in non-primary cortical pathways for processing speech and music sounds (Kell et al., 2018; Wang et al., 2022). In the current study, deep neural networks were also used as an investigative tool to infer relationships between behavioral responses to natural auditory scenes and nonlinear abstractions of these scenes extracted by learned models. This approach extends the correlations inferred

from low-level acoustic features and sheds light on the balance between recognizing discrete sound events and perceiving global characteristics of a natural scene.

Purpose of Present Study

The present study aimed to evaluate the contributions of high-level global properties and low-level acoustic features to natural auditory scene perception. Participants listened to 200 auditory scenes and made a series of global property judgments on them. Additionally, we conducted an acoustic analysis on all 200 scenes with the goal of understanding how these features are related to global processing of auditory scenes. We predicted there would be a general consistency on all eight global property ratings of each auditory scene across participants, which was measured using intraclass correlations of each rating scale. We conducted two separate factor analyses on the average global property ratings and acoustic measures of each scene to determine the number of factors that characterize the variability found within scene judgments and within the array of acoustic features. An additional eight multiple linear regression analyses were also conducted to predict performance on the global property rating task based on the acoustic features of the scenes. We did not originally plan on conducting this analysis; however, we decided it was more helpful for directly testing the relationships between the average global property ratings and acoustic measures of each scene.

Finally, we deploy two state-of-the-art deep neural networks, both carefully tailored to map the intricate acoustic signals onto high-dimensional embeddings. The first model is optimized to pinpoint specific sound events within a scene—whether it be the chirping of a bird, the chatter of human voices, or the distant bark of a dog. The second model is trained to identify the general setting of a scene, be it a bustling kitchen or a quiet office. This analysis aims to address a central question in auditory perception: Is the broader context of an auditory scene captured first, subsequently aiding in the recognition of specific sound events (global-to-local processing)? Or do we first detect individual sounds and then integrate them to make sense of the scene as a whole (local-to-global processing)?

METHODS

Ethics Statement

All procedures were approved by the University of Nevada, Las Vegas (UNLV) Institutional Review Board. All de-identified experimental data, analysis techniques, protocols, and stimuli are available on this project's Open Science Framework repository (<https://osf.io/zj4xe/>).

Auditory Scene Collection and Database

We recorded and processed 200 auditory scenes from various locations across the United States we believed to represent typical environments humans are exposed to, such as parks, classrooms, hiking trails, city streets, forests, and cafes. There are multiple recordings in similar but distinct settings (e.g., different cafes, parks, etc.). The scenes differ in the number of sound sources, but each scene contains more than one source (e.g., talking, wind, a bird chirping). Using a standardized recording procedure, we placed a Zoom Q8 camcorder (Zoom North America, Inc., Hauppauge, NY) on a tripod and made one-minute recordings at each location, noting various aspects of the scene, such as the date, time of day, cardinal direction the camcorder was pointing, temperature (°F), sounds observed, and any additional notes about the recording. After each recording session, the field notes were digitized into a spreadsheet for ease of organization and file identification. We then listened to each recording and confirmed

all sounds identified in the field notes were heard in the recording. Next, we edited each minute-long recording into a four-second-long version which best characterized the scene location and included more than one sound object. Our collection of auditory scenes, including detailed descriptions of each scene can be found on our project's OSF repository.

Participants

68 English-speaking adults (48 female) aged 18–25 (mean = 21.19 years) with no known hearing, visual, or neurological deficits were recruited from the UNLV participant pool and across the United States. Participants from the participant pool were reimbursed with course credit and participants external to the university volunteered for no compensation. In total, 142 participants were excluded from this study because they did not speak English, did not complete the experiment, did not have normal hearing, had any type of severe neurological or psychiatric disorder (e.g., schizophrenia, bipolar disorder, stroke, traumatic brain injury), or failed the headphone check, compliance check, and/or attention check (see below for criteria and descriptions of tasks).

Stimuli

We selected a total of 200 auditory scenes for this experiment based on common heuristics for exploratory factor analysis sample sizes (Pearson & Mundform, 2010). Stimuli consisted of 200 naturalistic auditory scenes originating from our database of acoustic scenes. Each scene was four seconds in length and matched for RMS amplitude. A linear on-ramp from zero amplitude was imposed on the first and last 10 msec of each sound clip to avoid introducing artifacts due to abrupt sound onsets and offsets (Gregg et al., 2014, 2017; Gregg & Samuel, 2008, 2009).

Procedure

Participants were provided a link to complete the experiment online via Qualtrics (Qualtrics, Provo, UT). Experiment links can be found on our project's OSF repository. Informed consent was obtained online from each participant before they began the experiment. Participants were asked to complete the study on a desktop or laptop computer using headphones and while in a quiet environment. In total, the experiment took 60–120 minutes to complete. The experiment consisted of four sections: 1) a headphone check, 2) the global property rating task, 3) a compliance and attention check, and 4) a demographic questionnaire.

Headphone Check. Because this was an online study of auditory perception, we tested each participant's sound quality by administering a headphone check. This test consists of 6 trials of a 3-AFC intensity discrimination task (Woods et al., 2017). Participants were asked to indicate which tone had the lowest volume by selecting one of three button options labeled "Tone 1", "Tone 2", or "Tone 3." Any participants who did not correctly answer five out of the six trials were excluded from the study.

Global Property Rating Task. Each participant was asked to judge all 200 scenes on four different global properties on a Likert scale ranging from 1 (lowest extreme) to 7 (highest extreme) (see Figure 1 for full descriptions of each rating scale provided to participants). The global properties chosen for this experiment were based on the work of Greene, Oliva, and colleagues (Greene & Oliva, 2009a, 2009b; Oliva & Torralba, 2001, 2006) and were properties we thought might be perceived in the auditory domain. Although we were unsure if participants would be able to hear some of the properties, such as season and temperature, we wanted to have a good number of properties whose ratings could be submitted into the factor

Open vs. Enclosed						
How OPEN vs. CLOSED is this scene? Please rate this on a scale ranging from 1 (extremely open) to 7 (extremely closed).						
○	○	○	○	○	○	○
Extremely OPEN	Moderately OPEN	Slightly OPEN	Neither OPEN nor CLOSED	Slightly CLOSED	Moderately CLOSED	Extremely CLOSED
Outdoor vs. Indoor						
How OUTDOOR vs. INDOOR is this scene? Please rate this on a scale ranging from 1 (extremely outdoor) to 7 (extremely indoor).						
○	○	○	○	○	○	○
Extremely OUTDOOR	Moderately OUTDOOR	Slightly OUTDOOR	Neither OUTDOOR nor INDOOR	Slightly INDOOR	Moderately INDOOR	Extremely INDOOR
Natural vs. Human-Influenced						
How NATURAL vs. HUMAN-INFLUENCED is this scene? Please rate this on a scale ranging from 1 (extremely natural) to 7 (extremely human-influenced).						
○	○	○	○	○	○	○
Extremely NATURAL	Moderately NATURAL	Slightly NATURAL	Neither NATURAL nor HUMAN-INFLUENCED	Slightly HUMAN-INFLUENCED	Moderately HUMAN-INFLUENCED	Extremely HUMAN-INFLUENCED
Temperature						
How WARM vs. COLD is this scene? Please rate this on a scale ranging from 1 (extremely warm) to 7 (extremely cold) .						
○	○	○	○	○	○	○
Extremely WARM	Moderately WARM	Slightly WARM	Neither WARM nor COLD	Slightly COLD	Moderately COLD	Extremely COLD
Season						
What season does this scene sound like?						
○	○	○	○	○	○	○
Winter	Between Winter and Spring	Spring	Between Spring and Summer	Summer	Between Summer and Fall	Fall
Transience						
How much change is happening in this scene? Please rate this on a scale ranging from 1 (no change) to 7 (extreme amount of change).						
○	○	○	○	○	○	○
No Change	Very Little Change	Slight Amount of Change	Moderate Amount of Change	Medium-High Amount of Change	High Amount of Change	Extreme Amount of Change
Navigability						
How difficult or easy would it be to navigate through this scene? Please rate this on a scale ranging from 1 (extremely difficult) to 7 (extremely easy).						
○	○	○	○	○	○	○
Extremely Difficult	Moderately Difficult	Slightly Difficult	Neither Difficult or Easy	Slightly Easy	Moderately Easy	Extremely Easy
Sparseness						
How much sound is in this scene? Please rate this on a scale ranging from 1 (very little sound) to 7 (high amount of sound).						
○	○	○	○	○	○	○
No Sound	Very Little Sound	Slight Amount of Sound	Moderate Amount of Sound	Medium-High Amount of Sound	High Amount of Sound	Extreme Amount of Sound

Figure 1. Global Property Rating Scales.

analysis. Participants were allowed to listen to each scene as many times as they needed to make each of the four judgments. The eight global property judgments were pseudo-randomized across participants using a Latin square design, with each global judgment type only appearing in each possible position once (see Table 1 for order of questions in each condition). Participants were randomly assigned to one of eight condition groups. Groups 1–4 (n = 32) made the following global property judgments on each scene: Open vs. Enclosed, Outdoor vs. Indoor, Natural vs. Human-Influenced, and Temperature, and Groups 5–8 (n = 36) made judgments on each scene’s Season, Transience, Navigability, and Sparseness.

Compliance and Attention Check. We used a set of questions from Mehr et al. (2018) to ensure participants were adequately attending to the experimental task. The following question was

Table 1. The order of rating scales questions in each condition.

Group	Order of Rating Scale Questions			
1	Open vs. Enclosed	Outdoor vs. Indoor	Natural vs. HI	Temperature
2	Outdoor vs. Indoor	Open vs. Enclosed	Temperature	Natural vs. HI
3	Natural vs. HI	Temperature	Open vs. Enclosed	Outdoor vs. Indoor
4	Temperature	Natural vs. HI	Outdoor vs. Indoor	Open vs. Enclosed
5	Season	Transience	Navigability	Sparseness
6	Transience	Season	Sparseness	Navigability
7	Navigability	Sparseness	Season	Transience
8	Sparseness	Navigability	Transience	Season

Note. HI = Human-Influenced.

dispersed throughout the global property rating task: 1) “What color is the sky? Please answer this incorrectly, on purpose, by choosing RED instead of blue.”, with the response options of “Green,” “Red,” “Blue,” or “Yellow.” The correct response option (“Red”) was changed upon each presentation (e.g., the correct response was only presented in each answer slot once). Any participant who did not select this response option was excluded.

Upon completion of the rating task, participants were asked the following compliance questions:

- 1) “People are working on this task in many different places. Please tell us about the place you worked on this task. Please answer honestly.” The response options for this question were: “I worked on this study in a very noisy place, I worked on this study in a somewhat noisy place, I worked on this study in a somewhat quiet place, or I worked on this study in a very quiet place.” Any participant who answered with “I worked on this study in a very noisy place” or “I worked on this study in a somewhat noisy place” was excluded.
- 2) “Please tell us if you had difficulty loading the sounds. Please answer honestly.” The response options for this question were “Yes” or “No.” Any participant who responded with “Yes” was excluded.
- 3) “How carefully did you complete this experiment? Please answer honestly. The response options for this question were: “Not at all carefully,” “Slightly carefully,” “Moderately carefully,” “Quite carefully,” or “Very carefully.” Any participant who answered with “Not at all carefully,” “Slightly carefully,” or “Moderately carefully” were excluded.

Demographic Questionnaire. Lastly, participants completed a demographics questionnaire which asked about their health history and engagement with music. Additional questions were asked about participants’ auditory environment (e.g., time spent in various environments). Data collected from this questionnaire has not been included in the analyses reported here.

Acoustic Analysis

To quantitatively gauge how low-level information may be used to understand auditory scenes, an extensive acoustic analysis was conducted on all 200 auditory scenes. The acoustic

analyses chosen for this study have been used in various prior studies (Ballas, 1993; Gygi et al., 2007; Houtgast & Steeneken, 1985; Leech et al., 2009; Slaney, 1995) and were executed in MATLAB (MATLAB version 9.10 (R2021a)). A description of each acoustic measure is listed below.

Envelope-based Intensity and Rhythm Measures. (1) Long-term RMS/Pause corrected RMS, which indicates the amount of silence present within each auditory scene; (2) number of peaks, where a peak is defined as a point in the vector that has a greater amplitude than the previous point by at least 80% of the range of amplitudes present in the vector; (3) number of bursts, showing an increase in amplitude of at least 4 dB lasting at least 20 msec (Ballas, 1993); (4) total duration; and (5) burst duration/total duration, a measure of how “rough” the envelope is.

Autocorrelation Pitch Statistics. (1) Number of peaks; (2) maximum peak; and (3) standard deviation (SD) of the peaks. In this autocorrelation function, the peaks express periodicities in the waveform. The distribution of periodicities across various frequencies as well as the strength of a periodicity are evaluated in this measure.

Correlogram-based Pitch Measures. (1) mean pitch; (2) median pitch; (3) SD pitch; (4) maximum pitch; (5) mean pitch salience; and (6) maximum pitch salience. This measure evaluates pitch and pitch salience by autocorrelating in sliding 16 msec time windows.

Moments of the Spectrum. (1) mean; (2) SD; (3) skew; and (4) kurtosis. This measures the distribution of energy related to the overall timbre of the scene.

RMS Energy in Octave-wide Frequency Bands. Gygi et al. (2007) used frequency bands ranging from 63–16,000 Hz. This measures the distribution of energy separately for different frequencies.

Spectral Shift in Time Measures. (1) Centroid mean; (2) centroid SD; (3) mean; (4) SD; and (5) maximum centroid velocity. The measures of the centroid mean and SD are established on sequential 50-msec time windows throughout the entirety of the waveform, while the measure of spectral centroid velocity is determined by calculating the overall change in the spectral centroid across sliding 50-msec time windows.

