





Perception of dynamic multi-speaker auditory scenes under different modes of attention

Stephanie Graceffo , David F. Little, Emine Merve Kaya, Mounya Elhilali *

Laboratory for Computational Audio Perception, Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA

ABSTRACT

Attention is not monolithic; rather, it operates in multiple forms to facilitate efficient cognitive processing. In the auditory domain, attention enables the prioritization of relevant sounds in an auditory scene and can be either attracted by elements in the scene in a bottom-up fashion or directed towards features, objects, or the entire scene in a top-down fashion. How these modes of attention interact and whether their neural underpinnings are distinct remains unclear. In this work, we investigate the perceptual and neural correlates of different attentional modes in a controlled “cocktail party” paradigm, where listeners listen to the *same* stimuli and attend to either a spatial location (operationalized here as a feature-based auditory attribute), a speaker (object-based), or the entire scene (global or free-listening) while detecting deviations in pitch of a voice in the scene. Our findings indicate that object-based attention is more perceptually effective than feature-based or global attention. Furthermore, object-based and spatial-based attention engage distinct neural mechanisms and are differentially modulated by bottom-up salience. Notably, while bottom-up salience aids in the initial segregation of auditory objects, it plays a reduced role in object tracking once attention has been voluntarily allocated. In addition, decoding the stimulus envelope from the EEG data revealed a source-sampling scheme in the global attention mode that is not present in the object or spatial modes. Overall, the study shows that the perception of the same acoustic scene differs according to the listening task, guided by an interaction between top-down and bottom-up processes.

1. Introduction

To cope with limited cognitive resources, the brain deploys attention as an adaptive mechanism to sort and prioritize sensory information. These processes direct focus towards relevant stimuli while filtering out distractions. In the auditory domain, attention enables us to navigate complex sound environments by making sense of competing sources that impinge on our ears. The cocktail party effect (Cherry, 1957; Haykin and Chen, 2005) illustrates this ability, allowing us to follow a conversation of interest despite background noise and select sounds of interest. A long-standing debate centers on *where* in the sensory hierarchy that selection is enacted and *what* unit of representation it targets. One influential hypothesis argues for an object-centered view: Attention locks onto a perceptual object (e.g., a single voice) and strengthens *all* its features as a coherent package, allowing effortless tracking across pitch fluctuations or movement of the sound across space (Shinn-Cunningham, 2008; Kerlin et al., 2010; Mesgarani and Chang, 2012). A competing account emphasizes feature-selective mechanisms, proposing that attention can be steered towards low level attributes such as spatial direction or fundamental frequency, with ‘objecthood’ emerging only after an attended feature has been isolated (Zatorre et al., 1999; Petkov et al., 2004). In the auditory system, spatial location is not represented as a primary sensory dimension but is instead computed from acoustic cues such as interaural time and level differences (ITDs, ILDs). Accordingly, in this study we operationalize

spatial attention as a form of feature-based auditory attention, reflecting selection over derived acoustic attributes rather than a distinct spatial hierarchy as typically defined in vision. This framework allows us to treat spatial and feature-based attention as related mechanisms while distinguishing their functional roles. Reconciling these views requires clarifying whether feature based and object based selection are separable processes or simply different entry points into a shared hierarchy.

Evidence from vision underscores the complexity of this issue. Spatial, feature, and object attention each leave distinct neural fingerprints (Treue and Trujillo, 1999; Saenz et al., 2002; Liu, 2011), yet they interact in non trivial ways. Object-centric theories predict a spread of attention, whereby focusing on one feature of an object automatically boosts its unattended attributes (O’Craven et al., 1999; Schoenfeld, 2003). By contrast, more recent work shows that feature-based selection can *suppress* other object features when they are task irrelevant (Wegener and Ehn, 2008; Freeman and Macaluso, 2014). To reconcile these seemingly contradictory results, current theories of visual processing suggest that feature- and object-based attentional mechanisms coexist in the visual pathway and operate as parallel filters that can cooperate or compete depending on task demands (Kravitz and Behrmann, 2011; Mayer and Vuong, 2012; Wegener et al., 2014; Geigerman et al., 2016).

* Corresponding author.

E-mail address: mounya@jhu.edu (M. Elhilali).

Auditory research mirrors this debate but with less consensus. Early PET and EEG studies reported little evidence for feature specific gain, concluding that attention targets integrated auditory objects (e.g. an entire voice) rather than low level cues like frequency or spatial hemifield (Zatorre et al., 1999; Petkov et al., 2004). Other studies including fMRI, MEG and intracranial recordings demonstrated robust enhancement of tonotopic and spatial maps when listeners attend to a single attribute, supporting a genuinely feature-based mechanism in both primary and non primary cortex (Ahveninen et al., 2006; Krumbholz et al., 2007; Da Costa et al., 2013). Crucially, suppression of an unattended feature which is diagnostic of independent feature filters has yet to be unambiguously demonstrated in audition (Paltoglou et al., 2009). Thus, two interpretations remain viable: (i) feature attention is simply an early manifestation of object-centered processing, or (ii) feature and object selection are distinct but cooperative processes that modulate auditory cortex in parallel. A critical test, therefore, is to present identical acoustic scenes while instructing listeners to adopt object, feature, or scene level goals, and to compare both perceptual efficacy and neural signatures across these modes. Because the acoustic input is held constant across tasks, any differences in behavioral or neural responses must arise from listeners' attentional strategies rather than from stimulus structure.

Attention without explicit goals, sometimes labeled global, free-listening or non-selective attention, is yet another form of attention that has received far less scrutiny. Visual work suggests that global attention yields rapid "gist" extraction of scene layout before finer analysis (Navon, 1977; Oliva and Torralba, 2007) and recruits early, low spatial frequency channels (Hegde, 2008). In addition, mismatch negativity studies show that deviant sounds can be detected even when they are task irrelevant, implying a parallel global monitor (Näätänen et al., 2007). Whether that monitoring filter integrates the whole scene into a unitary representation or, alternatively, samples individual sources in a serial fashion is still debated. These competing models make distinct neural predictions: a serial sampling mechanism should produce temporally shifting source-specific signatures, whereas a spatial anchoring mechanism should yield stable patterns tied to overall scene layout. One hypothesis is that, absent task goals, attention cycles through salient sources, guided by bottom-up conspicuity and exploratory drive (Kaya and Elhilali, 2014). A counter hypothesis posits a spatial scaffold: even in global mode, attention defaults to stable spatial anchoring, enhancing events that follow the prevailing layout (Böhm et al., 2013). Distinguishing these models is vital for understanding everyday listening, where listeners frequently drift between focused and diffuse states.

Another unresolved dimension concerns top-down/bottom-up interplay. When a listener locks onto a target, does increasing the intrinsic salience of a competing stream break that lock (a capture model), or are salient distractors actively suppressed once selection is established (a gating model)? Conversely, in global listening, is salience beneficial because it guides the sampling schedule or is it detrimental because simultaneous salience bursts create mutual masking? Prior studies rarely manipulate salience and attentional modes orthogonally, leaving the field with contradictory predictions.

To sift through these competing views, the current study employs identical, spatially dynamic "cocktail party" scenes and asks them to adopt three attentional modes in separate blocks (Fig. 1A): (1) monitor the entire mixture (*global*), (2) attend to any speech on the right (*feature/spatial*), or (3) track the male voice wherever it appears (*object*). Three voices (a male and two females) constitute the scene and all three move dynamically in space but are never concurrently present in one spatial location at the same time. All narrations are in German, a language unfamiliar to our listeners hence engaging the auditory system at a more sensory level and relying less on linguistic processing. Using unfamiliar speech also minimizes semantic learning or memory accumulation across repeated stimulus presentations, ensuring that performance differences reflect attentional mode rather than stimulus

familiarity. Targets (pitch shift) at carefully positioned moments in the scene facilitate monitoring of participants' attentional focus. This design allows us to examine explicit hypotheses:

- H1 - Dominance: Object-based attention will outperform feature-based attention, suggesting an object first hierarchy. A counter hypothesis is that feature-based attention will be equally effective once the feature is diagnostic and sufficient for segregation and tracking, implying shared mechanisms.
- H2 - Global sampling: Global attention operates via serial source sampling and neural markers will cycle across voices. A counter hypothesis is that global attention preserves a spatial map and that neural activity will remain anchored to spatial layout of the entire scene rather than individual streams.
- H3 - Salience modulation: Bottom-up salience boosts performance primarily in global mode but contributes little once selective attention is engaged. A counter hypothesis is that highly salient events capture attention even within selective streams, resetting the attentional focus.

We evaluate these competing accounts using converging behavioral metrics, stimulus reconstruction decoding and spectro-temporal EEG patterns. The analysis employs stimulus-reconstruction techniques ("decoding") that examines how different attentional modes modulate the encoding of acoustic streams, as well as classification approaches that capture how overall neural activity patterns distinguish among attentional states. This approach provides a framework to delineate when and how auditory attention privileges objects, features, or exploratory sampling in a dynamic setting mimicking real-world listening.

2. Results

2.1. Different modes of attention yield distinct behavioral performances

Even though the stimuli are exactly the same, behavioral responses reveal that the three modes of attention are not equally effective. Fig. 1B shows a clear ordinal pattern in hit rates, with global attention resulting in the lowest performance ($HR = 0.695 \pm 0.012$), followed by spatial attention ($HR = 0.768 \pm 0.019$), and object attention with the highest performance ($HR = 0.824 \pm 0.013$). A Friedman test comparing the effect of hit rates on the three attention modes reveals a statistically significant difference in hit rate and attention modes ($F = 55.2, p = 1.01e-12$). A Wilcoxon signed-rank test with Bonferroni corrections between hit rates shows a significant difference between global and object ($W = 100, p = 9.57e-12$), global and spatial ($W = 395, p = 7.05e-7$), and object and spatial ($W = 520, p = 8.24e-4$) modes.

