# Schema learning for the cocktail party problem

Kevin J. P. Woods[a,b] and Josh H. McDermott[a,b,1]

[a]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139; and [b]Program in Speech and Hearing Bioscience and Technology, Division of Medical Sciences, Harvard University, Boston, MA 02115

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved February 21, 2018 (received for review February 2, 2018)

**The cocktail party problem requires listeners to infer individual sound sources from mixtures of sound. The problem can be solved only by leveraging regularities in natural sound sources, but little is known about how such regularities are internalized. We explored whether listeners learn source "schemas"—the abstract structure shared by different occurrences of the same type of sound source—and use them to infer sources from mixtures. We measured the ability of listeners to segregate mixtures of time-varying sources. In each experiment a subset of trials contained schema-based sources generated from a common template by transformations (transposition and time dilation) that introduced acoustic variation but preserved abstract structure. Across several tasks and classes of sound sources, schema-based sources consistently aided source separation, in some cases producing rapid improvements in performance over the first few exposures to a schema. Learning persisted across blocks that did not contain the learned schema, and listeners were able to learn and use multiple schemas simultaneously. No learning was evident when schema were presented in the task-irrelevant (i.e., distractor) source. However, learning from task-relevant stimuli showed signs of being implicit, in that listeners were no more likely to report that sources recurred in experiments containing schema-based sources than in control experiments containing no schema-based sources. The results implicate a mechanism for rapidly internalizing abstract sound structure, facilitating accurate perceptual organization of sound sources that recur in the environment.**

auditory scene analysis | perceptual learning | implicit learning | statistical learning

Sounds produced by different sources sum in the air before entering the ear, requiring the auditory system to infer sound sources of interest from a mixture (the "cocktail party problem") (1–9). Because many different sets of source signals could generate an observed mixture, the problem is inherently ill-posed. In the real world, however, constraints on the generation of sound mean that assumptions can be made about which components of sound energy came from the same source, enabling us to correctly infer source structure much of the time. Understanding human listening abilities thus requires understanding these assumptions and how they are acquired.

Some of the assumptions guiding scene analysis may be rather general. For example, frequency components appearing at integer multiples of a common "fundamental frequency" are usually heard as arising from the same source (10–12), as are sounds that begin and end at the same time (13, 14) and sound patterns that repeat (15–17). These grouping cues are believed to reflect constraints on sound generation that are common across natural sources (1) and thus likely apply across a wide range of sounds and contexts.

Other cues to perceptual organization might apply only to particular contexts. Natural sources often produce sounds that are patterned consistently across occurrences (as in animal vocalizations, spoken words, or sung melodies), resulting in an abstract time-varying structure shared by a subset of sound events. Internalizing this recurring structure might be expected to aid scene analysis, but unlike more generic grouping cues, which could be internalized over evolution or by a general learning process operating on all auditory input, source-specific structure would have to be learned upon the appearance of a new sound source.

Although auditory memory has been argued to have lower capacity than visual memory (18), human listeners clearly acquire rich knowledge of sound structure from listening. Many documented examples fall under the rubric of "statistical learning," in which humans internalize aspects of the sound input distribution, such as transition probabilities between sound elements (19–21) or correlations between sound properties (22). Such learning is thought to be important for both speech (23) and music (24–26) perception. Specific recurring sound structures, typically noise samples, can also be learned (27–29). Such learning is apparently often implicit (24–26, 30).

The ability to learn the structure of sound sources suggests that such knowledge might be used for scene analysis, and source-specific structures used for this purpose are often termed "schemas" in the scene-analysis literature (1, 31–36). A role for learned schemas has been suggested by prior findings that listeners are better able to extract highly familiar voices (e.g., one's spouse) (37), familiar languages (38), well-known melodies (39–42), and words (43). However, because these sources were already familiar to listeners before the experiments, the underlying learning process has remained opaque. Open issues include the rapidity with which schemas can be learned and used in scene analysis, the specificity of the learned representation, whether schemas can be learned in the presence of multiple sources, whether learning is dependent on attention to schema exemplars, and whether listeners must be aware of what they are learning. Also, because prior work has largely been confined to familiar structures in speech and music, it has been unclear if schema learning is a general phenomenon in audition.

The experiments presented here were designed to reveal the process of learning a new schema. Our approach was to have listeners perform source-separation tasks on synthetic stimuli that traversed a pattern over time and to test if performance improved for targets derived from a particular pattern (the schema) that appeared intermittently over the course of the experiment. We employed this general approach in three separate experimental

## Significance

The "cocktail party problem" is encountered when sounds from different sources in the world mix in the air before arriving at the ear, requiring the brain to estimate individual sources from the received mixture. Sounds produced by a given source often exhibit consistencies in structure that might be useful for separating sources if they could be learned. Here we show that listeners rapidly learn the abstract structure shared by sounds from novel sources and use the learned structure to extract these sounds when they appear in mixtures. The involvement of learning and memory in our ability to hear one sound among many opens an avenue to understanding the role of statistical regularities in auditory scene analysis.

Author contributions: K.J.P.W. and J.H.M. designed research; K.J.P.W. performed research; K.J.P.W. analyzed data; and K.J.P.W. and J.H.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: jhm@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1801614115/-/DCSupplemental.

Published online March 21, 2018.

paradigms with different types of stimuli that varied in complexity but never contained familiar source structure. In paradigm 1, listeners discriminated four-tone melodies presented concurrently with "distractor" tones, or comparable stimuli composed of noise bursts (Fig. 1). In paradigm 2, listeners had to attentively track one of two concurrent sources that changed stochastically over time in pitch and the first two formants (spectral peaks that determine vowel quality) (Figs. 2–4). In paradigm 3, pitch and formant contours were extracted from recorded speech and resynthesized to produce a stimulus that contained the pitch and formant contours of an actual speech utterance but that was not intelligible. Listeners heard a mixture of two such utterances followed by a probe utterance and were asked if the probe utterance was contained in the mixture (Fig. 5). Audio demonstrations of all stimuli can be found online at mcdermottlab.mit.edu/schema_learning/.

In each of these experimental paradigms, sources generated from a common schema recurred over the course of the testing session. These schema-based sources never appeared on consecutive trials and were transformed each time to avoid exact replication and to mimic the variation that occurs in real-world sound sources. We then compared performance for sources derived from a common schema with that for sources derived from schemas that did not recur during an experiment. The results show that rapidly acquired memories contribute substantially to source separation.

## Results

**Paradigm 1. Detection of Discrete-Tone Melodies.** To explore schema learning with a relatively simple stimulus, we presented listeners with a six-tone "mixture" composed of a four-note melody with two additional distractor tones, followed by a four-tone "probe" melody in isolation (Fig. 1*A*). Listeners were asked if the isolated probe melody was contained in the mixture that preceded it. The probe melody was always transposed away from the melody in the mixture by up to an octave, and listeners were told that this would be the case. The transposition required listeners to extract the structure of the melody and prevented them from performing the task based on glimpsed features (e.g., note fragments) of the mixture. On trials where the correct response was "no" (the probe was not contained in the mixture), the probe was altered by changing the middle two tones, with the first and last notes of the melody retaining the relative positions they had in the mixture. As a consequence, the task could not be performed based on these outer tones alone. The tone onsets and durations in the probe melody were unaltered on these "foil" trials so that the task also could not be performed by recognizing temporal patterns alone. Because we wanted to explore the learning of novel structure, melodies were not confined to a musical scale or metrical grid; pitch and timing values were drawn from continuous uniform distributions so that there was no conventional musical structure.

A schema-based melody appeared in the mixture on every other trial (Fig. 1*B*) and on half of those trials also appeared as the four-tone probe. Although the recurring schema could thus occur in isolation (as the probe), the alternating-trials design meant that a schema-based probe never immediately preceded a mixture containing that schema, preventing immediate priming. The non–schema-based trials for each participant consisted of trials drawn randomly from the schema-based sets for other participants (one from each of the other sets), so that schema- and non–schema-based stimuli were statistically identical when pooled across participants. As a consequence, any difference in performance between schema- and non–schema-based trials must reflect learning of the schema.

Because pilot experiments indicated that learning effects might be rapid, it seemed desirable to run large numbers of participants on relatively short experiments. The number of participants required was beyond our capacity to run in the laboratory, and so we instead recruited and ran participants online using Amazon's Mechanical Turk service. To mitigate concerns about sound quality, we administered a headphone-screening procedure to detect participants disregarding our instructions to wear headphones (44). Evidence that the quality of data obtained online can be comparable to that from the laboratory was obtained by comparing performance between online and in-laboratory participants, described below (experiment S2).

***Schema learning of melodies (experiment 1).*** Listeners performed 100 trials of this task (taking ~10–15 min to complete). If exposure to a recurring melodic structure can help listeners detect it when it is embedded among distractors, then we might expect performance on schema-based trials to exceed that on non–schema-based trials over the course of the experiment.

As shown in Fig. 1*C*, performance improved over time for both schema- and non–schema-based sources [$F(1,159) = 5.06$, $P = 0.026$] but was better for schema- than for non–schema-based trials [$F(1,159) = 7.69$ $P = 0.0062$]. Because the schema- and non–schema-based stimuli were statistically identical when pooled across participants, the performance benefit for schema-based trials indicates that participants learned and applied the structure of the schema. There was no interaction between trial type and time [$F(1,159) = 0.97$ $P = 0.32$], perhaps because the learning of the schema was relatively rapid (we return to this issue in paradigm 2). However, the difference between schema and non-schema performance was significant in the second half of the session [$t(159) = 2.61$, $P = 0.0098$] but not during the first half of the session [$t(159) = 1.47$, $P = 0.14$]. The above-chance performance even in the absence of a schema indicates some ability to match the probe stimulus to the mixture despite the transposition. However, this ability is augmented by the acquired knowledge of the tones that are likely to belong together (the schema).

Because schema learning should, at a minimum, produce an improvement in performance late in the experiment, in most subsequent experiments we test for learning by comparing performance between conditions in the second half of trials within an experiment. Figures accordingly plot results binned into a small number of time bins, typically two (to maximize power).

***Schema learning is persistent (experiment 2).*** To test whether knowledge of the schema could be retained over time, we conducted a second experiment in which exposure to a schema was followed by a middle block in which the schema-based melody was totally absent, which in turn was followed by blocks featuring the original schema or a new schema (Fig. 1*D*). The interrupting block contained 20 trials (~3 min). When melodies based on the original schema returned in the final block, they showed a performance benefit compared with the new schema [$t(189) = 3.49$, $P < 0.001$]. This benefit suggests that effects of exposure early in the experiment persisted across the interrupting middle block.

