Auditory Object Formation in Temporally Complex Acoustic Scenes

Berfin Bastug (berfin.bastug@esi-frankfurt.de)
Max Planck School of Cognition, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
Wilhelm-Wundt-Institute of Psychology, Leipzig University, Leipzig, Germany
Ernst Strüngmann Institute for Neuroscience in Cooperation with Max Planck Society, Frankfurt am Main, Germany

Yue Sun (yue.sun@esi-frankfurt.de)
Ernst Strüngmann Institute for Neuroscience in Cooperation with Max Planck Society, Frankfurt am Main, Germany

Erich Schröger (schroger@uni-leipzig.de)
Max Planck School of Cognition, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
Wilhelm-Wundt-Institute of Psychology, Leipzig University, Leipzig, Germany

David Poeppel (dp101@nyu.edu)
Max Planck School of Cognition, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
Department of Psychology, New York University, New York, NY, USA

Abstract

The auditory system decomposes boundary-less sensory input into meaningful units through Auditory Scene Analysis (ASA) (Bregman, 1990). Repetition helps listeners segregate overlapping sounds and identify distinct auditory objects (McDermott et al., 2011). Previous studies suggest that repeated units in noisy or ambiguous contexts can eventually be perceived as stable auditory objects (Barczak et al., 2018; McDermott et al., 2011), though the behavioral dynamics of this process remain unclear. We investigated this build-up process using 'tone cloud' stimuli. By manipulating repetition strength and unit duration of tone cloud units, we created auditory analogues of the motion coherence paradigm. Participants completed repetition detection and sensorimotor synchronization tasks, allowing us to examine how the accumulation of sensory evidence supports the emergence and stabilization of auditory objects. Results reveal sigmoidal, quasi-categorical performance in both tasks. In detection, performance improves earlier for shorter durations. Interestingly, In synchronization, performance converges across durations, showing that once an object emerges, it can be tracked equally well regardless of duration. Our results suggest a categorical shift in perception, with stabilization occurring after sufficient repetition.

Keywords: auditory perception, auditory objects, repetition detection, synchronization

Introduction

The auditory system must decompose sensory input into perceptually meaningful units (Bregman, 1994). Repetition plays a crucial role in this process by helping listeners detect distinct auditory objects (Barascud et al., 2016; Barczak et al., 2018; McDermott et al., 2011; Winkler et al., 2009). McDermott et al. (2011) found that repetition improves object identification in noisy mixtures, with performance reaching ceiling after the third repetition. Barczak et al. (2018) provided complementary neurophysiological evidence, showing that neural oscillations in thalamocortical circuitry phase-align to repeating objects and gradually reach statistical significance level after the third cycle. Their findings can be summarized by a two-step process: build-up of sensory evidence followed by stabilization, such that additional evidence no longer influences perception. However, the behavioral signature of this dynamic process remains largely unexplored.

Here, we combined elements from the two aforementioned experiments to investigate auditory object formation. Our stimuli were tone clouds—randomly generated clusters of 50-ms tones lacking explicit boundary cues—similar to those used by Barczak et al. (2018). We adopted McDermott et al. (2011)'s target/distractor approach, manipulating repetition strength by varying the proportion of repeated (target) to regenerated (distractor) tones. This created a continuum from fully repeated to continuous sound sequences, forming an auditory analogue of motion coherence tasks in vision (Shadlen & Kiani, 2013; Shadlen & Newsome, 1996). To perceive repetition, participants had to group repeated tones into stable auditory objects. This design allowed us to explore the minimum amount of sensory evidence needed to trigger the perception of a repeated object in complex, unstructured acoustic scenes.

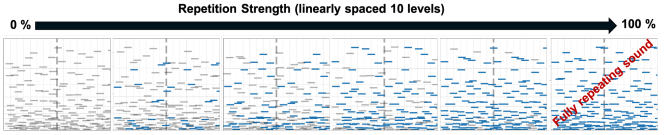

Figure 1. Illustration of tone clouds and repetition strength manipulation. Time-frequency grids were used to generate tone clouds with varying ratios of repeated (blue) and regenerated (gray) tones.

Methods

Participants completed two psychophysical experiments: repetition detection and sensorimotor synchronization (SMS). In the detection task, they judge