Incorrect responses in this experiment stem from two sources of false positives: control trials and distraction trials. Control trials do not contain a target, while distraction trials contain a target in the unattended stream. This latter case occurs when subjects are performing the object task and there is a target in the female voice, or subjects are performing the spatial task and there is a target on the left side. Fig. 1C shows that false positive responses to control trials (solid bars) are prominently smaller for spatial and object attention ($FP = 0.042 \pm 0.006$ and $FP = 0.056 \pm 0.007$, respectively) compared to global attention ($FP = 0.263 \pm 0.018$). Wilcoxon signed-rank tests with Bonferroni corrections between false positive rates reveal a significant difference for global and spatial ($W = 184, p = 7.25e-6$), and global and object ($W = 174.5, p = 7.8e-4$), but not for object and spatial modes ($W = 189.5, p = 0.096$). In object and spatial attention tasks, distraction trials contain targets in the unattended stream (striped bars in Fig. 1C). The false positive rate for the spatial mode ($FP = 0.282 \pm 0.016$) is significantly higher ($W = 85.5, p = 1.72e-6$) than that of the object mode ($FP = 0.200 \pm 0.014$). Furthermore, comparing distraction to control rates for each mode confirms significantly higher

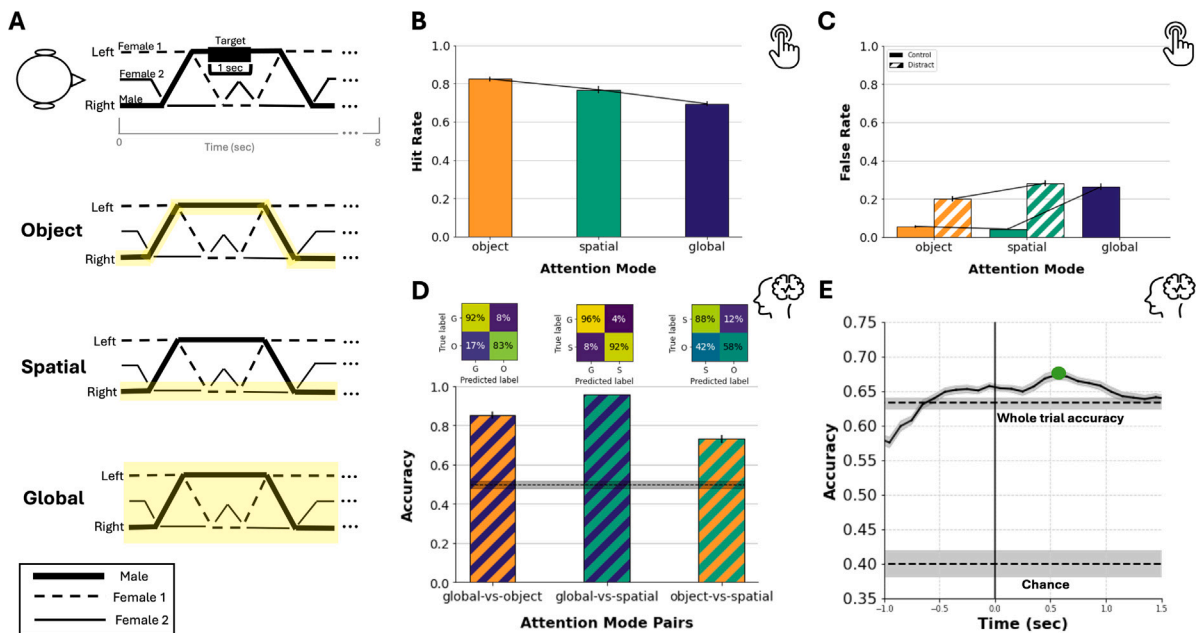


Fig. 1. (A) Setup of the cocktail-party stimulus used in this experiment. The stimuli consist of book sections narrated by three German speakers. The three speakers start at the left, center, and right, with occasional shifts in each direction. Targets are 1 s long segments of natural speech manipulated to have a pitch outside the pitch range of regular speech (3 semitones lower for male targets and 3 semitones higher for female targets). **Lower** Subjects listened to each auditory scene three times: Object (top) by attending to the male speaker regardless of spatial location, spatial (middle) by attending to the voices on the right, and global (bottom) by attending to the entire scene. (B) Target detection performance during each of the three attention modes of the behavioral experiment. Global attention represents free-listening, providing a baseline to evaluate selective attention. Directing attention to an acoustic spatial location (right side) and object (male speech) results in progressively increasing hit rates. Bars represent average hit rates for each condition and error bars represent standard error of the mean. Statistical significance is reported as (*), $p < 0.05$, (**), $p < 0.01$, (***) $p < 0.001$, (NS) Not Significant. The icon on the upper-right corner highlights that results reported in figure are based on behavioral responses. (C) False positive responses are elicited from both control trials (trials with no targets) and distraction trials (trials with a target in the unattended stream). (D) Neural responses from subjects elicit different signatures under each attention mode that are distinct enough for successful classification. Average accuracy for pairwise classifiers is reported across different cross-validation runs. The gray line represents the accuracy of chance classifier obtained from randomized labels. Confusion matrices (shown above each bar) illustrate class-wise prediction performance for each pairwise classifier. Values indicate row-normalized percentages (true labels on rows, predicted labels on columns). The icon on the upper-right corner highlights that results reported in figure are based on neural responses. (E) Three-way classification of a 0.75-second window beginning 0.25 s before the target onset shows a slight boost in accuracy compared to classifying across the whole trial.

distraction false rates (object mode ($W = 162.0$, $p = 6.47e-12$), spatial mode ($W = 19.0$, $p = 1.64e-13$)). Overall, the higher hit rate and lower distraction rate for the object task suggest that object-based attention is more perceptually effective than spatial attention for scenes that involve dynamic spatial configurations. Furthermore, the significantly higher distraction rates compared to control rates in both spatial and object attention allude to a failure of attentional deployment to the correct target rather than a failure in segregation.

Behavioral sensitivity across the three attention modes – measured using d' – follows the general trend of hit rates. Global attention has the lowest value ($d' = 1.31 \pm 0.826$), followed by spatial ($d' = 1.42 \pm 0.845$), then object attention ($d' = 1.84 \pm 0.850$). A Friedman test comparing the effect of all three attention modes on d' reveals that there is a statistically significant difference in d' and attention modes ($F = 24.11$, $p = 5.812e-6$). Wilcoxon signed-rank test with Bonferroni corrections between d' values reveals a significant difference in d' values for global and object ($W = 574$, $p = 3.29e-5$) and for object and spatial ($W = 549$, $p = 1.77e-5$) but not for global and spatial modes ($W = 1225$, $p = 0.617$). The higher d' for object attention compared to spatial attention supports the reasoning that auditory object tracking across spatial locations is facilitated by object-based attention but hindered during spatial-based attention due to the dynamic spatial nature of the scene.

The same analyses are replicated for behavioral responses obtained concurrently with neural recordings (albeit with a smaller sample size) compared to the behavioral only experiment. Comparison of the two responses using permutation tests (see Methods) shows no significant

difference between hit rates (object $p = 0.589$, spatial $p = 0.162$, global 0.300), false positive rates (object $p = 0.958$, spatial = 0.791 , global = 0.038), or d' (object $p = 0.855$, spatial = 0.958 , global = 0.061), demonstrating that behavioral effects are consistent across the two samples.

No significant effects of task order were observed. Specifically, comparisons between spatial-first and object-first groups revealed no significant difference between hit rates (object: $U = 830.5$, $p = 0.638$; spatial: $U = 843.0$, $p = 0.555$), false positive rates (object: $U = 664.0$, $p = 0.052$; spatial: $U = 667.5$, $p = 0.266$), or d' (object: $U = 945.5$, $p = 0.590$, spatial: $U = 869.0$, $p = 0.400$), indicating that behavioral effects were independent of task order.

2.2. Distinct neural signatures underlie different modes of attention

Neural responses to dynamic scenes under different modes of attention result in different EEG signatures, clearly distinguishable from one another (Fig. 1D). A classifier is trained on EEG responses from trials where subjects correctly detect the target, reasoning that these trials reflect proper attentional engagement aligned with task demands for each block. Classification accuracies for pairwise comparisons are highest for global vs. spatial ($95.9 \pm 0.10\%$, $AUC = 0.993$), followed by global vs. object ($85.3 \pm 1.8\%$, $AUC = 0.956$), and finally object vs. spatial ($73.2 \pm 2.1\%$, $AUC = 0.840$). All pairwise classifiers performed significantly above chance (Wilcoxon signed-rank test on subject-level accuracies: global vs. spatial $p = 4.49e-6$; global vs. object $p = 2.21e-5$; object vs. spatial $p = 9.10e-4$). Chance performance,

derived using a label-shuffling procedure, was $49.8 \pm 2.0\%$. Confusion matrices revealed asymmetric error patterns across comparisons. Global and spatial attention were classified with high recall in both the global vs. spatial and global vs. object contrasts ($> 90\%$), whereas object vs spatial showed substantially lower object recall (58%), indicating increased confusion between these two attention modes. Furthermore, a three-way classifier comparing the three modes of attention yields an average accuracy of $63.2 \pm 0.72\%$ across the entire trial (chance accuracy $40.0 \pm 1.87\%$), demonstrating that the three attentional states are neurally separable. One of the challenges in the experimental design is that each trial follows a distinct spatial layout, complicating the alignment of analyses across trials. Nonetheless, aligning the analysis near onset of target events reveals a slight, yet significant boost in classification accuracy about 500 ms after target onset relative to whole-trial accuracy ($68.8 \pm 0.64\%$ relative to $63.2 \pm 0.72\%$, t-test $p = 3.22e - 5$).

2.3. Scalp topographies support distinct neural activity patterns under different modes of attention

Fig. 2A depicts heatmaps in the three attentional modes during hit trials allowing us to examine possible underpinnings of differences between attention tasks. Average scalp topographies reveal distinct patterns in the three tasks. The object attention condition shows the expected frontal and bilateral occipital activation though accompanied by a weak temporal activation; while spatial attention reveals a stereotypical parietal-occipital suppression contra-lateral to the attended location with moderate frontal activation. In contrast, the global mode yields a very strong frontal enhancement with general posterior parietal suppression, likely to enhance auditory processing. An analysis of neural oscillations that may drive these topographic differences is performed across different frequency bands and anatomically distinct channel clusters. A one-way ANOVA comparing theta activity in frontal channels across the three attention modes reveals a significant main effect ($F(2, 46) = 6.47, p = 0.003$). Post-hoc paired t-tests with Bonferroni correction yields a significantly higher theta power in global vs. spatial ($t(23) = 3.67, p = 0.004$), object vs. spatial ($t(23) = 2.90, p = 0.024$) and global vs object condition ($t(23) = 3.57, p = 0.005$). In addition, spatial attention shows a significantly higher parietal-occipital alpha power relative to both object ($t(23) = 3.59, p = 0.005$) and global ($t(23) = 2.72, p = 0.037$) modes. Furthermore, subject-level correlations with the group topography were predominantly positive (Spatial Hit: 0.19 ± 0.26 ; Spatial False Positive Distract: 0.23 ± 0.38), with 17 out of 24 and 19 out of 24 subjects respectively showing positive similarity to the group heatmap. These results indicate that the group-level spatial pattern reflects a population-level trend rather than being driven by only a few individuals. The substantial variability across subjects is expected given the low signal-to-noise ratio of single-subject EEG data.

