Attention modulates neural representations of acoustic, categorical, and identity features in a task-dependent manner

Varis, O[1] - Muukkonen, I[1,2] - Wikman, P[1,3,4,5]

1 Department of Psychology, PO Box 21 FI-00014 University of Helsinki, Finland

2 Department of Brain and Cognition, PO box 03711, KU Leuven, Leuven, Belgium

3 Advanced Magnetic Imaging Centre, PO Box 11000 FI-00076, Aalto University, Espoo, Finland

4 Neuroscience Center, HiLife, PO Box 21 FI-00014 University of Helsinki, Finland

5 Helsinki Collegium For Advanced Studies, PO Box 21 FI-00014 University of Helsinki, Finland

Corresponding author: Onnipekka Varis (onnipekka.varis@helsinki.fi)

Acknowledgment: This work was supported by The Research Council of Finland (grant number #1348353). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We want to thank M. Salmikivi, W. Vikatmaa, S. Rossow, and L. Lehtimäki for helping with data acquisition, as well as Ville Laaksonen and Jaakko Kauramäki for help with the study design.

Data Availability Statement: All preprocessed fMRI data, anonymized to protect participant privacy, and the corresponding analysis code will be deposited on the Open Science Framework (OSF).

Abstract

Aims:
The human auditory system represents sounds at multiple levels, from low-level acoustic features to abstract category- and object-level information. Although selective auditory attention enables listening in complex soundscapes, it remains unclear which representational levels are modulated by attention and how this depends on scene structure. Using fMRI, representational similarity analysis (RSA), and cross-experiment decoding, this study examined how attention modulates neural representations of sound features under different competitive listening conditions.

Methods:
Twenty participants completed an fMRI experiment involving three selective attention conditions: one sound presented alone, three overlapping sounds from different categories, and three overlapping sounds from the same category. Neural activation patterns were compared to models of multiple low-level acoustic and abstract sound features using RSA, with covariance between models statistically removed to isolate unique representations. Attentional modulation was quantified by contrasting representations of attended and ignored sounds' features, and cross-experiment decoding was used to assess object-identity representations.

Conclusions:
Low-level acoustic features were robustly represented in early auditory cortex, whereas abstract sound features were broadly represented along temporal and frontal regions of the auditory ventral stream. Attentional modulation was task-dependent: when competing sounds could not be distinguished by category, attention enhanced low-level acoustic feature processing, whereas in cross-category scenes attention primarily targeted abstract category-level representations. Object-identity representations were modulated by attention across both scene types, with category-specific differences. These findings demonstrate that auditory attention flexibly targets the representational level most informative for resolving competition in complex soundscapes.

Introduction

The human auditory system integrates different low-level sound features (e.g., pitch, amplitude) into abstract sound features, which represent complex and meaningful information, such as the category of a sound or the identity of a speaker (Theunissen & Elie, 2014). These operations have been suggested to rely on neural processing in the auditory ventral stream (AVS) where sound features are processed along a gradient from simple to abstract (Hickok & Poeppel, 2007; Rauschecker & Scott, 2009; Rauschecker & Tian, 2000). Specifically, the primary auditory cortex (A1) differentiates low-level features such as pitch (Saenz & Langers, 2014), while the belt areas surrounding it are sensitive to band pass noise bursts - sounds that contain a range of frequencies within a defined bandwidth (Wessinger et al., 2001). Following the belt regions, the superior temporal gyrus (STG) shows intermediate representations that bridge low-level and abstract sound features (Giordano et al., 2023). Further along this pathway, both the superior temporal sulcus (STS), anterior portions of the STG and the inferior frontal gyrus (IFG) exhibit category-selective responses to complex sound classes, particularly vocalizations and speech (Alho et al., 2006; Belin et al., 2000; Latinus & Belin, 2011; Liakakis et al., 2011; Sharda & Singh, 2012).

While the cortical progression from low-level to abstract sound feature processing is relatively well established, much less is known about how attention enables selective processing of naturalistic sound objects in the presence of other competing sounds. A central question in auditory attention research pertains to which sound features are selectively enhanced by attention, and where in the auditory hierarchy these effects emerge. fMRI studies using simple tones have shown that attention can modulate activity already in primary auditory areas in a frequency-specific manner, suggesting that attention may act as a low-level filter (Costa et al., 2013; Paltoglou et al., 2009; Riecke et al., 2017). However, it is not yet clear whether such effects generalise to other low-level sound features such as amplitude variability or harmonicity.

A limitation of studies using simple stimuli such as tones is that these stimuli do not possess abstract sound features that are present in natural sounds. Indeed, studies using overlapping natural sounds with rich abstract properties (e.g., overlapping

speech) have generally not observed reliable low-level attentional effects in early auditory areas. Instead, attentional modulation tends to emerge at later representational stages, typically in STG, STS, and frontal regions (Kiremitçi et al., 2021; J. O'Sullivan et al., 2019; Puschmann et al., 2024; Renvall et al., 2021; Wikman et al., 2021). For example, Kiremitçi et al. (2021) showed that in a two-speaker cocktail-party task, attention enhanced articulatory and semantic—but not spectral—feature representations, with effects localised primarily to higher-order auditory regions. This preferential enhancement of higher-level articulatory and semantic information over low-level spectral coding suggests that, in complex natural scenes, attention often operates at the level of auditory objects or streams rather than individual low-level features.

An important constraint on the generalizability of previous work on auditory attention is that it has largely focused on scenes containing either overlapping simple sounds or overlapping speech. The studies that have investigated attentional modulation in more heterogeneous soundscapes suggest attentional modulation is category-specific and emerges in distinct temporal regions depending on the category of the attended sound (Hausfeld et al., 2018; Wikman et al., 2025) with speech in particular showing privileged attentional modulation in supratemporal cortex (Renvall et al., 2021). Moreover, attentional selection appears to rely on different neural mechanisms depending on the structure of the scene: in scenes where the overlapping sounds are from the same category, attention primarily enhances object-level representations, whereas in cross-category scenes it engages category-level distinctions (Pedersini et al., 2025; Wikman et al., 2025). Thus, attention may flexibly target both object- and feature-level representations depending on the soundscape (See also Shamma et al., 2011; Wikman et al., 2021).

Another challenge in studying the neural basis of auditory processing is that different sound features are inherently correlated with each other (Giordano et al., 2013; Theunissen & Elie, 2014). Hence, to dissociate the role each sound feature plays in neural activity, the covariance of between sound features must be taken into account. This problem has been partially solved in prior research by selecting or creating auditory stimuli with matched low-level profiles across abstract categories (e.g., Petkov et al., 2008; Sharda & Singh, 2012). While such methods help isolate higher-

level effects, they also reduce the ecological validity of the sounds, as they are by design highly similar in their low-level features. Another approach involves using statistical methods to remove covariances between models of different sound features (Giordano et al., 2013; Kiremitçi et al., 2021; Leaver & Rauschecker, 2010). This strategy has, however, not yet been adopted to delineate which sound features are modulated by attention in category-heterogenous scenes.

The central aim of this study was to delineate the level of attentional modulation (low-level acoustic features/abstract features) in complex scenes containing natural sounds. For this, we analysed data from an fMRI paradigm (previously reported in Wikman et al., 2025) where participants were presented with sounds from three natural sound categories: speech, instrument sounds, and animal sounds in three separate experiments (Figure 1). In the One Object Alone (1OA) experiment participants were asked to attend to one sound, which was played in isolation. In the three objects across categories (3OA) experiment, participants selectively attended to one sound from a soundscape which contained a sound from each of three different sound categories (e.g., Man, Trumpet, Bird). And finally, in the three objects within category (3OW) experiment, participants selectively attended to one sound from a soundscape consisting of three sounds from the same category (e.g., Bird, Monkey, Whale). Using the same approach as Giordano et al. (2013), we modelled low-level and abstract sound features of the stimuli with representational similarity analysis (RSA; Kriegeskorte et al., 2008; Kriegeskorte & Kievit, 2013) and removed covariances between the sound feature models. We then computed representational similarities between sound-feature models and neural activation patterns. Attentional modulation of a given sound feature was quantified as the difference in its representational strength between attended and distractor conditions (Figure 4). We finally used cross-experiment decoding, to evaluate for each category separately which brain regions store information that distinguish different individuals of the category more precisely than other category individuals and whether attention boosts such representations.

In the listening task (1OA), robust representations of low-level sound features were observed in early auditory cortex, whereas abstract sound features were broadly represented across temporal and frontal regions. During selective attention, when

competing sounds belonged to the same category (3OW), representations of low-level sound features were selectively enhanced for attended sounds. In contrast, when sound objects could be distinguished based on category-level information (3OA), attention primarily operated at an abstract category level rather than enhancing low-level feature processing. Importantly, object-identity representations were modulated by attention in both types of sound scenes, with the direction and strength of modulation differing across sound categories. These results highlight the dynamic and task-dependent nature of attention – i.e., attention modulates different levels of hierarchical processing based on which level provides the most distinct contrast between the overlapping sounds and operates differently based on the category of the attended sound.

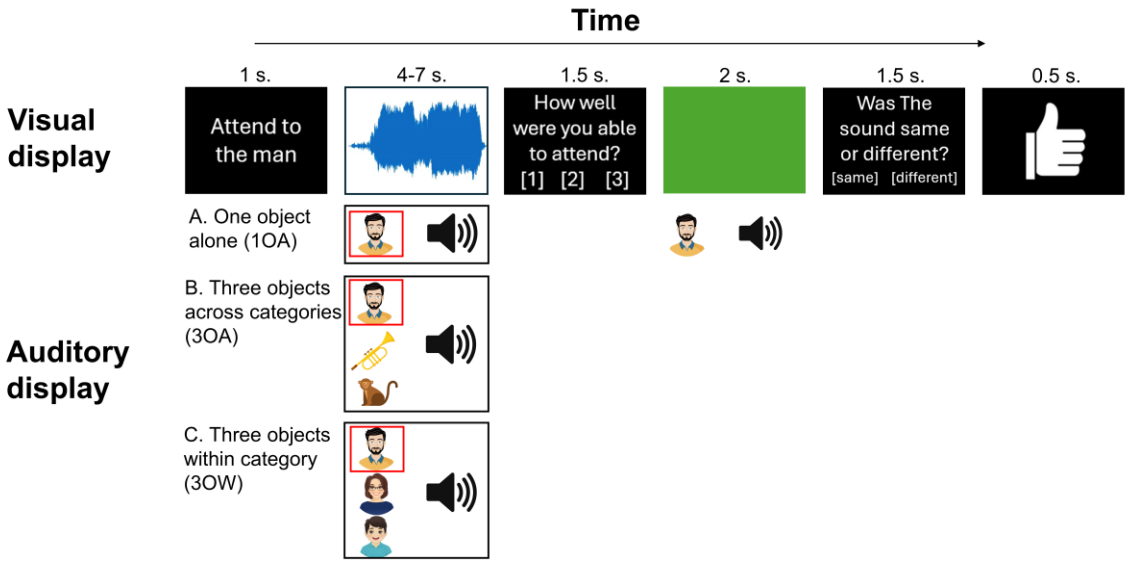

Figure 1. *A schematic representation of the trial structure and experiments*. Every trial started with an instruction which oriented the participant to attend to a specific subcategory. After the instruction the participant was presented with stimuli. A) In the One object alone (1OA) experiment, the participant was presented with only one auditory stimulus. B) In the three objects across categories (3OA) experiment the participant was presented with one auditory stimulus from each of the three categories. C) In the three objects within category (3OW) experiment the participant was presented with three auditory stimuli from the same category. The visual stimulus was also presented regardless of the experiment. After the stimulus presentation, the participant rated how well they were able to selectively attend to the target stimulus. In 25% of the trials participants completed a match-to-sample task. In the match-to-sample task the participant was presented with a short sample of either the target stimulus or another stimulus from the same subcategory. Afterwards the participant determined whether the sample was from the target stimulus or from a different stimulus. Finally, the participant was given visual feedback on whether the match-to-sample task was answered correctly or incorrectly.