Modulation Spectrum Statistics. Proposed by Houtgast and Steeneken (1985), the modulation spectrum displays intermittent temporal variations in the envelope of a scene. This measure “divides the signal into frequency bands approximately one critical band wide, extracts the envelope in each band, filters the envelope with low-frequency bandpass filters (upper f_0 ranging from 1–32 Hz), and determines the power at that frequency. The result is a plot of the depth of modulation-by-modulation frequency. The statistics measured here will be the height and frequency of the maximum point in the modulation spectrum, as well as the number, mean, and variance of bursts in the modulation spectrum (using the burst algorithm described above; Gygi et al., 2007, p. 846).

Spectral Flux Statistics. Spectral flux evaluates how much change occurs in the spectrum at various frequency bands. This measure can potentially show salient changes in energy, which can be deduced as moments in time where a change in energy may capture the observer’s attention, further influencing their perception of the scene.

Computational Models

Neural Network Models. Two computational models are tested to infer relationships between human judgments of global properties and processing granularity of the scenes: An event-based model which emphasizes a more local analysis of sound events in a scene versus a setting-based model which emphasizes a more global analysis of the setting of the scene. In the first model, we employ a state-of-the-art event-based deep neural network, hereafter referred to as M_{event} , based on the YAMnet architecture (Howard et al., 2017). The model was trained on the Audioset-YouTube database (Gemmeke et al., 2017), consisting of 2 million diverse audio samples taken from YouTube and containing a wide range of soundscapes spanning the broad classes of “Human sound,” “Animal sounds,” “Natural sounds,” “Music,” “Sounds of things,” “Source-ambiguous sounds,” and “Channel, environment, and background.” M_{event} was trained to recognize 521 unique audio event classes. The model is a convolutional neural network consisting of a first 1D convolutional layer, followed by 13 layers of depth-wise separable convolutions then two fully connected linear layers. The computational complexity as well as processing time constants increase as the input signal is analyzed through the different model layers, akin to the increased complexities and tuning observed along the hierarchy in sensory biological systems (Sharpee et al., 2011). For the current study, we focused on the first 14 processing layers of the model, excluding the final two fully connected layers. The 200 scenes used in the behavioral study were analyzed through the event-based model by first resampling each scene waveform at 16 kHz and normalizing within the range of $[-1.0, +1.0]$. Each waveform is then mapped using a short-time Fourier transform with a 25 msec Hann window and 10 msec hop, then a 64-bin mel-spectrogram spanning 125–7500 Hz was computed. Finally, the input was segmented into 1-second segments, each sampled into 96 frames, to match the trained architecture of YAMnet. Latent representations at each layer of the model were then concatenated across time segments to reconstruct the mapping over the entire scene duration. The embeddings at each layer were then integrated over time for each scene and yielded a matrix $[K_e, F_e]^I$, where K_e represents the number of kernels in each layer, F_e represents frequency and I is the layer index.

In parallel, the same scenes were analyzed through a separate, state-of-the-art setting-based model. This model, hereafter referred to as $M_{setting}$, consisted of a Resnet-like architecture (Hu et al., 2020) and was trained to classify different acoustic scenes. The model was trained on the TAU Urban Acoustic Scenes 2020 Mobile dataset (Heittola et al., 2020) to identify 10 unique scene settings: “airport,” “shopping mall,” “street pedestrian,” “metro station,” “public square,” “street traffic,” “tram,” “bus,” “metro,” and “urban park.” Each of the 200 scenes used in the behavioral study were analyzed through this $M_{setting}$ model by resampling each scene at 44.1 kHz then mapping the time waveform using a 2048 Fast Fourier Transform (FFT) point process with 46 msec window and 23 msec hop interval. Next, a 128 log Mel filter bank was derived augmented with log Mel delta and delta-delta features (Rabiner & Schafer, 2010). Each input tensor was normalized along the frequency dimension within the range $[0, 1]$. The model analyzes the input spectrogram using two paths focusing on different spans of the frequency axis, with the first path focusing on the lower frequency Mel bins (0–63) and the second path on the higher frequency Mel bins (64–127). These partial spectrograms are each analyzed through 17 convolutional layers. The final outputs are then concatenated and processed further by 2 convolutional layers, then 2 fully connected layers and a final SoftMax layer which results in the scene classification (Figure 2). In the current study, we extracted the latent representations from the first 17 layers by concatenating the embeddings in the two branches along frequency, and the 2 subsequent convolutions layers and excluded the final fully connected layers following the same analysis procedure using for M_{event} . The embeddings in each

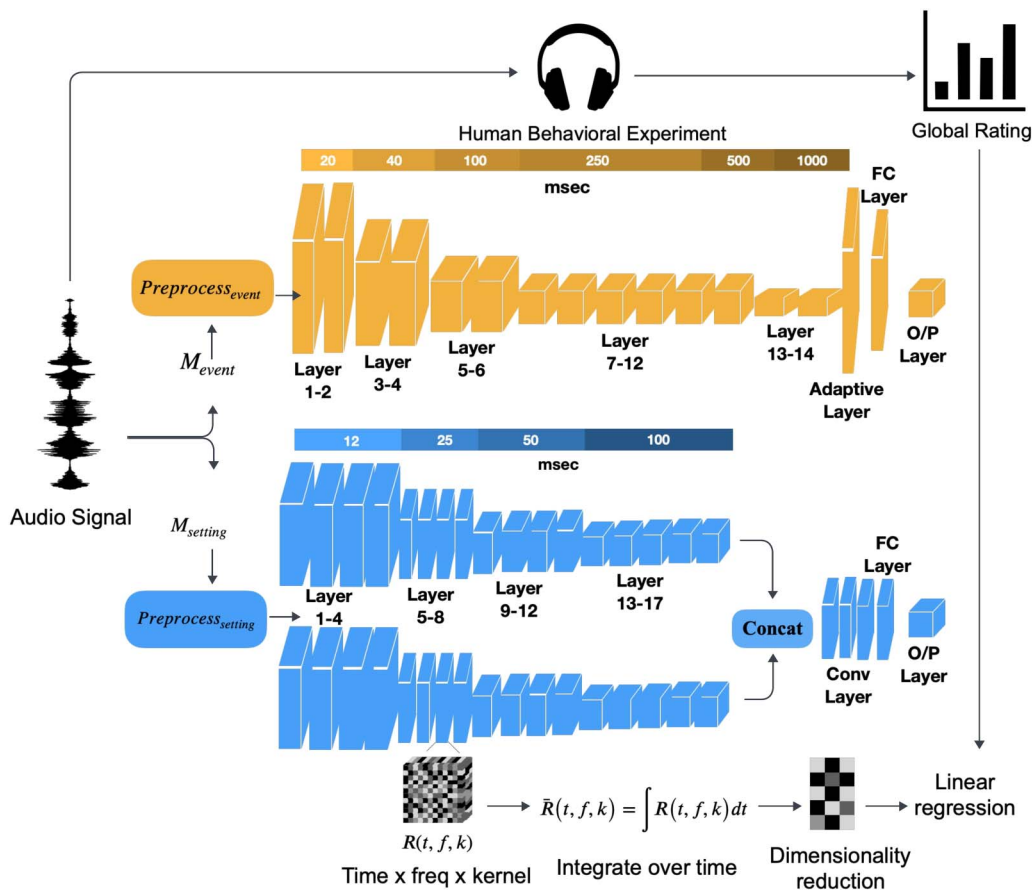


Figure 2. Note. A scene used in the behavioral study is given as input to both the event-based (M_{event}) model (top, yellow) and setting-based model ($M_{setting}$) model (bottom, Blue). The scene is analyzed through both models yielding abstract embeddings along different layers of each model. The representations from each layer and each model are extracted following the schematic in the bottom of the figure. High-dimensional tensors along Time \times Frequency \times Kernel are first integrated over time, then reduced in dimensionality. A linear regression is then performed to contrast the relationship between model embeddings and human judgments of global scene properties. The time constants (in msec) above each model reflect the different sampling rates at which the convolutional processes operate for each layer in each model.

layer for each scene were then integrated over time resulting in a two-dimensional tensor with dimensions $[K_s, F_s]^l$, where K_s represents the number of kernels, F_s represents frequency and l is the layer index.

Dimensionality Reduction. Each network's layer-wise latent representation for every scene was subsequently subjected to Principal Component Analysis (PCA; Wold et al., 1987). This step served the dual purpose of dimensionality reduction for further processing and the regulation of dominant dimensions, allowing a controlled and meaningful comparisons of activities across the two models. We opted for a reduced dimensionality of $D = 136$ following a selection process guided by two criteria: (1) capturing a minimum of 90% of the data variance across all embeddings, and (2) maintaining the same number of dimensions across layers and models for all global properties. This chosen value was employed in all subsequent analyses to evaluate the informative content of each layer and each model in relation to the behavioral scores. In a separate analysis, the embeddings from each layer of each model were

concatenated together after the initial PCA (with $D = 136$). This tensor was then subjected to a second dimensionality reduction using PCA (with $D = 136$) to assess the global contribution of each model to the behavioral scores.

Regression Analysis. Each of the eight empirically derived global properties was correlated with model representations using an independent linear regression. The variability in the human ratings for each property (averaged across subjects) across the 200 scenes was analyzed relative to: (a) the average embeddings for the two models ($M_{setting}$ and M_{event}) and (b) the embeddings per layer for each of the models. We analyzed all correlation trends using R^2_{adj} to control for the effect of multiple predictors in reflecting the goodness of fit with behavioral data.

To compare the correlation trends across different layers for the two models, a curve fitting procedure was performed to the adjusted R^2_{adj} using a bi-quadratic polynomial function, which used only statistically significant correlation values (p 's < 0.05). Since the number of layers in each model was not matched (14 for M_{event} versus 17 for $M_{setting}$), the absolute label of the layers was converted to a relative scale between $[0, 1]$ where 0 is the shallowest layer ($l = 1$) for both models and 1 is the deepest layer for both models.

RESULTS

Inter-Rater Reliability

To evaluate inter-rater reliability of ratings made by participants on each of the eight global property scales, intra-class correlation (ICC) coefficients and their 95% confidence intervals were calculated using IBM SPSS statistical software version 28 (see Table 2; IBM SPSS statistical software (Version 28)). We used a two-way random effects model based on average ratings to assess consistency across participants. The ICCs for all global property scales were statistically significant (all p -values $< .001$) and ranged from good to excellent (0.758–0.980; Koo & Li, 2016).

Accuracy of Global Property Judgments

Percent correct was calculated to determine participant accuracy on judging the following global properties: Outdoor vs. Indoor, Temperature, and Season (see Figure 4). To calculate

Table 2. Inter-rater reliability as measured by intra-class correlations (ICCs).

	ICC	95% CI		F value	F Test with True Value 0		
		Lower Bound	Upper Bound		<i>df</i> 1	<i>df</i> 2	<i>p</i> value
Sparseness	0.968	0.962	0.974	31.428	199	6965	$< .001$
Transience	0.944	0.933	0.955	17.919	199	6965	$< .001$
Season	0.758	0.707	0.804	4.135	199	6965	$< .001$
Navigability	0.787	0.742	0.827	4.667	199	6965	$< .001$
Open vs. Enclosed	0.925	0.909	0.939	13.288	199	6169	$< .001$
Outdoor vs. Indoor	0.977	0.972	0.981	43.402	199	6169	$< .001$
Natural vs. Human-Influenced	0.980	0.975	0.984	49.211	199	6169	$< .001$
Temperature	0.801	0.760	0.839	5.032	199	6169	$< .001$

Note. Results of ICC(2, k). Two-way random effects model, consistency definition, average measures. *df* = degrees of freedom.

these scores, we referred to the metadata associated with each scene, which is included in our database (see Figure 3 for the distribution of scenes in each global property). For the Temperature property, we placed scenes into seven bins of 15 degrees each (e.g., bin 1 = 15–30°F, bin 2 = 31–46°F ... bin 7 = 111–126°F) ranging from 15°F to 126°F. Next, we calculated each participant’s percent correct score based on their Temperature ratings to obtain an average score. Participants were correct 12.33% of the time on their ratings of Temperature, which is close to chance performance (14.29%).

Next, we categorized scenes according to the Season they were recorded in as it corresponds to the Season rating scale (e.g., Winter, Between Winter and Spring, etc., see Figure 1). We then calculated each participant’s percent correct score based on their Season ratings to obtain an average score, which revealed participants performed at 25.28%, which is above chance (14.29%). Finally, we categorized scenes as outdoor vs. indoor based on their recording location. We then dichotomized the outdoor vs. indoor scale and calculated percent correct based on participant ratings. Participants performed at 82.30%, which is well above chance (50%) on their ratings of Outdoor vs. Indoor. These results suggest participants can determine the season of a scene and whether it was recorded indoors or outdoors above chance, but not the temperature of a scene.

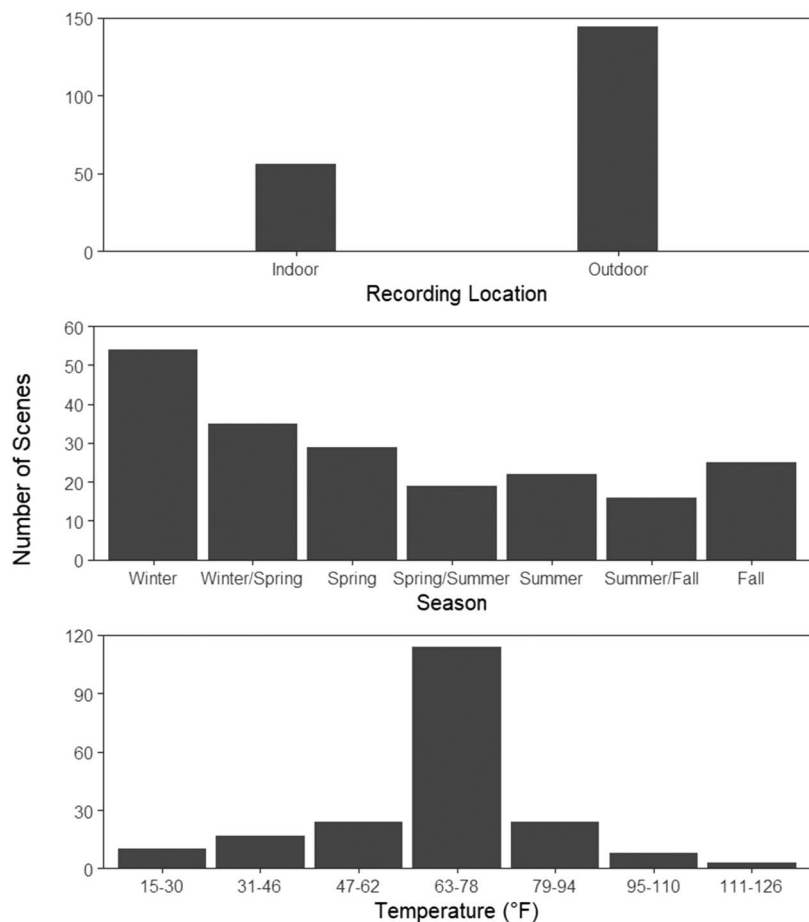


Figure 3. Distribution of Scenes. *Note.* Bar graphs representing the distribution of scenes in each of the following global properties: Outdoor vs. Indoor, Season, and Temperature.