To further compare the overall patterns across all channels, a pairwise correlation of different attention modes using bootstrapping is performed (see Methods). Fig. 2A reveals that comparing heatmaps derived during correct hit trials yields correlations that are not significantly different than zero, indicating a low similarity and supporting that neural activity is unique for each mode of attention. This analysis is then replicated to probe parallels between *distraction trials* in selective attention modes (object and spatial) and correct trials in global attention. An interesting trend emerges when comparing the distraction trials for object and spatial modes to their respective hit responses in selective tasks and to hit responses in the global attention task. Fig. 2B shows that the object false positive distract trials are significantly correlated with the global hit trials with 95% CI [0.32, 0.90] but not with the object hit trials, 95% CI [-0.30, 0.83]. Notably, the average scalp topography for the object false positive distract trials are qualitatively aligned with the global hit trials with stronger frontal activation and more posterior parietal suppression. Similarly, Fig. 2C shows that the spatial false positive distract trials are significantly

correlated with the global hit trials with 95%CI [0.20, 0.87], but not with the spatial hit trials, 95% CI [-0.30, 0.77]. Furthermore, the average topography for the spatial false positive distract trials (Fig. 2C inset) also more closely resembles that of the global hit trials (Fig. 2A inset) than it does the spatial hit trials (Fig. 2A inset), particularly in terms of strong frontal activation. In the case of spatial distract trials, we still note a mild contra-lateral suppression.

2.4. Different attention modes follow distinct temporal patterns influenced by scene dynamics

To explore temporal dynamics of distinct attention modes, behavioral and neural responses as a function of time are examined relative to the spatial dynamics of the scene in each trial. Fig. 3A shows the hit rate as a function of time after the switching point (the time at which the speakers change directions, moving from one location in space to another) and reveals different trends of build-up as the scene unfolds. Rates at every time t denote the average hit rate of targets that are occurring at that time. For targets that occur “early” in the trial (after trial onset or following an early spatial shift), the hit rate build-up lasts over 100’s of milliseconds (up to 1.2 s after the shift) before stabilizing. Surprisingly, targets that happen after a “late” shift do not show this pattern; instead, all three attention modes show a stable hit rate level over time.

Quantifying this buildup trend by estimating the angle of the slope of the hit rate curve using a linear fit confirms that build-up angles are most prominent for the early shifts, but there is virtually no build-up observed for the late shifts (Fig. 3B). A two-way ANOVA to examine the effects of attention mode and switch time on hit rate confirms a significant main effect of attention mode on hit rate ($F = 148.9, p = 1.0e-43$), indicating that performance differed across attention modes. There is also a significant main effect of switch time on hit rate ($F = 448.0, p = 2.51e-58$), with higher performance for the late switches. However, there is no significant interaction between attention mode and switch time ($F = 0.62, p = 0.538$), suggesting that the effect of switch time on performance is consistent across tasks.

It is worth noting that although all modes have stabilized, the object hit rate is still larger than the spatial hit rate, which are both larger than the global hit rate, once again illustrating that object attention is stronger than spatial attention, and both types are stronger than global attention.

Examining the neural responses around the time of scene switches also reveals different trends in brain responses to the three attentional modes, reflecting different effects of attention reorienting. Fig. 3C shows classification accuracy of neural responses under the three tasks as a function of time, aligned near the onset of the switch. The results show that a three-way classifier is generally stable before the onset of the switch (average classification accuracy for 1 s before is $62.8 \pm 4.0\%$ for early switches and $65.1 \pm 2.5\%$ for late switches). In contrast, the classification accuracy drops to $54.8 \pm 1.2\%$ after early switches while it remains stable at $62.3 \pm 2.7\%$ for late switches. This pattern aligns closely with that observed in behavioral responses and suggests that attention reorienting occurs more prominently earlier in the trial as voices shift in the scene before stabilizing as the spatial configuration changes become part of the regularity in the scene.

2.5. Distinct temporal dynamics characterize selective attention to complex auditory environments under object and spatial modes

To further investigate how attention unfolds over time, we decode the stimulus envelope from the neural data, and examine the decoding correlation (between actual and reconstructed envelopes) as a function of time for each selective attention mode. This analysis focuses on the object and spatial tasks as they both have a clear notion of “attended” versus “unattended” perceptual streams. Separate decoders are trained for each voice and each spatial location in the stimulus and are grouped

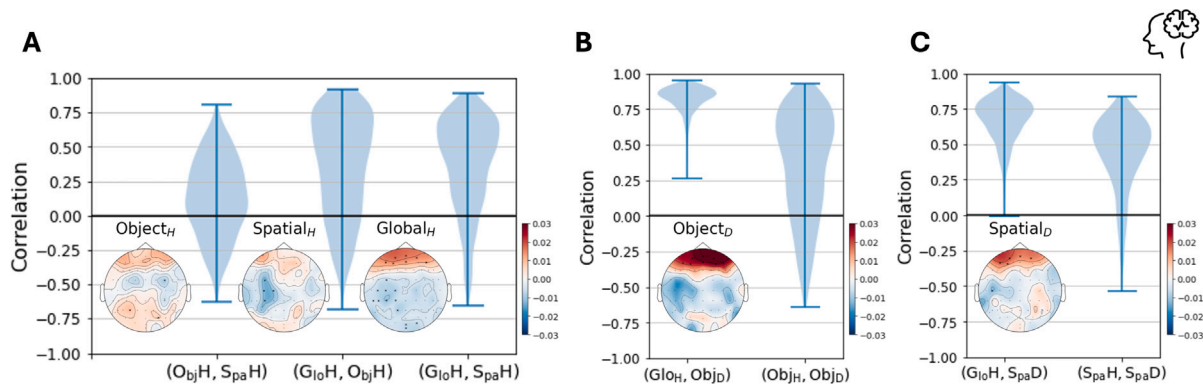


Fig. 2. (A) *Hit Responses*: Correlations between topographies for hit trials for attention mode pairs are not significantly different from zero indicating low similarity between pairs. The inserted topographic maps of the object, spatial, and global attention modes qualitatively illustrate these differences as well. (B) *Object False Positive Responses*: Object false positive and global hit responses are significantly correlated, while object hit and object false positive responses are not significantly correlated. This can also be observed qualitatively when comparing the scalp topographies of the object false positive to those of the global hit and object hit in A. (C) *Spatial False Positive Responses*: Spatial false positive and global hit responses are significantly correlated, while spatial hit and spatial false positive responses are not significantly correlated. This can also be observed qualitatively when comparing the scalp topographies of the spatial false positive to those of the global hit and spatial hit in A.

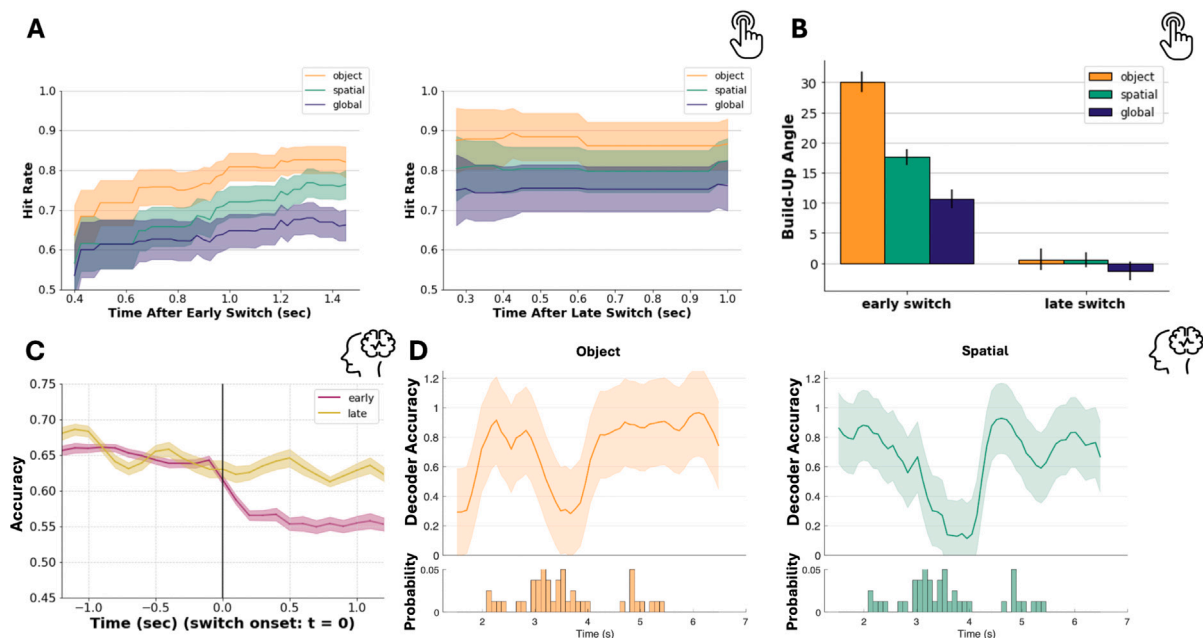


Fig. 3. (A) Time course of hit rate for early (before or after the first speaker switch) attention reorientation (*Left*) and for late (after the second or third speaker switch) attention reorientation (*Right*). (B) Build-up angle of the time course for early and late shifts. Angle is found by fitting a line at the build-up time window. (C) 3-way classification as a function of time from switch onset for the early and late switch (D) Decoding accuracy for Object with probability distribution for times where speakers converge on the same side (*Left*). Decoding accuracy for Spatial with probability distribution for times where speaker moves away from the attended (*right*) side (*Right*).

into attended stream decoders (male left and male right for the object task; male right, female 1 right, and female 2 right for the spatial task) and unattended stream decoders (see Section 4.4.6). The decoding accuracy is denoted as 1 if the attended output exceeds the unattended output and 0 otherwise.