Methods

Participants

Twenty healthy native Finnish-speaking, right-handed individuals (8 females, age range 19–32 years, mean age 22.7, SD = 3.68) participated in the fMRI session. All participants had normal or corrected-to-normal vision and hearing, with no history of neurological or developmental disorders. To assess handedness, the Edinburgh Handedness inventory (Oldfield, 1971) was used. Written consent was obtained from each participant prior to their involvement. For their time, participants received monetary compensation at a rate of 15 €/hour. The study received approval from the Ethical Review Board in the Humanities and Social and Behavioural Sciences at the University of Helsinki and was conducted in alignment with the Declaration of Helsinki.

Stimuli

The experiment was run using Presentation software (version 24.0; Neurobehavioral Systems, Berkeley, CA, USA), which controlled both auditory and visual stimulation. Sounds were delivered binaurally via Sensimetrics S14 earphones (Sensimetrics, Malden, MA, USA), and the volume was individually adjusted for each participant at the start of the session to approximately 85 dB. Visual material was displayed on a screen and viewed through a mirror mounted on the head coil. Participants responded using an fMRI-compatible button box. The auditory stimulus set consisted of sound clips from three object categories—speech, animal, and instrument sounds—each further divided into six subcategories (Table 1). Speech sounds were extracted from spoken dialogues recorded in a sound proof studio (See Leminen et al., 2020). Instrument sounds were created from speech clips produced by the same talkers as the speech stimuli, with a different talker assigned to each instrument, such that the resulting instrument stimuli followed speech-like temporal patterns rather than forming musical sequences. Animal sounds were collected from free internet sources. Each subcategory included eight exemplars, yielding a total of 144 unique sound clips (3 categories × 6 subcategories × 8 exemplars). Clip durations varied between 4 and 7 seconds, with a mean length of 5.3 seconds. Visual stimuli were derived from all 214 auditory clips used in the study, including both

training items (see Pre-trial) and non-targets from the match-to-target tasks (see trial structure). For each sound, we generated a scrolling waveform video in MATLAB, showing the waveform of the sound in blue on a white background (See Figure 1). The videos were presented independently of their corresponding auditory stimuli. The stimulus generation procedure is described in more detail in Wikman et al. (2025).

Table 1. Sound categories and subcategories

Speech	Instrument sounds	Animal sounds
Male1	Baritone guitar	Monkey
Male2	12-string guitar	Husky
Female1	Trumpet	Seal
Female2	Saxophone	Cat
Boy	Grand organ	Chick
Girl	Bebop organ	Whale

Trial structure

The trial structure (Figure 1) can be divided into four parts. 1.) In the beginning of each trial, the participant was visually instructed (1 sec.) to selectively attend to a specific stimulus (target stimulus). The target stimulus was either a specific auditory subcategory (e.g., “Attend to the boy”) or the visual stimulus (“Attend to the video”).

2) Following the instruction, the participant was presented with experimental stimuli (4 – 7 sec.). The presented stimuli always included a visual stimulus and auditory stimuli which varied based on the experiment (See experiments).

3) After the stimulus presentation, the participant was asked to rate how well they had been able to selectively attend to the target stimulus. The rating was conducted by pressing a button with an appropriate option: 1 (not at all), 2 (approximately 50 % of the time) or 3 (the whole time). The time window for the self-rating was 1.5 seconds long.

4) Lastly, to ensure that the participant attended to the target stimulus, on some of the trials the participant was asked to perform a match-to-sample task. To alert the participant of the match-to-sample task the background of the visual display was changed from black to green. Subsequently a short (2 sec.) additional stimulus was played. The additional stimulus was either a sample of the target stimulus (50 % probability) or a sample of another stimulus from the target subcategory (50 % probability). After the additional stimulus was played, the participants had to decide whether the additional stimulus was a sample from the target stimulus or from a different stimulus by pressing one of two buttons (1.5 sec. response window). Finally, the participant received feedback on their performance. The match-to-sample task occurred on 25 % of trials where the target stimulus was auditory and on 100 % of the trials where the target stimulus was visual. Results from behavioural data are reported elsewhere (Wikman et al., 2025).

Experiments

The participants completed three experiments in total. The experiments differed in the presentation of auditory stimuli and the selection of the target stimulus. In the One Object Alone (1OA) experiment, only one auditory stimulus was presented during stimulus presentation, and the auditory stimulus always served as the target. Visual videos were presented concurrently but were never task-relevant, and participants were instructed to ignore them. In each 1OA trial, a different auditory stimulus was presented, allowing for presentation of all 144 auditory stimuli in isolation (See Experimental session). The order of the stimulus presentation was pseudorandomized individually for each participant.

In the Three Objects Across categories (3OA) and Three Objects Within Category (3OW) experiments, three overlapping auditory stimuli were presented on each trial. In 3OA, the three auditory stimuli were drawn from different sound categories (e.g., Male1, Trumpet, Monkey), whereas in 3OW all three sounds belonged to the same category (e.g., Male1, Female1, Boy). On 75 % of trials, the target was an auditory stimulus. On these auditory-target trials, participants were instructed to selectively attend to the target sound while ignoring the other two sounds (distractor sounds) and the concurrently presented visual stimulus. To facilitate attentional orienting,

the target sound began 250–500 ms before the onset of the distractor sounds. On the remaining 25 % of trials, the target was the visual stimulus, and participants were instructed to selectively attend to the video and ignore all auditory stimuli. The stimuli were pseudorandomized into the trials individually for each participant. The pseudorandomization was done in such a way that the following parameters were always fulfilled: 1) Each subcategory acted as the target stimulus twice and as a distractor 4 times per run. 2) No two trials used the same three auditory stimuli – i.e., each trial was acoustically unique. 3) All 144 experimental stimuli were used one time per run: 108 stimuli were presented during the auditory trials and 36 during the visual trials. See Wikman et al., (2025) for further details on stimulus pseudorandomization.

Experimental session

Each run of the 1OA experiment included 36 trials and lasted approximately 6 minutes. In the 3OA and 3OW experiments, each run consisted of 36 auditory-target trials and 12 visual-target trials (48 trials in total), with a duration of approximately 8 minutes per run. For each experiment, participants completed 4 runs, resulting in 12 runs across the three experiments. All participants completed the 1OA task first. Thereafter, the order of the 3OA and 3OW experiments was counterbalanced across participants: half completed 3OA second and 3OW third, and the other half completed the experiments in the opposite order. Between runs, participants were given short breaks and an opportunity to talk with the researchers. After completing all functional runs, anatomical images were acquired. The total scanning session lasted approximately 100 minutes.

Pre-trial

Participants practiced all experiments outside of the scanner before the session. Each experiment was practiced for two runs in total. Initially, each experiment was practiced with slower response times. After that, the experiments were practiced with the same pace as the actual paradigm. To mitigate potential learning effects, the stimuli used in the pre-trial were distinct from the stimuli used in the actual paradigm.

Data acquisition

Functional imaging data were obtained using a 3T MAGNETOM Skyra whole-body scanner (Siemens Healthcare, Erlangen, Germany). A 32-channel head coil, with who channels removed to provide a clear view of the stimuli, was used. A T2* echo-planar imaging sequence was used to acquire 45 continuous oblique axial slices (TR = 1050 ms, TE 30 ms, flip angle = 75°, field of view = 21 cm, slice thickness = 2.7 mm, in-plane resolution = 2.7 mm × 2.7 mm × 2.7 mm). Each session included 12 runs: the first 4 runs (1OA experiment) produced approximately 360 volumes per run, while the final 8 runs (3OA & 3OW experiments) yielded about 570 volumes per run. In total, roughly 6000 volumes were collected per participant. After functional imaging, a high-resolution T1-weighted anatomical scan was preformed (TE = 3.3 ms, TR = 2530 ms, voxel matrix = 256 × 256, in-plane resolution = 1 mm × 1 mm × 1 mm) for co-registration purposes.

Sound feature analysis

All 144 main experimental auditory stimuli were analysed for low-level features using MATLAB R2025a. Each stimulus was characterized along five low-level features.

Frequency content (FC) describes how strongly different frequencies contribute to a sound. FC was obtained by applying a fast Fourier transform (FFT) to the sound's waveform, yielding a vector of energy amplitudes across frequency bins.

All other features were calculated within 40 ms windows using a 20 ms step to capture temporal dynamics. For these features, window-wise values were summarised into a single scalar per feature and stimulus. Before feature computation, stereo signals were converted to mono, z-scored for amplitude normalization, and each frame was multiplied by a Hann window to reduce spectral edge artifacts. Low-energy frames were excluded using an amplitude threshold to prevent silent segments from biasing the estimates.

Frequency content standard deviation (FCSD) indexes the temporal variability of a sound's spectral profile. The spectral centroid for each window was calculated as the power-weighted average frequency, and FCSD was defined as the standard deviation

of these centroid values across time (Supplementary Figure 1). *Amplitude standard deviation* (AMSD) captures variability in the sound's amplitude across time. AMSD was derived by taking the standard deviation of the root-mean-square amplitude across windows (Supplementary Figure 2). *Harmony to noise ratio* (HNR) characterizes the degree of periodicity in a sound, indicating how harmonic or noisy it is. For each window, a normalized autocorrelation was computed, the largest positive-lag peak within adaptive frequency bounds was identified, and HNR was defined as

$$HNR = 10\log_{10}\left(\frac{r_{max}}{1-r_{max}}\right),$$

where r_{max} is the peak autocorrelation value. The resulting values were averaged across valid windows (Supplementary Figure 3). Spectral entropy quantified the average spectral complexity of each sound and was computed using the MIRtoolbox (Lartillot et al., 2008) function mirentropy, then averaged across all valid windows (Suppelementary Figure 4). The low-level sound features were selected based previous research (Giordano et al., 2013; Leaver & Rauschecker, 2010; Wikman et al., 2025).

Sound feature models

After computing the low-level sound features for each stimulus, we averaged the feature values within each subcategory (e.g., trumpet; Table 1). RDMs were then derived separately for each sound feature. For FC, which is a vector-valued feature (the magnitude spectrum across frequency bins), the RDM was constructed using $1 - r$, where r is the Pearson correlation between the mean FC vectors of each pair of subcategories. For the remaining four scalar features (FCSD, AMSD, HNR, Entropy), each subcategory was described by a single value, and the RDMs were constructed using the Euclidean distance between all pairs of subcategory means. In total, five low-level feature models were generated (Figure 2). To operationalize abstract sound features, four RDMs were constructed that divide the stimuli into categories. More specifically, we created a Category-model that equally differentiates each category, a Speech-model that differentiates speech from non-speech, an Inanimate-model that

differentiates instrument sounds from animate sounds (animals and speech), and an Animal-model that differentiates animal sounds from non-animal sounds (Figure 2).