***Multiple schemas can be learned simultaneously (experiment 3).*** The persistence of a learned schema might allow multiple schemas to be learned and used concurrently. To examine this possibility, we conducted a third experiment in which two different schemas alternated, again interspersed with non-schema trials (each schema thus appeared on every fourth trial). The experiment was lengthened to 200 trials to present each schema 50 times, as before. As shown in Fig. 1*E*, the results suggest a learning effect for each schema similar to what we saw in the previous experiments [pooled schema-based trials vs. non–schema-based trials in second half of experiment; $t(87) = 2.98$, $P < 0.005$]. These results suggest that multiple schemas can be learned and used at the same time.

***Schema learning occurs without isolated exposure to the schema (experiment S1).*** To test whether the learning effects were dependent on exposure to the schema in isolation via its presence in the probe stimulus, we conducted an experiment which did not present isolated probes. Instead, listeners were presented with two mixtures and judged whether they contained the same melody. Performance was

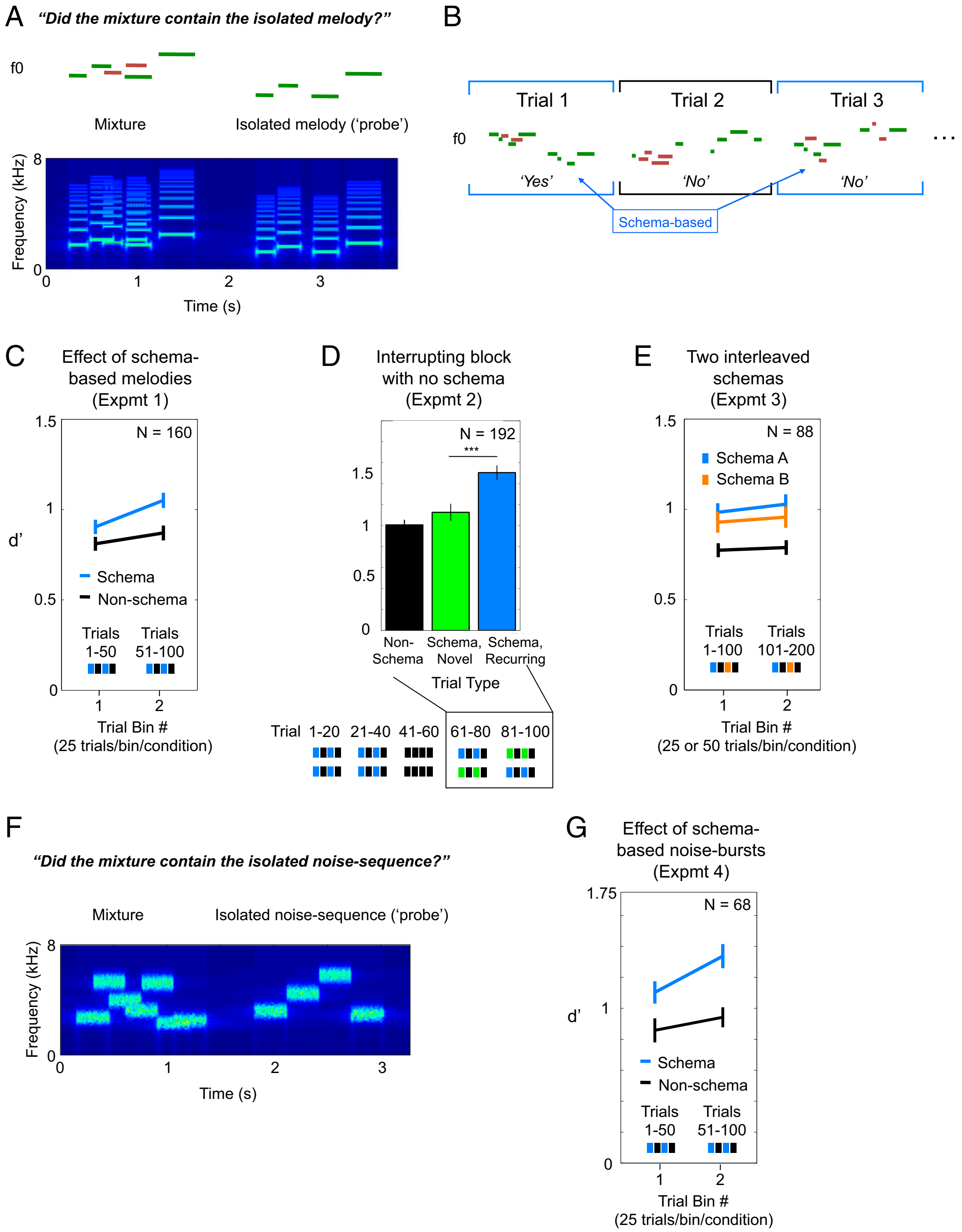

**Fig. 1.** Schema learning in melody segregation (paradigm 1). (*A*) Schematic of the trial structure (*Upper*) and a spectrogram of a sample stimulus (*Lower*). A target melody (green line segments) was presented concurrently with two distractor notes (red line segments), followed by a probe melody (green line segments). Listeners judged whether the probe melody matched the target melody in the mixture. The probe melody was transposed up or down in pitch by a random amount. (*B*) Schematic of the basic experiment structure. On every other trial the target melody was generated from a common schema. On schema-based trials, the melody in the mixture was drawn from the schema 50% of the time, while the probe was always drawn from the schema. (*C*) Results of experiment 1: recognition of melodies amid distractor tones with and without schemas ($n$ = 160). Error bars throughout this figure denote the SEM. (*D*) Results of experiment 2: effect of an intervening trial block on learned schema ($n$ = 192). Listeners were exposed to a schema, then completed a block without the schema, and then completed two additional blocks, one containing the original schema and one containing a new schema. The order of the two blocks was counterbalanced across participants. (*Lower*) The two rows of the schematic depict the two possible block orders. (*Upper*) The data plotted are from the last two blocks. (*E*) Results of experiment 3: effect of multiple interleaved schemas ($n$ = 88). Results are plotted separately for the two schemas used for each participant, resulting in 25 and 50 trials per bin for the schema and non-schema conditions, respectively. (*F*) Spectrogram of a sample stimulus from experiment 4. Stimulus and task were analogous to those of experiment 1, except that noise bursts were used instead of tones. (*G*) Results of experiment 4: recognition of noise-burst sequences amid distractor bursts, with and without schemas ($n$ = 68).

low overall for this experiment, presumably because there were twice as many opportunities to make streaming errors (Fig. S1). However, the schema learning effect persisted, with better performance for schema-based than for non–schema-based trials in the second block [$t\ (39) = 3.10, P < 0.005$]. Listeners thus appear to be able to detect and learn recurring structure even when it does not occur in isolation.

***Schema learning occurs for atypical sound sources (experiment 4).*** To test whether comparable phenomena would occur for sound sources that were even less typical of musical sources, we conducted an analogous experiment with sequences of noise bursts. Unlike the tones, the noise bursts were aperiodic and lacked a pitch in the usual sense but nonetheless instantiated patterns of frequency variation that were recognizable to human listeners (45). As shown in Fig. 1*G*, listeners again performed better for stimuli generated from a common schema [second block; $t(67) = 3.56$, $P < 0.001$]. It thus appears that there is some generality to the ability to learn recurring patterns and use this knowledge to improve the extraction of those patterns from mixtures with other sounds.

**Paradigm 2. Attentive Tracking of Smooth Pitch-Formant Contours.** To further explore the generality of this phenomenon, we next turned to a task and stimulus we had originally developed to study auditory attentive tracking (46). Synthetic voices were generated that varied in several speech-relevant feature dimensions (pitch and the first two formants: f0, F1, F2) according to independent, randomly generated trajectories. Listeners were presented with mixtures of two such time-varying sources and were cued beforehand to attend to one of them (with the starting portion of one voice). We measured listeners' ability to track this cued voice by subsequently presenting them with the tail end of one of the voices; their task was to judge whether this probe segment belonged to the cued voice (Fig. 2*A*).

Trajectories in a mixture were required to cross each other at least once in each feature dimension, so listeners could not perform the task simply by attending to a high or low value of one of the features. Although the trajectories of each source were continuous, task performance is not critically dependent on complete continuity, as it is robust to the insertion of intermittent gaps in the stimuli (46). Moreover, despite the continuity, the task is effortful for human listeners, and success depends in part on accurate tracking of the cued voice as it varies throughout the mixture (46). One other difference between paradigms 1 and 2 was that the probe in paradigm 2 consisted only of the ending portion of a source (unlike paradigm 1, in which the probe had the same form as the target melody). As a result, listeners never experienced a source trajectory in the absence of a concurrent source, providing another test of whether schemas can be learned and used even when sources never occur in isolation.

Given that we ran experiments online to obtain sufficient sample sizes, it is natural to wonder whether data quality was comparable to that in experiments run in the laboratory. To address this issue, we compared performance on this attentive tracking task between online and in-laboratory participants (experiment S1). We chose to perform this comparison for the attentive tracking paradigm because it seemed most likely to suffer in a subject pool that was less motivated, as might be expected in an online environment. However, we found that performance was similar between online and in-laboratory participants once online participants were screened for headphone use (Fig. S2). This result gave us some confidence in our online testing procedures.

***Schema learning extends to pitch-formant contours (experiment 5).*** To first test for the basic schema learning effect with this task, we ran an experiment in which the cued voice on every other trial was derived from a common schema trajectory (Fig. 2*B*). These schema-based sources were not exact replicas of each other but were related by time dilation and transposition, as might occur in natural sound sources, such as prosodic patterns in speech. Trials in which the target was not schema-based had targets drawn from the sets of schema-based targets presented to other participants (*Methods*), so that when pooled across subjects the distribution of schema-based and non–schema-based targets was identical. To better explore the time course of any learning effect, we ran a longer experiment than we did for paradigm 1 (168 trials; ∼35–40 min, which we divided into two time bins for analysis with maximum power but also plot in six bins to provide a sense of the dynamics over time).