Fig. 3D shows the average decoding accuracy over time for the object and spatial conditions. For the Object condition (*left*), the probability histogram below the curve represents the probability of one speaker starting to switch towards the side of the auditory scene where another speaker already is (speaker convergence). This condition captures increased segregation difficulty given that multiple speakers are likely to share the same spatial location. We first note that overall decoder accuracy starts reasonably low near the beginning of the trial indicating reduced phase-locked responses to the male speaker (from

either left or right side) very early in the scene. Throughout the trial, when the probability of speaker convergence is low (i.e., greater spatial separation), the decoder accuracy yields a high performance near 80%. This performance drops sharply between 3–4.2 s to around 37% when there is a high probability of one speaker moving towards the side where another speaker already is, leading to less spatial separation between speakers. Spearman rank correlation between decoding accuracy and the speaker convergence probability distribution reveals a strong negative correlation ($r_s = -0.684$), indicating that accuracy tends to decrease as the speakers converge and spatial separation between speakers decreases. A permutation test (1000 iterations) confirms that this relationship is statistically significant ($p = 0.001$).

For the Spatial condition (*right*), the histogram probability represents the likelihood of a speaker starting to switch from the attended

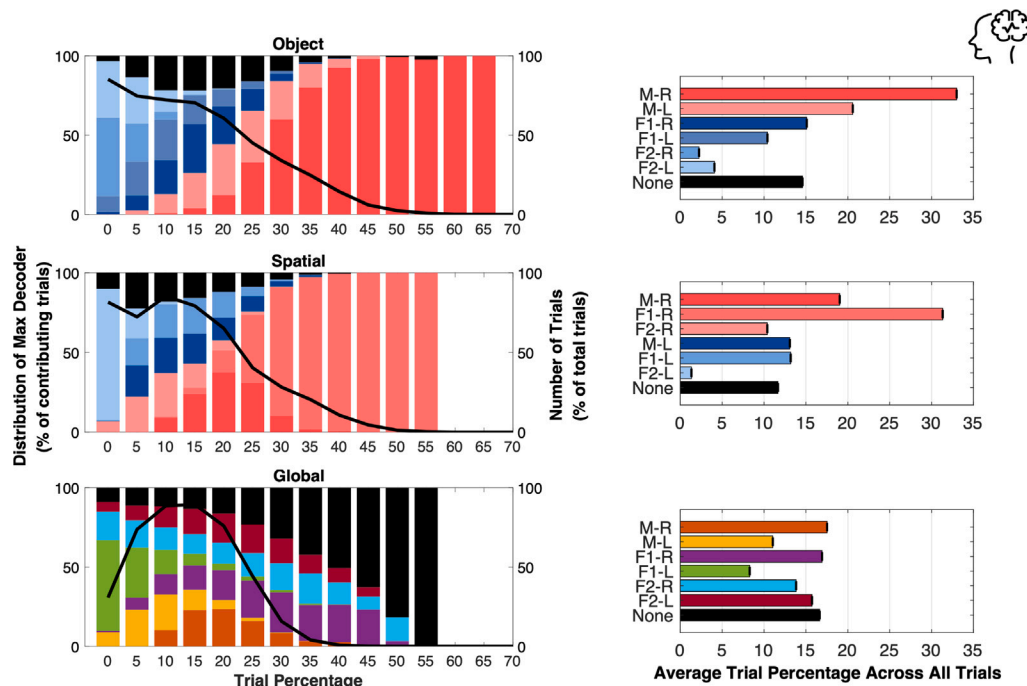


Fig. 4. Left Panel: Distribution of trial percentages during which each decoder is dominant. Right panel: Average trial percentage for each decoder across all trials. Both panels: Top: Object, Middle: Spatial, Bottom: Global. For object and spatial plots, red represents decoders in the attended stream and blue represents decoders in the unattended stream. For all plots, black represents no dominant decoder. M-R: male-right, M-L: male-left, F1-R: female1-right, F1-L: female1-left, F2-R: female2-right, F2-L: female2-left.

(right) side of the auditory scene (attended-side departure). This distribution is most relevant to the spatial mode as it reflects disruption in spatial stability of the scene. At the beginning of the trial, the decoder accuracy starts reasonably high given that participants are immediately cued to the direction to attend to. This cueing likely underlies increased encoding of all voices in the attended direction though segregation of overlapping voices may still need to unfold over time as seen in the buildup effects of Fig. 3A. The accuracy starts dropping around 2.2 seconds as the probability a speaker leaves the attended side increases. The decoder accuracy reaches its lowest point, below 20%, just before 4 s, when the attended-side departure probability is highest. From 4–4.5 s, as the probability decreases, the accuracy increases. A second peak in switch probability occurs around 4.5–5.5 s which again causes a dip in decoder accuracy. Again, Spearman rank correlation between the decoding accuracy and the attended-side departure probability distribution reveals a strong negative correlation ($r_s = -0.302$), indicating that accuracy tends to decrease as the probability that a speaker switches away from the attended side increases. A permutation test (1000 iterations) confirms that this relationship is statistically significant ($p = 0.037$).

2.6. Global attention operates using a source sampling strategy by examining different voices as the scene unfolds

In contrast with the selective attention modes where an attended and unattended stream emerges, the global attention task requires subjects to attend to the scene as a whole. To gain insight into neural mechanisms underlying this task, we examine decoders trained on each speaker in the scene (male, female 1, female 2) and each side (right, left) for each of the 3 attentional modes. For each analysis window, the decoder with the highest correlation is denoted as winner.

Fig. 4 (Left Panel) shows the distribution of trial time percentages during which each decoder is considered the dominant decoder. A decoder is determined to be the dominant decoder at a given time point if it has the maximum value among all decoders, exceeds a predefined

threshold, and surpasses the next-highest decoder by a minimum margin (see 4.4.6). The x-axis (Trial Percentage) represents the percentage of a trial where a given decoder is the dominant decoder. For example, if a trial is 8 s, $x = 25$ means 25% of the trial, or 2 s of the trial. The right y-axis (Number of Trials) represents the percentage of the total number of trials that actively contribute to each time point (black line). For example, in the object condition, at $x = 25$, 44% of all trials have at least one decoder that is dominant for 25% of the trial. The left y-axis (Distribution of Max Decoder) represents the proportion of trials (as a percent of the contributing trials, right y-axis) in which each decoder dominates at each point in time. For example, in the object condition, at $x = 25$, 44% of all trials contribute, and the dominant decoder is the Male-Right for 33% of the contributing trials, Male-Left for 32% of the contributing trials, Female1-Right for 14% of the contributing trials, Female2-Right for 6% of the contributing trials, and the dominant decoder could not be determined for the remaining 15% of the time; the remaining two decoders (Female1-Left, Female2-Left) are not the dominant decoders in any trial for this amount of time. Fig. 4 (Right Panel) shows the average trial time percentage for each decoder across all trials. For the object and spatial modes, the shades of red indicate decoders in the attended stream, and the shades of blue indicate decoders in the unattended stream. Because there are no attended and unattended decoders in the global mode, each decoder is assigned a different color. In all three modes, the black indicates that a dominant decoder could not be determined.

In the object and spatial modes, the subjects' attention tends to be on the attended decoders (the two male decoders for the object mode, and the three right decoders for the spatial mode) more often than the unattended decoders. This is clearly indicated by the prominence of the red-shaded decoders in the top and middle panels. In the case of global attention, an interesting phenomenon emerges with global attention showing no consistent preference for any individual decoder. Fig. 4, bottom row shows that global attention tends to switch among the six decoders, implying that subjects tend to sample each sound source in the auditory scene, with no one voice (or decoder) occupying more than 18% of any given trial and most decoders being dominant

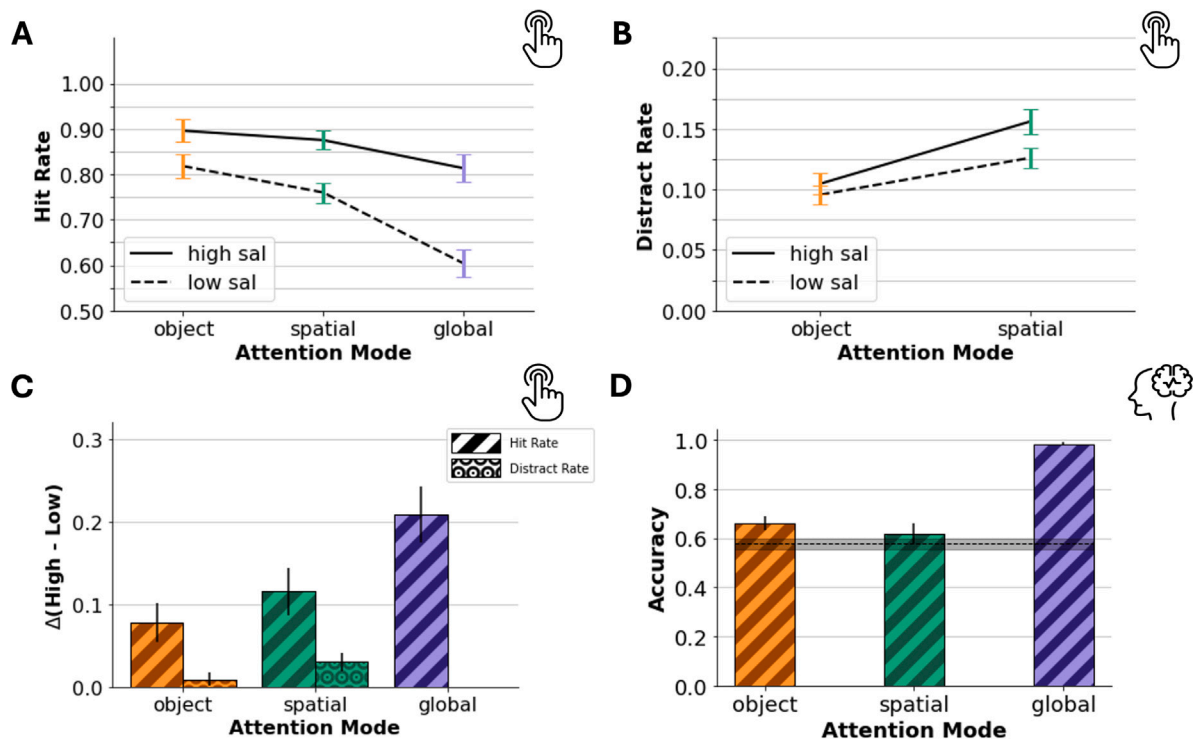


Fig. 5. (A) Hit rates for high and low salience targets under each mode of attention. (B) Distract rates for high and low targets under object and spatial modes of attention. (C) The greatest difference in hit rate occurs for the global mode, while the greatest difference in distract rate occurs for the spatial mode. (D) Neural signature classification of high and low salience targets for each attention mode. Similar to the behavioral responses, classification accuracy is greatest for the global mode.

between 8% and 17% of each trial. This uniform distribution is consistent with a source-sampling pattern for the global mode of attention, suggesting that participants do not maintain a sustained preference for a single stream and instead alternate across multiple sound sources. Importantly, decoder assignments were restricted to time points with sufficiently strong and well-separated correlation values (See Methods 4.4.6), reducing the likelihood that the observed switching reflects low-confidence or noise-driven fluctuations.