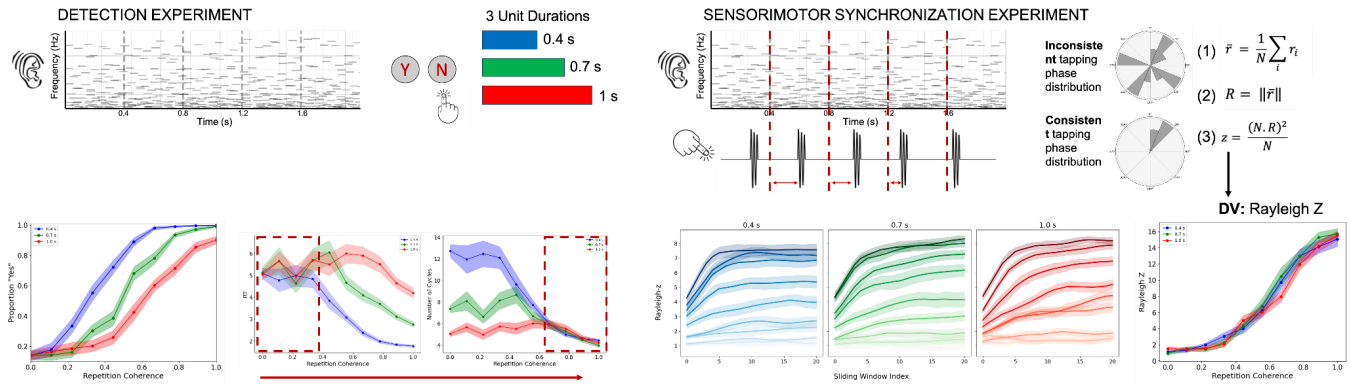

Figure 2. Schematic illustrations and behavioral results of the two experiments. Left: Repetition detection task with corresponding behavioral results. Right: Sensorimotor synchronization (SMS) task with behavioral outcomes. Illustrations include example "good" and "bad" trials mapped on a unit circle, alongside the formula used to calculate phase alignment.

whether a sound sequence contained repeated patterns. In addition to repetition strength, we also manipulated unit duration to better understand the nature of evidence required for stable object formation.

Detection provides a single time-point decision, capturing the outcome but not the dynamics of decision. To address this, participants also performed the SMS experiment, in which they tapped in synchrony with the repeating pattern. Tapping behavior served as a real-time proxy for perceptual object formation and offered continuous behavioral insight into the build-up and stabilization of auditory objects.

Results

We show sigmoidal, quasi-categorical performance across repetition levels in both experiments. In detection, an interaction between unit duration and repetition strength is observed, with shorter durations showing earlier performance improvement. We also observe a strategy shift. When repetition is unclear, participants appear to rely on fixed-duration timing; when repetition becomes clear, they switch to cycle-based timing. Interestingly, the interaction between unit duration and repetition disappears in decision time analysis. Regardless of unit duration, participants wait for the same number of cycles (~ 4) to make a decision.

In the SMS, sigmoidal curves converge across unit durations, eliminating the interaction effect, paralleling the detection time results. Within-trial progression analysis revealed three patterns: (1) tap alignment improves and stabilizes when an object emerges, (2) tap alignment remains flat when no object emerges, and (3) intermediate conditions show gradual tap alignment increases, reflecting incremental refinement of the object representation. The endpoint of the trial-wise analysis clarifies why curves converge: once a stable object representation emerges, it can be tracked equally well across all durations.

Discussion

Accurate detection of repetition depends on repetition strength, which interacts with unit duration, but the decision speed remains unaffected. Once repetition is clearly perceivable, participants require the same cycle number to make a judgment, regardless of how long or information-rich the unit is. Also, this cycle number is in line with pre-existing results in the literature. The SMS experiment reveals a two-stage object formation process: when repetition is detectable, performance gradually builds up before reaching a saturation point, suggesting a categorical perceptual shift in strong repetition conditions, in which the additional evidence no longer enhances performance. In contrast, intermediate conditions show continuous refinement of object representation. The underlying neural mechanisms that govern this transition from time-based to cycle-based timing strategies—and the perceptual shift from gradual accumulation to stabilization—remain open questions for future research.

References

Barascud, N., Pearce, M. T., Griffiths, T. D., Friston, K. J., & Chait, M. (2016). Brain responses in humans reveal ideal observer-like sensitivity to complex acoustic patterns. *Proceedings of the National Academy of Sciences*, *113*(5), E616–E625. https://doi.org/10.1073/pnas.1508523113

Barczak, A., O'Connell, M. N., McGinnis, T., Ross, D., Mowery, T., Falchier, A., & Lakatos, P. (2018). Top-down, contextual entrainment of neuronal oscillations in the auditory thalamocortical circuit. *Proceedings of the National Academy of Sciences*, *115*(32). https://doi.org/10.1073/pnas.1714684115

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press. https://doi.org/10.7551/mitpress/1486.001.0001

Bregman, A. S. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press.

McDermott, J. H., Wrobleski, D., & Oxenham, A. J. (2011). Recovering sound sources from embedded repetition. *Proceedings of the National Academy of Sciences*, *108*(3), 1188–1193. https://doi.org/10.1073/pnas.1004765108

Shadlen, M. N., & Kiani, R. (2013). Decision Making as a Window on Cognition. *Neuron*, *80*(3), 791–806. https://doi.org/10.1016/j.neuron.2013.10.047

Shadlen, M. N., & Newsome, W. T. (1996). Motion perception: Seeing and deciding. *Proceedings of the National Academy of Sciences*, *93*(2), 628–633. https://doi.org/10.1073/pnas.93.2.628

Winkler, I., Denham, S. L., & Nelken, I. (2009). Modeling the auditory scene: Predictive regularity representations and perceptual objects. *Trends in Cognitive Sciences*, *13*(12), 532–540. https://doi.org/10.1016/j.tics.2009.09.003