Performance over the course of the experiment is shown in Fig. 2*C*. Overall task performance was again significantly better for trials whose targets were based on a common schema [main effect of trial type, $F(1,85) = 16.2$, $P = 0.0001$; repeated-measures ANOVA]. Moreover, although there was a general improvement over the course of the experiment [$F(1,85) = 10.1$, $P = 0.002$], performance on schema-based target trials improved more than did performance for regular trials, yielding a significant interaction of trial type and time [$F(1,85) = 12.5$, $P = 0.0007$; repeated-measures ANOVA], again driven by a performance difference in the second half of trials [$t(85) = 5.14$, $P < 10^{-5}$]. These results suggest that performance can be facilitated by recurring structure even when sources vary in multiple dimensions and never appear in isolation.

***Learning effect is not explained by cues and probes (experiment 6).*** One potential explanation for the learning effects in this task is that listeners learn something about the relationship between the cues and probes for the schema-based trajectories rather than from the trajectory itself. Although time dilation of schema-based trajectories resulted in the associated cues and probes not having a fully consistent relation to one another, we nonetheless guarded against any such strategy when generating stimuli by matching the distributions of distances between the cues and the correct and incorrect probes (*Methods*). To ensure that these measures indeed prevented listeners from performing the task with the cue and probe alone, we ran a second experiment in which the mixture on the last one-third of trials was replaced by noise (Fig. 2*D*). In a pilot experiment we found that when the relationship between the cues, correct probes, and incorrect probes was not controlled, performance remained significantly above chance during the noise block for both types of trials [non-schema trials: $t(87) = 4.41$, $P < 0.000$; schema trials: $t(87) = 3.22$, $P < 0.0018$], demonstrating the effectiveness of this control experiment (and the necessity of controlling the stimuli).

Replicating experiment 5, superior performance for schema-based trajectories was apparent over the first two-thirds of the experiment [$F(1,132) = 8.35$, $P = 0.005$] (Fig. 2*D*). However, performance fell to chance levels once the mixtures were replaced by noise [$t$ tests vs. chance: schema-based trajectories, $t(132) = 0.71$, $P = 0.48$; non–schema-based trajectories, $t(132) = 0.06$, $P = 0.96$], with no difference in performance between schema-based and non–schema-based trajectories [$t(132) = 0.46$, $P = 0.65$]. It thus appears that listeners cannot perform the attentive tracking task based on the cue and probe alone, and that the benefit of schema-based trajectories is not due to learning cue–probe relationships for these trajectories.

***Rapid learning evident via crowdsourcing (experiment 7).*** In experiments 5 and 6, the schema-based sources seemed to have elicited different levels of performance from the outset of the experiment (Fig. 2 *C* and *D*). Because the balancing of the stimulus sets was intended to prevent an intrinsic difference in difficulty between conditions, we considered the possibility that learning might be occurring on a fast timescale. To test this, we pooled the data from experiments 4 and 5 (these experiments were identical for the first 56 trials) and examined performance over smaller bins of seven trials per bin rather than 14. A power analysis indicated that additional participants would be required to discover possible effects at this timescale, and so an additional 183 participants were run on a shorter, 56-trial version of the attentive tracking paradigm (experiment 7).

With a resolution of seven trials per bin, it is evident that performance at the outset of the experiment did not differ

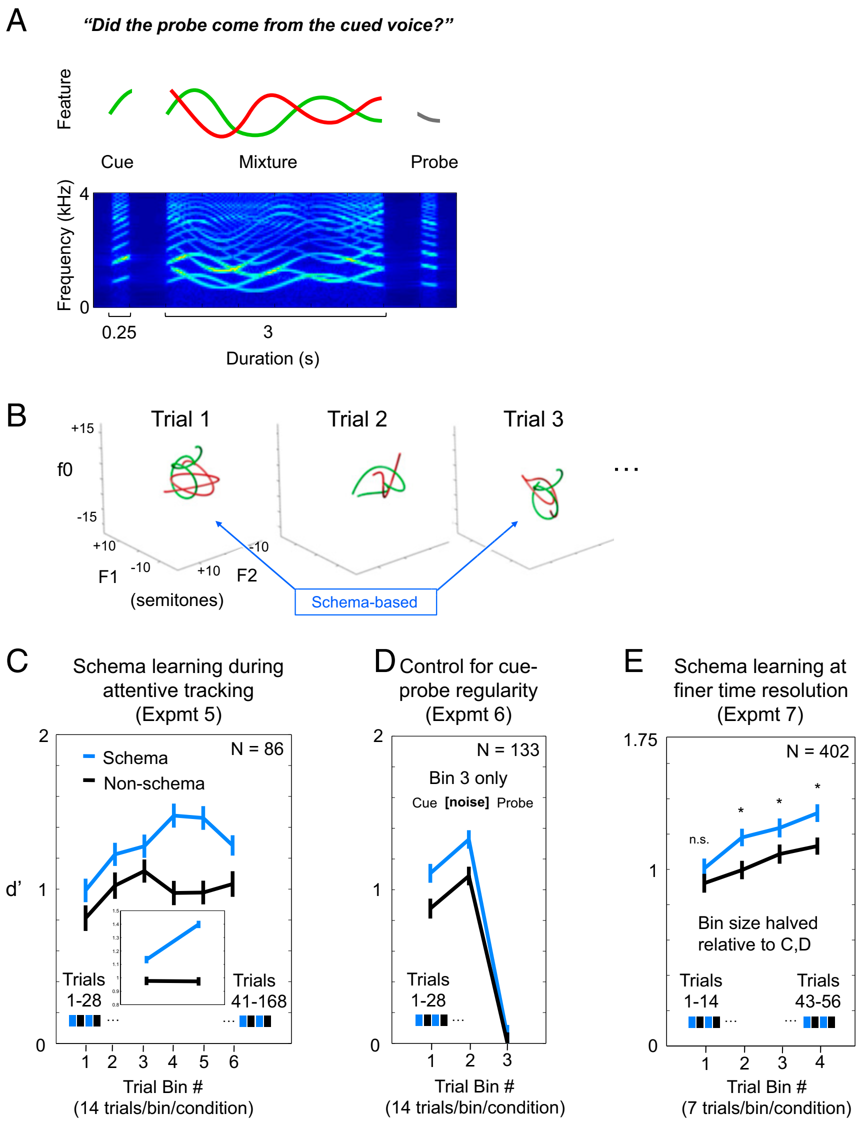

**Fig. 2.** Schema learning in attentive tracking of synthetic voices (paradigm 2). (*A*) Schematic of the trial structure (*Upper*) and spectrogram of an example stimulus (*Lower*). A target voice (green curve) was presented concurrently with a distractor voice (red curve). Both voices varied smoothly but stochastically over time in three feature dimensions: f0, F1, and F2 (the fundamental frequency and first two formants; for clarity the schematic only shows variation in a single dimension). Voices in a mixture were constrained to cross at least once in each dimension. Listeners were cued beforehand with the initial portion of the target voice. Following the mixture, listeners were presented with a probe stimulus that was the ending portion of one of the voices and judged whether this probe came from the target. (*B*) Schematic of the experiment structure. On every other trial the target voice was generated from a common schema. Voices are depicted in three dimensions. f0, F1, and F2 are plotted in semitones relative to 200, 500, and 1,500 Hz, respectively. (*C*) Results of experiment 5: effect of schemas on attentive tracking ($n = 86$). The *Inset* denotes results with trials binned into 42 trials per condition to maximize power for an interaction test (reported in text). Error bars throughout this figure denote the SEM. (*D*) Results of experiment 6: a control experiment to ensure listeners could not perform the task with cues and probes alone ($n = 146$). In the last one-third of trials, the voice mixture was replaced with noise. (*E*) Schema learning on a finer time scale ($n = 402$). Data from the first 56 trials of experiments 5 and 6 were combined with new data from experiment 7 and replotted with seven trials per bin. The finer binning reveals similar performance at the experiment's outset, as expected. n.s., not significant. *$P < 0.05$.

between conditions [first time bin: $t(401) = 0.92$, $P = 0.36$] (Fig. 2*E*) but that performance differences emerge quickly (significant differences for all other time bins: $P < 0.05$ in all cases). The ability to observe this rapid learning was facilitated by the fact that the experiments were run online, which allowed us to efficiently test a relatively large number of listeners ($n = 402$). This observation provides some confirmation that the stimuli in the schema-based and non–schema-based conditions do not differ in their intrinsic difficulty; it is only the presence of other schema-based stimuli that boosts performance. The results are suggestive of a learning effect occurring over relatively small numbers of exposures.

***Schema learning need not be explicit.*** The presence of a learning effect raises the question of whether participants are aware of what they are learning. To address this issue, after finishing the task, participants were asked if they had noticed repetition in the cued voice. The proportion of "yes" responses from experiment 5 (the longest experiment run with this paradigm) is shown in Fig. 3 along with responses from two control experiments: one that contained no schema-based sources (control experiment 1) and another in which every cued voice was schema-based (control experiment 2, in which we expected participants to notice the recurring source structure). Participants did not report repetition in experiment 5 any more often than in the control experiment with no schema-based sources (z = 0.98, $P = 0.33$). This result suggests that the general similarity across stimuli caused listeners to report hearing recurring structure whether or not it was actually there. However, when the schema was made obvious by inserting it into every trial of the experiment (control experiment 2), listeners reported recurring structure more than in the other two experiments (experiment 5: z = 2.04, $P$ = 0.021; control experiment 1: z = 2.68, $P$ = 0.0037). Overall, these results suggest that in our main experiment listeners were not fully aware of the schema-based similarities that facilitated performance, indicating that schema learning need not occur explicitly.

***Multiple feature dimensions are concurrently learned (experiment 8).*** The use of three feature dimensions (f0, F1, and F2) raises the question of whether the schema that is learned by listeners incorporates all three available dimensions. To address this issue, we conducted an experiment that was identical to experiment 5 except that the formant trajectories of the schema-based stimuli were randomized from trial to trial starting halfway through the experiment (Fig. 4*A*). If the schema that is learned is mainly a function of the f0 trajectory, listeners should continue to show a schema benefit in the second half of the experiment. If instead all three feature dimensions are important to the schema, the learning effect might weaken or disappear altogether once the formant trajectories are randomized.

As shown in Fig. 4*B*, a schema benefit was obtained in the first half of the experiment [$t(85) = 2.00$, $P = 0.049$] but not in the second half of the experiment [$t(85) = 0.63$, $P = 0.53$]. Moreover, when the data from the second half of experiment 5 (in which schema-based formants continued throughout) was compared with that of the present experiment, there was a significant interaction between the experiment and the effect of trial type [$F(1,170) = 13.0$, $P < 0.0005$; data pooled across bins 4–6]. This result suggests that listeners learned the recurring structure in the formants in addition to the f0 and that schemas are thus not limited to pitch variation.