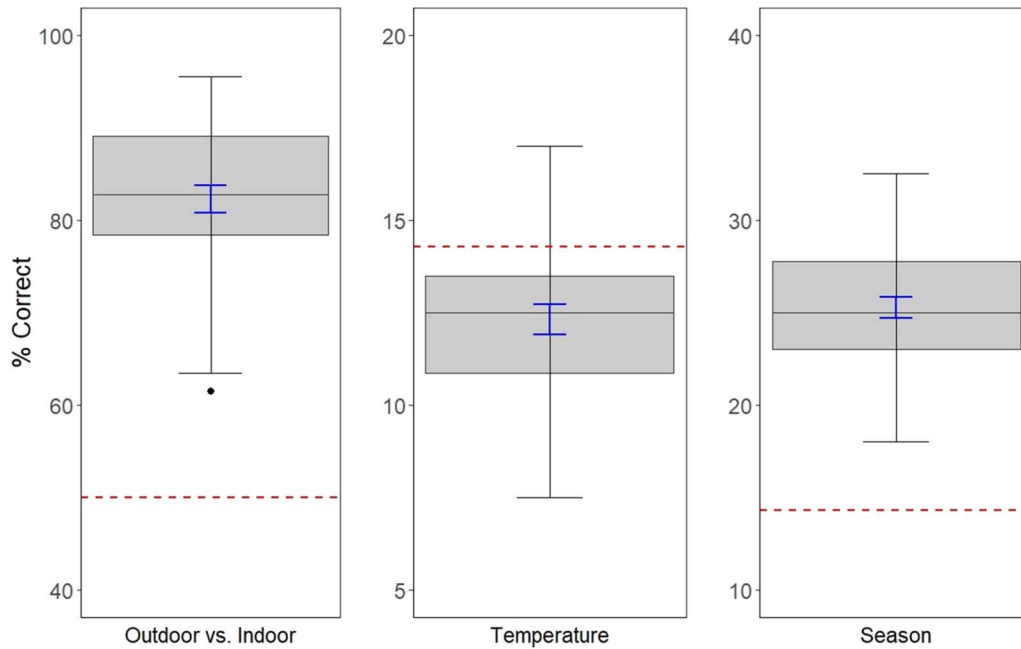


Figure 4. Accuracy of Global Property Judgments. *Note.* Boxplots showing the average percent correct scores for judgments made by participants on each of the following global properties: Outdoor vs. Indoor, Temperature, and Season. There is one outlier on the Outdoor vs. Indoor rating. The blue error bars represent standard error. The dashed red line represents the level of chance for each global property.

Correlations Between Global Properties

Pearson correlations between average global property ratings of each scale are reported in Table 3. Overall, there are several moderate, strong, and very strong correlations between global property rating scales, which justifies their use in the exploratory factor analysis to determine the underlying factor structure of auditory scene perception.

Table 3. Summary of Means, Standard Deviation, and Correlations Among Average Global Property Ratings.

Variable	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	7	8
1. Natural vs. Human-Influenced	4.55	1.61	—							
2. Open vs. Enclosed	3.05	0.95	.61***	—						
3. Outdoor vs. Indoor	3.15	1.55	.63***	.94***	—					
4. Temperature	3.81	0.47	.41***	.44***	.35***	—				
5. Navigability	4.45	0.54	-.25***	-.17*	-.12	-.35***	—			
6. Transience	3.62	0.73	.52***	.14*	.30***	-.03	.24***	—		
7. Season	4.17	0.56	.45***	.07	.17*	.03	.24***	.73***	—	
8. Sparseness	4.12	0.82	.35***	.05	.22**	-.07	.32***	.87***	.73***	—

Note. Correlations between average global property ratings for all auditory scenes (*n* = 200). * *p* < .05, ** *p* < .01, *** *p* < .001.

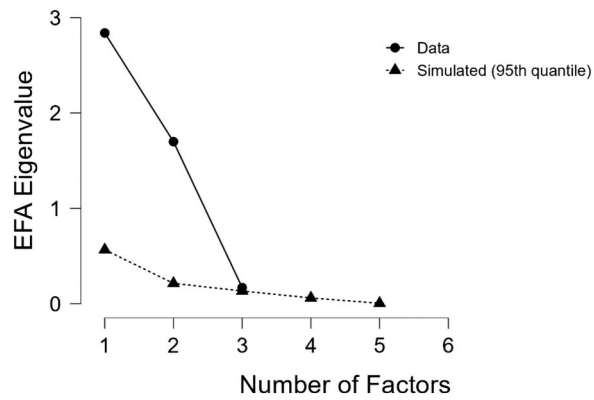


Figure 5. Scree plot of eigenvalues revealing a 2-factor model.

Exploratory Factor Analyses

To evaluate the dimensionality of scene perception, we submitted the average global property ratings of each scale and all 35 acoustic measures to two separate exploratory factor analyses (EFA) using JASP statistical software version 16.1 (JASP Team, 2022).

Exploratory Factor Analysis: Global Properties. The average global property ratings of each scale (Naturalness, Openness, Sparseness, Navigability, Temperature, Outdoor vs. Indoor, Season, Transience) were entered into an EFA, with maximum likelihood factor extraction and Oblimin (oblique) rotation. The Kaiser-Meyer-Olkin test revealed sufficient sampling adequacy for the EFA, $KMO = 0.69$. Bartlett’s test of sphericity indicated the correlation structure of the variables was adequate for the EFA as well, $\chi^2(28) = 1253.84, p < .001$. Upon visual inspection of the scree plot as well as a parallel analysis (see Figure 5), a two-factor solution was revealed and accounts for 64.0 % of the variance in the data. Table 4 displays the variables and factor loadings, with loadings less than |.40| excluded for clarity.

Table 4. Factor loadings of global property scales.

Factor Loadings	Factor 1	Factor 2
Transience	0.94	
Sparseness	0.93	
Season	0.78	
Navigability	0.37	
Open vs. Enclosed		1.01
Outdoor vs. Indoor		0.93
Natural vs. Human-Influenced		0.58
Temperature		0.46

Note. Extraction method: maximum likelihood; Rotation method: Oblimin (oblique).

Factor 1. After rotation, Factor 1 consisted of four variables that accounted for 33% of the variance in the model. The global property variables which loaded onto this factor were Transience (0.94), Sparseness (0.93), Season (0.78). Although Navigability (0.37) also loaded onto this factor, its interpretation should be made with caution as its loading is below the 1.401 threshold.

Factor 2. Factor 2 consisted of four variables that accounted for 31% of the variance in model. The global property variables which loaded onto this factor were Openness (1.01), Outdoor vs. Indoor (0.93), Natural vs. Human-Influenced (0.58), and Temperature (0.46).

Exploratory Factor Analysis: Acoustic Variables. All 35 acoustic measures (envelope-based intensity and rhythm measures, autocorrelation pitch statistics, correlogram-based pitch measures, moments of the spectrum, RMS energy in octave-wide frequency bands, spectral shift in time measures, modulation spectrum statistics) were submitted to a separate factor analysis using maximum likelihood factor extraction and Oblimin (oblique) rotation. Table 5 displays the variables and factor loadings for the rotated factors for the final model, with loadings less than 1.401 excluded for clarity.

The Kaiser-Meyer-Olkin (KMO) test revealed sufficient sampling adequacy for the final EFA analysis, $KMO = 0.71$. Bartlett's test of sphericity indicated the correlation structure of the variables was adequate for EFA as well, $\chi^2(595) = 7490.33, p < .001$. Upon visual inspection of the scree plot as well as a parallel analysis (see Figure 6), a seven-factor solution was revealed and accounts for 57% of the variance in the data.

Factor 1. After rotation, Factor 1 consisted of four variables that accounted for 10% of the variance in the model. One of the RMS energy measures of octave-wide frequency bands centered at 8000 Hz (0.89), two Moments of the Spectrum measures, the centroid (0.87) and standard deviation (0.81), and the mean pitch (0.61) all loaded onto this factor.

Factor 2. Factor 2 consisted of four variables which accounted for 10% of the variance in the model. Three Moments of the Spectrum measures, the standard deviation (1.00), maximum (0.90), and mean (0.86), as well as one modulation statistics measure, the maximum peak (0.49) loaded onto this factor.

Factor 3. Factor 3 consisted of three variables which accounted for 8% of the variance in the model. All these variables loaded negatively onto the factor, and they include two Moments of the Spectrum measures, the skew (-0.91) and kurtosis (-0.86), and the mean Spectral Flux (-0.43).

Factor 4. Factor 4 consisted of two variables which accounted for 8% of the variance in the model. Both variables were measures of the autocorrelation, the mean peak (1.02) and standard deviation of peaks (0.98).

Factor 5. Factor 5 consisted of seven variables which accounted for 8% of the variance in the model. Two variables also loaded onto other factors; the mean pitch (0.41) also loaded onto Factor 1, and the spectral flux mean (0.45) also loaded onto Factor 3. In addition, the maximum peak of the autocorrelation (0.62), maximum pitch salience (0.62), mean pitch salience (0.58), and the maximum peak of spectral flux (0.48) loaded onto this factor as well. One measure of RMS energy in octave-wide frequency bands centered at 250 Hz loaded negatively on this factor (-0.43).

Table 5. Factor loadings of acoustic variables.

Factor Loadings	Factor						
	1	2	3	4	5	6	7
RMS in band $f_c = 8000$ Hz	0.89						
Moments of the Spectrum (Centroid)	0.87						
Moments of the Spectrum (SD)	0.81						
Mean Pitch	0.61				0.41		
Spectral Velocity (SD)		1.00					
Spectral Velocity (Maximum)		0.90					
Spectral Velocity (Mean)		0.86					
Modulation Statistics (Max Peak)		0.49					
Moments of the Spectrum (Skew)			-0.91				
Moments of the Spectrum (Kurtosis)			-0.86				
Spectral Flux (Mean)			-0.43		0.45		
Mean Peak in Autocorrelation				1.02			
SD of Peaks in Autocorrelation				0.98			
Maximum Peak in Autocorrelation					0.62		
Maximum Pitch Salience					0.62		
Mean Pitch Salience					0.58		
Spectral Flux (Maximum Peak)					0.48		
RMS in band $f_c = 250$ Hz					-0.43		
Pause-Corrected RMS Amplitude						0.97	
Overall RMS Amplitude						0.96	
Total Number of Bursts							0.90
Moments of the Spectrum (# of Peaks)							0.86

Note. Extraction method: maximum likelihood; Rotation method: Geomin (oblique); Loadings less than |.40| are not displayed.

Factor 6. Factor 6 consisted of two variables which accounted for 7% of the variance in the model. Both variables were measures of RMS amplitude: the pause-corrected RMS (0.97) and overall RMS (0.96).

Factor 7. Factor 7 consisted of two variables which accounted for 6% of the variance in the model. Both variables were measures of the envelope: burst duration/total duration (0.90) and number of bursts (0.86).

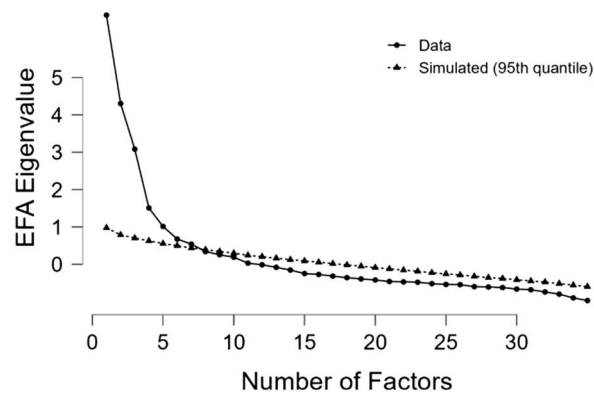


Figure 6. Scree plot of eigenvalues revealing a 7-factor model.

Multiple Linear Regression Analyses

Next, eight multiple linear regression analyses were calculated to predict average ratings on each global property scale based on the acoustic measures. Overall, each global property rating scale was significantly predicted by at least one acoustic variable (see Table 6).

Natural vs. Human-Influenced Regression. The first regression was calculated to predict ratings on the Natural vs. Human-Influenced scale based on all 35 acoustic variables and was statistically significant, $R^2 = 0.57$, $R^2_{adj} = 0.47$, $F(35, 162) = 6.09$, $p < .001$. The significant acoustic predictors were pause- corrected/overall RMS amplitude ($\beta = 0.16$, $p < .05$), mean pitch ($\beta = -0.44$, $p < .001$), standard deviation of pitch ($\beta = -0.26$, $p < .05$), RMS energy in octave-wide frequency bands centered at 500 Hz ($\beta = 0.22$, $p < .05$), the spectral flux standard deviation ($\beta = -0.25$, $p < .05$), and the range of peaks in the autocorrelation ($\beta = -0.14$, $p = .05$).