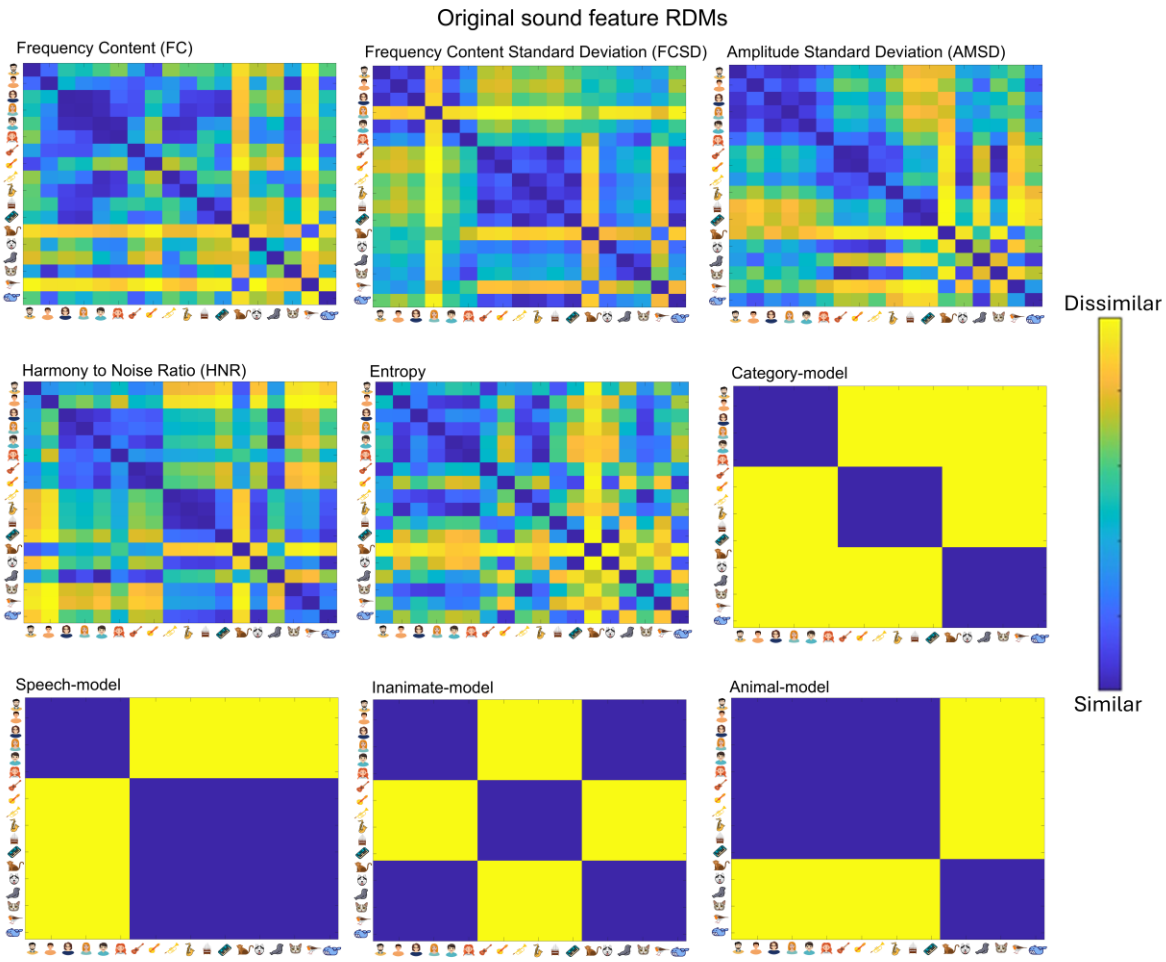

Figure 2. *Original sound feature representational dissimilarity matrices*. Here displayed are all 9 original RDMs. Yellow colours indicate high dissimilarity between two subcategories for the given feature, whereas blue colours indicate high similarity.

Next, to examine the covariance between the sound features, a Spearman correlation between each sound feature RDM was calculated (Supplementary Figure 5A). As can be seen from Supplementary Figure 5A, there exists substantial covariance between the sound feature RDMs. Without controlling for this covariance, it would be impossible to attribute any possible results to a specific sound feature.

To address this issue, covariances between the models were statistically removed. Each RDM was rank transformed to enable a "Spearman regression," and the shared variance between models was explained using general linear model. The residuals

from these regression models—representing the variance unique to each model—were retained for further analyses. These residualised RDMs are referred to as residualised models. Three sets of residualised models were created. For *Only Low-level models* all abstract RDMs were residualised out of each low-level RDM. This procedure eliminated correlations with abstract models while preserving correlations among the low-level features (Supplementary figure 5B). *Unique* models were obtained by residualising out all other models from each low-level feature model, thereby removing covariance with both abstract and other low-level models (Supplementary figure 5C). Importantly, some *Unique* models correlated only moderately with their original counterparts (e.g., r = 0.73 between *Unique* FC and *Original* FC), indicating partial deviation from the original model structure. Finally, to create *Only abstract* models, all low-level models were residualised out of each abstract model, eliminating correlations between abstract and low-level models (Supplementary figure 5D). See supplementary figures 6 – 8 for all residualised models.

Preprocessing and first-level fMRI data analysis

fMRI data was preprocessed using FEAT tool from FSL (FMRIB's Sofware Library, http:// www.fmrib.ox.ac.uk/fsl). Preprocessing included registration of fMRI volumes to the participant's high-resolution structural image using FLIRT, motion correction using MCFLIRT (Jenkinson et al., 2002), slice timing correction (interleaved), non-brain removal using BET (Smith, 2002) and high-pass temporal filtering using a cut-off of 100 seconds. The fMRI volumes were then transformed into the Freesurfer (Fischl, 2012) average surface space (fsaverage) using the Freesurfer function mri_vol2surfs, transforming the three-dimensional voxels into two dimensional vertices, without smoothening.

For the first level analyses, a GLM was fit to the time series of each vertex in each run. The GLM included a separate regressor for each trial. The GLM also included nuisance regressors for the 6 movement dimensions, to remove movement specific noise. This procedure resulted in trial specific spatial parametric maps.

Multivariate pattern analyses

Representational similarity analysis

The goal of the RSA analysis was to construct RDMs from neural activity patterns that capture how differently brain regions represent distinct experimental conditions. For each experiment separately, all trials in which participants listened (1OA) or attended to (3OA & 3OW) the same subcategory were grouped into conditions, resulting in 18 (one per subcategory) conditions per experiment. Each subcategory was listened to (1OA) or attended to (3OA, 3OW) eight times in total per experiment (two trials per run across four runs; see Experiments), yielding eight exemplars per condition.

Neural similarity between conditions was quantified using a linear support vector machine (SVM; Chang & Lin, 2011). The analyses were completed using 8 mm searchlights (Kriegeskorte et al., 2006) with the help of Surfing toolbox (Oosterhof et al., 2011) and predefined regions of interest (ROIs; see Regions of interest). The classifier was trained and tested with a leave-one-run-out cross-validation scheme: three runs served as training data (2 trials × 3 runs × 2 conditions = 12 exemplars for training), and one run was reserved for testing (2 trials × 1 run × 2 conditions = 4 exemplars for testing). This procedure was repeated four times so that each run was used once as the test set, and mean accuracy across folds represented the classification performance between each condition pair. Repeating this procedure for all possible condition pairs produced an RDM of decoding accuracies which index how similar a given brain regions activation pattern was for different ***attended*** subcategories (e.g., how similarly a region activated when *Male1* versus *Female1* was attended or listened to; Attended RDM).

For the selective attention experiments (3OA, 3OW), additional RDMs were created to capture the representational structure of distractor sound processing (See figure 3). The same decoding procedure was applied, but trials were grouped into conditions by using a shared distractor subcategory rather than attended subcategory. Because each auditory trial contained one target- and two distractor sounds, the distractor conditions initially contained twice as many exemplars as the attended ones. To ensure comparability, the distractor trials were randomly restricted to an equal number of exemplars as the attended trials. The resulting

RDMs reflect how similarly cortical regions responded to different ***distractor*** subcategories (e.g., when Male1 versus Female1 served as distractors; Distractor RDM).

Next, neural representations of sound features were derived. For the 1OA task this was done by computing spearman correlations between sound feature models and the *Attended RDMs*. For selective attention tasks (3OA & 3OW), we also calculated spearman correlations between the *distractor RDMs* and sound feature models. To isolate the effects of attentional modulation, these correlations were then subtracted from the corresponding correlations between *attended sound RDMs* and sound feature RDMs. That is, contrasts were calculated between the sound feature representations of attended and distractor sounds (Figure 4).

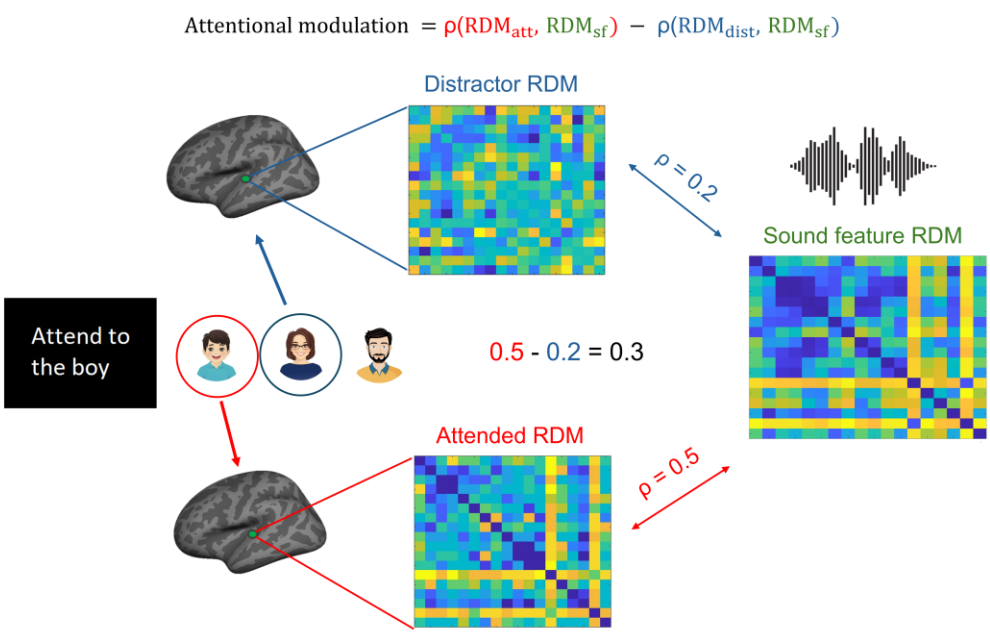

Figure 3. *Schematic representation of the analysis pipeline for the selective attention tasks.* Representational dissimilarity matrices (RDMs) were computed separately for attended and distractor sounds. Attended RDMs were generated by grouping trials with a common target sound into conditions and computing neural pattern similarities across these conditions. The same procedure was applied to distractor sounds to produce corresponding distractor RDMs. Sound feature RDMs capture the similarity structure of specific low-level or abstract sound features. Attentional modulation for each feature was quantified as the difference between its neural correspondence under attended versus distractor conditions. In other words, it reflects how strongly the neural representational structure aligns with a given sound feature when attention is directed toward that sound compared to when it is ignored.

Object identity analysis

The analyses described above establish whether neural activity patterns represent low-level acoustic features and abstract category-level distinctions (e.g., speech vs. animal sounds). To further assess representations at a finer level of abstraction, we conducted a within-category differentiation analysis to determine whether neural patterns distinguish individual sound objects within a category–i.e., object identity (e.g., Male1 vs. Male2). This level of analysis could not be addressed using the same RSA approach, as RSA required aggregation at the subcategory level to ensure a sufficient number of exemplars for reliable similarity estimation.

To index object identity representations, we implemented a cross-experiment classification scheme (Figure 4). A linear SVM classifier was first trained on trials from the One object alone (1OA) experiment, where sounds from a given subcategory were presented in isolation (e.g., Monkey vs. Bird trials; Figure 4A). The trained classifier was then tested under three conditions: 1) held-out 1OA trials, 2) trials from the selective-attention experiments in which the same sounds were attended, and 3) trials in which those sounds were present but acted as distractors (Figure 4B). For the held-out 1OA trials, classification followed the same leave-one-run-out cross-validation scheme described above. In contrast, for cross-experiment decoding, the classifier was trained once using all 1OA trials for the given two subcategories and tested once on all corresponding 3OA/3OW trials in which that subcategory served either as the attended target or as a distractor. This procedure yielded 16 1OA exemplars for training and 16 3OA/3OW exemplars for testing per classification.

To quantify attentional modulation of object-level representations, we contrasted the classification accuracy between attended and distractor conditions in the selective-attention experiments. Specifically, distractor classification accuracy was subtracted from attended classification accuracy to form an attentional modulation index. Positive values of this index indicate that neural activation patterns for attended sounds more closely matched the patterns elicited by the same sounds when presented alone (1OA), relative to when they were heard but not attended to. For each category, decoding accuracies were averaged across all unique within-category

object pairs, and this mean value was interpreted as the strength of the object identity representation (Figure 4C).