2.7. Global attention benefits most from bottom-up attention compared to selective modes

Dividing the targets into high and low salience events reveals that global attention shows the most gain from the salience level of the target compared to selective attention. Fig. 5A shows hit rates for high versus low salience targets for each mode of attention. Selective modes of attention strongly boost perception of low-salience targets compared to the global mode of attention, while high-salience targets are detected to a similar degree under all modes of attention. The hit rate in the global mode for low salience targets is only around 60%, but increases significantly for high salience targets around 82%.

This nonlinear behavior of target salience under different attention modes is confirmed with a two-way repeated-measures ANOVA on hit rates across attention mode and salience state and reveals that there is a statistically significant interaction between attention mode and salience state ($F = 6.57$, $p = 0.002$). Simple main effects show that both attention modes and salience states have a statistically significant effect on the hit rate ($F = 25.42$, $p = 7.92e-10$ for attention mode and $F = 53.71$, $p = 9.90e-10$ for salience state). A Friedman test conducted across attention modes for each salience level indicates significant effects of attention mode for each salience level (high salience: $\chi^2(2) = 10.65$, $p = 0.005$; low salience: $\chi^2(2) = 32.04$, $p = 1.10e-7$). Bonferroni-corrected Wilcoxon signed-rank tests are then used to examine pairwise differences between attention modes for each salience level. For high

salience targets, hit rates do not differ significantly between object and spatial ($W = 198.0$, $p = 1$) or between spatial and global ($W = 130.5$, $p = 0.067$), but object and global show a significant difference ($W = 97.5$, $p = 0.028$). For low salience targets, all pairwise comparisons are significant: object and spatial ($W = 228.5$, $p = 0.042$), object and global ($W = 46.5$, $p = 1.85e-7$), and spatial and global ($W = 113.0$, $p = 6.54e-5$). Within each attention mode, Wilcoxon signed-rank tests comparing high- and low-salience targets confirm significantly higher hit rates for high-salience targets across all attention modes (object: $W = 198.5$, $p = 2.24e-02$; spatial: $W = 152.5$, $p = 1.11e-04$; global: $W = 183.0$, $p = 4.33e-06$, Bonferroni-corrected)

Looking at the distract trials (Fig. 5B), false positive distract rates are slightly higher for high salience targets compared to low salience targets, though this trend is more evident in the spatial mode than the object mode. A two-way repeated-measures ANOVA across attention mode and salience state on the false positive distract rate reveals a statistically significant interaction between attention mode and salience state ($F = 5.63$, $p = 0.020$), suggesting the effect of salience state on distract rate varied across attention modes. Simple main effects show that attention mode has a significant effect on the distract rate ($F = 20.8919$, $p = 6.8924e-6$), implying distract rates differ for each task, with object having a lower distract rate. Similarly, salience state also has a significant effect on the distract rate ($F = 4.5829$, $p = 0.0330$), with lower distraction rates for low salience targets. Wilcoxon signed-rank tests for each mode indicate that there is no significant difference between high- and low-salience distractors for the object mode ($W = 797.0$, $p = 0.384$), but there is for the spatial mode ($W = 552.5$, $p = 0.008$). Furthermore, Wilcoxon signed-rank tests for each salience level indicate a significant difference between attention modes for both high-salience ($W = 318.0$, $p = 6.51e-6$) and low-salience ($W = 438.0$, $p = 0.001$) distractors. Fig. 5C shows the difference between high and low salience target hit rates for each attention mode, as well as the difference between high and low salience target distract rates for each

attention mode and captures the larger difference in hit rate observed in the global mode.

Using the neural responses to classify the salience of each target as high or low salience reveals the same benefit of high salience targets in the global attention. Fig. 5D shows the results of high/low salience classification from the neural data of correct trials. The greatest neural signature discrimination between high and low salience targets occurs for the global mode ($98.3 \pm 2.7\%$) compared to the object mode ($66.0 \pm 1.2\%$) and spatial mode ($61.9 \pm 1.2\%$), mirroring the trend observed in the behavioral data depicted in Fig. 5C. Interestingly, object and spatial attention performance is only slightly though still significantly above a baseline chance classifier ($57.7 \pm 2.35\%$). This suggests that salience may not be sufficiently distinguishable in object and spatial attention but plays a significant role in driving performance in global attention.

To verify that classifier performance was not driven by a small subset of subjects, we examined classification accuracy at the individual level. Because several participants had too few trials in one salience category, a full within-subject decoding analysis was not feasible. Instead, we evaluated subject-wise classification performance using leave-one-subject-out cross-validation. This analysis showed that subjects' accuracies were broadly distributed and did not contain extreme outliers, indicating that no single participant or small group disproportionately influenced the group-level results. These findings confirm that the salience classification patterns observed for each attentional mode reflect population-level trends rather than being driven by atypical individuals.

3. Discussion

Auditory attention allows listeners to make sense of real-world environments by selectively enhancing relevant sounds while suppressing irrelevant ones. This capacity has long been studied through paradigms such as the “cocktail-party” effect, where listeners are able to track a voice of interest despite simultaneous competing sources (Cherry, 1953; Haykin and Chen, 2005). Previous research has shown that attention can be directed towards specific auditory features such as spatial location (Mondor and Zatorre, 1995; Ahveninen et al., 2006), towards coherent auditory objects like speech streams (Best et al., 2008; Ding and Simon, 2012), or more diffusely in global or free-listening modes that lack task-driven focus (Näätänen et al., 2007). However, whether these modes reflect distinct neural mechanisms or a unified attentional system operating at different levels remains a key unresolved question in auditory neuroscience.

The present study directly addresses this issue by contrasting object-based, feature-based (spatial), and global attention while holding the acoustic input constant. This allows us to isolate attentional strategy from stimulus-driven factors. Empirical findings indicate that these attentional modes are not only behaviorally distinct but also engage separable neural mechanisms, supporting the idea that auditory attention is implemented by multiple, functionally dissociable systems. These findings extend and clarify a range of earlier observations from the neuroimaging literature on spatial tuning and speech tracking (Altmann et al., 2008; Degerman et al., 2008; Petkov et al., 2004; Fritz et al., 2003) to models of attention that posit feature and object selection as coexisting yet distinct processes (Kravitz and Behrmann, 2011; Mayer and Vuong, 2012).

In line with our first hypothesis (H1), object-based attention yielded the strongest behavioral performance, with significantly higher hit rates and lower distraction compared to spatial attention. These findings replicate and reinforce earlier work showing that attention directed towards auditory objects results in enhanced perception and cortical tracking (Kerlin et al., 2010; Mesgarani and Chang, 2012). What is notable from the findings is that these benefits were observed despite the fixed acoustic input hence ruling out confounds related to acoustic variability or task difficulty and confirming that the attentional

mode itself shapes perception and behavior. EEG analyses revealed distinct neural signatures for each mode: Object attention produced bilateral activation in temporal areas, while spatial attention elicited the characteristic parietal alpha lateralization associated with spatial focus (Banerjee and Snyder, 2011; Frey and Mainy, 2014). These results suggest that even when feature-based attention is effective, object-based tracking provides more robust engagement, particularly when the auditory scene is dynamic and sources move in space. Yet, object-based attention did not always outperform feature-based attention. In the case of salient targets, spatial and object tasks yielded statistically similar detection rates. This result argues that when bottom-up cues are strong and diagnostic, feature-based selection can match the effectiveness of object tracking. Such flexibility aligns with visual attention models showing that strong feature salience can drive efficient selection even in the absence of object continuity (Frey and Mainy, 2014; Wegener et al., 2014). However, under lower-salience conditions, the difference between the two modes was more pronounced, suggesting that object attention confers resilience when perceptual demands are high or cues are ambiguous. This hierarchical relationship mirrors the idea proposed in auditory scene analysis that proto-objects may form from early feature grouping and serve as the substrate for object-based attention (Shinn-Cunningham, 2008; Bregman, 1990).

It is worth noting that although in this work we have only tested feature-based attention to acoustic locations, neuroimaging results demonstrating attentional modulation for a variety of features suggest that similar effects could possibly be seen for directed attention to other features. This is especially the case considering that there does not appear to be feature-dependent differences in feature-based attention in vision, and that visual and auditory attention seem to share many common mechanisms (Scholl, 2001; Busse and Roberts, 2005). One distinction from vision, however, is in the treatment of the spatial dimension. Space-based attention has traditionally been treated as a separate form of selective attention in vision (Soto and Blanco, 2004; Kravitz and Behrmann, 2011), though it has been suggested that it could be unified under the same framework as feature-based attention (Maunsell and Treue, 2006). However, there is little evidence supporting space as a special feature in audition. Studies that have investigated the effect of attention to frequency and space have suggested the two features operate under the same fundamental process (Krumbholz et al., 2007; Maddox and Shinn-Cunningham, 2012). It is also worth considering that space in audition is derived from neural computations on signals reaching the two ears, in a similar manner to pitch or other acoustic features. Even if spatial attention differed significantly from feature-based attention, the current results still demonstrate that auditory attention can operate in distinct modes, with spatial or feature-based attention differing from global and object-based attention.