Open vs. Enclosed Regression. The regression predicting ratings on the Open vs. Enclosed scale based on all acoustic variables was statistically significant, $R^2 = 0.57$, $R^2_{adj} = 0.48$, $F(35, 162) = 6.21$, $p < .001$. The significant acoustic predictors were the Moments of the Spectrum centroid ($\beta = 0.39$, $p < .05$), pause-corrected RMS amplitude ($\beta = 5.96$, $p < .05$), overall RMS amplitude ($\beta = -6.05$, $p < .05$), mean pitch ($\beta = -0.50$, $p < .001$), standard deviation of pitch ($\beta = -0.20$, $p < .05$), spectral flux standard deviation ($\beta = -0.30$, $p < .05$), range of peaks in the autocorrelation ($\beta = -0.23$, $p < .05$), RMS energy in octave-wide frequency bands centered at 250 Hz ($\beta = 0.18$, $p < .05$), and 1000 Hz ($\beta = -0.18$, $p < .05$).

Outdoor vs. Indoor Regression. The regression predicting ratings on the Outdoor vs. Indoor scale based on all acoustic variables was statistically significant, $R^2 = 0.58$, $R^2_{adj} = 0.49$, $F(35, 162) = 6.33$, $p < .001$. The significant acoustic predictors were the Moments of the Spectrum centroid ($\beta = 0.35$, $p < .05$), pause-corrected RMS amplitude ($\beta = 4.89$, $p < .05$), overall RMS amplitude ($\beta = -4.94$, $p < .05$), mean pitch ($\beta = -0.57$, $p < .001$) and standard deviation of pitch ($\beta = -0.26$, $p < .05$), mean pitch salience ($\beta = 0.32$, $p < .05$), spectral flux standard deviation ($\beta = -0.37$, $p < .001$), maximum peak in spectral flux ($\beta = 0.20$, $p = .05$), maximum peak in the autocorrelation ($\beta = -0.23$, $p < .05$), the range of peaks in the autocorrelation ($\beta = -0.17$, $p < .05$), RMS energy in octave-wide frequency bands centered at 1000 Hz ($\beta = -0.14$, $p = .05$) and 2000 Hz ($\beta = 0.13$, $p = .05$).

Temperature Regression. A regression was calculated to predict ratings on the Temperature scale based on all acoustic variables and was statistically significant, $R^2 = 0.33$, $R^2_{adj} = 0.19$,

Table 6. Significant acoustic predictors for each global property rating scale.

Variable	β	R ²	Variable	β	R ²
Natural vs. Human-Influenced		0.57	Temperature		0.33
Pause-Corrected/Overall RMS Amplitude	0.16		Moments of the Spectrum (SD)	0.48	
Pitch (Mean)	-0.44		Moments of the Spectrum (# of Peaks)	-0.34	
Pitch (SD)	-0.26		Pause-Corrected/Overall RMS Amplitude	3.68	
Spectral Flux (SD)	-0.25		Spectral Velocity (Mean)	-0.6	
Autocorrelation (Range of Peaks)	-0.14		Spectral Velocity (SD)	0.86	
RMS in band $f_c = 500$ Hz	0.22		Navigability		0.41
Openness		0.57	Pause-Corrected/Overall RMS Amplitude	-0.17	
Moments of the Spectrum (Centroid)	0.39		Season		0.56
Pause-Corrected RMS Amplitude	5.96		Moments of the Spectrum (Skew)	-0.8	
Overall RMS Amplitude	-6.05		Moments of the Spectrum (Kurtosis)	0.46	
Pitch (Mean)	-0.50		Transience		0.78
Pitch (SD)	-0.20		Moments of the Spectrum (Skew)	-0.8	
Spectral Flux (SD)	-0.30		Moments of the Spectrum (Kurtosis)	0.43	
Autocorrelation (Range of Peaks)	-0.23		Moments of the Spectrum (# of Peaks)	0.12	
RMS in band $f_c = 250$ Hz	0.18		Pitch (Mean)	-0.22	
RMS in band $f_c = 1000$ Hz	-0.18		Pitch Salience (Mean)	0.23	
Outdoor vs. Indoor		0.58	RMS in band $f_c = 500$ Hz	0.19	
Moments of the Spectrum (Centroid)	0.35		RMS in band $f_c = 1000$ Hz	0.2	
Pause-Corrected RMS Amplitude	4.89		Sparseness		0.87
Overall RMS Amplitude	-4.94		Moments of the Spectrum (Skew)	-0.97	
Pitch (Mean)	-0.57		Moments of the Spectrum (Kurtosis)	0.51	
Pitch (SD)	-0.26		Moments of the Spectrum (# of Peaks)	0.09	
Pitch Salience (Mean)	0.32		Pitch (Mean)	-0.15	
Spectral Flux (SD)	-0.37		Pitch (Maximum)	0.13	
Spectral Flux (Max Peak)	0.2		Pitch Salience (Mean)	0.27	
Autocorrelation (Max Peak)	-0.23		Spectral Flux (Mean)	-0.21	
Autocorrelation (Range of Peaks)	-0.17		Spectral Flux (SD)	-0.11	
RMS in band $f_c = 1000$ Hz	-0.14		RMS in band $f_c = 500$ Hz	0.1	
RMS in band $f_c = 2000$ Hz	0.14		RMS in band $f_c = 1000$ Hz	0.12	

$F(35, 162) = 2.32, p < .001$. The significant predictors were the Moments of the Spectrum standard deviation ($\beta = 0.48, p < .05$) and number of peaks ($\beta = -0.34, p < .001$), pause-corrected/overall RMS amplitude ($\beta = 3.68, p < .05$), spectral velocity mean ($\beta = -0.60, p < .05$), and spectral velocity standard deviation ($\beta = 0.86, p < .05$).

Navigability Regression. The regression predicting ratings on the Navigability scale based on all acoustic variables was statistically significant, $R^2 = 0.41, R_{adj}^2 = 0.28, F(35, 162) = 3.18, p < .001$. The only significant predictor was pause-corrected/overall RMS amplitude ($\beta = -0.17, p < .05$).

Season Regression. A regression was calculated to predict ratings on the Season scale based on all acoustic variables and was statistically significant, $R^2 = 0.56, R_{adj}^2 = 0.47, F(35, 162) = 5.99, p < .001$. The significant predictors were the Moments of the Spectrum skew ($\beta = -0.80, p < .05$) and kurtosis ($\beta = 0.46, p < .05$), as well as RMS energy in octave-wide frequency bands centered at 1000 Hz ($\beta = 0.15, p < .05$).

Transience Regression. A regression was calculated to predict ratings on the Transience scale based on all acoustic variables and was statistically significant, $R^2 = 0.78, R_{adj}^2 = 0.73, F(35, 162) = 16.28, p < .001$. The significant predictors were the Moments of the Spectrum skew ($\beta = -0.80, p < .001$), kurtosis ($\beta = 0.43, p < .05$), and number of peaks ($\beta = 0.12, p < .05$), mean pitch ($\beta = -0.22, p < .05$), mean pitch salience ($\beta = 0.23, p < .05$), as well as RMS energy in octave-wide frequency bands centered at 500 Hz ($\beta = 0.19, p < .001$) and 1000 Hz ($\beta = 0.20, p < .001$).

Sparseness Regression. A regression was calculated to predict ratings on the Sparseness scale based on all acoustic variables and was statistically significant, $R^2 = 0.87, R_{adj}^2 = 0.84, F(35, 162) = 31.51, p < .001$. The significant predictors were the Moments of the Spectrum skew ($\beta = -0.97, p < .001$), kurtosis ($\beta = 0.51, p < .001$), and number of peaks ($\beta = 0.09, p < .05$), mean pitch ($\beta = -0.15, p < .05$), maximum pitch ($\beta = 0.13, p < .05$), mean pitch salience ($\beta = 0.27, p < .001$), spectral flux mean ($\beta = -0.21, p < .05$), spectral flux standard deviation ($\beta = -0.11, p < .05$), as well as RMS energy in octave-wide frequency bands centered at 500 Hz ($\beta = 0.10, p < .001$) and 1000 Hz ($\beta = 0.12, p < .001$).

Computational Modeling Results

Overall, when evaluating the predictive power of aggregate embeddings from the setting versus event models, adjusted R^2 values are consistently higher for $M_{setting}$ across all global properties ($M_{setting}$ average aggregate $R_{adj}^2 = 0.75, M_{event}$ average aggregate $R_{adj}^2 = 0.59$). Figure 7A illustrates the adjusted R^2 values for aggregate representations from each model for individual global properties reported empirically. The regression values vary greatly across global properties for both models with explainable variance exceeding 80% for global properties like Transience ($R_{adj}^2 = 0.89, p < .001$) and Sparseness ($R_{adj}^2 = 0.90, p < .001$) for the $M_{setting}$ model and as low as close to 40% for properties like Navigability ($R_{adj}^2 = 0.41, p < .001$) and Season ($R_{adj}^2 = 0.45, p < .001$) for the M_{event} model.

An important element that stands out from this analysis is the *relative* improvement in explanatory power for the $M_{setting}$ model relative to the M_{event} model across all global properties. The results reveal that high difference in explanatory power for properties like Transience ($\Delta R_{adj}^2 = 31.8\%$), Navigability ($\Delta R_{adj}^2 = 30.0\%$), Sparseness ($\Delta R_{adj}^2 = 29.8\%$), and Season ($\Delta R_{adj}^2 = 29.5\%$). This difference in explanatory power across both models is slightly decreased for the global properties of Outdoor vs. Indoor ($\Delta R_{adj}^2 = 14.1\%$), Open vs. Enclosed

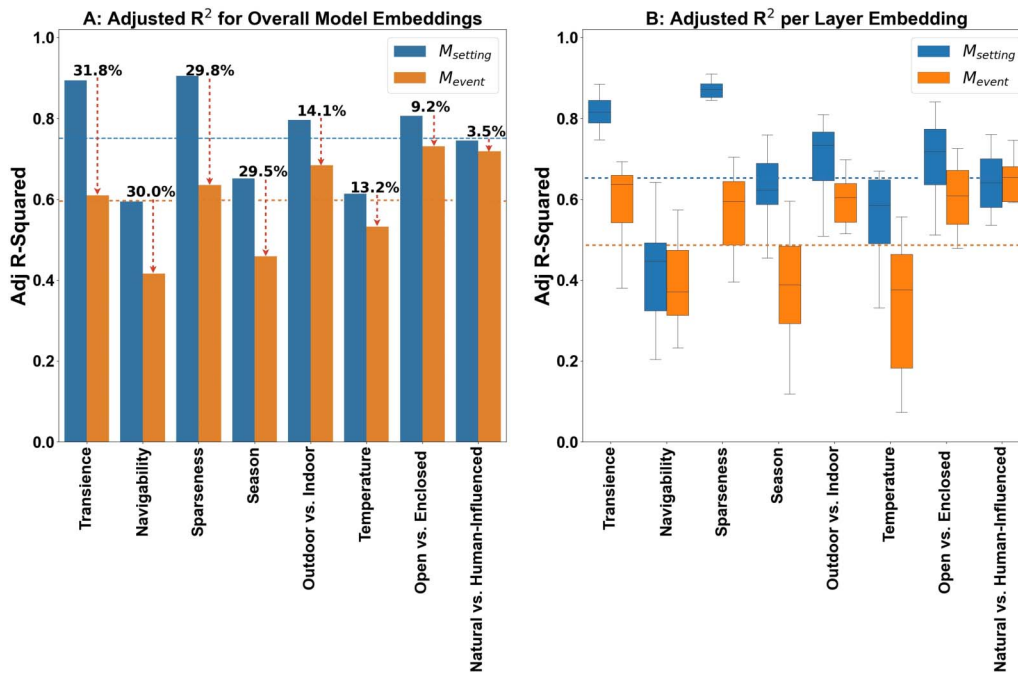


Figure 7. Note. (A) The bar graph depicts R_{adj}^2 values for each model across aggregate embeddings, with individual bars representing the correlation strength for each global property. The average adjusted R^2 value for aggregate model embeddings across all global properties is shown as horizontal dashed lines. (B) R_{adj}^2 distributions for the $M_{setting}$ and M_{event} models quantified for each layer of each model per global property. The average R_{adj}^2 value for layer wise model embeddings across all global properties is shown as horizontal dashed line.

($\Delta R_{adj}^2 = 13.2\%$), and Temperature ($\Delta R_{adj}^2 = 9.2\%$), and even more severely reduced for the Natural vs. Human-Influenced ($\Delta R_{adj}^2 = 3.5\%$) global property. These differences may underlie contrasting representations in both models that can predict variability in human ratings of different properties across scenes and further support the divergence in factor loadings between properties like Transience, Navigability, Sparseness, and Season, as well as the other 4 properties (Outdoor vs. Indoor, Open vs. Enclosed, Temperature, Natural vs. Human-Influenced), which is in line with effects observed in the EFA reported in Table 4.

Looking closely at the variability of model embeddings across individual layers, Figure 7B reveals that—on average—layer-wise embeddings of the $M_{setting}$ model are higher than those from the M_{event} model ($M_{setting}$ layer-wise average $R_{adj}^2 = 0.65$, M_{event} average layer-wise $R_{adj}^2 = 0.48$). Nevertheless, the goodness of fit of individual layers to the behavioral ratings varies greatly across global properties. On the one hand, we note that global properties like Transience and Sparseness have a tight variance of for adjusted R^2 ($M_{setting}$ standard deviation across layers: Transience, $\sigma_{R^2_{adj}} = 0.07$, Sparseness, $\sigma_{R^2_{adj}} = 0.06$). This variance visibly increases for properties such as Navigability ($\sigma_{R^2_{adj}} = 0.14$), Open vs. Enclosed ($\sigma_{R^2_{adj}} = 0.11$), and Temperature ($\sigma_{R^2_{adj}} = 0.11$). On the other hand, the variability of regression fits across layers is generally higher for the M_{event} model. Properties such as Natural vs. Human-Influenced ($\sigma_{R^2_{adj}} = 0.13$) and Outdoor vs. Indoor ($\sigma_{R^2_{adj}} = 0.14$) reveal the lowest variability, while properties such as Navigability ($\sigma_{R^2_{adj}} = 0.17$) and Temperature ($\sigma_{R^2_{adj}} = 0.16$) have much higher variability across layers.

Figure 8 looks closely at the trend of correlations for each global property across the layers of both models and again underscores the general advantage of the setting model especially in explaining variability across global properties like Transience, Sparseness, and Season.