Finally, we examined whether observed object identity representations could be explained by systematic differences in low-level features. If object-pair classification accuracies are significantly associated with low-level acoustic distances, this would indicate that the classification performance is at least partly driven by low-level feature processing (or attentional modulation) rather than pure abstract object-identity representations. To assess this, we derived a general low-level distance model by computing the first principal component across all five low-level sound-feature RDMs. This component captured shared variance in low-level pairwise distances across features. We then computed Spearman correlations between within-category classification RDMs and the corresponding low-level distance estimates for each category (Figure 4D). The analysis was conducted using ROIs (see Regions of interest).

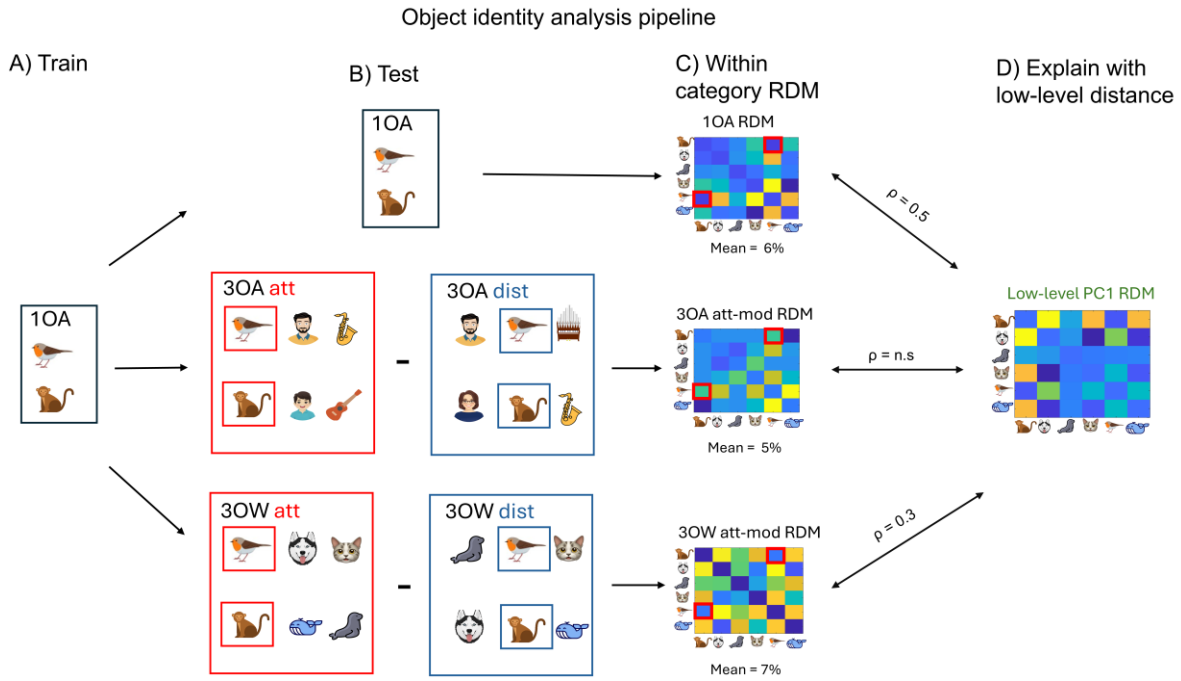

Figure 4. *Cross-experiment classification analysis for object identity representations.* A) A linear support vector machine (SVM) classifier was trained on trials from the one-object-alone (1OA) experiment, in which sounds from individual subcategories were listened to in isolation (e.g., Monkey vs. Bird trials). B) The trained classifier was tested on three conditions: held-out 1OA trials, selective-attention trials (3OA, 3OW) in which the same subcategories were attended, and trials in which the same subcategories acted as distractors. Classification performance was computed separately for attended and distractor conditions. C) Object-identity representation strength was defined as the mean classification accuracy across all unique within-category subcategory pairs. Attentional modulation was quantified as the difference between attended and distractor classification accuracies. D) To assess whether object-identity classification could be explained by low-level acoustic differences, a general low-level distance model was derived as the first principal component across all five low-level sound-feature representational dissimilarity matrices (RDMs). Pairwise distances from this model were correlated with within-category decoding RDMs using Spearman correlation, testing whether classification accuracy was associated low-level differences.

Regions of interest

We used five ROIs for both hemispheres that were selected based on theoretical relevance (Table 3). These ROIs included key regions of the auditory ventral stream. The ROIs were defined by combining subregions from the human connectome project atlas (Glasser et al., 2016; see Table 2 & Figure 7).

Table 2. Selected ROIs and the corresponding ROIs from the human connectome atlas

Region of interest	Human connectome atlas regions of interest
Primary auditory cortex (A1)	A1
Belt areas	MBelt + PBelt + LBelt
Superior temporal gyrus (STG)	A4 + A5 + TA2 + STGa
Superior temporal sulcus (STS)	STSda + STSdp + STSva + STSvp
Inferior frontal gyrus (IFG)	45

Inference of statistical significance

For the searchlight analysis, the resulting correlation maps were Fisher Z transformed and thresholded with a minimum Z value of 3.1. The Z-value maps where then smoothed on the cortical surface using a full-width at half-maximum of 5 mm^2. Statistical significance in the searchlight analysis was assessed using cluster-based permutation testing ($Z \geq 3.1$, $p < 0.05$, corrected), as implemented in FreeSurfer's *mri_glmfit-sim* tool (1,000 sign-flipping iterations). Only spatially contiguous clusters exceeding the cluster-size threshold derived from the null distribution were retained.

For the RSA ROI analysis, we tested whether the observed correlations and attentional modulations significantly differed from 0 using a two-sided t-test. The same approach was taken with the object identity analysis, where significant deviation from 0 for both the classification means and the low-level RDM correlations were tested using a two-sided t-test. P-values were corrected for multiple comparisons across all experiments and ROIs using the Benjamini–Hochberg false discovery rate (FDR) procedure with $\alpha = 0.05$.

Finally, noise ceilings were computed separately for each ROI and experiment. For the 1OA experiment, noise ceilings were estimated using the standard RSA noise-ceiling procedure with Spearman correlations (Nili et al., 2014). For the selective-attention experiments, we estimated noise ceilings for the attentional modulation index. For each ROI and experiment, we computed a group-level modulation template as the difference between the mean attended and mean distractor RDMs. For each subject, we then correlated both the attended and distractor RDM vectors with this template and defined the subject's modulation as the difference between

these correlations. Lower and upper bounds were obtained using leave-one-subject-out versus inclusive group templates, respectively.

Results

Sound feature representations during isolated listening

Whole-brain correlates of low-level sound features

In the 1OA experiment, participants listened to one sound at a time, requiring active listening rather than selective attention. Figure 5 shows correlations between low-level sound feature RDMs and searchlight-based neural activation patterns during this task. As shown in Figure 5A, *Original* low-level sound-feature models exhibited broad correlations across temporal and frontal regions. This pattern is unlikely to reflect low-level processing, as such effects are generally thought to be confined to the auditory cortex.

Consistent with this expectation, when using *Only low-level* models that were statistically independent of the abstract sound-feature models (see Supplementary Figure 5), correlations became restricted to overlapping auditory cortical regions, specifically A1, Belt areas and the STG (Figure 5B). This indicates that the widespread temporal and frontal correlations observed for the *Original* models were driven by shared variance with abstract sound features.

Finally, when using *Unique* models that were statistically independent of one another, low-level sound features were more precisely localized and exhibited feature-specific spatial differences. Specifically, FC, FCSD, and AMSD correlated primarily with A1 and Belt areas, whereas HNR and entropy correlated with the left and right STG, respectively (Figure 5C).

Together, these results emphasize the necessity of covariance control between sound-feature models. Whereas *Original* models failed to dissociate low-level and abstract representations, and *Only low-level* models localized all low-level features mainly to overlapping regions, *Unique* models revealed meaningful spatial differentiation across the processing of different low-level features. Accordingly, all subsequent low-level results in the main text are reported using *Unique* models.

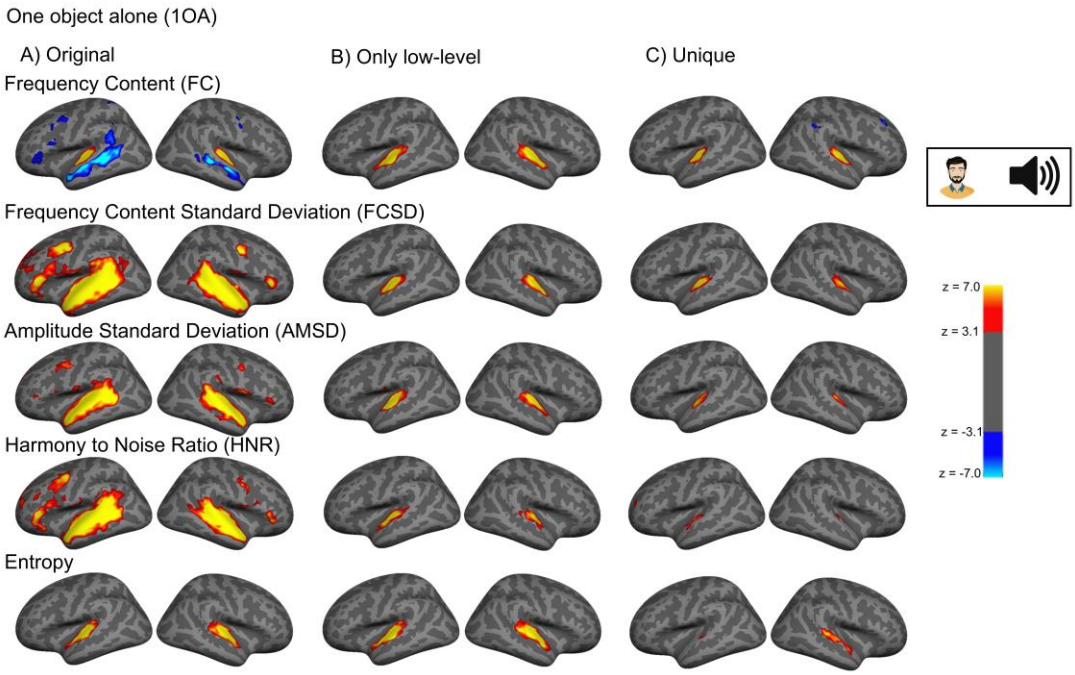

Figure 5. *Correlates of low-level sound features during the 1OA experiment.* This figure illustrates significant correlations between low-level sound feature RDMs (Figure 2) and searchlight-based activation pattern RDMs from the 1OA experiment. Brighter colours indicate stronger correlations, with warm colours indicating the presence of positive correlations and cold colours denoting negative correlations. Each row corresponds to a specific low-level sound feature, while each column represents a different version of the residualised models. A) column displays the *Original* models, B) column displays the *Only low-level* models, and C) column displays the *Unique* models. A schematic trial from the 1OA experiment is shown in the top right corner. Statistical overlays are displayed on FreeSurfer fsaverage inflated surfaces, where light gray indicates gyri and dark gray denotes sulci. Only lateral surfaces are shown here. All statistical parametric maps use an initial threshold of $Z \geq 3.1$, with permutation-based cluster-size correction at $p < 0.05$.

Whole-brain correlates of abstract sound features

Original abstract sound-feature RDMs showed strong correlations with neural activation patterns across widespread temporal and frontal regions (Figure 6A). Category and Speech models correlated broadly across temporal and frontal cortex, while Inanimate- and Animal-models showed widespread correlations encompassing STG, STS, IFG, and premotor cortex (PMC).