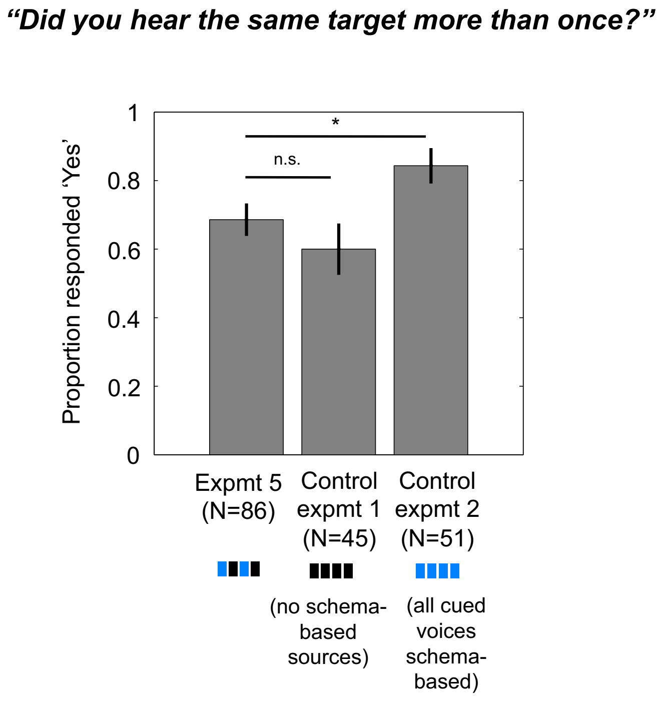

**Fig. 3.** Evidence for implicit learning. Following experiment 5 and separate control experiments, participants were asked if they had noticed a recurring structure in the cued voice. In experiment 5, schema-based sources occurred on every other trial. For comparison, control experiment 1 contained no schema-based sources, while control experiment 2 contained schema-based sources on every trial. Error bars denote the SEM, derived from bootstrap.

***Generalization from the learned schema (experiment 9).*** To further explore the nature of the learned schema, we investigated whether the schema benefit varied with the similarity of a sound to the schema's central tendency, taking advantage of the stimulus variation inherent to the experimental design. We ran an experiment otherwise like experiment 5 but with twice as many participants, allowing analyses of subsets of trials based on similarity to the schema. To analyze non-schema trials, we evaluated similarity as the mean of the correlation in each dimension between the canonical schema (i.e., the one at the center of the dilation/transposition range) and the non-schema target. To analyze schema trials, we considered variants in the middle of the dilation range separately from those at the extremes of this range. In both cases we tested whether there was a performance benefit for the one-third of targets most similar to the canonical schema compared with the one-third of targets least similar to the schema.

As shown in Fig. 4*C*, the non-schema target stimuli that were most similar to the schema produced similar performance to the schema-based target stimuli [$t(171) = 1.28$, $P = 0.2$, with data pooled over both time bins], unlike the least similar targets [$t(171) = 2.95$, $P = 0.004$]. Moreover, the two groups of non-schema targets produced different performance levels [$t(171) = 2.95$, $P = 0.0037$ in the first of the two time bins, with a nonsignificant trend when pooled across both bins: $t(171) = 1.67$, $P = 0.098$]. Due to the design of the experiment (in which schema trials alternated with non-schema trials), this difference could have been driven by immediate priming by the preceding stimulus rather than the learned schema. However, analysis of the schema-based stimuli revealed a similar result: The more canonical schema-based stimuli (the middle of the range of time-warped variants) produced better performance than the schema variants that were most dilated/compressed [$t(171) = 1.96$, $P = 0.051$; data pooled across both time bins] (Fig. 4*D*). In addition, the most compressed/dilated schema-based stimuli produced better performance than the non-schema stimuli [$t1(171) = 1.98$, $P = 0.049$]. As expected given the results of experiment 8, both pitch and formants contributed to the schema-similarity boost; the same analysis on f0, F1, and F2 separately did not reveal significant effects, either in the pooled data or in any single time bin ($P > 0.2$ in all cases). The results suggest a graded effect of the schema, perhaps because listeners learn and represent the underlying stimulus distribution to some extent.

**Paradigm 3. Segregation of Speech-Like Utterances.** The stimuli used in paradigms 1 and 2 (discrete tone or noise patterns and continuous voice-contours) are arguably simplistic compared with natural stimuli such as speech. To test for schema learning in a more naturalistic stimulus, we extracted pitch and formant contours from utterances in a large speech corpus and resynthesized them with the same method used to synthesize the smooth stochastic trajectories in paradigm 2. We reproduced only the voiced regions of speech and left silent gaps in place of consonants, so that the resulting stimulus was not intelligible but otherwise traversed the space of pitch and formants in a realistic manner. Listeners were presented with a mixture of two utterances followed by an isolated probe utterance and had to decide if this isolated utterance had also appeared in the mixture (Fig. 5*A*). The probe utterance not only was transposed away from the mixture (as in paradigm 1) but also was time-warped (i.e., compressed or dilated) by up to ±25% (*Methods*). The transposition and dilation again required listeners to extract the structure of the utterance and prevented them from performing the task by recognizing local features glimpsed during the mixture.

***Schema learning of feature trajectories from real speech (experiment 10).*** Would the schema-learning effects in the first two paradigms also manifest with this more realistic stimulus? We conducted a 100-trial experiment (~10–15 min) in which schema-based targets appeared on alternate trials (Fig. 5*B*). As with paradigms 1 and 2, the non–schema-based stimuli for a participant were drawn from the schema-based trials for other participants, so that schema- and

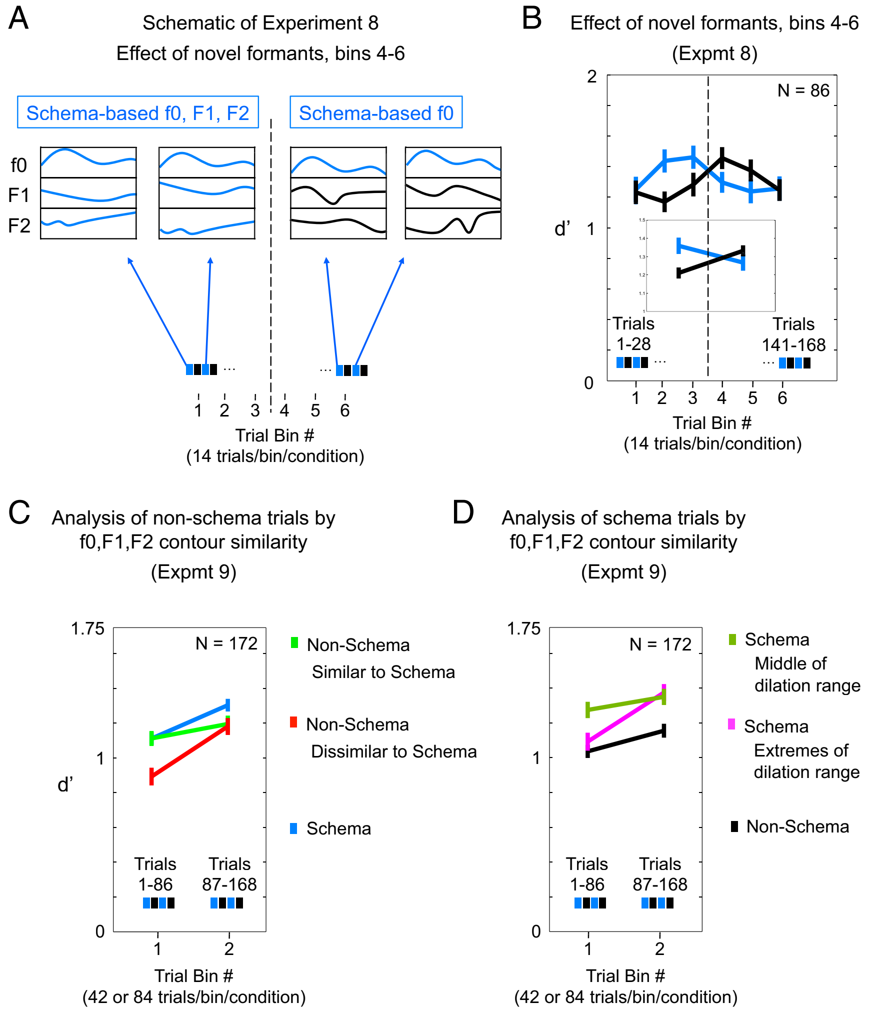

**Fig. 4.** Dependence of schema benefit on multiple dimensions and similarity to the schema. (*A*) Schematic of the structure of experiment 8. On each trial, listeners were cued to a target voice, heard a target-distractor voice pair, and judged if a subsequent probe was from the end of the target or the distractor (paradigm 2). Schema-based trials alternated with non–schema-based trials, but the formant trajectories on schema-based trials were randomized halfway through the experiment. (*B*) Results of experiment 8: effect of multiple dimensions on schema learning ($n$ = 86). The *Inset* denotes results with trials binned into 42 trials per condition to maximize power. Error bars throughout this figure denote the SEM. (*C*) Results of subdividing non-schema trials from experiment 9 ($n$ = 146). Performance was computed separately for non-schema trials whose feature trajectories were most and least correlated with those of the schema. (*D*) Results of subdividing schema trials from experiment 9. Performance was computed separately for schema trials in the middle and extremes of the dilation/transposition range.

non–schema-based trials were statistically identical when pooled across participants.

As shown in Fig. 5*C*, a benefit was rapidly obtained from the recurring schema, with performance on schema-based trials again exceeding non–schema-based trials during the second half of the experiment [$t(92) = 2.94$, $P = 0.0041$]. These results demonstrate rapid schema learning for natural feature trajectories that have more complex generative constraints than the stimuli used in paradigms 1 and 2, raising the possibility that fast and flexible schema learning could help us hear behaviorally relevant sources in the real world.