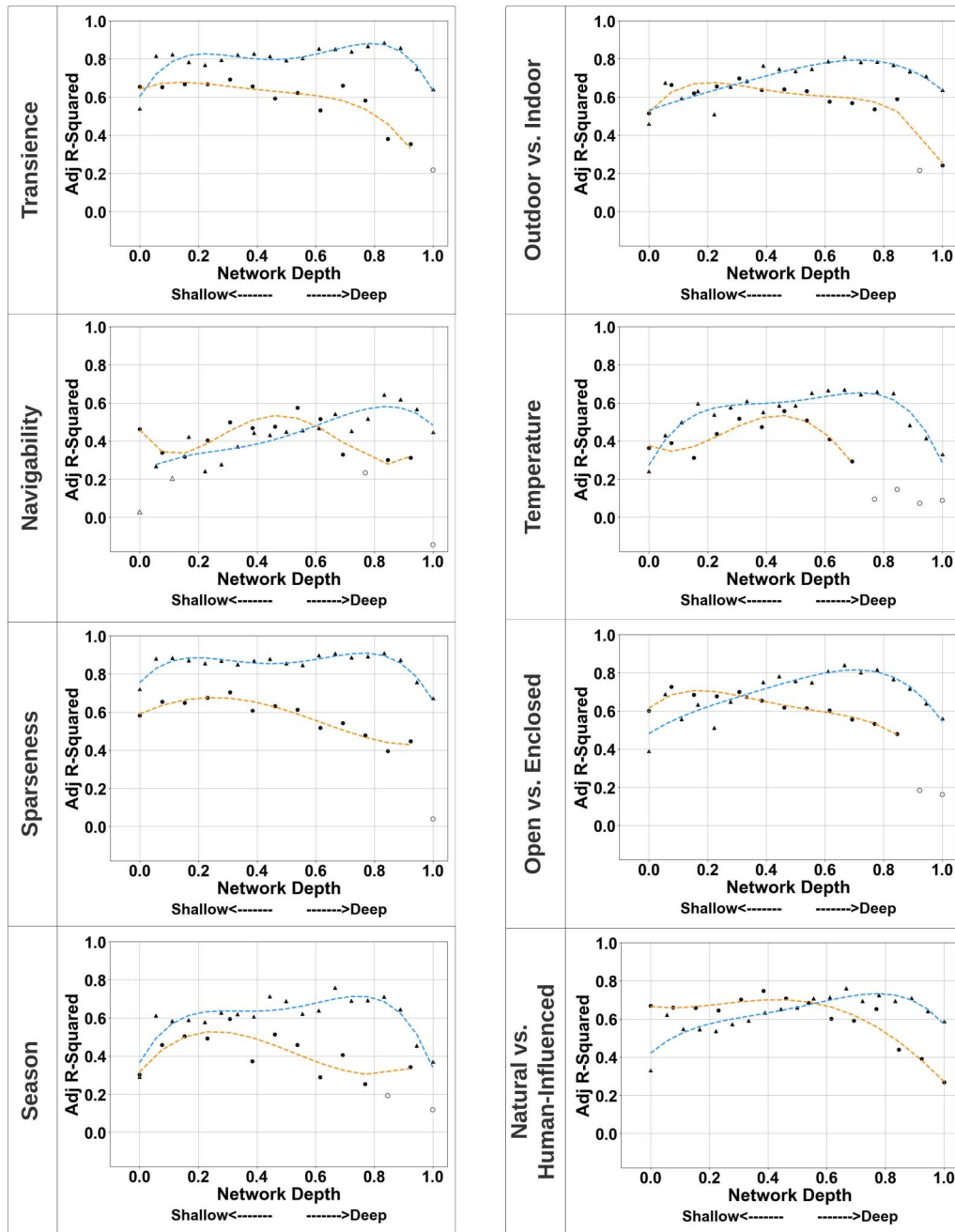


Figure 8. Note. Correlation between global properties and neural network hidden representations across all layers for M_{event} (in orange) and $M_{setting}$ (in blue) networks. The y-axis signifies the adjusted R^2 values, and the x-axis showcases the normalized relative depth from the shallowest (0) to the deepest layer (1).

Furthermore, we note differences in predictive power across the network layers for both models, with the $M_{setting}$ model revealing correlation peak at deeper points of the network relative to the M_{event} model across global properties. The curve fit for $M_{setting}$ model shows a peak corresponding to relative depth of 0.77 for Transience, 0.83 for Navigability, 0.77 for Sparseness, 0.72 for Season, 0.72 for Outdoor vs. Indoor, 0.72 for Temperature, 0.72 for Open vs. Enclosed, and 0.77 for Natural vs. Human-Influenced. Conversely, the M_{event} model appears to lose predictive power in the deeper layers of the model across all properties and

shows a peak corresponding to relative depth of 0.15 for Transience, 0.46 for Navigability, 0.23 for Sparseness, 0.23 for Season, 0.23 for Outdoor vs. Indoor, 0.46 for Temperature, 0.15 for Open vs. Enclosed, and 0.46 for Natural vs. Human-Influenced. Here, a relative depth of 0 represents the shallowest layer and 1 represents the deepest layer. The normalization facilitates a direct comparison between the two models, which have differing layer counts (see *Methods*). These peaks are likely driven by great variability in analysis time constants and accumulation of nonlinearities across the model layers from shallower to deeper processing stages (Figure 2).

GENERAL DISCUSSION

Here, we investigated the contributions of high-level global information and low-level acoustic features to auditory scene perception. Participants listened to complex, real-world auditory scenes and made judgments on a series of global properties (Sparseness, Transience, Season, Navigability, Open vs. Enclosed, Outdoor vs. Indoor, Natural vs. Human-Influenced, Temperature). We found high between-participant consistency on ratings of all eight global property rating scales (indicated by the significant intra-class correlations). This particular result provides preliminary evidence for the ability to perceive auditory scenes from a global perspective, which is consistent with findings in the visual domain (Greene & Oliva, 2009a). A variety of acoustic measures were useful in predicting each of the global property ratings, though some of the acoustic-global relationships were non-linear. Additionally, the results of the computational modeling analysis provide preliminary evidence that individual sound events within a scene may not be the primary drivers of global judgments. This supports the notion that the scene's overarching acoustic structure and its contextual setting largely influence the perceptual judgments of these global ratings.

These results are consistent with the hypothesis that global scene properties serve as high-level dimensions to inform a scene's layout, function, and constancy, allowing observers to rapidly understand the scene without needing to identify individual objects that are present (Greene & Oliva, 2009a, 2009b, 2010; Ross & Oliva, 2010). Using both behavioral and computational methods, prior studies have demonstrated that observers can more quickly and accurately categorize visual scenes into a global category (e.g., an open environment) than a basic level category (e.g., a waterfall; Greene & Oliva, 2009a), use global information to perform basic-level categorization tasks (Greene & Oliva, 2009b), and adapt to global properties of visual scenes (Greene & Oliva, 2010). Additional electrophysiological studies have indicated the P2 event related potential (ERP) as a neural marker for global scene properties (Harel et al., 2016) and have revealed that global scene information is extracted in early ERP components (P1, N1, and P2; Hansen et al., 2018). The mounting evidence for the use of global scene properties in the visual domain provides a promising foundation for future behavioral, electrophysiological, neuroimaging, and computational studies in the auditory domain. Although our study only measured eight global properties, many more may be uncovered and investigated by future research to provide a well-rounded understanding of how people interpret complex auditory scenes. For example, future studies can evaluate how well people can identify the setting of a scene compared to the objects within it. Additionally, it would be useful to measure how quickly and in what order global scene judgments are made (e.g., is openness perceived prior to temperature?) and whether observers can adapt to these properties.

Dimensionality of Auditory Scene Perception

The results of our study indicate a high amount of dimensionality reduction along the auditory pathway when we listen to auditory scenes. Dimensionality reduction has been demonstrated

in the perception of environmental sounds (Gygi et al., 2007), timbre (Grey, 1977), musical tonality (Krumhansl, 1979; Shepard, 1982; Toivainen et al., 1995), and rhythm (Desain & Honing, 2003; Jacoby & McDermott, 2017), suggesting this is a common feature of auditory processing. The multiple linear regressions calculated to predict performance on the global property rating scales based on the acoustic properties of each scene revealed each rating scale was significantly predicted by at least one acoustic variable. This finding, along with the difference in the number of reduced factors in the exploratory factor analyses (2 global property factors vs. 7 acoustic factors), suggest global variables may be processed at a higher level of the auditory pathway where acoustic features have been abstracted out of or have nonlinear relationships with global variables. Although the EFA on the global properties yielded a factor structure which mirrors the judgments made by our groups of participants (i.e., the global properties that loaded onto Factor 1 correspond to judgments made by Groups 1–4, and those that loaded onto Factor 2 correspond to judgments made by Groups 5–8), the results of our modeling analysis suggest this is not purely an artifact of our study design. When we compared the explanatory power for all global properties between the setting- and object-based models (Figure 8), we found a high difference in power for the global properties that loaded onto Factor 1 (Navigability, Transience, Sparseness, and Season) while a much smaller difference was found for properties that loaded onto Factor 2 (Outdoor vs. Indoor, Open vs. Enclosed, Temperature, and Natural vs. Human-Influenced). These differences in explanatory power across models suggest there are underlying differences in the representations in both models, further supporting the results of the EFA. The transformation of low-level acoustic information into high-level global representations of auditory scenes could occur similarly to processing along the ventral visual stream, where low-level information about visual objects (e.g., an object's geometric shape, position in space, etc.) culminates into high level representations of visual objects which allow for object recognition (DiCarlo et al., 2012). Additionally, common neural population codes have been shown to have highly nonlinear relationships, suggesting nonlinear transformation is a common feature of sensory processing (De & Chaudhuri, 2023). The results of the modeling analysis provide additional evidence for this conclusion, revealing the best predictive power across all global properties is achieved after several transformations of the time-frequency input, which ultimately highlights the critical role of information abstraction that underlies the processing hierarchy in these deep models (Kell & McDermott, 2019; Yamins & DiCarlo, 2016). A similar abstraction is a hallmark of sensory pathways, and particularly at the level of auditory cortex whereby hierarchical mappings and functional specializations appear to facilitate various complex auditory tasks (Bizley & Cohen, 2013; de Heer et al., 2017; Kumar et al., 2007; Okada et al., 2010).

Investigating responses to auditory scenes along the auditory pathway will be essential to our understanding of how the auditory system integrates low-level acoustic features of auditory scenes into high-level global representations of scenes. The auditory system functions hierarchically, showing tuning specificity for simple stimuli and acoustic features, such as pitch (Bendor & Wang, 2005; Norman-Haignere et al., 2013; Patterson et al., 2002), frequency (Da Costa et al., 2011; Humphries et al., 2010), spatial cues (Higgins et al., 2017; Rauschecker & Tian, 2000; Stecker et al., 2005), and spectral and temporal modulations (Barton et al., 2012; Chi et al., 2005; Santoro et al., 2014; Schönwiesner & Zatorre, 2009) in primary auditory areas as well as tuning specificity for more complex stimuli, such as noise bursts (Kaas & Hackett, 2000), vocalizations (Belin et al., 2000; Petkov et al., 2008; Rauschecker & Tian, 2000), speech (Mesgarani et al., 2014; Norman-Haignere et al., 2015; Overath et al., 2015; Scott et al., 2000), song (Norman-Haignere et al., 2022), and music (Boebinger et al., 2021) in non-primary auditory areas. The increase in response complexity along the auditory pathway

suggests sound features are abstracted from combinations of more simple responses, such as the acoustic features of sounds, which parallels findings in the visual system (Carandini et al., 1999; Cumming & DeAngelis, 2001; De Valois & De Valois, 1980; DiCarlo et al., 2012; Gegenfurtner & Kiper, 2003; Horwitz & Hass, 2012; Hubel & Wiesel, 1962; Movshon et al., 1978; Tootell et al., 1988, 1998).

Neural Pathways for Scene Processing

The existence of global properties is supported by behavioral, computational (Greene & Oliva, 2009a, 2009b, 2010; Ross & Oliva, 2010), and neural (Hansen et al., 2018; Harel et al., 2016) studies in the visual domain as well as the results of our behavioral study. This raises important questions regarding the neural pathways allowing for global processing of auditory scenes.

The computations underlying auditory perception are suggested to occur along two parallel processing streams analogous to the dorsal and ventral streams in the visual domain (Goodale & Milner, 1992; Milner & Goodale, 2006; Mishkin et al., 1983), allowing us to identify where a sound is coming from and also identify what we are listening to (Alain et al., 2001; Griffiths, 2008; Lomber & Malhotra, 2008; Rauschecker, 1998; Rauschecker & Tian, 2000). The dorsal auditory stream affords the ability to localize auditory stimuli in space (Rauschecker, 2012; Rauschecker & Scott, 2009) and map sounds onto motor-based representations involved in speech production (Hickok & Poeppel, 2004), while the ventral auditory stream affords identification and semantic processing of auditory stimuli (including speech).

The ventral auditory stream originates in the core auditory fields A1 and R, continues to the anterolateral and middle lateral belt regions, and terminates in the ventrolateral prefrontal cortex (vIPFC; Kaas & Hackett, 2000; Rauschecker & Tian, 2000). Neurons in the core prefer lower-level sound features such as frequency and intensity, while neurons in the anterolateral belt respond to vocalizations, frequency-modulated sweeps, and band-passed noise (Chang et al., 2010; Rauschecker et al., 1995; Rauschecker & Tian, 2000, 2004; Tian & Rauschecker, 2004; Tian et al., 2001).