Using *Only abstract* models—i.e., abstract feature models residualised with respect to low-level sound feature models—revealed several systematic changes (Figure 6B). Firstly, correlations for the Category- and Speech- and Animal-models were slightly reduced in the dorsal A1, consistent with this region being primarily sensitive to low-level acoustic features rather than abstract sound properties (Saenz & Langers,

2014). Second, correlations for the Inanimate-model weakened substantially following residualisation, becoming largely restricted to anterior STG, whereas correlations for the *Only abstract* Animal model strengthened and became more spatially extensive. These opposing effects are likely explained by changes in covariance between abstract models introduced by residualisation rather than by changes related to covariance with low-level features. This is because the inanimate-model weakened and the Animal-model strengthened prominently in the left IFG and bilateral STS, which are regions implicated in the processing of speech rather than low-level features [source] and accordingly the residualisation reduced the correlation between the Inanimate- and Speech-models (0.06 to −0.08), while increasing the correlation between the Animal- and Speech-models (0.06 to 0.14; Supplementary Figure 5A, C).

Although residualising low-level sound features yielded mixed effects across abstract models, the main text primarily reports results *Only abstract* models to dissociate abstract from low-level sound-feature processing.

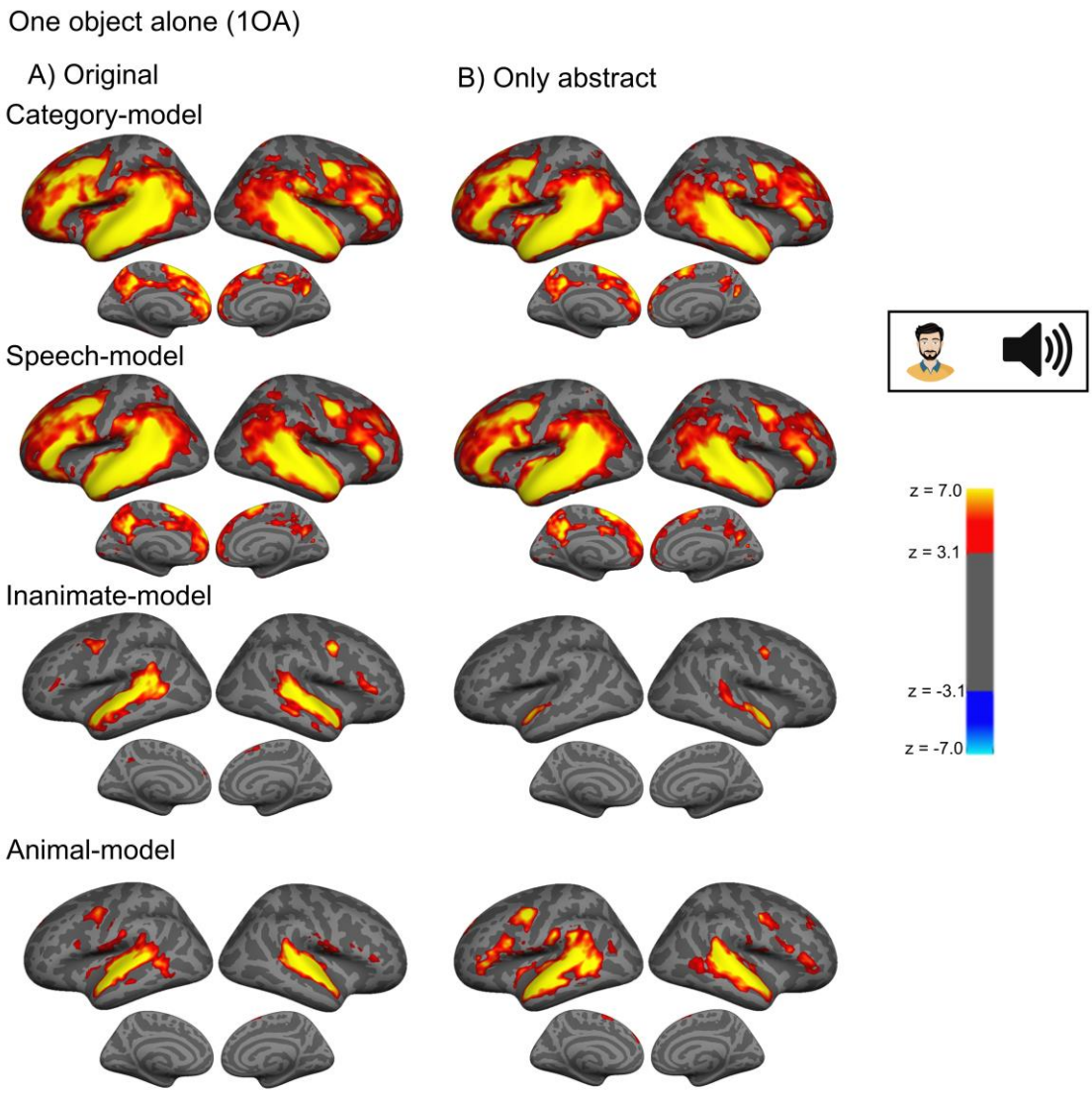

Figure 6. *Correlates of abstract sound features during the 1OA experiment.* This figure illustrates significant correlations between abstract sound feature RDMs (Figure 2) and searchlight-based activation pattern RDMs from the 1OA experiment. Rows correspond to different abstract sound feature models and columns represent different versions of the residualised models. Column A) shows results for the Original models, and Column B) shows results for the Only abstract models. Brighter colours indicate stronger correlations, with warm colours denoting positive correlations and cold colours denoting negative correlations. Lateral and medial are shown.

Progression of sound feature processing along the auditory ventral stream

To more closely examine the progression of sound feature representations, we conducted the same analysis as above within predefined ROIs (Figure 7, middle; table 2). For visualization purposes, we show mean RDMs across participants within each left hemisphere ROI from the 1OA experiment (Figure 7, top). This progression shows a visible transformation from activity patterns that dissociate monkey & bird

sounds (acoustic outliers) to areas that distinguish speech sounds from other sounds (abstract feature). To test this qualitative observation, we correlated the ROI RDMs with the sound-feature RDMs (Figure 7A–B). Here, we report results for the *Unique* and *Only abstract* models, which maximize interpretability. Results for the *Original* and *Only low-level* models are provided in Supplementary Figure 9.

Low-level sound-feature models correlated bilaterally with activation patterns in A1, Belt, and STG (Figure 7A). In contrast, abstract sound-feature models showed strongest correlations in higher-order auditory regions, with bilateral STG, STS, and left IFG exhibiting particularly strong correlations ($r > 0.5$) with the Speech model (Figure 7B), indicating robust discrimination of speech sounds from other sound categories. Together, these results replicate the well-established observation that representations of sound features become progressively more abstract along the auditory ventral stream (Rauschecker & Scott, 2009).

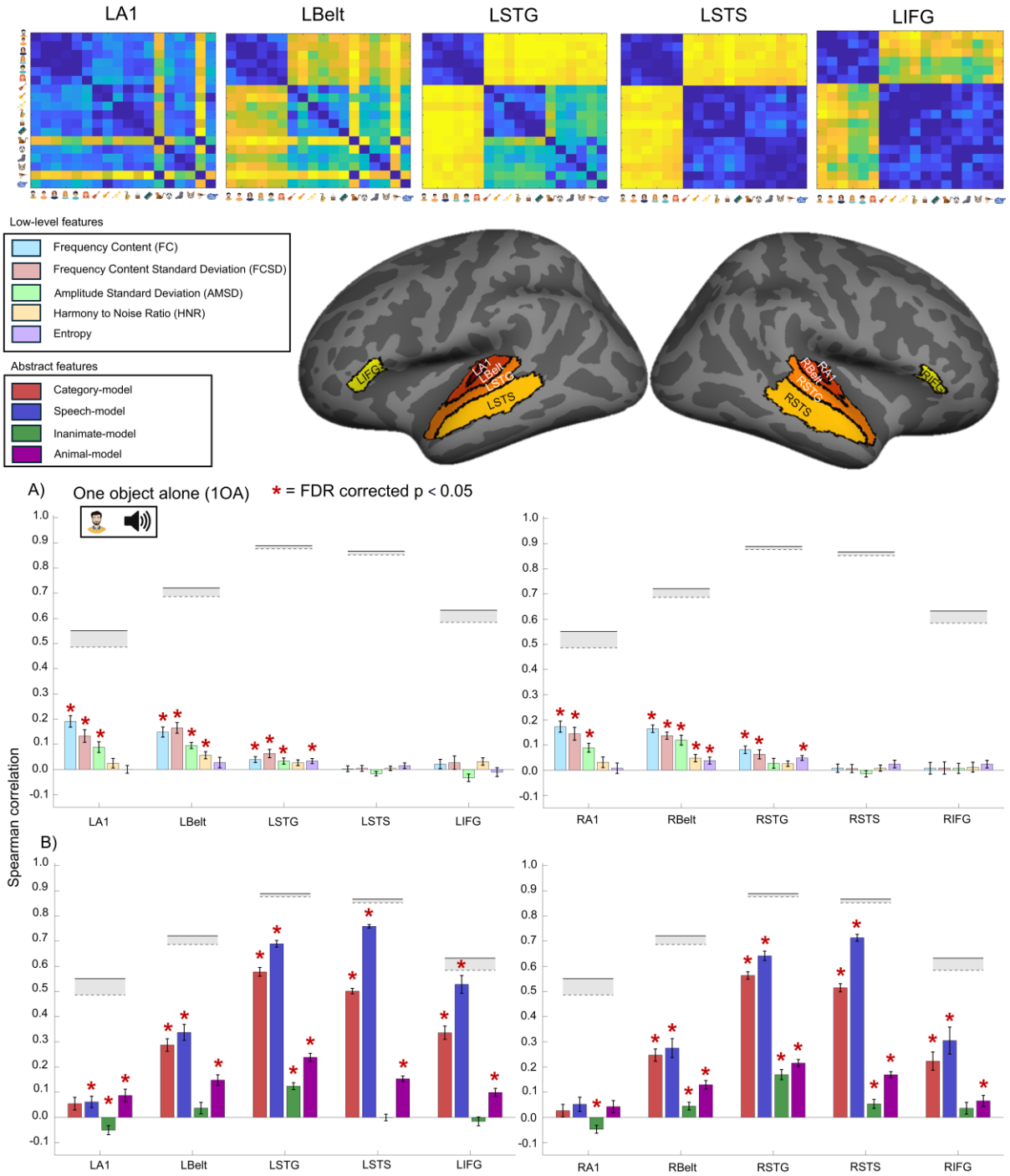

Figure 7. *Region of interest analysis for the One object alone experiment.* The top panel shows mean representational dissimilarity matrices (RDMs) from the 1OA experiment for each left-hemisphere ROI, illustrating activity patterns along the auditory ventral stream. ROI locations are shown on the cortical surfaces below the RDMs. The bottom panels present Spearman correlations between sound-feature RDMs and ROI RDMs. To maximize interpretability, results from the *Unique* and *Only abstract* models are shown. A) Panels display correlations for low-level sound features, and B) panels display correlations for abstract sound features. Left and right columns show left- and right-hemisphere ROIs, respectively. Red asterisks indicate correlations that significantly deviate from zero (FDR-corrected $p < 0.05$). Error bars represent ±1 SEM. Grey bars denote estimated noise ceilings, with dashed and solid lines indicating lower and upper bounds, respectively.

Attentional modulation of sound feature processing

To examine attentional modulation of sound-feature processing, we applied the same analysis pipeline as in the listening task separately to attended and distractor objects during the selective-attention tasks (3OA/3OW). Attentional modulation was quantified by subtracting distractor correlations from attended correlations, indexing the extent to which features of sounds were more strongly represented when the sounds objects were attended than when they were ignored (Figure 4).