***Schema-based distractors (experiment 11).*** It seemed of interest to test effects on performance when schema-based sources appeared as the distractor instead of the target. However, paradigms 1 and 2 did not provide a clear means to study this: In paradigm 1 the distractors were not of the same form as the target melodies (being a pair of two tones rather than a four-note sequence), while in paradigm 2 the need to control cue–probe relationships made it methodologically challenging to implement experiments with similar distractors. The mixture–probe task of paradigm 3 was well suited to address this question, so we conducted an additional experiment in which the nontarget utterance in the mixture (on alternate trials) was generated from a common schema (Fig. 5*D*). This experiment was run for 200 trials rather than 100 trials as in experiment 7, providing the listener with considerable exposure to the schema by the end of the experiment. The results (Fig. 5*E*) show that the schema-based distractor nonetheless had no detectable effect on performance [$F(1,201) = 1.85$, $P = 0.18$; no significant difference for any time bin], providing evidence that a recurring schema is less likely to be internalized if it does not occur in the attended source.

## Discussion

Sources in auditory scenes often produce sound with some degree of consistency. We explored the conditions in which this consistency might be learned and used to guide scene analysis. We tested if listeners would obtain a source-separation benefit from the recurrence of a source under transformations such as transposition and time dilation, which produce acoustically distinct variants that share abstract structure. Such a benefit would

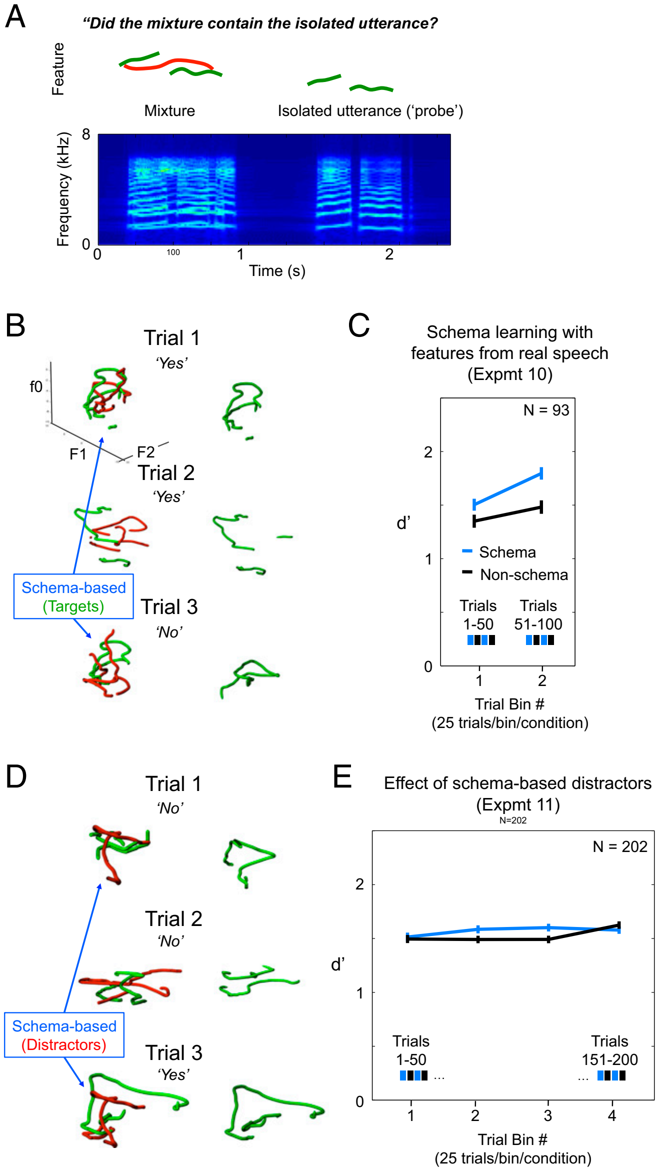

**Fig. 5.** Schema learning in the segregation of resynthesized speech utterances (paradigm 3). (*A*) Schematic of the trial structure (*Upper*) and a spectrogram of an example stimulus from experiments 10 and 11 (*Lower*). A target utterance (green curve) was presented concurrently with a distractor utterance (red curve) and was followed by a probe utterance (second green curve). The utterances were synthesized from the pitch and formant contours of speech excerpts. For clarity the schematic only shows variation in a single dimension. Because only the first two formants were used, and because unvoiced speech segments were replaced with silence, the utterances were unintelligible. Listeners judged whether the probe utterance had also appeared in the mixture. When this was the case, the probe utterance was transposed in pitch and formants from the target utterance in the mixture and was time-dilated or compressed. (*B*) Schematic of the structure of experiment 10. On every other trial the target utterance was generated from a common schema. Utterances are depicted in three dimensions. (*C*) Results of experiment 10: effect of schemas on the segregation of speech-like utterances ($n = 89$). Error bars here and in *E* denote the SEM. (*D*) Schematic of structure of experiment 11. On every other trial, the distractor utterance was generated from a common schema. (*E*) Results of experiment 11: effect of schema-based distractors ($n = 202$).

imply learning of the abstract structure of the recurring source (a schema). Using several types of stimuli and tasks, we found that source separation benefitted from the presence of a schema and that the learning of the schema was rapid, persistent, and could occur without listeners noticing. Schema learning occurred even when sources never appeared in isolation (experiments S1 and 5–9). We also found that learning was at least somewhat specific to attended sources: No benefit was obtained when the schema appeared in the distractor source. The benefit of the learned schema generalized to some extent to sounds sufficiently similar to the schema, with a greater benefit for sounds most similar to the schema. Overall, our results demonstrate that abstract source structure can be rapidly internalized to facilitate accurate segregation, demonstrating one way recurring structure in the auditory environment could aid the separation and selection of sound sources from auditory scenes.

**Prior Evidence for Schemas in Scene Analysis.** Prior work showing that source-separation tasks are aided by memory typically used highly familiar sound sources to which listeners have considerable prior exposure, such as familiar voices (37), melodies (39–41), and linguistic content (43). However, because these sources were already familiar to participants, the learning process could not be studied. Our results reveal schema learning in action and place constraints on the underlying mechanisms.

To our knowledge, the only prior evidence of abstract structure benefitting source separation for previously unfamiliar sources came from a study using randomly generated melodies (31). In that study, listeners judged if two target melodies were the same or different; one of the melodies was interleaved with a distractor melody, forming a mixture, and was transposed to a different pitch range (so that discrimination required comparing pitch intervals). Performance was better when the mixture appeared second (preceded by the isolated melody), suggesting that segregation could be guided by abstract melodic structure when primed immediately beforehand. However, in this study melodies did not recur across trials, and so the effect of recurring structure was evident only within a trial. Our results are distinct from these effects in showing that schemas can be learned if they recur intermittently over the course of an experiment and that their effects are persistent over time, do not seem to require explicit knowledge of the target, and need not reflect immediate priming. We also find that schemas extend to nonmelodic stimuli.

**Relation to Other Examples of Learning in Audition.** A number of studies have demonstrated auditory learning of unfamiliar sounds such as noise samples or tone patterns over the course of an experimental session (28, 29) or of statistical properties of the input (19–26). Practical effects of such learning may be found in adaptation to unfamiliar speakers (47) or foreign accents (48). However, these studies presented stimuli in isolation and did not assess the effect of learning on source separation. Before our work, it was thus unclear whether source structure could be learned in the presence of concurrent sources and whether any such learning would benefit scene analysis. Our work used unfamiliar stimuli in combination with sound-segregation tasks to show that rapid learning aids source separation.

The learning evident in our experiments bridges two lines of prior learning research. As in experiments where specific noise samples are learned (27, 28), our learning effects had some specificity to the learned source, producing a performance advantage from schema-based over non–schema-based stimuli, even though they were similarly generated and shared many of the same statistical properties. However, as in statistical learning paradigms, listeners appear to learn aspects of the distribution associated with the variants of a schema (19–26), with a greater benefit for stimuli most similar to the central tendency of the schema.

Previously described examples of schema-based source separation have been thought to rely on explicit knowledge, as when we follow a familiar melody in music (1, 40). However, like many other examples of perceptual learning (24–26, 30), our effects appear to be largely implicit in at least some of the experiments.

As in some instances of visual implicit learning (49), learning nonetheless appears to be somewhat limited to task-relevant stimuli. We thus demonstrate that source structure can be learned amid concurrent sources but perhaps only when attention is directed to it.

**Relation to Effects of Short-Term Source Repetition.** Another relevant line of previous work involves effects of sources that exhibit regular or repeating structure. For example, cyclically repeating patterns in ambiguous streams of tones are known to group over time to form a single auditory stream (16). Repetition also causes random noise sources to be segregated from mixtures (17). These phenomena are distinct from those that we studied here in that the recurring structure is exact, occurs on a short timescale, causes the repeating elements to group together, and shows no evidence of learning (i.e., retention over periods where the source is not present). That said, one can envision a continuum between the conditions of these previous studies (back-to-back and exact repetition) and those of the present study (abstract recurrence across intervening stimuli), and it remains to be seen whether there is any relation between the underlying mechanisms.

**What is Learned?** Our listeners evidently learned something about the recurring structure in each experiment. Because sources were transposed and time-dilated/compressed over the course of learning, the recurring schemas were not individuated by particular feature values. Our results suggest that listeners instead learned something about the way the source's features changed over time. Experiment 8 demonstrated that the learned schema can incorporate variation in formants as well as pitch, since the performance benefit for schema-based stimuli was eliminated when formant trajectories were randomized in schema-based trials. Experiment 9 showed that the learned schema provided a benefit to non-schema targets that were sufficiently similar to the schema (provided the similarity metric included both pitch and formants) and an added benefit to schema targets in the middle of the range of possible time-dilated variants. Overall, the results indicate that learning can occur over a range of task-relevant features and that the effect of the schema is graded, providing the greatest benefit to stimuli most similar to the canonical schema.

The recurring structures in our experiments were abstract, since the schema always appeared transposed or dilated/compressed to varying degrees (transformations inspired by the variation that occurs in speech and music). It would be interesting to further explore the nature of the learned representation by testing the transfer of learning across different transformations (e.g., time reversal or rotation of trajectories in feature-space) and to explore limits to the types of abstract structure that can be learned (e.g., by exposing listeners to different types of source transformations during learning). There are also likely limits to the contexts in which they can be utilized. For example, listeners are known to have difficulty detecting even a highly familiar melody if it is perfectly interleaved with a set of distractor tones (31, 39). Understanding how schemas interact with other constraints on source separation is thus another important topic for future research.