Many neurophysiological and neuroimaging studies have identified cortical regions with selectivity for distinct aspects and categories of auditory input (e.g., voices, environmental sounds, music, etc.). Pitch is an important perceptual feature of many sounds, including speech, music, and environmental sounds, and it allows us to identify voices, segregate and organize sounds in complex auditory scenes, and convey musical structure and emotion. Pitch-selectivity has been demonstrated in anterolateral regions of A1 (Bendor & Wang, 2005, 2006, 2010; Norman-Haignere et al., 2013, 2015; Penagos et al., 2004; Patterson et al., 2002) and speech-selective regions have been localized further along the auditory pathway along the middle and anterior superior temporal sulcus (STS) and superior temporal gyrus (STG; Belin et al., 2000, 2002; Boebinger et al., 2021; Chang et al., 2010; Davis & Johnsrude, 2003; Mesgarani et al., 2014; Norman-Haignere et al., 2015, 2022; Pernet et al., 2015; Yi et al., 2019). Music-selective populations have been observed in areas anterior and posterior to A1, while speech-selective populations have been observed in areas lateral to A1, which suggests music and speech representations begin to diverge in non-primary auditory cortex and are potentially processed in partially distinct pathways along the ventral stream. A recent study identified a song-selective (i.e., music with singing) population co-located with music and speech selective regions (Norman-Haignere et al., 2022), suggesting multiple neural populations exist that are selective for particular aspects of music (e.g., singing), and further demonstrates the complexity of processing within the auditory cortex.

Selectivity for animal sounds, voices, human-made sounds, and tools have been localized to middle temporal gyrus (MTG), STS, and STG (Bethmann & Brechmann, 2014; Lee et al., 2009; Sharda & Singh, 2012; Zhang et al., 2015, 2020), and a recent human electrocorticography (ECoG) and fMRI study found selectivity for a variety of environmental sounds (e.g., whistling, telephone dial, wind chimes, car horn, splashing water, etc.) in posterior primary auditory cortex, which may suggest a third stream exists in the auditory cortex for environmental sound processing (Norman-Haignere et al., 2022). The processing of environmental sounds may not emerge until later along the ventral stream, potentially closer to or in the vIPFC, where processing may reflect post-sensory processes such as auditory attention, working memory, the meaning of sounds, and integration of multisensory information and aid in our perception of many categories of sound stimuli (Cohen et al., 2009; Lee et al., 2009; Ng et al., 2014; Plakke & Romanski, 2014; Poremba et al., 2003; Romanski et al., 2005; Russ, Ackelson et al., 2008; Russ, Orr, & Cohen, 2008).

There is evidence of both object-selective and scene-selective regions (Epstein & Baker, 2019) along the ventral visual stream. Epstein and Kanwisher (1998) identified the parahippocampal place area (PPA), a region of the cortex which responds more strongly to pictures of scenes (e.g., houses) than objects (e.g., bodies, faces) during both active perception and mental imagery of visual scenes. More recent studies have highlighted the role of PPA in recognition of non-visual information as well, such as descriptions of famous places (Aziz-Zadeh et al., 2008) and audio descriptions of spatial information (Häusler et al., 2022). Although not located in the ventral visual stream, the medial place area (MPA) and occipital place area (OPA) have demonstrated roles related to visually guided navigation and map-based navigation, respectively (Epstein & Kanwisher, 1998; Dilks et al., 2013, 2022; Nakamura et al., 2000; O'Craven & Kanwisher, 2000). An additional fMRI study identified a series of distributed cortical networks which show tuning specific to various scene categories (e.g., navigation, social interaction, human activity, motion-energy, texture, non-human animals, civilization, natural environment), demonstrating the complexity of scene processing in the human cerebral cortex (Çelik et al., 2021).

Many object-specific areas have been identified as well; some areas respond most to faces (fusiform and occipital face areas; Haxby et al., 2000; Kanwisher et al., 1997), shapes (posterior fusiform and lateral occipital complex (LOC); Malach et al., 1995; Grill-Spector & Malach, 2004), or bodies (fusiform and extrastriate body area; Downing et al., 2001; Peelen & Downing, 2005). A replication of Dilks et al. (2013) provided further evidence of a double dissociation in scene and object processing in the OPA and LOC. Transcranial magnetic stimulation (TMS) delivered to OPA impaired the recognition of scenes while TMS delivered to LOC impaired recognition of objects (Wischniewski & Peelen, 2021). Scene and object processing has also been explored using intracranial electroencephalography (iEEG), which offers high anatomical and temporal resolution. Vlcek et al. (2020) collected iEEG data as epileptic patients viewed images containing both objects and scenes. Their results support the roles of the PPA and LOC in scene and object processing, respectively, as well as scene-selective areas MPA and OPA. This scene network was shown to extend to regions involved in processing scene novelty (anterior temporal lobe regions, including the hippocampus and parahippocampal gyrus). Additional object-selective areas were identified, including areas selective for tool use (intraparietal sulcus, supramarginal gyrus and middle temporal cortex; Vidal et al., 2010) and object recognition (inferior frontal gyrus and perirhinal cortex; Bar et al., 2001; Clarke & Tyler, 2014; Nakamura et al., 2000).

Future neuroimaging studies or invasive neurophysiological studies will be necessary to evaluate how complex auditory scenes are processed along auditory-specific pathways. An additional research avenue could investigate the potential relationship between the ventral

visual stream and auditory scene processing; it is possible that visual areas may aid in the perception and representations of auditory scenes. Additionally, computational modeling studies could better our understanding of the neural and computational processes contributing to auditory scene processing. Most models of auditory scene analysis focus on the segregation of two auditory stimuli (e.g., tones, noise bursts, speech, foreground/background) into perceptual streams (Elhilali & Shamma, 2008; Krishnan et al., 2014; Ma, 2011), but these models are limited and do not explain how auditory objects and scenes are identified or understood. One model of the auditory system assessed the recognition and understanding of synthesized sound textures (i.e., temporally homogenous sounds such as a rainstorm or a choir of crickets; McDermott & Simoncelli, 2011); however, future studies are necessary to evaluate how auditory objects and scenes are processed from early to late stages in the auditory system.

Limitations

There are potential limitations of this study. Participants were not provided with examples of scenes at low, medium, and high levels of each global property scale, similar to Greene and Oliva (2009a, 2009b). Including examples of scenes at either extreme of each global property could provide participants with a better understanding of the task and should thus be included in future studies. The lack of examples could have made the scales harder to interpret; however, we are confident participants used each scale consistently since our measures of inter-rater reliability showed high agreement (see Table 2). Although ratings of global properties showed high agreement, it is possible that participants could not accurately rate some scales, such as temperature. Future studies should investigate other global variables that can be perceived more accurately in the auditory domain. It is also important to note the rating scale for Season did not include an option for “Between Fall and Winter,” which could have affected the results obtained from this scale.

There are many factors that may have influenced the degree to which participants used the global properties of scenes and/or objects within the scenes to make their judgments. The scene duration (4 seconds) could have led participants to rely on objects and the overall setting of each scene rather than global properties to make their judgments. We were not confident participants would be able to extract as much information from shorter durations (e.g., 1 or 2 seconds) compared to a longer duration of 4 seconds. However, if this experiment were to be repeated in the future with shorter scene durations, it is possible that participants could make judgments based more on the global properties instead of individual objects within the scenes. Additionally, the inter-rater reliability scores and factor structure could differ from what is reported here if shorter scene durations were to be used.

Further, this study is correlational, and used a relatively small group of 200 scenes. In the future, a more powerful design could use a higher number of scenes that vary in setting and types of objects present within each scene. Finally, this study was conducted online; although we included a headphone check and attention check, we did not have control over distractions, or the quality of headphones used by participants.

Conclusions

In summary, our results provide preliminary evidence for the ability to perceive auditory scenes from a global perspective. Additionally, the results of both the behavioral experiment and computational modeling analysis suggest a high degree of dimensionality reduction along the auditory pathway wherein global properties of scenes are processed at a high level and the acoustic features of scenes are processed at a low level. Examining the role of global properties as well as individual sound events in our perception of auditory scenes is essential to gain a

finer understanding of the underlying processes which construct our representations of the auditory environment.

ACKNOWLEDGMENTS

The authors wish to thank Dr. Jessica Nave-Blodgett, Chastity Balagtas, Jared Leslie, Sarah Flood, Lucy Pineda-Roman, Lan Nguyen, and Dr. Karli M. Nave for their assistance with collecting and editing auditory scenes for our database. They also thank Dr. Colleen M. Parks and Dr. Erin E. Hannon for their expertise and feedback on the methods and writing of this manuscript, as well as Rodica R. Constantine for her scientific insights, feedback, and general support of this project. Finally, the authors thank the reviewers whose valuable comments and suggestions helped to improve the quality of this manuscript.

FUNDING INFORMATION

Margaret A. McMullin was supported by the Department of Defense through the National Defense Science & Engineering Graduate (NDSEG) Fellowship Program. This work was partially supported by ONR N00014-23-1-2050.

AUTHOR CONTRIBUTIONS

M. A. M.: Conceptualization and design; Data collection; Formal analysis; Visualization; Writing – original draft; Writing – review & editing. R. K.: Conceptualization and design; Formal analysis; Visualization; Writing – original draft; Writing – review & editing. N. C. H.: Data cleaning and analysis; Writing – review & editing. B. G.: Data analysis; Writing – review & editing. M. E.: Conceptualization and design; Formal analysis; Supervision; Writing – original draft; Writing – review & editing. J. S. S.: Conceptualization and design; Formal analysis; Supervision; Writing – original draft; Writing – review & editing.

DATA AVAILABILITY STATEMENT

The data and stimuli are available at OSF: <https://osf.io/zj4xe/>.

REFERENCES

- Alain, C., Arnott, S. R., Hevenor, S., Graham, S., & Grady, C. L. (2001). "What" and "where" in the human auditory system. *Proceedings of the National Academy of Sciences*, 98(21), 12301–12306. <https://doi.org/10.1073/pnas.211209098>, PubMed: 11572938
- Aziz-Zadeh, L., Fiebach, C. J., Naranayan, S., Feldman, J., Dodge, E., & Ivry, R. B. (2008). Modulation of the FFA and PPA by language related to faces and places. *Social Neuroscience*, 3(3–4), 229–238. <https://doi.org/10.1080/17470910701414604>, PubMed: 18979378
- Ballas, J. A. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology: Human Perception and Performance*, 19(2), 250–267. <https://doi.org/10.1037/0096-1523.19.2.250>, PubMed: 8473838
- Bar, M., Tootell, R. B., Schacter, D. L., Greve, D. N., Fischl, B., Mendola, J. D., Rosen, B. R., & Dale, A. M. (2001). Cortical mechanisms specific to explicit visual object recognition. *Neuron*, 29(2), 529–535. [https://doi.org/10.1016/S0896-6273\(01\)00224-0](https://doi.org/10.1016/S0896-6273(01)00224-0), PubMed: 11239441
- Barton, B., Venezia, J. H., Saberi, K., Hickok, G., & Brewer, A. A. (2012). Orthogonal acoustic dimensions define auditory field maps in human cortex. *Proceedings of the National Academy of Sciences*, 109(50), 20738–20743. <https://doi.org/10.1073/pnas.1213381109>, PubMed: 23188798
- Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Cognitive Brain Research*, 13(1), 17–26. [https://doi.org/10.1016/S0926-6410\(01\)00084-2](https://doi.org/10.1016/S0926-6410(01)00084-2), PubMed: 11867247
- Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, 403(6767), 309–312. <https://doi.org/10.1038/35002078>, PubMed: 10659849
- Bendor, D., & Wang, X. (2005). The neuronal representation of pitch in primate auditory cortex. *Nature*, 436(7054), 1161–1165. <https://doi.org/10.1038/nature03867>, PubMed: 16121182
- Bendor, D., & Wang, X. (2006). Cortical representations of pitch in monkeys and humans. *Current Opinion in Neurobiology*, 16(4), 391–399. <https://doi.org/10.1016/j.conb.2006.07.001>, PubMed: 16842992
- Bendor, D., & Wang, X. (2010). Neural coding of periodicity in marmoset auditory cortex. *Journal of Neurophysiology*, 103(4), 1809–1822. <https://doi.org/10.1152/jn.00281.2009>, PubMed: 20147419