Attentional modulation of sound feature processing in cross category soundscapes

In the 3OA experiment the participant was tasked with attending to one sound out of a soundscape which consisted of three sounds from different categories (e.g., Male1, Monkey, Trumpet). During this experiment attentional modulation of low-level sound features was observed only when using the *Original* models (supplementary figure 9A), which share variance with abstract feature models and are therefore unlikely to reflect genuine attentional modulation of low-level feature processing. Consistent with this interpretation, neither the searchlight nor the ROI analyses showed significant positive attentional modulation of low-level sound features when using residualised models (Supplementary figures 9, 12, 13 & Figure 10A). These results indicate that attention did not modulate low-level feature processing during the 3OA task.

In contrast, significant attentional modulation of abstract sound-feature processing was observed in higher-order temporal and frontal regions (Figure 8A). This effect was most pronounced for the Category and Speech models, which showed robust attentional modulation in STG, STS, left IFG, and premotor cortex (PMC). These abstract attentional effects were independent of low-level feature covariance (Figure 8B) and were also found in the ROI analysis (Figure 9B).

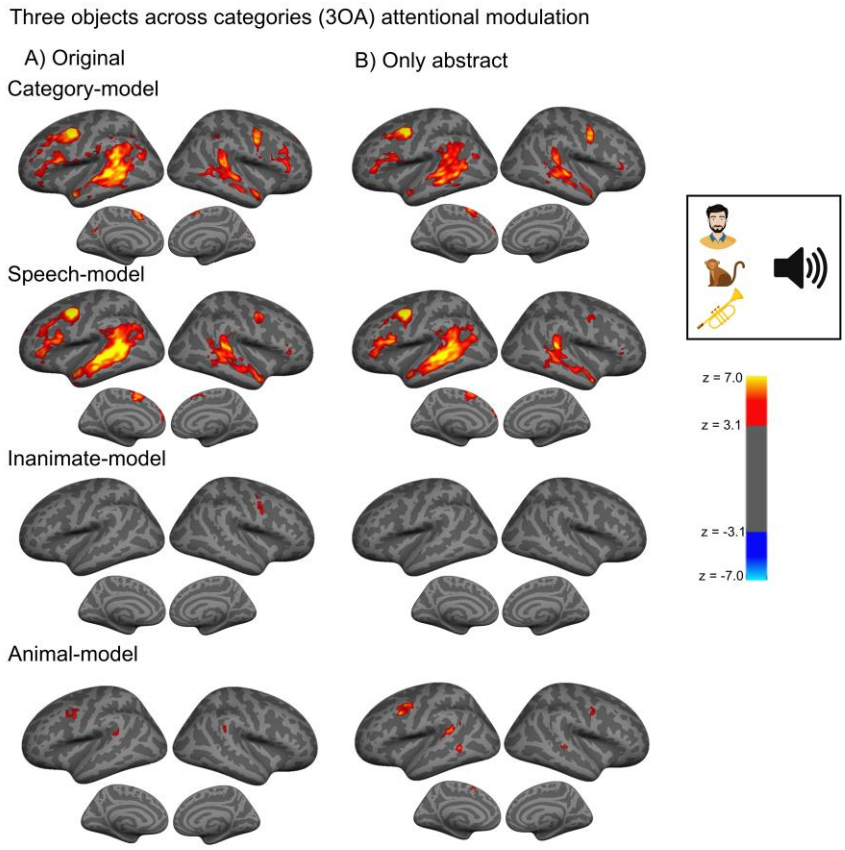

Figure 8. *Whole-brain attentional modulation of abstract sound feature processing during the 3OA experiment.* This figure shows searchlight-based attentional modulation effects for abstract sound feature models during the Three Objects Across Categories (3OA) task. Panels A–D display results for the different models. Attentional modulation was quantified as the difference between correlations for attended and distractor sounds. Warm colours indicate stronger representations of attended relative to distractor sounds, whereas cold colours indicate the opposite.

Attentional modulation of sound feature processing in within category soundscapes

In the 3OW experiment, participants attended to one sound within a soundscape consisting of three sounds from the same category (e.g., Male1, Female1, Boy). Whole-brain searchlight analyses revealed primarily small and weak non-auditory cortical clusters of attentional modulation, with a limited effect for FCSD in the right middle STG (Supplementary Figure 10). Overall, attentional modulation of low-level sound features was not robustly detected from the searchlight analysis.

In contrast, the ROI analysis revealed clear attentional modulation of low-level feature processing (Figure 9C). Attention modulated the processing of FC in the left A1 and Belt areas. HNR and entropy showed attentional modulation in left STG, whereas FCSD was modulated in right STG and left STS. Importantly, unlike the

other effects, the FCSD modulation in left STS was not present when using the *Original* or *Only low-level* models (Supplementary Figures 12 & 13), suggesting that this effect likely reflects an artefact from covariance removal rather than modulation of a genuine low-level sound feature.

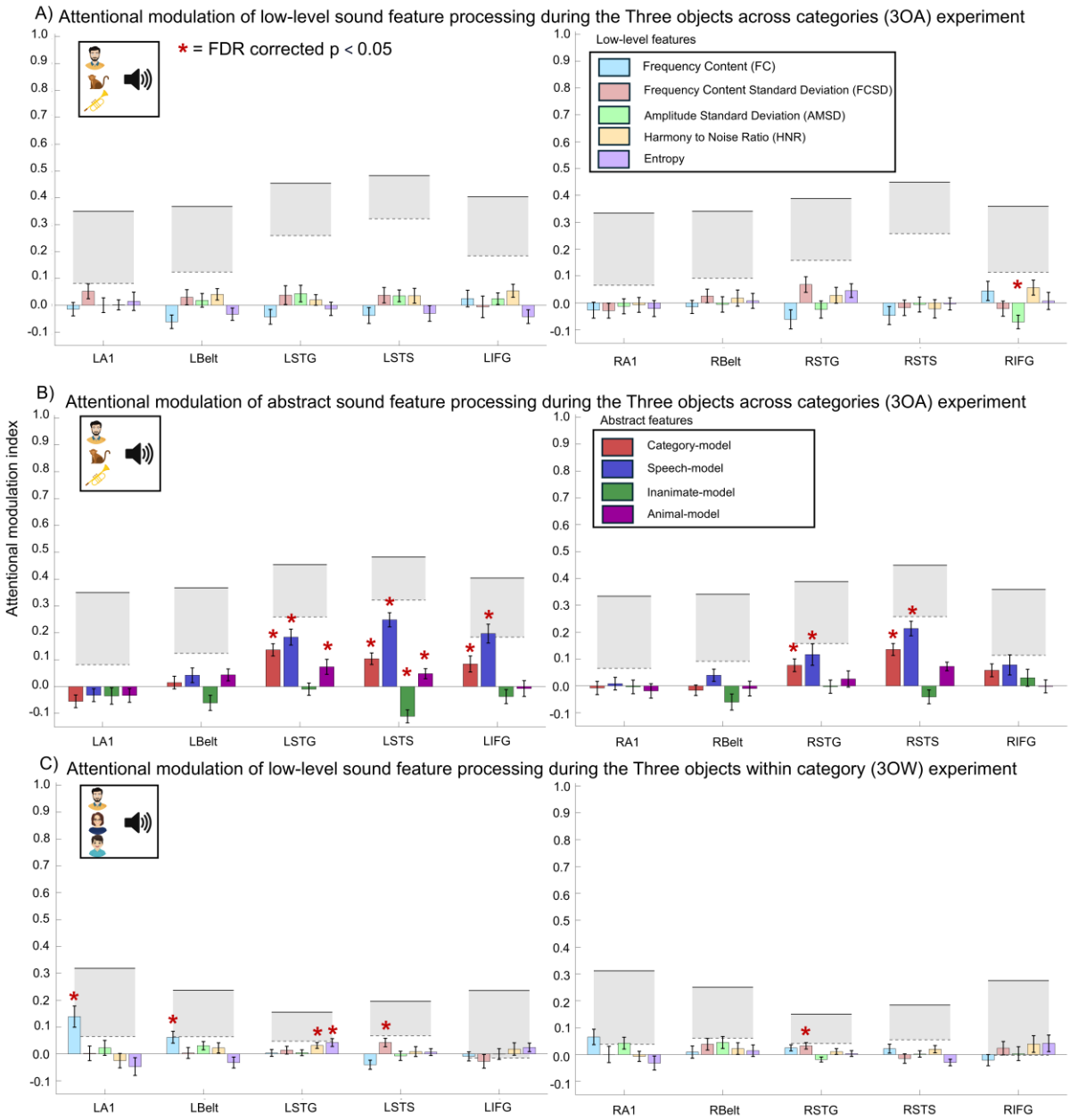

Figure 9. *ROI analysis of attentional modulation of sound feature processing during selective attention.* Attentional modulation indices (Figure 4) are shown for low-level and abstract sound features. To maximize interpretability, results are shown for *Unique* and *Only abstract* models. Panels A and B show results from the Three Objects Across Categories (3OA) experiment for low-level A) and abstract (B) sound features, respectively. Panel C shows results from the Three Objects Within Category (3OW) experiment for low-level sound features. Bars represent mean attentional modulation indices for each ROI, shown separately for left and right hemispheres. Red asterisks indicate attentional modulation values that significantly deviate from zero (FDR-corrected p < 0.05). Error bars

represent ±1 SEM. Grey bars denote estimated noise ceilings, with dashed and solid lines indicating lower and upper bounds, respectively.

Within-category differentiation of object identity

Above, we have shown that attention modulates abstract, category-level, feature processing (e.g., is the attended sound Speech) to separate sounds from different categories in the 3OA task. Likewise, we have shown that attention modulates low-level feature processing (e.g., is the attended sound high in frequency content) to separate sounds from the same category in the 3OW experiment. Importantly, in the 3OW experiment, attended and distractor sounds always belonged to the same category, making category-level distinctions unavailable for separating the overlapping sounds. Under these conditions, attentional selection may instead operate at the level of another abstract feature: object identity (e.g., is the attended sound Male1).

To test this hypothesis, we implemented a cross-experiment decoding approach (Figure 4). A classifier was trained on 1OA trials, in which individual sounds were presented in isolation. The classification accuracy was first evaluated on held-out 1OA trials to establish baseline object-identity representations. The classifier was then applied to the selective-attention experiments, with classification accuracy assessed separately for attended and distractor sounds. This approach allowed us to test whether sound identity could be classified more accurately when the sounds were attended compared to when they were unattended. To determine whether any observed classification could be explained by low-level acoustic differences rather than abstract object identity, we additionally looked for associations between the classification accuracy and low-level acoustic distance measures (See methods, Object identity analysis).

Decoding of object identity during isolated listening

Figure 10 shows the results of the cross-experiment object-identity decoding analysis. When sounds were presented in isolation (1OA), reliable object-identity decoding was observed across auditory regions, providing a baseline for identity

representations. For speech, instrument, and animal sounds, classification accuracy was significantly above chance in A1, Belt, and STG.

To assess whether these identity representations reflected abstract identity rather than low-level acoustic separability, decoding performance was related to low-level acoustic distance measures. For instrument and animal sounds, decoding accuracy correlated significantly with low-level acoustic distance across most ROIs (with the exception of right A1 for instruments and right STS for animals), indicating that identity decoding in these categories was at least partly driven by low-level sound feature differences. In contrast, no significant correlations with low-level acoustic distance were observed for speech sounds, suggesting that speech identity decoding was largely independent of low-level acoustic feature processing.

Attentional modulation of object identity processing

In the 3OA experiment, attentional modulation of object identity was observed for speech (Left STG, STS, IFG & Right STS) and animal sounds (Left Belt areas, STG & STS). These effects did not correlate with low-level acoustic distance, indicating that attentional modulation of identity representations in this experiment was not explained by attentional modulation of low-level sound features. This result is expected, as no attentional modulation of low-level feature processing was observed for this experiment (Figure 9A).

In the 3OW experiment, attentional modulation of object identity was observed for speech sounds in right STG and STS, and for animal sounds in left Belt and STG. Of these effects, however, only the speech effect in right STG survived FDR correction. Seeming attentional modulation of animal object identity was also observed in right Belt and STG, but these effects were accompanied by significant correlations with low-level acoustic distance, meaning that they relied at least partially on attentional modulation of low-level sound feature processing.