**Primitive vs. Schema-Based Scene Analysis.** Schema-based scene analysis in audition has historically been contrasted with "primitive" scene analysis, in which putatively universal grouping cues are mediated by processes that do not require attention or memory (1, 8, 35, 50). For example, sequential sounds that are similar (e.g., in pitch or timbre) often arise from the same source in the world and tend to group perceptually over time (7, 35, 51). However, because schema-based scene analysis has not been studied extensively in the laboratory, little is known about the underlying mechanisms, and the extent to which they are distinct from those of primitive scene analysis has been unclear. The methodology introduced here should enable future progress on these issues.

It is possible that the schemas that are learned in our experiments affect perception in much the same way as putatively primitive grouping cues (e.g., pitch differences between talkers). This notion could be tested by comparing the neural or behavioral consequences of schema-driven segregation with those of segregation via other cues (e.g., pitch differences). For instance, it could be diagnostic to examine whether a learned schema affects the ability to discriminate the temporal relationship between elements of the schema and another concurrent source, which is often taken as a signature consequence of streaming (43, 52, 53).

The effect of the learned schema may also be to alter the interaction of streaming and attention. It could be that a learned schema makes it easier to attend to a source conforming to the schema, explaining the better performance on our tracking task, for instance. Alternatively, if memory is complementary to attention, then schema learning might serve to reduce the attentional resources that would otherwise be required to segregate a recurring source from others. These possibilities could be disentangled by measuring attentional selection [for instance, by asking listeners to detect perturbations to sources (46)] before and after schema learning.

However, should the tasks we used even be considered to involve streaming? All the stimuli involve discriminating sound sources embedded in mixtures with other sounds, but the stimuli were relatively short. As such, they are distinct from the long sequences of alternating tones commonly used to study streaming (35, 51). Such stimuli notably exhibit a "buildup of streaming" in which the likelihood of hearing two streams increases with time (3, 53, 54). Although the stimuli we used do not evoke this particular phenomenon, they nonetheless require sound energy to be grouped over time. As such, we conceive them as involving streaming in a more general sense of the term and view them as useful for understanding real-world scenarios in which sources do not repeat cyclically ad infinitum.

The rapidity of the learning effects shown here also raise the possibility that learning could influence all aspects of scene analysis, even those that are quite general in their applicability. Even evidence that newborns exhibit aspects of similarity-based streaming (55, 56) is consistent with learning from early experience. The difference between primitive and schema-based processes might thus be better described in terms of the scale and scope of learning: Primitive scene analysis could effectively be mediated by schema that are very general and that can be applied indiscriminately.

**Schema Learning May Be Ubiquitous in Audition.** In real-world auditory scenes, sources are sometimes unfamiliar, and recurring structure may occur only intermittently and concurrent with other sounds. Our results demonstrate that the auditory system can rapidly learn to utilize abstract source structure even in such challenging conditions. The robustness of this learning could allow schema-based scene analysis to occur across a much wider range of scenarios than previously imagined.

## Materials and Methods

All experiments were approved by the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology. On the initial page of the online task, participants read a disclaimer (as required by the MIT Committee for the Use of Humans as Experimental Subjects) and consented to participation in the experiment. Participants in the in-laboratory condition in experiment S1 signed a form indicating their consent to participate. Methods are described in *SI Materials and Methods*.

**ACKNOWLEDGMENTS.** We thank the members of the J.H.M. laboratory for helpful comments on the manuscript. This work was supported by a McDonnell Foundation Scholar Award (to J.H.M.), National Science Foundation Grant BCS-1454094, NIH Grant 1R01DC014739-01A1, and NIH Training Grant T32DC000038.

1. Bregman AS (1990) *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA).
2. Bronkhorst AW (2000) The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acust United Acust* 86: 117–128.
3. Carlyon RP (2004) How the brain separates sounds. *Trends Cogn Sci* 8:465–471.
4. Bee MA, Micheyl C (2008) The cocktail party problem: What is it? How can it be solved? And why should animal behaviorists study it? *J Comp Psychol* 122:235–251.
5. McDermott JH (2009) The cocktail party problem. *Curr Biol* 19:R1024–R1027.
6. Shinn-Cunningham BG (2008) Object-based auditory and visual attention. *Trends Cogn Sci* 12:182–186.
7. Shamma SA, Micheyl C (2010) Behind the scenes of auditory perception. *Curr Opin Neurobiol* 20:361–366.
8. Snyder JS, Gregg MK, Weintraub DM, Alain C (2012) Attention, awareness, and the perception of auditory scenes. *Front Psychol* 3:15.
9. Middlebrooks JC, Simon JZ, Popper AN, Fay RR (2017) *The Auditory System at the Cocktail Party* (Springer, New York).
10. Moore BCJ, Glasberg BR, Peters RW (1986) Thresholds for hearing mistuned partials as separate tones in harmonic complexes. *J Acoust Soc Am* 80:479–483.
11. Hartmann WM, McAdams S, Smith BK (1990) Hearing a mistuned harmonic in an otherwise periodic complex tone. *J Acoust Soc Am* 88:1712–1724.
12. Micheyl C, Oxenham AJ (2010) Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings. *Hear Res* 266:36–51.
13. Darwin CJ (1981) Perceptual grouping of speech components differing in fundamental frequency and onset-time. *Q J Exp Psychol Sect A* 33:185–207.
14. Darwin CJ, Ciocca V (1992) Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component. *J Acoust Soc Am* 91:3381–3390.
15. Kidd G, Jr, Mason CR, Deliwala PS, Woods WS, Colburn HS (1994) Reducing informational masking by sound segregation. *J Acoust Soc Am* 95:3475–3480.
16. Bendixen A, Denham SL, Gyimesi K, Winkler I (2010) Regular patterns stabilize auditory streams. *J Acoust Soc Am* 128:3658–3666.
17. McDermott JH, Wrobleski D, Oxenham AJ (2011) Recovering sound sources from embedded repetition. *Proc Natl Acad Sci USA* 108:1188–1193.
18. Cohen MA, Horowitz TS, Wolfe JM (2009) Auditory recognition memory is inferior to visual recognition memory. *Proc Natl Acad Sci USA* 106:6008–6010.
19. Saffran JR, Johnson EK, Aslin RN, Newport EL (1999) Statistical learning of tone sequences by human infants and adults. *Cognition* 70:27–52.
20. Creel SC, Newport EL, Aslin RN (2004) Distant melodies: Statistical learning of nonadjacent dependencies in tone sequences. *J Exp Psychol Learn Mem Cogn* 30: 1119–1130.
21. Aslin RN, Newport EL (2012) Statistical learning: From acquiring specific items to forming general rules. *Curr Dir Psychol Sci* 21:170–176.
22. Stilp CE, Rogers TT, Kluender KR (2010) Rapid efficient coding of correlated complex acoustic properties. *Proc Natl Acad Sci USA* 107:21914–21919.
23. Saffran JR, Newport EL, Aslin RN (1996) Word segmentation: The role of distributional cues. *J Mem Lang* 35:606–621.
24. Tillmann B, Bharucha JJ, Bigand E (2000) Implicit learning of tonality: A self-organizing approach. *Psychol Rev* 107:885–913.
25. Pearce MT, Ruiz MH, Kapasi S, Wiggins GA, Bhattacharya J (2010) Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *Neuroimage* 50:302–313.
26. Rohrmeier M, Rebuschat P (2012) Implicit learning and acquisition of music. *Top Cogn Sci* 4:525–553.
27. Kaernbach C (2004) The memory of noise. *Exp Psychol* 51:240–248.
28. Agus TR, Thorpe SJ, Pressnitzer D (2010) Rapid formation of robust auditory memories: Insights from noise. *Neuron* 66:610–618.
29. Leek MR, Watson CS (1988) Auditory perceptual learning of tonal patterns. *Percept Psychophys* 43:389–394.
30. Perruchet P, Pacton S (2006) Implicit learning and statistical learning: One phenomenon, two approaches. *Trends Cogn Sci* 10:233–238.
31. Bey C, McAdams S (2002) Schema-based processing in auditory scene analysis. *Percept Psychophys* 64:844–854.
32. Haykin S, Chen Z (2005) The cocktail party problem. *Neural Comput* 17:1875–1902.
33. Darwin CJ (2008) Listening to speech in the presence of other sounds. *Philos Trans R Soc Lond B Biol Sci* 363:1011–1021.
34. Alain C, Bernstein LJ (2008) From sounds to meaning: The role of attention during auditory scene analysis. *Curr Opin Otolaryngol Head Neck Surg* 16:485–489.
35. Moore BCJ, Gockel HE (2012) Properties of auditory stream formation. *Philos Trans R Soc Lond B Biol Sci* 367:919–931.
36. Bronkhorst AW (2015) The cocktail-party problem revisited: Early processing and selection of multi-talker speech. *Atten Percept Psychophys* 77:1465–1487.
37. Johnsrude IS, et al. (2013) Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychol Sci* 24:1995–2004.
38. Cooke M, Garcia Lecumberri ML, Barker J (2008) The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *J Acoust Soc Am* 123:414–427.
39. Dowling WJ (1973) The perception of interleaved melodies. *Cognit Psychol* 5:322–337.
40. Dowling WJ, Lung KM-T, Herrbold S (1987) Aiming attention in pitch and time in the perception of interleaved melodies. *Percept Psychophys* 41:642–656.
41. Devergie A, Grimault N, Tillmann B, Berthommier F (2010) Effect of rhythmic attention on the segregation of interleaved melodies. *J Acoust Soc Am* 128:EL1–EL7.
42. Szalárdy O, et al. (2014) The effects of rhythm and melody on auditory stream segregation. *J Acoust Soc Am* 135:1392–1405.
43. Billig AJ, Davis MH, Deeks JM, Monstrey J, Carlyon RP (2013) Lexical influences on auditory streaming. *Curr Biol* 23:1585–1589.
44. Woods KJP, Siegel MH, Traer J, McDermott JH (2017) Headphone screening to facilitate web-based auditory experiments. *Atten Percept Psychophys* 79:2064–2072.
45. McDermott JH, Lehr AJ, Oxenham AJ (2008) Is relative pitch specific to pitch? *Psychol Sci* 19:1263–1271.
46. Woods KJP, McDermott JH (2015) Attentive tracking of sound sources. *Curr Biol* 25: 2238–2246.
47. Nygaard LC, Pisoni DB (1996) Learning voices. *J Acoust Soc Am* 99:2589–2603.
48. Reinisch E, Holt LL (2014) Lexically guided phonetic retuning of foreign-accented speech and its generalization. *J Exp Psychol Hum Percept Perform* 40:539–555.
49. Jiang Y, Chun MM (2001) Selective attention modulates implicit learning. *Q J Exp Psychol A* 54:1105–1124.
50. Fritz JB, Elhilali M, David SV, Shamma SA (2007) Auditory attention–Focusing the searchlight on sound. *Curr Opin Neurobiol* 17:437–455.
51. Moore BCJ, Gockel H (2002) Factors influencing sequential stream segregation. *Acta Acust United Acust* 88:320–333.
52. Micheyl C, Hunter C, Oxenham AJ (2010) Auditory stream segregation and the perception of across-frequency synchrony. *J Exp Psychol Hum Percept Perform* 36: 1029–1039.
53. Thompson SK, Carlyon RP, Cusack R (2011) An objective measurement of the build-up of auditory streaming and of its modulation by attention. *J Exp Psychol Hum Percept Perform* 37:1253–1262.
54. Roberts B, Glasberg BR, Moore BCJ (2008) Effects of the build-up and resetting of auditory stream segregation on temporal discrimination. *J Exp Psychol Hum Percept Perform* 34:992–1006.
55. McAdams S, Bertoncini J (1997) Organization and discrimination of repeating sound sequences by newborn infants. *J Acoust Soc Am* 102:2945–2953.
56. Winkler I, et al. (2003) Newborn infants can organize the auditory world. *Proc Natl Acad Sci USA* 100:11812–11815.
57. Peer E, Brandimarte L, Samat S, Acquisti A (2017) Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *J Exp Soc Psychol* 70:153–163.
58. Peer E, Vosgerau J, Acquisti A (2014) Reputation as a sufficient condition for data quality on Amazon mechanical Turk. *Behav Res Methods* 46:1023–1031.
59. Dowling WJ, Fujitani DS (1971) Contour, interval, and pitch recognition in memory for melodies. *J Acoust Soc Am* 49:524–531.
60. Klatt DH (1980) Software for a cascade/parallel formant synthesizer. *J Acoust Soc Am* 67:971–995.
61. Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS (1993) DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STIRecon Tech Rep N* 93:27403.
62. Kawahara H, et al. (2008) Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, Piscataway, NJ), pp 3933–3936.
63. Mustafa K, Bruce IC (2006) Robust formant tracking for continuous speech with speaker variability. *IEEE Trans Audio Speech Lang Process* 14:435–444.
64. Buhrmester M, Kwang T, Gosling SD (2011) Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspect Psychol Sci* 6:3–5.
65. Crump MJC, McDonnell JV, Gureckis TM (2013) Evaluating Amazon's mechanical Turk as a tool for experimental behavioral research. *PLoS One* 8:e57410.
66. Paolacci G, Chandler J (2014) Inside the Turk understanding mechanical Turk as a participant pool. *Curr Dir Psychol Sci* 23:184–188.