- Bethmann, A., & Brechmann, A. (2014). On the definition and interpretation of voice selective activation in the temporal cortex. *Frontiers in Human Neuroscience*, *8*, Article 499. <https://doi.org/10.3389/fnhum.2014.00499>, PubMed: 25071527
- Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147. <https://doi.org/10.1037/0033-295X.94.2.115>, PubMed: 3575582
- Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, *14*(10), 693–707. <https://doi.org/10.1038/nrn3565>, PubMed: 24052177
- Boebinger, D., Norman-Haignere, S. V., McDermott, J. H., & Kanwisher, N. (2021). Music-selective neural populations arise without musical training. *Journal of Neurophysiology*, *125*(6), 2237–2263. <https://doi.org/10.1152/jn.00588.2020>, PubMed: 33596723
- Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. MIT Press. <https://doi.org/10.7551/mitpress/1486.001.0001>
- Carandini, M., Heeger, D. J., & Anthony Movshon, J. (1999). Linearity and gain control in V1 simple cells. In P. S. Ulinski, E. G. Jones, & A. Peters (Eds.), *Models of cortical circuits* (pp. 401–443). Springer. https://doi.org/10.1007/978-1-4615-4903-1_7
- Çelik, E., Keles, U., Kiremitçi, İ., Gallant, J. L., & Çukur, T. (2021). Cortical networks of dynamic scene category representation in the human brain. *Cortex*, *143*, 127–147. <https://doi.org/10.1016/j.cortex.2021.07.008>, PubMed: 34411847
- Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, *13*(11), 1428–1432. <https://doi.org/10.1038/nn.2641>, PubMed: 20890293
- Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *Journal of the Acoustical Society of America*, *118*(2), 887–906. <https://doi.org/10.1121/1.1945807>, PubMed: 16158645
- Clarke, A., & Tyler, L. K. (2014). Object-specific semantic coding in human perirhinal cortex. *Journal of Neuroscience*, *34*(14), 4766–4775. <https://doi.org/10.1523/JNEUROSCI.2828-13.2014>, PubMed: 24695697
- Cohen, M. A., Horowitz, T. S., & Wolfe, J. M. (2009). Auditory recognition memory is inferior to visual recognition memory. *Proceedings of the National Academy of Sciences*, *106*(14), 6008–6010. <https://doi.org/10.1073/pnas.0811884106>, PubMed: 19307569
- Cumming, B. G., & DeAngelis, G. C. (2001). The physiology of stereopsis. *Annual Review of Neuroscience*, *24*, 203–238. <https://doi.org/10.1146/annurev.neuro.24.1.203>, PubMed: 11283310
- Da Costa, S., van der Zwaag, W., Marques, J. P., Frackowiak, R. S. J., Clarke, S., & Saenz, M. (2011). Human primary auditory cortex follows the shape of Heschl's gyrus. *Journal of Neuroscience*, *31*(40), 14067–14075. <https://doi.org/10.1523/JNEUROSCI.2000-11.2011>, PubMed: 21976491
- Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *Journal of Neuroscience*, *23*(8), 3423–3431. <https://doi.org/10.1523/JNEUROSCI.23-08-03423.2003>, PubMed: 12716950
- De, A., & Chaudhuri, R. (2023). Common population codes produce extremely nonlinear neural manifolds. *Proceedings of the National Academy of Sciences*, *120*(39), Article e2305853120. <https://doi.org/10.1073/pnas.2305853120>, PubMed: 37733742
- de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, *37*(27), 6539–6557. <https://doi.org/10.1523/JNEUROSCI.3267-16.2017>, PubMed: 28588065
- Desain, P., & Honing, H. (2003). The formation of rhythmic categories and metric priming. *Perception*, *32*(3), 341–365. <https://doi.org/10.1068/p3370>, PubMed: 12729384
- De Valois, R. L., & De Valois, K. K. (1980). Spatial vision. *Annual Review of Psychology*, *31*, 309–341. <https://doi.org/10.1146/annurev.ps.31.020180.001521>, PubMed: 7362215
- DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434. <https://doi.org/10.1016/j.neuron.2012.01.010>, PubMed: 22325196
- Dilks, D. D., Julian, J. B., Paunov, A. M., & Kanwisher, N. (2013). The occipital place area is causally and selectively involved in scene perception. *Journal of Neuroscience*, *33*(4), 1331–1336a. <https://doi.org/10.1523/JNEUROSCI.4081-12.2013>, PubMed: 23345209
- Dilks, D. D., Kamps, F. S., & Persichetti, A. S. (2022). Three cortical scene systems and their development. *Trends in Cognitive Sciences*, *26*(2), 117–127. <https://doi.org/10.1016/j.tics.2021.11.002>, PubMed: 34857468
- Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, *293*(5539), 2470–2473. <https://doi.org/10.1126/science.1063414>, PubMed: 11577239
- Elhilali, M., & Shamma, S. A. (2008). A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation. *Journal of the Acoustical Society of America*, *124*(6), 3751–3771. <https://doi.org/10.1121/1.3001672>, PubMed: 19206802
- Epstein, R. A., & Baker, C. I. (2019). Scene perception in the human brain. *Annual Review of Vision Science*, *5*, 373–397. <https://doi.org/10.1146/annurev-vision-091718-014809>, PubMed: 31226012
- Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601. <https://doi.org/10.1038/33402>, PubMed: 9560155
- Gegenfurtner, K. R., & Kiper, D. C. (2003). Color vision. *Annual Review of Neuroscience*, *26*, 181–206. <https://doi.org/10.1146/annurev.neuro.26.041002.131116>, PubMed: 12574494
- Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., & 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)* (pp. 776–780). IEEE. <https://doi.org/10.1109/ICASSP.2017.7952261>
- Geisler, W. S., & Perry, J. S. (2011). Statistics for optimal point prediction in natural images. *Journal of Vision*, *11*(12), Article 14. <https://doi.org/10.1167/11.12.14>, PubMed: 22011382
- Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25. [https://doi.org/10.1016/0166-2236\(92\)90344-8](https://doi.org/10.1016/0166-2236(92)90344-8), PubMed: 1374953
- Greene, M. R., & Oliva, A. (2009a). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, *20*(4), 464–472. <https://doi.org/10.1111/j.1467-9280.2009.02316.x>, PubMed: 19399976
- Greene, M. R., & Oliva, A. (2009b). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*(2), 137–176. <https://doi.org/10.1016/j.cogpsych.2008.06.001>, PubMed: 18762289

- Greene, M. R., & Oliva, A. (2010). High-level aftereffects to global scene properties. *Journal of Experimental Psychology: Human Perception and Performance*, 36(6), 1430–1442. <https://doi.org/10.1037/a0019058>, PubMed: 20731502
- Gregg, M. K., & Samuel, A. G. (2008). Change deafness and the organizational properties of sounds. *Journal of Experimental Psychology: Human Perception and Performance*, 34(4), 974–991. <https://doi.org/10.1037/0096-1523.34.4.974>, PubMed: 18665739
- Gregg, M. K., & Samuel, A. G. (2009). The importance of semantics in auditory representations. *Attention, Perception, & Psychophysics*, 71(3), 607–619. <https://doi.org/10.3758/APP.71.3.607>, PubMed: 19304650
- Gregg, M. K., Irsik, V. C., & Snyder, J. S. (2014). Change deafness and object encoding with recognizable and unrecognizable sounds. *Neuropsychologia*, 61, 19–30. <https://doi.org/10.1016/j.neuropsychologia.2014.06.007>, PubMed: 24937187
- Gregg, M. K., Irsik, V. C., & Snyder, J. S. (2017). Effects of capacity limits, memory loss, and sound type in change deafness. *Attention, Perception, & Psychophysics*, 79(8), 2564–2575. <https://doi.org/10.3758/s13414-017-1416-4>, PubMed: 28856615
- Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America*, 61(5), 1270–1277. <https://doi.org/10.1121/1.381428>, PubMed: 560400
- Griffiths, T. D. (2008). Sensory systems: Auditory action streams? *Current Biology*, 18(9), R387–R388. <https://doi.org/10.1016/j.cub.2008.03.007>, PubMed: 18460320
- Grill-Spector, K., & Malach, R. (2004). The human visual cortex. *Annual Review of Neuroscience*, 27, 649–677. <https://doi.org/10.1146/annurev.neuro.27.070203.144220>, PubMed: 15217346
- Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27), 10005–10014. <https://doi.org/10.1523/JNEUROSCI.5023-14.2015>, PubMed: 26157000
- Gygi, B., Kidd, G. R., & Watson, C. S. (2007). Similarity and categorization of environmental sounds. *Perception & Psychophysics*, 69(6), 839–855. <https://doi.org/10.3758/BF03193921>, PubMed: 18018965
- Gygi, B., & Shafiro, V. (2010). Development of the Database for Environmental Sound Research and Application (DESRA): Design, functionality, and retrieval considerations. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010, Article 654914. <https://doi.org/10.1155/2010/654914>
- Hansen, N. E., Noesen, B. T., Nador, J. D., & Harel, A. (2018). The influence of behavioral relevance on the processing of global scene properties: An ERP study. *Neuropsychologia*, 114, 168–180. <https://doi.org/10.1016/j.neuropsychologia.2018.04.040>, PubMed: 29729276
- Harel, A., Groen, I. I. A., Kravitz, D. J., Deouell, L. Y., & Baker, C. I. (2016). The temporal dynamics of scene processing: A multifaceted EEG investigation. *ENEURO*, 3(5), Article ENEURO.0139-16.2016. <https://doi.org/10.1523/ENEURO.0139-16.2016>, PubMed: 27699208
- Häusler, C. O., Eickhoff, S. B., & Hanke, M. (2022). Processing of visual and non-visual naturalistic spatial information in the “parahippocampal place area”. *Scientific Data*, 9(1), Article 147. <https://doi.org/10.1038/s41597-022-01250-4>, PubMed: 35365659
- Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, 4(6), 223–233. [https://doi.org/10.1016/S1364-6613\(00\)01482-0](https://doi.org/10.1016/S1364-6613(00)01482-0), PubMed: 10827445
- Heittola, T., Mesaros, A., & Virtanen, T. (2020). TAU urban acoustic scenes 2020 mobile, evaluation dataset. *Zenodo*. <https://doi.org/10.5281/zenodo.3685828>
- Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, 92(1–2), 67–99. <https://doi.org/10.1016/j.cognition.2003.10.011>, PubMed: 15037127
- Higgins, N. C., McLaughlin, S. A., Rinne, T., & Stecker, G. C. (2017). Evidence for cue-independent spatial representation in the human auditory cortex during active listening. *Proceedings of the National Academy of Sciences*, 114(36), E7602–E7611. <https://doi.org/10.1073/pnas.1707522114>, PubMed: 28827357
- Horwitz, G. D., & Hass, C. A. (2012). Nonlinear analysis of macaque V1 color tuning reveals cardinal directions for cortical color processing. *Nature Neuroscience*, 15(6), 913–919. <https://doi.org/10.1038/nn.3105>, PubMed: 22581184
- Houtgast, T., & Steeneken, H. J. M. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *Journal of the Acoustical Society of America*, 77(3), 1069–1077. <https://doi.org/10.1121/1.392224>
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv*. <https://doi.org/10.48550/arXiv.1704.04861>
- Hu, H., Yang, C.-H. H., Xia, X., Bai, X., Tang, X., Wang, Y., Niu, S., Chai, L., Li, J., Zhu, H., Bao, F., Zhao, Y., Siniscalchi, S. M., Wang, Y., Du, J., & Lee, C.-H. (2020). Device-robust acoustic scene classification based on two-stage categorization and data augmentation. *arXiv*. <https://doi.org/10.48550/arXiv.2007.08389>
- Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *Journal of Physiology*, 160(1), 106–154. <https://doi.org/10.1113/jphysiol.1962.sp006837>, PubMed: 14449617
- Humphries, C., Liebenthal, E., & Binder, J. R. (2010). Tonotopic organization of human auditory cortex. *NeuroImage*, 50(3), 1202–1211. <https://doi.org/10.1016/j.neuroimage.2010.01.046>, PubMed: 20096790
- IBM Corp. (2020). *IBM SPSS Statistics for Windows, version 27.0*. IBM Corp.
- Iyer, N., Thompson, E. R., Simpson, B. D., Brungart, D., & Summers, V. (2013). Exploring auditory gist: Comprehension of two dichotic, simultaneously presented stories. *Proceedings of Meetings on Acoustics*, 19(1), Article 050158. <https://doi.org/10.1121/1.4800507>
- Jacoby, N., & McDermott, J. H. (2017). Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Current Biology*, 27(3), 359–370. <https://doi.org/10.1016/j.cub.2016.12.031>, PubMed: 28065607
- JASP Team. (2022). JASP (version 0.16.1)[Computer software]. <https://jasp-stats.org/>
- Kaas, J. H., & Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences*, 97(22), 11793–11799. <https://doi.org/10.1073/pnas.97.22.11793>, PubMed: 11050211
- Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11), 4302–4311. <https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997>, PubMed: 9151747
- Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3), 630–644. <https://doi.org/10.1016/j.neuron.2018.03.044>, PubMed: 29681533