In contrast to the selective modes, global attention served as a baseline in our paradigm and emerged not as passive monitoring but as an active, dynamic strategy. Decoding of stimulus envelopes revealed that listeners attending globally did not maintain a combined percept of the scene as a whole but rather cycled through individual sources. The analysis of decoding performance supported the hypothesis H2 that no decoder dominated the trial and the system appeared to sample streams serially. This “source-sampling” behavior provides new empirical support for theoretical claims that global attention involves exploratory scanning guided by salience and expectation (Oliva and Torralba, 2007; Kaya and Elhilali, 2014). These findings challenge the notion that global attention reflects a degraded or noisy version of selective listening. Instead, it appears to be a qualitatively distinct attentional state, particularly suited for detecting unexpected events across a complex scene. We note that a noise-driven account of this pattern cannot be fully excluded. However, decoder dominance was assigned only under conservative confidence thresholds and verified via bootstrap resampling (Methods), making the observed uniform

distribution more consistent with active source sampling than with unstable tracking failure.

This interpretation is further supported by the temporal dynamics of behavior and EEG responses. Following early spatial shifts in speaker location, performance in all attentional tasks initially declined but quickly recovered, particularly for object attention. These build-up effects were limited to early shifts. Later in the trial, attention had stabilized, and further spatial changes no longer disrupted performance. These findings align with previous work showing that object formation takes time but is resilient to spatial perturbation once it is established (Cusack et al., 2004; Micheyl et al., 2005). Importantly, we interpret these findings cautiously, not as evidence of object formation per se, but of rapid object updating or re-binding within a stable perceptual framework. The transient drop in accuracy after early shifts may reflect reorientation within an established object map rather than full re-segregation.

While the dynamics of selective and global attention were largely shaped by task goals, the influence of bottom-up salience introduced a different dimension and revealed that the relationship between bottom-up salience and attention mode was highly nonlinear. In support of hypothesis H3, high-salience targets significantly improved detection in the global task, but had much smaller effects under selective attention. This suggests that once attention is allocated to a particular object or spatial stream, the system is less influenced by salience in competing sources. This pattern is consistent with studies showing suppression of unattended salient stimuli during focused tasks (Wegener and Ehn, 2008; Taya and Adams, 2009; Freeman and Macaluso, 2014). Our neural classification of salience echoed this behavioral pattern: high- vs. low-salience targets were most discriminable in global attention (98%), but only marginally so in object and spatial tasks (~62%). These findings add to a growing body of literature suggesting that bottom-up salience is modulated by top-down task constraints, and further suggest that the influence of salience is greatest when attentional goals are weakly specified or diffuse.

Taken together, these findings clarify long-standing ambiguities in the auditory attention literature and offer a framework for understanding how attentional systems interact in real-world listening. By testing object, feature, and global attention in a tightly controlled yet naturalistic context, we provide evidence for distinct processing modes with different temporal dynamics, neural signatures, and sensitivity to salience. These results unify previous psychoacoustic, imaging, and electrophysiological findings and suggest that attentional mechanisms are flexibly deployed based on task demands and scene statistics. Future work should investigate how these modes interact and transition, whether they are instantiated by overlapping or independent cortical networks, and how they contribute to auditory perception in more complex, multi-modal environments.

4. Materials and methods

4.1. Experimental paradigm

Participants in both behavioral and neural experiments were presented with sound mixtures consisting of simultaneously played German audiobook narration extracts by three voices (a young adult female, an old adult female, and a young adult male). The three voices moved spatially throughout the duration of the trial, following a pattern explained next. In test trials, a pitch shift of one of the voices was introduced at some point during the trial. The experiments engaged the participants in three consecutive tasks divided into three experimental blocks:

1. In the first block (global attention), participants were asked to attend to the sound mixtures and indicate whether *any* of the voices had a change in pitch. Behavioral responses were recorded at the end of each trial as two-alternative forced-choice Yes/No response in both the behavioral-only and neural experiments.

2. In the second block (spatial attention), participants were instructed to attend only to voices in the right ear and flag any change in pitch. Detection of pitch changes in voices on the left would be considered a false alarm.
3. In the final block (object attention), participants were instructed to pay attention to the male voice regardless of its spatial location and flag the presence of pitch changes, while ignoring the female voices.

All participants performed the global condition as their first block and then were randomly assigned the spatial and object block in a counter-balanced fashion. Critically, the *same* 50 trials were presented in all three blocks (with trial order randomized in each block); therefore allowing careful control over acoustic parameters across attentional conditions. Counter-balancing the order of the spatial and object blocks helped minimize any learning or familiarity effects that might arise from repeated exposure to the same stimuli. Participants performed a short training block before the experiment to ensure they understood the global task. Subjects had no prior knowledge of the spatial/object tasks until the end of the global block. The choice of German narrations was deliberate, as participants in both behavioral and neural experiments did not speak any German, hence relying primarily on the acoustic characteristics of the scene without any linguistic or semantic processing and reducing the influence of semantic memory across blocks. Both behavioral and neural experiments took place inside a soundproof booth and stimuli were delivered with headphones (Sennheiser HD595).

4.1.1. Stimulus design

The 3 voices were manually extracted from public domain Librivox German book narrations (Male: <https://librivox.org/ein-vade-mecum-fur-den-hrn-sam-gotht-lange-pastor-in-laublingen-by-gotthold-ephraim-lessness/>, Female 1: <https://librivox.org/menschen-im-krieg-by-andrea-s-latzko/>, Female 2: <https://librivox.org/das-letzte-maerchen-by-paul-keller/>) recorded at 22050 Hz. From these narrations, 81 male segments were extracted from chapters 2 and 5, 61 female 1 segments were extracted from chapter 1, and 58 female 2 segments were extracted from chapter 10. The overall pitch range of the speakers was approximately A3-D4. Segments were chosen to have a prosody to sound like spoken single sentences, with no regard to meaning of the words spoken or whether the segment contains actual sentences. Listeners need continuous speech to follow speakers in a busy auditory scene, so segments were manually processed to reduce silent periods, including narrator pauses and words spoken quietly. The 50 trials were constructed by selecting one unique segment for each narrator such that the length difference between the three segments would be less than 300 ms.

The task was to detect a segment with a change in pitch. Fifty trials with unique sentences were constructed, 10 of which contained no pitch altered segments (control trials). The target specification in the remaining trials was as follows: 10 male-right, 15 male-left, 10 female-right, 5 female-left. The target was a 1 s long segment of the speech that was altered to have slightly higher (if female) or lower (if male) pitch. The right target segments were in the absolute right for the duration of the target, but the left targets could be anywhere between center and absolute left.

4.1.2. Target construction

Targets were 1 s long segments of voice manipulated to have a modified pitch. Pitch manipulation was performed by time dilation for male targets and compression for female targets with a phase vocoder. Targets in the male voice were 3 semitones lower than the natural speech, while targets in the female 1 voice were 3 semitones higher. Male targets never appeared before the first direction change of the male from right to left. Female 1 targets could appear on the left side before the direction change. There were no female 2 targets. Target segment onsets and offsets, as well as speech immediately prior to and following the target were smoothed by a 10 ms ramp to avoid abrupt transitions in sound. All manipulations preserved the temporal structure and intelligibility of the original recordings.

4.1.3. Spatial parameters

Each trial consisted of the 3 voices narrations played simultaneously over headphones. Each speaker started at a specific spatial location at the start of every trial: The male always started on the right, female 1 always started on the left, and the female 2 always started in the center. After a brief period, the speakers began switching locations in space. Once one voice reached the new location, the speaker remained there for a few seconds while the voice that was originally at that location moved to the opposite side. This movement pattern happened 1–3 times in each trial.

In order not to bias the audio towards any one direction, the voice movements were constructed such that there would be no gaps of speech at absolute right or absolute left at any point during the trial. Voices never overlapped in any direction except for the brief moments when one voice was reaching a direction just as another speaker was leaving. Additionally, no voice lingered in the center after stimulus onset: Voices were either moving between right and left spatial locations or entirely at the right or left sides. Stimulus length varied between approximately 5–10 s. The speed at which a voice moved between the right and left directions was constant, with a voice taking 1.2 s to move from one side to the other. The direction shifts started at a random time after the first second, and the number of shifts in each trial depended on trial length.

To simulate voice positions in 3-D space, a head-related transfer function (HRTF) was applied. The HRTF was recorded on a mannequin (Neumann KU 100) under the same conditions in which human HRTFs are recorded. The NH172 HRTF was used from the ARI HRTF database. Trajectories for each voice were constructed between -90 degrees (left) and 90 degrees (right) denoting their position at each point in time. Spatial dynamics of the scenes were increased by adding jitter to the trajectory when speakers had stable position at left, center or right. That is, instead of a straight trajectory, a sinusoid with a period of 5 spanning 50 degrees was inserted. At the time of the direction switch, the trajectory moved from the middle of the left or right jitter (65 or -65 degrees) in a linear line lasting 1.2 s. The female 1 and male voices were permitted to move between the left and right sides, while female 2 only moved between the right side and center. Female 2 moved at the same speed as the other two speakers, but the movements took only as long as necessary to fill in gaps in the absolute right.

The spatial treatment of the voices resulted in 3 stereo signals: male-right, male-left, female1-right, female1-left, female2-right, female2-left. These tracks were combined together resulting in a rich 3D scene that was presented to participants.

4.2. Participants

All behavioral and neural experimental procedures were approved by the Johns Hopkins Homewood IRB; and all participants gave informed consent and were compensated. 90 subjects (55% females, 18–31 years of age) participated in the behavioral experiments. 18 subjects had a negative D' and were removed from further analysis due to inability to perform the task. 24 subjects (58% females, 18–31 years of age) participated in the neural experiments. No participants from the behavioral-only experiment took part in the neural experiment. The same 50 trials used in the behavioral experiment were presented to the subjects while neural data was collected via EEG.