# Supporting Information

Woods and McDermott 10.1073/pnas.1801614115

## SI Materials and Methods

**General Experimental Procedures.** Online experiments were run using two crowdsourcing services, Amazon's Mechanical Turk and Prolific Academic (both known to produce high-quality data in behavioral experiments) (57). Participants were restricted to those who had previously completed 50 or more tasks online with a work acceptance rate of 90% or higher (58) and who were located in the United States or Canada. Participants on Prolific Academic were restricted to those who had never used Mechanical Turk. On the initial page of the online task, participants read a disclaimer (as required by the MIT Committee for the Use of Humans as Experimental Subjects), consented to participation in the experiment, and were informed of the need for headphones. We obtained demographic information and implemented a short behavioral screening test to ensure that participants were listening through headphones as instructed (44). At this stage, any participants failing the headphone screening were released from the experiment with partial pay; participants who went on to complete the full task received a bonus amount. Of online participants, 64.6% passed the headphone screening and continued to our main task. The number of trials varied by experiment; the longest, experiment 4, had 168 trials and ran ~35 min.

***In-laboratory experiments.*** One condition of experiment S1 was conducted in the laboratory to compare performance with that obtained online. After reading and signing a form indicating their consent to participate, participants were seated at a desk in a sound-attenuated chamber and listened over headphones (Sennheiser HD280) to sounds produced by a Mac Mini computer.

***Feedback.*** Feedback on each trial was provided in all experiments.

### Paradigm 1.

***Stimulus generation.*** Stimuli were composed of complex tones containing the first 10 harmonics with –10 dB per octave rolloff and starting phases set to zero. A half-Hanning window (10 ms) was applied to the onset and offset of each tone. Four-note melodies were generated by drawing four pitch values from a continuous uniform distribution from 0 to 12 semitones (ST). This constrained the total range of a melody to a maximum of an octave. Consecutive notes in a melody were required to differ by at least one ST; this constraint was enforced by repeatedly sampling each note of a melody until one was obtained that differed sufficiently from the previous note. The temporal pattern of the melody was generated by drawing tone durations from a continuous uniform distribution from 200 to 400 ms and drawing intertone intervals from a continuous uniform distribution from 10 to 200 ms.

On each trial, the isolated probe melody was transposed away from the melody in the mixture by up to an octave. This was achieved by drawing two values from a continuous uniform distribution from –6 to +6 (ST) and setting the mean pitch of each melody to 300 Hz plus the sampled values (as semitones relative to 300 Hz). The mixture and probe were separated by a silent gap of 300 ms.

On half of the trials the probe melody was a foil, i.e., different from the melody in the mixture. Foils were generated by changing the middle two notes, leaving the first and fourth notes unaltered. The two foil tones were drawn from the same distribution as the tones in the original melody but with the constraint that the pitch interval they formed with the preceding note (i.e., intervals 1 and 2 of the melody) had a different sign than in the original melody, altering its contour (59).

Two distractor tones were added to the first melody on each trial to form the mixture. These distractors were drawn from the same distribution as the melody tones (i.e., limited to a 12-ST range) and then were shifted to set their mean pitch equal to the mean pitch of the four-tone melody. The distractor tones were additionally constrained to differ by at least 4 ST from the target melody tones with which they co-occurred (to avoid energetic masking), by sampling distractor tones until this constraint was satisfied. The durations of the distractor tones and the time interval between them were drawn from the same distribution as the melody tones. The onset of the first distractor tone was drawn from a uniform distribution (400 ms in width, centered on the onset of the second note of the melody but excluding the center 100 ms, so that the onset was constrained to be at least 50 ms before or after the onset of the second note of the melody). The onset of the second distractor tone was generated according to the same procedure used for the notes of the melody.

For each participant, a particular melody was designated as the schema. Schema-based trials used random transpositions of this melody, as described above. Non–schema-based trials used melodies from schema-based trials for other participants (one per schema). In practice this was achieved by generating many sets of schema-based melodies in advance and then drawing from them as needed. The consequence of this was that the schema- and non–schema-based trials were statistically identical when pooled across participants, so that systematic differences in performance could only reflect learning.

We note that the stimuli were designed so that there were few obvious grouping cues by which to segregate the target melody from the distractor tones. As a consequence, baseline performance in the task was low (typically with d′ <1). Low baseline performance levels seemed desirable a priori, as they would leave room for a schema effect to exhibit itself. Performance was nonetheless substantially above chance, presumably because listeners have some ability to match the probe stimulus to the mixture despite the transposition and lack of grouping cues. This above-chance baseline performance was likely important in order for participants to maintain motivation and engagement with the task. In paradigms 2 and 3, stimulus continuity provided some basis on which to group the stimuli, and baseline performance was somewhat higher, as expected.

***Procedure: Experiment 1 (schema-based melody segregation).*** For each participant a schema was selected at random, and the corresponding set of schema-based trials was presented (with order randomized) in alternation with a randomly selected trial from each one of the other schema sets. Participants were never told (in this or subsequent experiments) that recurring sounds were a focus of study. Following our online screening process (detailed above) (44), participants were given instructions that they would hear two melodies separated by a brief silence and would have to judge if they were the same. It was explained that the same melody could be "shifted up or down so it starts on a different note," and examples of transposed melodies were presented (including incorrect foil examples). Finally, participants were told that distractor tones would occur on top of the first melody but that they should nonetheless try to do the task as before. Six example trials were provided alongside the correct response to each; these could be played repeatedly, while stimuli in the test section could be played only once. A button at the bottom of the training page progressed participants to the test section. Participants completed 100 trials: 50 schema-based trials and 50 non–schema-based trials, interleaved (with the order counterbalanced across participants).

***Procedure: Experiment 2 (interrupting block without schema).*** This experiment differed from experiment 1 in two respects. First, trials 41–60 were entirely non–schema-based. Second, the schema-based stimuli in either trials 61–80 or trials 81–100 were drawn from a novel schema (i.e., one not used in the rest of the schema-based trials within the experiment). The schema presented at the outset of the experiment (trials 1–40) thus disappeared for 20 trials for half the participants (~3 min) and for 40 trials for the other half of participants (~6 min). This experiment structure is depicted in the schematic in Fig. 1*D*.

***Procedure: Experiment 3 (two schemas per participant).*** In this experiment, two different schemas were selected at random for each participant and were presented in alternation so that each schema recurred on every fourth trial. This experiment was lengthened to 200 trials to present each schema a total of 50 times, as in the previous experiments.

***Procedure: Experiment S1 (schema learning occurs without isolated exposure to the schema).*** This experiment was identical to experiment 1, except that two distractor tones were added to the probe melody on each trial, generated with the same procedure used for the distractors added to the target melody on a trial.

***Procedure: Experiment 4 (schema learning with noise-burst sequences).*** This experiment was similar to experiment 1, but instead of pure-tone notes, sources consisted of noise bursts (Fig. 1*F*). The noise bursts were 4 ST wide (i.e., one-third of an octave) and 300 ms in duration, generated by applying a rectangular window to white noise in the frequency domain. The four noise bursts in the target sequence were presented back-to-back. Noise burst center frequencies were drawn from a uniform distribution of 36 ST around a center value. The center value was drawn from a uniform distribution one octave wide, centered at 1,000 Hz ± 6 ST. Consecutive bursts were constrained to have center frequencies at least 6 ST apart. Three distractor bursts were centered on the onset of the second, third, and fourth of the target sequence. Distractor bursts were drawn from the same distribution as the target bursts, also subject to the constraint that successive bursts had to be at least 6 ST apart. The probe sequence was transposed by a random amount uniformly distributed between −6 and +6 ST relative to the target sequence.