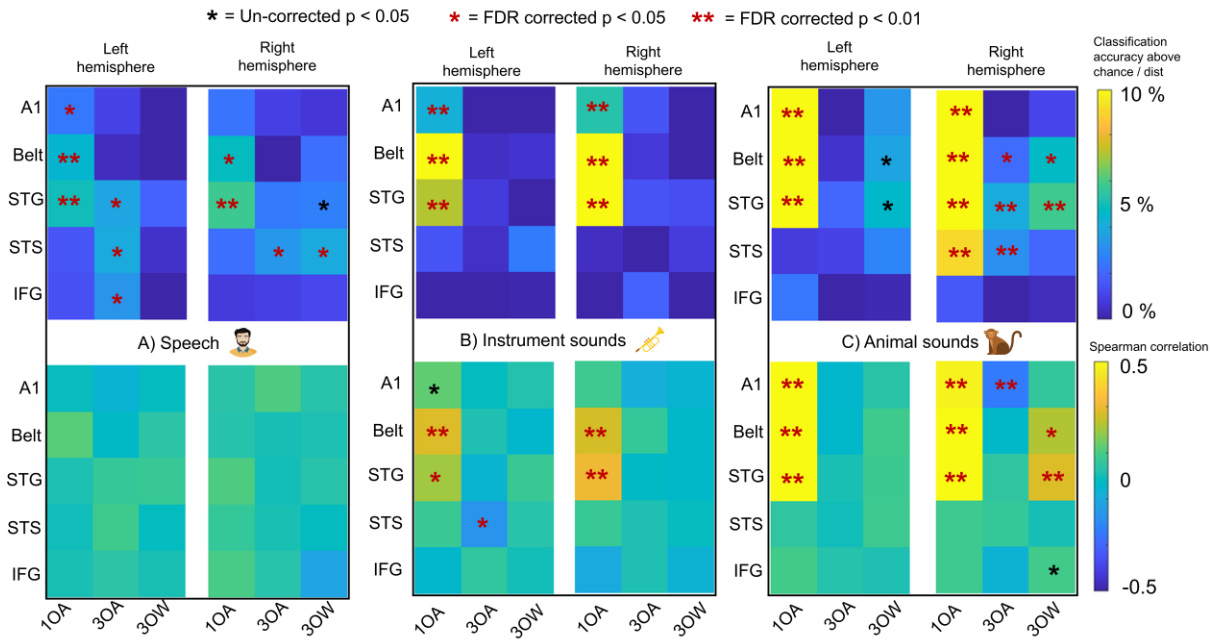

Figure 10. Decoding of object identity. Cross-experiment classification accuracy (top rows) are shown for speech A), instrument B), and animal C) sounds. The heatmaps display classification accuracies above chance for 1OA and above distractors (att > dist) for 3OA/3OW (in %). Bottom panel heatmaps show Spearman correlations between within-category decoding RDMs and low-level distance RDMs, assessing whether identty decoding can be explained by systematic acoustic differences. Asterisks indicate statistical significance (* uncorrected p < 0.05; red * FDR-corrected p < 0.05; ** FDR-corrected p < 0.01).

Discussion

The present study investigated how selective attention modulates auditory representations at different levels of abstraction in complex soundscapes. By combining representational similarity analysis, and cross-experiment decoding, we show that cortical attentional modulation is not fixed to a single representational stage. Instead, attention flexibly targets the representational level that is most informative for resolving competition, ranging from low-level acoustic features, abstract category-level features to and object identity within category.

Complexity of sound feature processing progresses along the auditory ventral stream

The 1OA experiment in this study closely resembled the experience of everyday listening to natural and isolated sounds and therefore allowed us to examine where different sound features are represented. The present study outlines five low-level sound features, which are processed at different levels of the auditory hierarchy independently of each other and the category of the natural sound being played. These findings are broadly consistent with previous work using similar acoustic features (Leaver & Rauschecker, 2010), RSA-based methods (Giordano et al., 2013) and model free decompositions of auditory cortex activity (Boos et al., 2021; Norman-Haignere et al., 2015). Namely, Norman-Haignere et al. (2015), identified six canonical response components; the first two captured low- and high-frequency energy in A1, closely paralleling the FC-related representations observed here, whereas another component reflected sensitivity to harmonic pitch structure, with maximal weights near anterior belt areas and middle STG, in line with the present HNR representations.

Together, these observations suggest a coarse hierarchical organization of low-level sound feature processing. Features that require relatively little transformation of the acoustic signal—such as amplitude variability, frequency content, and frequency variability—were primarily represented in early auditory cortex. In contrast, less intuitive HNR and spectral entropy were predominantly represented in later regions, including belt areas and the STG, suggesting that these features may be used to bridge acoustic and semantic information (Giordano et al., 2023). Indeed, Prior work has implicated HNR as a second-order processing feature in natural sounds (Lewis et al., 2009), whereas, research in the visual domain speculates that image entropy may be used to reduce sensory redundancy, with strongest effects observed in non-primary visual cortex (Muukkonen & Salmela, 2023), a pattern that appears homologous to the spectral entropy effects observed here.

Abstract sound features—most prominently the distinction between speech and non-speech—were represented broadly across temporal and frontal cortices (Figure 6), with particularly strong effects in STG, STS, and IFG, consistent with their established roles in speech processing (Alho et al., 2006; Belin et al., 2000; Latinus

& Belin, 2011; Liakakis et al., 2011; Petkov et al., 2008; Sharda & Singh, 2012). Notably, both speech–non-speech and animal–non-animal distinctions were also decodable from A1 (Figure 7B), even after controlling for covariance with multiple low-level acoustic features. As the A1 has been shown to exhibit task dependent plasticity (Atiani et al., 2009; Fritz et al., 2003), it is possible that these A1 effects arise from category-dependent task demands in sound encoding for the match-to-sample task. This interpretation is supported by the within-category object-identity analysis, which showed that object-level decoding correlated with acoustic distance for animal and instrument sounds—but not for speech—suggesting that different encoding mechanisms were engaged across categories: Whereas animal and instrument sounds were encoded using acoustic features, speech sounds were more likely encoded at the level of object identity or linguistic information (Bonte et al., 2014). Indeed, activity patterns in primary auditory cortex are modulated by speech-specific acoustic structure (Kiremitçi et al., 2021; Mesgarani et al., 2014) and by top down modulation arising from semantic context (Wikman et al., 2021). An additional possibility is that apparent category-level representations in A1 reflect covariance with low-level features not explicitly modelled in the present study. Thus, although category-level decoding was observed in A1, these effects are unlikely to reflect category abstraction *per se*. Instead, they may arise from task-dependent encoding strategies or from residual acoustic structure not fully captured by the present feature models.

Attentional modulation of sound feature processing is task dependent

Does attention modulate early auditory processing to selectively represent low-level features of the attended sound? The present results suggest that it does, however, in a task-dependent manner. More specifically, in the 3OW experiment, attention selectively modulated FC in A1 and belt areas, and HNR, entropy and FCSD in the STG (Figure 7E). These findings replicate earlier demonstrations of frequency specific attentional modulation for simple sounds (Costa et al., 2013; Paltoglou et al., 2009; Riecke et al., 2017), and extend them by showing that similar effects occur when selectively attending to natural sounds. Moreover, the results indicate that attentional modulation is not limited to the frequency content of the attended sound but can selectively enhance the processing of multiple independent low-level acoustic

features. Additionally, this modulation is not confined to primary auditory cortex, as reliable attentional effects were also observed in the STG.

This result contrasts with several prior studies that failed to observe attentional modulation of low-level feature processing or early auditory cortex activity. Puschmann et al. (2024) showed that BOLD responses in primary auditory areas were not explained significantly better by attended than unattended 10-minute spoken narratives, whereas attentional modulation emerged in higher-order regions such as STS. Using a similar paradigm with 10–15-minute narratives presented alone or in overlapping pairs, Kiremitçi et al. (2021) observed attentional modulation already in primary auditory cortex, but this modulation was selective for articulatory rather than spectral features. Discrepancies in the results between these studies and the present work may provide insight into attentional modulation. First, attentional modulation of low-level features may be weaker for speech than for other sound categories. This interpretation is supported by our finding that apparent object-identity representations were at least partially explained by attentional modulation of low-level feature processing for animal sounds, whereas no such effect was observed for speech (Figure 10). This is probably because speech-specific features such as articulation (Kiremitçi et al., 2021), semantic context (Wikman et al., 2021) or speaker identity (Mesgarani & Chang, 2012) may provide more efficient cues for selection. Second, attentional modulation of low-level feature processing may be amplified in soundscapes containing a larger number of overlapping sound objects, increasing competition at early sensory stages. Third, attentional modulation of low-level feature processing may be transient, such that during prolonged listening attention progressively shifts toward more abstract representations. This interpretation is consistent with spatiotemporal evidence showing that auditory attention initially modulates early sensory representations but rapidly reconfigures processing toward higher-order, object-level representations (Wikman et al., 2024) which may explain why sustained low-level modulation is difficult to detect during extended listening.

Electrophysiological work on speech (Dijkstra et al., 2015; Ding & Simon, 2012; Kerlin et al., 2010; J. A. O'Sullivan et al., 2015) and instruments (Hausfeld et al., 2021) has repeatedly shown that neural activity tracks the amplitude envelope of an

attended stream more strongly than that of an unattended stream. Recent fMRI results have localized this effect to primary auditory cortex and the right medial STS (Hausfeld et al., 2024), implying that envelope tracking may depend partly on low-level sound-feature processing. Although stronger envelope tracking has been argued to reflect top-down semantic predictions rather than genuine low-level modulation (Broderick et al., 2019), the present findings show that attention can modulate low-level feature processing directly. This suggests that genuine low-level mechanisms may also contribute to enhanced tracking of attended speech. Hausfeld et al. (2024) further dissociated the nature of envelope tracking across cortical regions: primary auditory areas tracked the continuous amplitude envelope of the attended stream, whereas the right STS tracked a more abstract, binary representation of speech presence (speaking vs. not speaking). The authors speculated that this pattern reflects a division of labour whereby primary auditory cortex may encode genuine low-level features, while the right STS represents higher-level, object-based aspects of the attended speaker. The present results provide empirical support for this interpretation, as attentional modulation of low-level feature processing is observed in primary auditory cortex, while the right STS expresses a representation of speaker identity that is largely independent of low-level features. The pronounced role of the right STS in decoding speaker identity has also been noted in other work (Bonte et al., 2014).

Importantly, modulation of low-level sound feature processing was not observed when the soundscape contained sounds from different categories (3OA). In those scenes, attention instead enhanced abstract category features represented in higher-level temporal and frontal regions with an especially strong effect in dissociating speech sounds from non-speech sounds (Renvall et al., 2021). This pattern suggests that auditory attention is adaptively guided by the representational level that best separates the competing sounds. When different sound categories are present, broad abstract distinctions help isolate the target, however when the sounds can't be distinguished based on category, attention resorts to modulation of low-level sound feature processing. Interestingly, however, even when the sounds were from different categories, the object identity representations of speakers and animals were modulated by attention (Figure 10). In other words, not only does attention gravitate neural activity patterns towards a broader abstract category template (Pedersini et

al., 2025; Wikman et al., 2025) but also contains sub-templates for the attended objects.

Conclusions

This study explored the neural representations of low-level and abstract sound features during listening and selective attention using multivariate fMRI analyses. The results revealed that:

1. Low-level sound features of natural sounds were robustly represented in early auditory areas, while abstract sound features are progressively encoded along the auditory ventral stream. This result aligns with influential models which suggest that sound feature complexity is processed hierarchically, with increasing abstraction along the ventral stream (Rauschecker & Scott, 2009).
2. Robust attentional modulation of low-level sound feature processing was observed in early auditory areas, a phenomenon which has not been previously demonstrated using natural sounds. This finding highlights the role of early auditory regions in selectively enhancing the processing of relevant acoustic information.
3. The extent of attentional modulation depends on how overlapping sounds can be distinguished. When sound objects could be separated based on category, attention operated at a higher level rather than enhancing low-level feature processing. This suggests that attentional modulation is task-dependent and dynamic – i.e., attention modulates different levels of hierarchical processing based on which level provides the most distinct contrast between the overlapping sounds.