- Kell, A. J., & McDermott, J. H. (2019). Deep neural network models of sensory systems: Windows onto the role of task constraints. *Current Opinion in Neurobiology*, *55*, 121–132. <https://doi.org/10.1016/j.conb.2019.02.003>, PubMed: 30884313
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>, PubMed: 27330520
- Krishnan, L., Elhilali, M., & Shamma, S. (2014). Segregating complex sound sources through temporal coherence. *PLoS Computational Biology*, *10*(12), Article e1003985. <https://doi.org/10.1371/journal.pcbi.1003985>, PubMed: 25521593
- Krumhansl, C. L. (1979). The psychological representation of musical pitch in a tonal context. *Cognitive Psychology*, *11*(3), 346–374. [https://doi.org/10.1016/0010-0285\(79\)90016-1](https://doi.org/10.1016/0010-0285(79)90016-1)
- Kumar, S., Stephan, K. E., Warren, J. D., Friston, K. J., & Griffiths, T. D. (2007). Hierarchical processing of auditory objects in humans. *PLoS Computational Biology*, *3*(6), Article e100. <https://doi.org/10.1371/journal.pcbi.0030100>, PubMed: 17542641
- Lee, J. H., Russ, B. E., Orr, L. E., & Cohen, Y. E. (2009). Prefrontal activity predicts monkeys' decisions during an auditory category task. *Frontiers in Integrative Neuroscience*, *3*, Article 16. <https://doi.org/10.3389/neuro.07.016.2009>, PubMed: 19587846
- Leech, R., Gygi, B., Aydelott, J., & Dick, F. (2009). Informational factors in identifying environmental sounds in natural auditory scenes. *Journal of the Acoustical Society of America*, *126*(6), 3147–3155. <https://doi.org/10.1121/1.3238160>, PubMed: 20000928
- Lomber, S. G., & Malhotra, S. (2008). Double dissociation of 'what' and 'where' processing in auditory cortex. *Nature Neuroscience*, *11*(5), 609–616. <https://doi.org/10.1038/nn.2108>, PubMed: 18408717
- Ma, L. (2011). *Auditory streaming: Behavior, physiology, and modeling* [Doctoral dissertation]. University of Maryland.
- Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., Ledden, P. J., Brady, T. J., Rosen, B. R., & Tootell, R. B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, *92*(18), 8135–8139. <https://doi.org/10.1073/pnas.92.18.8135>, PubMed: 7667258
- MATLAB. (2021). *Version 9.10.0.1851785 (R2010a)*. The MathWorks Inc.
- McDermott, J. H., & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, *71*(5), 926–940. <https://doi.org/10.1016/j.neuron.2011.06.032>, PubMed: 21903084
- Mehr, S. A., Singh, M., York, H., Glowacki, L., & Krasnow, M. M. (2018). Form and function in human song. *Current Biology*, *28*(3), 356–368. <https://doi.org/10.1016/j.cub.2017.12.042>, PubMed: 29395919
- Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006–1010. <https://doi.org/10.1126/science.1245994>, PubMed: 24482117
- Milner, D., & Goodale, M. (2006). *The visual brain in action* (2nd ed.). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198524724.001.0001>
- Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences*, *6*, 414–417. [https://doi.org/10.1016/0166-2236\(83\)90190-X](https://doi.org/10.1016/0166-2236(83)90190-X)
- Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978). Receptive field organization of complex cells in the cat's striate cortex. *Journal of Physiology*, *283*(1), 79–99. <https://doi.org/10.1113/jphysiol.1978.sp012489>, PubMed: 722592
- Nakamura, K., Kawashima, R., Sato, N., Nakamura, A., Sugiura, M., Kato, T., Hatano, K., Ito, K., Fukuda, H., Schormann, T., & Zilles, K. (2000). Functional delineation of the human occipito-temporal areas related to face and scene processing. A PET study. *Brain*, *123*(9), 1903–1912. <https://doi.org/10.1093/brain/123.9.1903>, PubMed: 10960054
- Ng, C.-W., Plakke, B., & Poremba, A. (2014). Neural correlates of auditory recognition memory in the primate dorsal temporal pole. *Journal of Neurophysiology*, *111*(3), 455–469. <https://doi.org/10.1152/jn.00401.2012>, PubMed: 24198324
- Norman-Haignere, S. V., Feather, J., Boebinger, D., Brunner, P., Ritaccio, A., McDermott, J. H., Schalk, G., & Kanwisher, N. (2022). A neural population selective for song in human auditory cortex. *Current Biology*, *32*(7), 1470–1484. <https://doi.org/10.1016/j.cub.2022.01.069>, PubMed: 35196507
- Norman-Haignere, S., Kanwisher, N., & McDermott, J. H. (2013). Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. *Journal of Neuroscience*, *33*(50), 19451–19469. <https://doi.org/10.1523/JNEUROSCI.2880-13.2013>, PubMed: 24336712
- Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, *88*(6), 1281–1296. <https://doi.org/10.1016/j.neuron.2015.11.035>, PubMed: 26687225
- O'Craven, K. M., & Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *Journal of Cognitive Neuroscience*, *12*(6), 1013–1023. <https://doi.org/10.1162/089892900051137549>, PubMed: 11177421
- Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I.-H., Saberi, K., Serences, J. T., & Hickok, G. (2010). Hierarchical organization of human auditory cortex: Evidence from acoustic invariance in the response to intelligible speech. *Cerebral Cortex*, *20*(10), 2486–2495. <https://doi.org/10.1093/cercor/bhp318>, PubMed: 20100898
- Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*, 145–175. <https://doi.org/10.1023/A:1011139631724>
- Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*, 23–36. [https://doi.org/10.1016/S0079-6123\(06\)55002-2](https://doi.org/10.1016/S0079-6123(06)55002-2), PubMed: 17027377
- Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, *18*(6), 903–911. <https://doi.org/10.1038/nn.4021>, PubMed: 25984889
- Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., & Griffiths, T. D. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron*, *36*(4), 767–776. [https://doi.org/10.1016/S0896-6273\(02\)01060-7](https://doi.org/10.1016/S0896-6273(02)01060-7), PubMed: 12441063
- Pearson, R. H., & Mundform, D. J. (2010). Recommended sample size for conducting exploratory factor analysis on dichotomous data. *Journal of Modern Applied Statistical Methods*, *9*(2). <https://doi.org/10.22237/jmasm/1288584240>
- Peelen, M. V., & Downing, P. E. (2005). Selectivity for the human body in the fusiform gyrus. *Journal of Neurophysiology*, *93*(1), 603–608. <https://doi.org/10.1152/jn.00513.2004>, PubMed: 15295012

- Penagos, H., Melcher, J. R., & Oxenham, A. J. (2004). A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. *Journal of Neuroscience*, *24*(30), 6810–6815. <https://doi.org/10.1523/JNEUROSCI.0383-04.2004>, PubMed: 15282286
- Pernet, C. R., McAleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E. G., Watson, R. H., Fleming, D., Crabbe, F., Valdes-Sosa, M., & Belin, P. (2015). The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage*, *119*, 164–174. <https://doi.org/10.1016/j.neuroimage.2015.06.050>, PubMed: 26116964
- Petkov, C. I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., & Logothetis, N. K. (2008). A voice region in the monkey brain. *Nature Neuroscience*, *11*(3), 367–374. <https://doi.org/10.1038/nn2043>, PubMed: 18264095
- Plakke, B., & Romanski, L. M. (2014). Auditory connections and functions of prefrontal cortex. *Frontiers in Neuroscience*, *8*, Article 199. <https://doi.org/10.3389/fnins.2014.00199>, PubMed: 25100931
- Poremba, A., Saunders, R. C., Crane, A. M., Cook, M., Sokoloff, L., & Mishkin, M. (2003). Functional mapping of the primate auditory system. *Science*, *299*(5606), 568–572. <https://doi.org/10.1126/science.1078900>, PubMed: 12543977
- Rabiner, L., & Schafer, R. (2010). *Theory and applications of digital speech processing*. Prentice Hall Press.
- Rauschecker, J. P. (1998). Parallel processing in the auditory cortex of primates. *Audiology & Neuro-otology*, *3*(2–3), 86–103. <https://doi.org/10.1159/000013784>, PubMed: 9575379
- Rauschecker, J. P. (2012). Ventral and dorsal streams in the evolution of speech and language. *Frontiers in Evolutionary Neuroscience*, *4*, Article 7. <https://doi.org/10.3389/fnevo.2012.00007>, PubMed: 22615693
- Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*(6), 718–724. <https://doi.org/10.1038/nn.2331>, PubMed: 19471271
- Rauschecker, J. P., & Tian, B. (2000). Mechanisms and streams for processing of “what” and “where” in auditory cortex. *Proceedings of the National Academy of Sciences*, *97*(22), 11800–11806. <https://doi.org/10.1073/pnas.97.22.11800>, PubMed: 11050212
- Rauschecker, J. P., & Tian, B. (2004). Processing of band-passed noise in the lateral auditory belt cortex of the rhesus monkey. *Journal of Neurophysiology*, *91*(6), 2578–2589. <https://doi.org/10.1152/jn.00834.2003>, PubMed: 15136602
- Rauschecker, J. P., Tian, B., & Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science*, *268*(5207), 111–114. <https://doi.org/10.1126/science.7701330>, PubMed: 7701330
- Romanski, L. M., Averbeck, B. B., & Diltz, M. (2005). Neural representation of vocalizations in the primate ventrolateral prefrontal cortex. *Journal of Neurophysiology*, *93*(2), 734–747. <https://doi.org/10.1152/jn.00675.2004>, PubMed: 15371495
- Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., ... Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770. <https://doi.org/10.1038/s41593-019-0520-2>, PubMed: 31659335
- Ross, M. G., & Oliva, A. (2010). Estimating perception of scene layout properties from global image features. *Journal of Vision*, *10*(1), Article 2. <https://doi.org/10.1167/10.1.2>, PubMed: 20143895
- Russ, B. E., Ackelson, A. L., Baker, A. E., & Cohen, Y. E. (2008). Coding of auditory-stimulus identity in the auditory non-spatial processing stream. *Journal of Neurophysiology*, *99*(1), 87–95. <https://doi.org/10.1152/jn.01069.2007>, PubMed: 18003874
- Russ, B. E., Orr, L. E., & Cohen, Y. E. (2008). Prefrontal neurons predict choices during an auditory same-different task. *Current Biology*, *18*(19), 1483–1488. <https://doi.org/10.1016/j.cub.2008.08.054>, PubMed: 18818080
- Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Computational Biology*, *10*(1), Article e1003412. <https://doi.org/10.1371/journal.pcbi.1003412>, PubMed: 24391486
- Saxe, A., Nelli, S., & Summerfield, C. (2021). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, *22*(1), 55–67. <https://doi.org/10.1038/s41583-020-00395-8>, PubMed: 33199854
- Schönwiesner, M., & Zatorre, R. J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences*, *106*(34), 14611–14616. <https://doi.org/10.1073/pnas.0907682106>, PubMed: 19667199
- Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, *123*(12), 2400–2406. <https://doi.org/10.1093/brain/123.12.2400>, PubMed: 11099443
- Sharda, M., & Singh, N. C. (2012). Auditory perception of natural sound categories—An fMRI study. *Neuroscience*, *214*, 49–58. <https://doi.org/10.1016/j.neuroscience.2012.03.053>, PubMed: 22522473
- Sharpee, T. O., Atencio, C. A., & Schreiner, C. E. (2011). Hierarchical representations in the auditory cortex. *Current Opinion in Neurobiology*, *21*(5), 761–767. <https://doi.org/10.1016/j.conb.2011.05.027>, PubMed: 21704508
- Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. *Psychological Review*, *89*(4), 305–333. <https://doi.org/10.1037/0033-295X.89.4.305>, PubMed: 7134331
- Slaney, M. (1995). *Auditory toolbox: A MATLAB toolbox for auditory modeling work* [Apple Tech. Rep. No. 45]. Apple Computer.
- Snyder, J. S., Gregg, M. K., Weintraub, D. M., & Alain, C. (2012). Attention, awareness, and the perception of auditory scenes. *Frontiers in Psychology*, *3*, Article 15. <https://doi.org/10.3389/fpsyg.2012.00015>, PubMed: 22347201
- Stecker, G. C., Harrington, I. A., & Middlebrooks, J. C. (2005). Location coding by opponent neural populations in the auditory cortex. *PLoS Biology*, *3*(3), Article e78. <https://doi.org/10.1371/journal.pbio.0030078>, PubMed: 15736980
- Tian, B., & Rauschecker, J. P. (2004). Processing of frequency-modulated sounds in the lateral auditory belt cortex of the rhesus monkey. *Journal of Neurophysiology*, *92*(5), 2993–3013. <https://doi.org/10.1152/jn.00472.2003>, PubMed: 15486426
- Tian, B., Reser, D., Durham, A., Kustov, A., & Rauschecker, J. P. (2001). Functional specialization in rhesus monkey auditory cortex. *Science*, *292*(5515), 290–293. <https://doi.org/10.1126/science.1058911>, PubMed: 11303104
- Toivainen, P., Kaipainen, M., & Louhivuori, J. (1995). Musical timbre: Similarity ratings correlate with computational feature space distances. *Journal of New Music Research*, *24*(3), 282–298. <https://doi.org/10.1080/09298219508570686>
- Tootell, R. B., Hadjikhani, N. K., Vanduffel, W., Liu, A. K., Mendola, J. D., Sereno, M. I., & Dale, A. M. (1998). Functional analysis of primary visual cortex (V1) in humans. *Proceedings of the National Academy of Sciences*, *95*(3), 811–817. <https://doi.org/10.1073/pnas.95.3.811>, PubMed: 9448245

- Tootell, R. B., Switkes, E., Silverman, M. S., & Hamilton, S. L. (1988). Functional anatomy of macaque striate cortex. II. Retinotopic organization. *Journal of Neuroscience*, *8*(5), 1531–1568. <https://doi.org/10.1523/JNEUROSCI.08-05-01531.1988>, PubMed: 3367210
- Vidal, J. R., Ossandón, T., Jerbi, K., Dalal, S. S., Minotti, L., Ryvlin, P., Kahane, P., & Lachaux, J.-P. (2010). Category-specific visual responses: An intracranial study comparing gamma, beta, alpha, and ERP response selectivity. *Frontiers in Human Neuroscience*, *4*, Article 195. <https://doi.org/10.3389/fnhum.2010.00195>, PubMed: 21267419
- Vlcek, K., Fajnerova, I., Nekovarova, T., Hejtmanek, L., Janca, R., Jezdik, P., Kalina, A., Tomasek, M., Krsek, P., Hammer, J., & Marusic, P. (2020). Mapping the scene and object processing networks by intracranial EEG. *Frontiers in Human Neuroscience*, *14*, Article 561399. <https://doi.org/10.3389/fnhum.2020.561399>, PubMed: 33192393
- Wang, L., Liu, H., Zhang, X., Zhao, S., Guo, L., Han, J., & Hu, X. (2022). Exploring hierarchical auditory representation via a neural encoding model. *Frontiers in Neuroscience*, *16*, Article 843988. <https://doi.org/10.3389/fnins.2022.843988>, PubMed: 35401085
- Wiesmann, S. L., & Vö, M. L.-H. (2022). What makes a scene? Fast scene categorization as a function of global scene information at different resolutions. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(8), 871–888. <https://doi.org/10.1037/xhp0001020>, PubMed: 35708933
- Wiesmann, S. L., & Vö, M. L.-H. (2023). Disentangling diagnostic object properties for human scene categorization. *Scientific Reports*, *13*(1), Article 5912. <https://doi.org/10.1038/s41598-023-32385-y>, PubMed: 37041222
- Wischnowski, M., & Peelen, M. V. (2021). Causal evidence for a double dissociation between object- and Scene-Selective regions of visual cortex: A preregistered TMS replication study. *Journal of Neuroscience*, *41*(4), 751–756. <https://doi.org/10.1523/JNEUROSCI.2162-20.2020>, PubMed: 33262244
- Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1–3), 37–52. [https://doi.org/10.1016/0169-7439\(87\)80084-9](https://doi.org/10.1016/0169-7439(87)80084-9)
- Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064–2072. <https://doi.org/10.3758/s13414-017-1361-2>, PubMed: 28695541
- Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 3485–3492). IEEE. <https://doi.org/10.1109/CVPR.2010.5539970>
- Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. <https://doi.org/10.1038/nn.4244>, PubMed: 26906502
- Yi, H. G., Leonard, M. K., & Chang, E. F. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron*, *102*(6), 1096–1110. <https://doi.org/10.1016/j.neuron.2019.04.023>, PubMed: 31220442
- Zhang, F., Wang, J.-P., Kim, J., Parrish, T., & Wong, P. C. M. (2015). Decoding multiple sound categories in the human temporal cortex using high resolution fMRI. *PLoS One*, *10*(2), Article e0117303. <https://doi.org/10.1371/journal.pone.0117303>, PubMed: 25692885
- Zhang, J., Zhang, G., Li, X., Wang, P., Wang, B., & Liu, B. (2020). Decoding sound categories based on whole-brain functional connectivity patterns. *Brain Imaging and Behavior*, *14*(1), 100–109. <https://doi.org/10.1007/s11682-018-9976-z>, PubMed: 30361945