4.3. EEG acquisition and preprocessing

EEG measurements were obtained using a 64-channel Biosemi Active Two system. EEG data were analyzed using MATLAB (Mathworks Inc, MA), with FieldTrip (Oostenveld et al., 2011) and Noisetools analysis tools and Python toolboxes and custom scripts. Neural signals were first down-sampled to 256 Hz, then demeaned. Next, bad channels were identified and interpolated using neighboring good channels. The data were detrended, and any remaining bad channels were again identified

and interpolated. Eyeblinks were eliminated using time-shifted PCA, and outliers were detected based on a weighted correlation structure and removed. Finally, the data were re-referenced by subtracting a weighted mean.

4.4. Data analysis

4.4.1. Behavioral responses

Correct responses (i.e. detect targets in test trials) and false alarms were obtained for each attentional mode and each subject. All results were corrected for multiple comparisons using Holm–Bonferroni correction to confirm statistical significance (Snedecor and Cochran, 1989) and results post correction were reported. Residuals were checked for normality using the Shapiro–Wilk test ($p < 0.05$). To assess potential task-order effects, behavioral performance metrics (hit rate, false positive rate, and sensitivity, d') were compared between participants who completed the spatial block first and those who completed the object block first using nonparametric Mann–Whitney U tests.

4.4.2. Behavioral responses in behavioral only vs. behavioral and neural experiments

Behavioral responses were obtained from two datasets: one collected in a behavioral-only experiment with a larger sample (90 subjects), and one collected concurrently with neural recordings (24 subjects). To assess whether differences in sample size influenced the results, permutation tests were conducted for hit rate, false positive rate, and d' . For each metric, 24 subjects were randomly sampled without replacement from the larger dataset 1000 times, and the mean was computed for each subsample. The resulting distribution of subsampled means was then compared to the mean of the smaller dataset. p -values were calculated as the proportion of sub-sampled means that were as or more extreme than the observed mean in the smaller dataset.

4.4.3. Saliency calculation

Target saliency was computed using a bottom-up attention model (Kaya and Elhilali, 2014) trained on an independent dataset of human saliency judgments from natural soundscapes. No data from the present experiment were used to train or tune the model. The model computes statistical predictions across a set of acoustic features, and saliency for each feature is quantified as the deviation from its predicted value at each point in time. To obtain a compact representation for each trial, we extracted the peak feature deviations occurring at the moment of the target, yielding one model-derived feature vector per target event. Listeners' responses played no role in computing these model features. Instead, subject responses were used only to define a binary detectability outcome, indicating whether the target was behaviorally detected. A trial was labeled “detected” if at least half of participants reported hearing the target and “missed” otherwise. These detectability labels are therefore behavioral ground truth and remain distinct from the bottom-up saliency features.

To categorize the model-derived saliency features into interpretable classes, a logistic regression classifier was trained to map the model features onto the detected vs. missed behavioral labels. The classifier output represents the estimated probability that a given target would be detected. Following the standard logistic threshold, trials with predicted probability ≥ 0.5 were assigned to a “high saliency” category and those below 0.5 to a “low saliency” category. We use the terms “high” and “low” saliency to describe these model-based categories, which reflect the extent to which the model predicts that a target's acoustic features are sufficient to make it perceptually detectable.

4.4.4. Neural heatmaps

Heatmaps were generated from the EEG data processed as outlined in Section 4.3. First, the EEG data for each trial was normalized

to zero-mean and unit-variance, and only the portions of the trial corresponding to the non-switch regions were retained for analysis. This normalization and selection process was repeated for all subjects and trials. To visualize group-level patterns, we used a bootstrap aggregation procedure. In each resample, approximately two-thirds of the subjects were randomly selected, and their trial-averaged EEG data were averaged together to form a bootstrap group-average sample. This process was repeated 1000 times, yielding 1000 resampled group-average EEG patterns, which were then used to construct the heatmaps via Python's MNE toolbox.

Channel significance was determined via bootstrap confidence intervals on the subject means. Subject-level means were first obtained by averaging across trials within each subject. For each channel, a 95% bootstrap confidence interval (1000 resamples) was computed across subjects. Channels whose confidence intervals did not include zero were considered significant.

Analysis of oscillatory activity across scalp channels used 5 anatomically distinct channel clusters following the grouping used in Wilroth et al. (2025). The analysis focused on "frontal", "bilateral temporal", "central", "parietal" and "occipital" groups. For each channel cluster, the average power spectrogram was derived using a 1-second window with 50% overlap. Power in specific frequency bands was then derived over the theta (3–7 Hz), alpha (8–14 Hz), beta (15–29 Hz) and gamma (30–125 Hz). For each band and channel cluster, a one-way repeated-measures ANOVA was conducted to examine the effect of attention mode on EEG activity. Post-hoc pairwise comparisons (using Bonferroni-corrected t-tests) were applied to explore individual differences between attention modes. Given the limited number of trials in each condition across bands, channels and subjects, other statistical tests using bootstrapping were also performed to confirm observed trends. Effects reported here were confirmed using both methods.

4.4.5. Neural classification of attentional modes

A logistic regression classifier was used for all classification analyses. Binary logistic regression was applied for pairwise mode classification (Fig. 1D) and for high/low salience target classification (5D). For the three-way attention-mode analyses (1E and 3C), we used the multinomial (softmax) extension of logistic regression, which allows logistic regression to handle multi-class classification.

For both attention-mode and salience classification, EEG data were preprocessed as described in Section 4.3. Next, Multiway Canonical Correlation Analysis (MCCA) (de Cheveigne and Di Liberto, 2018) was applied to the EEG data, reducing the original 64 channels to 30 canonical components. These components were then encoded into five frequency bands ($\delta = 1-2$, $\theta = 3-7$, $\alpha = 8-14$, $\beta = 15-29$, $\gamma = 30-100$ Hz). Finally, the average power within each band was computed to serve as the input features to the classifier. To evaluate classifier performance and assess stability, the model was trained using a grid search over regularization parameters, specifically, a range of values for the inverse regularization strength (C) and the elastic net mixing parameter (L1 ratio) using 5-fold cross-validation for model selection. To ensure robustness to data partitioning, this entire procedure was repeated ten times, each with a unique random seed controlling the stratified shuffling. This resulted in 10 independent classifier performance estimates, from which the mean and standard deviation were computed to assess overall classification accuracy and variability.

To evaluate classifier performance relative to chance, a null model was established using label-shuffled classification. For pairwise classification between groups, separate null models were computed for each of the three attention mode pairs and averaged to obtain a single baseline value (Fig. 1D). For high/low salience level classification within each attention mode, null models were computed independently for each mode and similarly averaged (Fig. 5D). For the three-way classification, a single null model was computed across all classes (Fig. 1E). These null models were used to quantify and compare classifier accuracy relative to chance.

4.4.6. Neural decoding

The stimulus envelope was reconstructed from neural recordings, in order to evaluate the tracking of speech envelopes under different attentional conditions (Aiken and Picton, 2008; Ding and Simon, 2012). The neural responses were processed according to the procedure in Section 4.3 then Python's MNE toolbox was used to fit decoding models to estimate the stimulus (envelope) from neural responses.

In this experiment, each trial had three voices (male, female 1, female 2) and two sides (right, left), for a total of six acoustic signals per trial. For each attentional mode (object, spatial, global), subject-specific decoders were estimated for each acoustic signal (male right, male left, female 1 right, female 1 left, female 2 right, female 2 left). A leave-one-out cross-validation was performed to find the optimal regularization parameter from a set of values ranging from 10^{-4} to 10^3 and optimal time lags ranging from 0 to 500 ms. To train the model, all trials but one were concatenated then normalized to zero-mean and unit-variance, while the remaining trial was normalized and reserved for testing.

This procedure resulted in 6 predicted stimulus envelopes for a given test trial. Pearson's correlation between predicted and actual envelopes was computed over 1-second windows, with varying resolutions (between 50% and 90% overlap depending on analysis). The analysis was performed for each attentional mode, subject, and trial stimulus, and quantitatively similar results were obtained for different window sizes and overlap values.

Correlation values over time were organized in three $N \times T \times D$ matrices (one for each attention mode) where N is the number of trials in the experiment, T is the time index sampled over different windows and D is the number of decoders (6 corresponding to 3 stereo stimuli). For each task, the mean across trials resulted in a correlation value for each window and each decoder. Mean correlations across decoders were compared in two ways: combining decoders into attended (foreground) and ignored (background) streams based on attentional condition; or computing the maximum correlation across all 6 decoders.

Attended vs. Ignored decoders:

The attended/ignored decoder grouping was performed as follows. For the object task, the attended stream (A_{obj}) and unattended stream (U_{obj}) decoders were defined by combining together decoder outputs of the male voice (attended stream) and combining decoder outputs of the female1 and female2 voices (ignored streams), following the rule:

$$A_{obj} = \max(\text{male-left, male-right})$$

$$U_{obj} = \max(\text{female1-left, female1-right, female2-left, female2-right})$$

For the spatial task, the attended stream decoder (A_{spa}) was defined over all voices on the right channel, while the unattended stream decoder (U_{spa}) was defined over all voices on the left channel, following the rule:

$$A_{spa} = \max(\text{male-right, female1-right, female2-right})$$

$$U_{spa} = \max(\text{male-left, female1-left, female2-left})$$

No such grouping was feasible for the global condition. For the object and spatial tasks, the maximum correlation was obtained for the attended and ignored streams, producing a $1 \times T$ vector of correlation values.

To analyze decoding accuracy, the classification rules described above were applied within a bootstrap sampling framework to reduce the impact of noise and account for experimental complexity and variability. For each iteration (1000 total), the trials were randomly sampled with replacement and averaged to form a representative trial. The attended and unattended groups were defined according to the specified rules, and each resampled trial was flagged as correct (1) if the attended output exceeded the unattended output, or incorrect (0), otherwise. Accuracy and variability were then computed by averaging across all iterations for each time window.

To evaluate the relationship between the accuracy and the probability distribution, Spearman rank correlation coefficient, a non-

parametric measure of monotonic association that does not assume normality, was computed. Because both the accuracy and probability distribution failed a normality test (Kolmogorov–Smirnov, $p < 0.05$), we assessed statistical significance using a permutation test rather than relying on standard correlation p -values. Specifically, we computed the observed Spearman correlation, then generated a null distribution by randomly permuting one variable 10,000 times and recalculating the correlation at each iteration. The p -value was computed as the proportion of permuted correlations that were as or more extreme than the observed value (two-tailed test). This procedure allowed us to determine whether the observed correlation was significantly different from chance.