References

Alho, K., Salmi, J., Koistinen, S., Salonen, O., & Rinne, T. (2015). Top-down controlled and bottom-up triggered orienting of auditory attention to pitch activate overlapping brain networks. *Brain Research*, *1626*, 136–145. https://doi.org/10.1016/j.brainres.2014.12.050

Alho, K., Vorobyev, V. A., Medvedev, S. V., Pakhomov, S. V., Starchenko, M. G., Tervaniemi, M., & Näätänen, R. (2006). Selective attention to human voice enhances brain activity bilaterally in the superior temporal sulcus. *Brain Research*, *1075*(1), 142–150. https://doi.org/10.1016/j.brainres.2005.11.103

Atiani, S., Elhilali, M., David, S. V., Fritz, J. B., & Shamma, S. A. (2009). Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields. *Neuron*, *61*(3), 467–480. https://doi.org/10.1016/j.neuron.2008.12.027

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

Bonte, M., Hausfeld, L., Scharke, W., Valente, G., & Formisano, E. (2014). Task-Dependent Decoding of Speaker and Vowel Identity from Auditory Cortical Response Patterns. *Journal of Neuroscience*, *34*(13), 4548–4557. https://doi.org/10.1523/JNEUROSCI.4339-13.2014

Boos, M., Lücke, J., & Rieger, J. W. (2021). Generalizable dimensions of human cortical auditory processing of speech in natural soundscapes: A data-driven ultra high field fMRI approach. *NeuroImage*, *237*, 118106. https://doi.org/10.1016/j.neuroimage.2021.118106

Broderick, M. P., Anderson, A. J., & Lalor, E. C. (2019). Semantic Context Enhances the Early Auditory Encoding of Natural Speech. *Journal of Neuroscience*, *39*(38), 7564–7575. https://doi.org/10.1523/JNEUROSCI.0584-19.2019

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, *2*(3), 27:1-27:27. https://doi.org/10.1145/1961189.1961199

Costa, S. D., Zwaag, W. van der, Miller, L. M., Clarke, S., & Saenz, M. (2013). Tuning In to Sound: Frequency-Selective Attentional Filter in Human Primary Auditory Cortex. *Journal of Neuroscience*, *33*(5), 1858–1863. https://doi.org/10.1523/JNEUROSCI.4405-12.2013

Dijkstra, K. V., Brunner, P., Gunduz, A., Coon, W., Ritaccio, A. L., Farquhar, J., & Schalk, G. (2015). Identifying the attended speaker using electrocorticographic (ECoG) signals. *Brain-Computer Interfaces*, *2*(4), 161–173. https://doi.org/10.1080/2326263X.2015.1063363

Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(29), 11854–11859. https://doi.org/10.1073/pnas.1205381109

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Fritz, J., Shamma, S., Elhilali, M., & Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience*, *6*(11), 1216–1223. https://doi.org/10.1038/nn1141

Giordano, B. L., Esposito, M., Valente, G., & Formisano, E. (2023). Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nature Neuroscience*, *26*(4), 664–672. https://doi.org/10.1038/s41593-023-01285-9

Giordano, B. L., McAdams, S., Zatorre, R. J., Kriegeskorte, N., & Belin, P. (2013). Abstract Encoding of Auditory Objects in Cortical Activity Patterns. *Cerebral Cortex*, *23*(9), 2025–2037. https://doi.org/10.1093/cercor/bhs162

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178. https://doi.org/10.1038/nature18933

Hausfeld, L., Disbergen, N. R., Valente, G., Zatorre, R. J., & Formisano, E. (2021). Modulating Cortical Instrument Representations During Auditory Stream Segregation and Integration With Polyphonic Music. *Frontiers in Neuroscience*, *15*. https://doi.org/10.3389/fnins.2021.635937

Hausfeld, L., Hamers, I. M. H., & Formisano, E. (2024). FMRI speech tracking in primary and non-primary auditory cortex while listening to noisy scenes. *Communications Biology*, *7*(1), 1217. https://doi.org/10.1038/s42003-024-06913-z

Hausfeld, L., Riecke, L., & Formisano, E. (2018). Acoustic and higher-level representations of naturalistic auditory scenes in human auditory and frontal cortex. *NeuroImage*, *173*, 472–483. https://doi.org/10.1016/j.neuroimage.2018.02.065

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402. https://doi.org/10.1038/nrn2113

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, *17*(2), 825–841. https://doi.org/10.1006/nimg.2002.1132

Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010). Attentional Gain Control of Ongoing Cortical Speech Representations in a "Cocktail Party". *Journal of Neuroscience*, *30*(2), 620–628. https://doi.org/10.1523/JNEUROSCI.3631-09.2010

Kiremitçi, I., Yilmaz, Ö., Çelik, E., Shahdloo, M., Huth, A. G., & Çukur, T. (2021). Attentional Modulation of Hierarchical Speech Representations in a Multitalker Environment. *Cerebral Cortex*, *31*(11), 4986–5005. https://doi.org/10.1093/cercor/bhab136

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. https://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. https://doi.org/10.3389/neuro.06.004.2008

Lartillot, O., Toiviainen, P., & Eerola, T. (2008). A Matlab Toolbox for Music Information Retrieval. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds), *Data Analysis, Machine Learning and Applications* (pp. 261–268). Springer. https://doi.org/10.1007/978-3-540-78246-9_31

Latinus, M., & Belin, P. (2011). Human voice perception. *Current Biology*, *21*(4), R143–R145. https://doi.org/10.1016/j.cub.2010.12.033

Leaver, A. M., & Rauschecker, J. P. (2010). Cortical Representation of Natural Complex Sounds: Effects of Acoustic Features and Auditory Object Category. *Journal of Neuroscience*, *30*(22), 7604–7612. https://doi.org/10.1523/JNEUROSCI.0296-10.2010

Leminen, A., Verwoert, M., Moisala, M., Salmela, V., Wikman, P., & Alho, K. (2020). Modulation of Brain Activity by Selective Attention to Audiovisual Dialogues. *Frontiers in Neuroscience*, *14*, 436. https://doi.org/10.3389/fnins.2020.00436

Lewis, J. W., Talkington, W. J., Walker, N. A., Spirou, G. A., Jajosky, A., Frum, C., & Brefczynski-Lewis, J. A. (2009). Human Cortical Organization for Processing Vocalizations Indicates Representation of Harmonic Structure as a Signal Attribute. *The Journal of Neuroscience*, *29*(7), 2283–2296. https://doi.org/10.1523/JNEUROSCI.4145-08.2009

Liakakis, G., Nickel, J., & Seitz, R. J. (2011). Diversity of the inferior frontal gyrus—A meta-analysis of neuroimaging studies. *Behavioural Brain Research*, *225*(1), 341–347. https://doi.org/10.1016/j.bbr.2011.06.022

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236. https://doi.org/10.1038/nature11020

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2014). Mechanisms of noise robust representation of speech in primary auditory cortex. *Proceedings of the National Academy of Sciences*, *111*(18), 6792–6797. https://doi.org/10.1073/pnas.1318017111

Muukkonen, I., & Salmela, V. R. (2023). *Entropy predicts early MEG, EEG and fMRI responses to natural images* (p. 2023.06.21.545883). bioRxiv. https://doi.org/10.1101/2023.06.21.545883

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLOS Computational Biology*, *10*(4), e1003553. https://doi.org/10.1371/journal.pcbi.1003553

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

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Oosterhof, N. N., Wiestler, T., Downing, P. E., & Diedrichsen, J. (2011). A comparison of volume-based and surface-based multi-voxel pattern analysis. *NeuroImage*, *56*(2), 593–600. https://doi.org/10.1016/j.neuroimage.2010.04.270

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., & Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex (New York, N.Y.: 1991)*, *25*(7), 1697–1706. https://doi.org/10.1093/cercor/bht355

O'Sullivan, J., Herrero, J., Smith, E., Schevon, C., McKhann, G. M., Sheth, S. A., Mehta, A. D., & Mesgarani, N. (2019). Hierarchical Encoding of Attended Auditory Objects in Multi-talker Speech Perception. *Neuron*, *104*(6), 1195-1209.e3. https://doi.org/10.1016/j.neuron.2019.09.007

Paltoglou, A. E., Sumner, C. J., & Hall, D. A. (2009). Examining the role of frequency specificity in the enhancement and suppression of human cortical activity by auditory selective attention. *Hearing Research*, *257*(1), 106–118. https://doi.org/10.1016/j.heares.2009.08.007

Pedersini, C. A., Muukkonen, I., & Wikman, P. (2025). *Attention modulates the geometry of auditory representations* (p. 2025.10.09.681368). bioRxiv. https://doi.org/10.1101/2025.10.09.681368

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

Puschmann, S., Regev, M., Fakhar, K., Zatorre, R. J., & Thiel, C. M. (2024). Attention-Driven Modulation of Auditory Cortex Activity during Selective Listening in a Multispeaker Setting. *Journal of Neuroscience*, *44*(15). https://doi.org/10.1523/JNEUROSCI.1157-23.2023

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

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

Renvall, H., Seol, J., Tuominen, R., Sorger, B., Riecke, L., & Salmelin, R. (2021). Selective auditory attention within naturalistic scenes modulates reactivity to speech sounds. *European Journal of Neuroscience*, *54*(10), 7626–7641. https://doi.org/10.1111/ejn.15504

Riecke, L., Peters, J. C., Valente, G., Kemper, V. G., Formisano, E., & Sorger, B. (2017). Frequency-Selective Attention in Auditory Scenes Recruits Frequency Representations Throughout Human Superior Temporal Cortex. *Cerebral Cortex*, *27*(5), 3002–3014. https://doi.org/10.1093/cercor/bhw160

Saenz, M., & Langers, D. R. M. (2014). Tonotopic mapping of human auditory cortex. *Hearing Research*, *307*, 42–52. https://doi.org/10.1016/j.heares.2013.07.016

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, *34*(3), 114–123. https://doi.org/10.1016/j.tins.2010.11.002

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

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–155. https://doi.org/10.1002/hbm.10062

Theunissen, F. E., & Elie, J. E. (2014). Neural processing of natural sounds. *Nature Reviews Neuroscience*, *15*(6), 355–366. https://doi.org/10.1038/nrn3731

Wessinger, C. M., VanMeter, J., Tian, B., Van Lare, J., Pekar, J., & Rauschecker, J. P. (2001). Hierarchical Organization of the Human Auditory Cortex Revealed by Functional Magnetic Resonance Imaging. *Journal of Cognitive Neuroscience*, *13*(1), 1–7. https://doi.org/10.1162/089892901564108

Wikman, P., Muukkonen, I., Kauramäki, J., Laaksonen, V., Varis, O., Petkov, C., & Rauschecker, J. (2025). Selective attention shapes neural representations of complex auditory scenes: The Roles of Object Identity and Scene Composition. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.0506-25.2025

Wikman, P., Sahari, E., Salmela, V., Leminen, A., Leminen, M., Laine, M., & Alho, K. (2021). Breaking down the cocktail party: Attentional modulation of cerebral audiovisual speech processing. *NeuroImage*, *224*, 117365. https://doi.org/10.1016/j.neuroimage.2020.117365

Wikman, P., Salmela, V., Sjöblom, E., Leminen, M., Laine, M., & Alho, K. (2024). Attention to audiovisual speech shapes neural processing through feedback-feedforward loops between different nodes of the speech network. *PLOS Biology*, *22*(3), e3002534. https://doi.org/10.1371/journal.pbio.3002534