Decoder switching:

A second analysis looked at all 6 decoders as a function of time across the different attentional tasks and evaluated the *duration* in which each of the decoders was the highest across each trial. The analysis computed the percentage of a trial each of the 6 decoders was maximum. 30% of trials were discarded from the analysis as they contained outlier counts for at least one decoder. Outliers were defined as values 1.5 times the inter-quartile range greater than the third quartile or less than the first quartile. To further mitigate noise in the data arising from the complex experiment paradigm, a bootstrap sampling method was implemented in which a subset of the trials from each attention mode were randomly selected and averaged together to form a representative trial for each mode; this process was repeated 1000 times to get a collection of representative trials. For each window in the trial, the decoder with the highest correlation was deemed to be the decoder to which the subject was attending at that point in time. If the top two correlations were within a certain threshold (0.2% of each other), or if the maximum correlation was below a certain threshold (0.01), no decoder was selected as the attended decoder. For each of these new sampled trials, once again the number of windows that each decoder was the maximum decoder were counted.

CRedit authorship contribution statement

Stephanie Graceffo: Writing – original draft, Visualization, Validation, Methodology, Formal analysis. **David F. Little:** Methodology, Conceptualization, Formal analysis. **Emine Merve Kaya:** Writing – original draft, Visualization, Formal analysis, Conceptualization. **Mounya Elhilali:** Writing – review & editing, Funding acquisition, Conceptualization, Methodology.

Code availability

EEG data analyses were performed in MATLAB with FieldTrip and Nisetoools analysis tools, Python toolboxes, and custom scripts. The software code that supports the findings of this study is available from the corresponding authors upon reasonable request.

Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

This work was supported in part by the Office of Naval Research, USA N00014-23-1-2086 and the National Science Foundation, USA 2444353-01.

Data availability

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

References

- Ahveninen, J., Jääskeläinen, I.P., Raij, T., Bonmassar, G., Devore, S., Hämäläinen, M., Levänen, S., Lin, F.-H., Sams, M., Shinn-Cunningham, B.G., Witzel, T., Belliveau, J.W., 2006. Task-modulated “what” and “where” pathways in human auditory cortex. *Proc. Natl. Acad. Sci.* 103 (39), 14608–14613.
- Aiken, S.J., Picton, T.W., 2008. Human cortical responses to the speech envelope. *Ear Hear.* 29 (2), 139–157.
- Altmann, C.F., Henning, M., Döring, M.K., Kaiser, J., 2008. Effects of feature-selective attention on auditory pattern and location processing. *NeuroImage* 41 (1), 69–79.
- Banerjee, S., Snyder, A., 2011. Oscillatory alpha-band mechanisms and the deployment of spatial attention to anticipated auditory and visual target locations: Supramodal or sensory-specific control mechanisms? *J. Neurosci.*
- Best, V., Ozmeral, E.J., Kopčo, N., Shinn-Cunningham, B.G., 2008. Object continuity enhances selective auditory attention. *Proc. Natl. Acad. Sci. USA.*
- Böhm, T.M., Shestopalova, L., Bendixen, A., Andreou, A.G., Georgiou, J., Garreau, G., Pouliquen, P., Cassidy, A., Denham, S.L., Winkler, I., 2013. The role of perceived source location in auditory stream segregation: Separation affects sound organization, common fate does not. *Learn. Percept.* 5 (Suppl. 2), 55–72.
- Bregman, A.S., 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound.* MIT Press, Cambridge, Mass.
- Busse, L., Roberts, K., 2005. The spread of attention across modalities and space in a multisensory object. *Proc. Natl. Acad. Sci. USA* 102 (51), 18751–18756.
- Cherry, E.C., 1953. Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* 25 (5), 975–979.
- Cherry, E.C., 1957. *On Human Communication.* MIT Press, Cambridge, MA.
- Cusack, R., Deeks, J., Aikman, G., Carlyon, R.P., 2004. Effects of location, frequency region, and time course of selective attention on auditory scene analysis, 30 (4), 643–656.
- Da Costa, S., van der Zwaag, W., Miller, L.M., Clarke, S., Saenz, M., 2013. Tuning in to sound: Frequency-selective attentional filter in human primary auditory cortex. *J. Neurosci.* 33 (5), 1858–1863.
- de Cheveigne, A., Di Liberto, G., 2018. Multiway canonical correlation analysis of brain signals. *NeuroImage.*
- Degerman, A., Rinne, T., Särkkä, A., Salmi, J., Alho, K., 2008. Selective attention to sound location or pitch studied with event-related brain potentials and magnetic fields. *Eur. J. Neurosci.* 27 (12), 3329–3341.
- Ding, N., Simon, J.Z., 2012. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci.* 109 (29), 11854–11859.
- Freeman, E., Macaluso, E., 2014. Fmri correlates of object-based attentional facilitation vs. suppression of irrelevant stimuli, dependent on global grouping and endogenous cueing. *Front. Integr. Neurosci.*
- Frey, J., Mainy, N., 2014. Selective modulation of auditory cortical alpha activity in an audiovisual spatial attention task. *J. Neurosci.*
- Fritz, J., Shamma, S., Elhilali, M., Klein, D., 2003. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neurosci.* 6 (11), 1216–1223.
- Geigerman, S., Verhaeghen, P., Cerella, J., 2016. To bind or not to bind, that's the wrong question: Features and objects coexist in visual short-term memory. *Acta Psychol.* 167, 45–51.
- Haykin, S., Chen, Z., 2005. The cocktail party problem. *Neural Comput.* 17 (9), 1875–1902.
- Hegde, J., 2008. Time course of visual perception: Coarse-to-fine processing and beyond. *Prog. Neurobiol.* 84 (4), 405–439.
- Kaya, E.M., Elhilali, M., 2014. Investigating bottom-up auditory attention. *Front. Hum. Neurosci.* 8, 327.
- Kerlin, J.R., Shahin, A.J., Miller, L.M., 2010. Attentional gain control of ongoing cortical speech representations in a “cocktail party”. *J. Neurosci.* 30 (2), 620–628.
- Kravitz, D.J., Behrmann, M., 2011. Space, object-, and feature-based attention interact to organize visual scenes. *Atten. Percept. Psychophys.* 73 (8), 2434–2447.
- Krumbholz, K., Eickhoff, S.B., Fink, G.R., 2007. Feature- and object-based attentional modulation in the human auditory “where” pathway. *J. Cogn. Neurosci.* 19 (10), 1721–1733.
- Liu, T., 2011. Constant spread of feature-based attention across the visual field. *Vis. Res.*
- Maddox, R., Shinn-Cunningham, B., 2012. Influence of task-relevant and task-irrelevant feature continuity on selective auditory attention. *J. Assoc. Res. Otolaryngol.* 13, 119–129.
- Maunsell, J.H., Treue, S., 2006. Feature-based attention in visual cortex, 29(6), 317–322.
- Mayer, K.M., Vuong, Q.C., 2012. The influence of unattended features on object processing depends on task demand. *Vis. Res.* 56, 20–27.
- Mesgarani, N., Chang, E.F., 2012. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485 (7397), 233–236.
- Micheyl, C., Tian, B., Carlyon, R.P., Rauschecker, J.P., 2005. Perceptual organization of tone sequences in the auditory cortex of awake macaques. *Neuron* 48 (1), 139–148.
- Mondor, T.A., Zatorre, R.J., 1995. Shifting and focusing auditory spatial attention. *J. Exp. Psychol. Hum. Percept. Perform.* 21 (2), 387–409.
- Näätänen, R., Paavilainen, P., Rinne, T., Alho, K., 2007. The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clin. Neurophysiol.* 118 (12), 2544–2590.

- Navon, D., 1977. Forest before Trees: The Precedence of Global Features in Visual Perception. Tech. Rep..
- O'Craven, K.M., Downing, P.E., Kanwisher, N., 1999. fMRI evidence for objects as the units of attentional selection. *Nature* 401 (6753), 584–587.
- Oliva, A., Torralba, A., 2007. The role of context in object recognition. *Trends Cogn. Sci.* 11 (12), 520–527.
- Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.M., 2011. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* 2011, 156869.
- 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. *Hear. Res.* 257 (1–2), 106–118.
- Petkov, C.I., Kang, X., Alho, K., Bertrand, O., Yund, E.W., Woods, D.L., 2004. Attentional modulation of human auditory cortex. *Nature Neurosci.* 7 (6), 658–663.
- Saenz, M., Buracas, G.T., Boynton, G.M., 2002. Global effects of feature-based attention in human visual cortex. *Nature Neurosci.* 5 (7), 631–632.
- Schoenfeld, M., 2003. Dynamics of feature binding during object-selective attention. *Proc. Natl. Acad. Sci.*
- Scholl, B., 2001. Objects and attention: The state of the art. *Cognition* 20 (1–2), 1–46.
- Shinn-Cunningham, B., 2008. Object-based auditory and visual attention. *Trends Cogn. Sci.* 12 (5), 182–186.
- Snedecor, G., Cochran, W., 1989. *Statistical Methods*. Iowa State University Press, Ames.
- Soto, D., Blanco, M., 2004. Spatial attention and object-based attention: A comparison within a single task. *Vis. Res.* 44 (1), 69–81.
- Taya, S., Adams, W.J., 2009. The fate of task-irrelevant visual motion: Perceptual load versus feature-based attention. *J. Vis.*
- Treue, S., Trujillo, J.C.M., 1999. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* 399 (6736), 575–579.
- Wegener, D., Ehn, F., 2008. Feature-based attention and the suppression of non-relevant object features. *Vis. Res.* 48 (27), 2696–2707.
- Wegener, D., Galashan, F.O., Aurich, M.K., Kreiter, A.K., 2014. Attentional spreading to task-irrelevant object features: Experimental support and a 3-step model of attention for object-based selection and feature-based processing modulation. *Front. Hum. Neurosci.* 8.
- Wilroth, J., Alickovic, E., Skoglund, M.A., Signoret, C., Rönnerberg, J., Enqvist, M., 2025. Improving tracking of selective attention in hearing aid users: The role of noise reduction and nonlinearity compensation. *Eneuro* 12 (2).
- Zatorre, R.J., Mondor, T.A., Evans, A.C., 1999. Auditory attention to space and frequency activates similar cerebral systems. *NeuroImage* 10 (5), 544–554.