**Paradigm 2.**

***Stimulus generation.*** We synthesized pairs of continuous vowels (using Klatt synthesis) (60) whose pitch and formant values intertwined as they evolved over their 3-s duration (Fig. 2*A*). The cue, presented to the listener before the mixture, consisted of the initial 0.25 s of the target voice. The probe was the last 0.25 s of either the target or distractor. Linear ramps (100 ms) were applied to the onset and offset of cues, probes, and mixtures. The target voice began 50 ms before the distractor voice in the mixtures (by zeroing out the first 50 ms of the distractor) to aid initial segregation. Trajectories in each feature were generated by bandpass filtering Gaussian noise (see below) and then scaling to cover a physiologically appropriate range. Feature means and SDs (expressed in semitones from the mean) were f0, $\mu = 238.9$ Hz, SD = 4.6 ST; F1, $\mu = 436.2$ Hz, SD = 3.5 ST; F2, $\mu = 1,456.9$ Hz, SD = 3.6 ST. For each target we generated potential distractors until one was found that met a set of criteria. First, distractor trajectories were selected that crossed the target at least once in all features during the middle 2.5 s of the mixture (0.25–2.75 s). The crossing of feature trajectories prevents the target and distractor from having distinguishing features, requiring listeners to track the target over time to maintain its identity (46). Second, the two voices in a mixture were constrained to never pass closer than a minimum distance of 5.5 ST in the feature space (Euclidean distance in f0 × F1 × F2), since prior work showed that listeners often fail to stream the voices correctly when they pass close together (46). Third, to ensure that target and distractor voices would not coincidentally end at similar points in feature space, we imposed an additional constraint on the process of selecting distractors: that the probe and foil have mean f0 values differing by more than 3 ST.

Like the other two paradigms used in this paper, paradigm 2 used a superset of stimuli that was composed of 43 stimulus sets, each based on one of 43 target trajectory schemas. Each of these 43 stimulus sets contained 84 target–distractor pairings: 42 with targets based on a common schema and 42 with unrelated targets. The unrelated targets were taken from the other schema sets in the superset (one from each). Since trajectories were drawn from filtered noise, individual targets could vary in properties that affect tracking (e.g., rate). However, by using one each of the other sets' schema targets as non–schema-based targets, we could help ensure that the schema-based trials would not differ intrinsically in difficulty from the non–schema-based trials when data were pooled across participants.

The stimuli in paradigm 2 were additionally constrained so that the distributions of distances from the cue to the correct probe and from the cue to the incorrect probe (foil) were similar (cue–probe, $\mu = 10.2$ ST, SD = 3.7 ST; cue–foil, $\mu = 10.6$ ST, SD = 3.8 ST) to minimize the chances that listeners would try to perform the task without tracking the target source. This was accomplished by drawing distractor voices for a given target until the cue–foil distance was less than two ST in all dimensions from the cue–probe distance of another target in the set. The distractor was then transposed in each dimension (shifted on a log-scale) until the mean feature values of its probe matched the required distance exactly. This was performed separately for the schema-generated and unrelated target–distractor pairs in each set, resulting in matched cue–probe and cue–foil distributions within the conditions and in each stimulus set as a whole.

Forty-three voices were randomly generated to serve as schema. The trajectory of each feature of each schema was generated by sampling a 4-s excerpt of Gaussian noise and filtering it (500 Hz sampling rate; lower cutoff, 0.25 Hz; upper cutoff, 1.25 Hz) by setting the amplitudes of frequencies outside this range to zero in the frequency domain. Variants were generated by time-dilating this trajectory: The slowest variant was generated by taking the first 2 s of the 4-s schema and time-dilating it to 3 s (the duration of all stimuli). The fastest variant was generated by taking the entire 4-s schema and time-compressing it to 3 s. Distractors were time-dilated or compressed by a factor drawn at random from the (uniform) distribution of dilation/compression factors imposed on the target, before pairing with a target. Mixtures were additionally transposed in f0, F1, and F2: Each time-dilated variant was randomly assigned to one of 42 transpositions, limited by physiologically motivated minimum and maximum values of f0, F1, and F2 [minimum and maximum feature values across sets after transposition: f0 = (60, 550); F1 = (175, 1,100); F2 = (550, 3,000)]. Consecutive schema-based targets (i.e., separated by a trial) were always shifted in the space by at least one-third of the total available range (mean = 4.1 ST, SD = 0.26 ST; Euclidean distance in three dimensions) to avoid targets on nearby trials starting out at conspicuously similar absolute values.

***Procedure: Experiment S2 (crowdsourcing validation).*** This experiment was a replication of our previously published paradigm (46), intended to probe the quality of our online participants relative to in-laboratory data on a task known to be attentionally demanding. All voices in this experiment were generated independently of each other, unlike the schema-based voices in experiments 4 and 5. The minimum distance between voices was 8 ST (Euclidean distance in the 3D feature space). Participants completed 72 trials in a single block.

***Procedure: Experiment 5 (schema-based attentive tracking).*** Participants completed 168 trials in which the cued voice on every other trial was based on a common schema. Following our online screening process (detailed above) (44), participants were introduced to the attentive tracking task (46) and were given two practice trials during which the stimulus could be played repeatedly. A button at the bottom of the training page progressed participants to the test section. Stimuli in the test section could be played only once. Here and in other experiments using this paradigm, participants with d′ values below 0.25 were excluded from analysis. To ensure that the stimulus sets remained balanced, these excluded participants were replaced with additional participants until each schema-based set was used an equal number of times.

***Implicitness query.*** After the last trial of experiment 4, a page appeared asking, "Did you notice if the target voices (the cued voices) were reused on different trials? (did you hear the same target more than once?)" The participant could select one of two options: "Yes, I remember hearing the same target more than once" or "No, I don't remember hearing the same target more than once."

***Procedure: Control experiments for implicitness query.*** Control experiment 1 presented all the non–schema-based trials from experiment 4 but followed it with the same query about whether the participant had noticed recurring structure in the cued voice. To test if responses would change if the recurring structure were made obvious, control experiment 2 presented only schema-based trials (i.e., the schema was present on every trial), again followed with the same query.

***Procedure: Experiment 6 (control for cue-probe regularities).*** Stimuli in this experiment were the first half of the stimuli used in experiment 4 except that the last third of trials replaced each mixture with noise. This noise was spectrally matched to the average power spectrum of all mixtures (a different sample of noise was used on each trial). Cues and probes were unchanged. For this experiment, the task page was modified to display the following: "Some of the sounds will have NOISE added. These may seem hard, but just try your best!" This experiment consisted of 84 trials, with the first two-thirds of the task (56 trials) indistinguishable from experiment 4 and the last one-third modified as described.

***Procedure: Experiment 7 (rapid learning on a finer timescale).*** This experiment consisted of 56 trials, identical to the beginning portions of experiments 4 and 5. The data in Fig. 2*E* include the 183 new participants from this shorter experiment, pooled with data from the first 56 trials of experiments 4 and 5.

***Procedure: Experiment 8 (role of formants in learned schema).*** This experiment was identical to experiment 5 except that the formant trajectories on schema-based trials in the last half of the experiment were randomly generated subject to the same constraints as non–schema-based trials: We applied the same generative parameters and requirements on target and distractor interaction (i.e., they crossed in all three dimensions and maintained the same minimum distance in the f0 × F1 × F2 feature space) and ensured that the cue–foil distance distribution with the random-formant stimuli matched that in the original stimuli with schema-based formants.

***Procedure: Experiment 9 (graded learning by similarity to the schema).*** We ran 172 participants on a replication of experiment 5 (using paradigm 2) to perform analyses on subgroups of stimuli.

**Paradigm 3.**

***Stimulus generation.*** For each of the 192 female speakers in the TIMIT speech corpus (61) the longest-duration sentence unique to that speaker was chosen and analyzed to extract time-varying pitch [f0, using STRAIGHT (62)] and formants [F1, F2, using the Mustafa and Bruce (63) formant tracker] as well as the ratio of voiced to unvoiced excitation (also measured using STRAIGHT). Feature tracks were extracted at a 1,000-Hz sampling rate. We first identified unvoiced speech segments as those regions of a speech signal longer than 10 ms in duration over which the ratio of voiced to unvoiced excitation fell below a threshold (selected by the experimenters as that below which the feature estimates became erratic). We then took the remaining segments as voiced, and interpolated the feature tracks across sub-segments in which the voicing ratio fell below threshold (because pitch and formant estimates are unstable when voicing is minimal or absent). Next, we smoothed the resulting feature tracks of each voiced segment with a 5 ms rectangular averaging window. Finally, the voiced segments were synthesized with Klatt synthesis, windowing each segment with 15 ms linear ramps at onset/offset. Then, for each speaker/sentence we looked for a 750-ms region that contained no voiced segments shorter than 120 ms and no unvoiced segments longer than 30 ms (we found that these parameters reliably identified continuously spoken phrases, without breaths or pauses). This left us with 188 750-ms utterances, each from a different female speaker (in four speakers or sentences, no suitable region was found). Feature tracks were resynthesized with the same method used in paradigm 2 (Klatt synthesis) (60).

Each trial consisted of a mixture of utterances (a target and a distractor) followed by an isolated probe utterance. The isolated probe utterance was transposed away from the target utterance in the mixture by up to eight ST in f0. This was achieved by drawing two values from a continuous uniform distribution from −4 to +4 (ST) and setting the mean pitch of each utterance to that number of semitones above or below 200 Hz. Formants were shifted by half the sampled distances in the same direction.

The distractor utterance in a mixture was chosen at random from the set of 188 utterances, excluding the schema utterance for that participant and the probe utterance for that trial. The dilation factor for the distractor was selected to match that of the nondistractor utterance (see below), so that the utterances in the mixture would have the same duration. The distractor was then transposed so that the two utterances in the mixture had the same mean pitch. On foil trials (where the probe utterance did not match the target), the probe was replaced by a different utterance drawn at random (excluding the schema utterance for that participant and the distractor(s) appearing on that trial). The mixture and probe were separated by a silent gap of 500 ms.

One schema was selected at random for each participant. Schema-based variants were generated by time dilation (50 variants from each utterance) imposed in steps of 1% up to −50% (i.e., the slowest utterances were half as fast as the original speech). Stimuli were not truncated following dilation, so that the duration of stimuli in this experiment ranged from 750 ms to 1,500 ms. On each schema-based trial, two different variants were selected at random to appear in the mixture and probe and were then transposed in their pitch and formants as described above (time dilation and transposition were applied independently and did not covary in this stimulus set). Time dilation and transposition were applied to the feature tracks before resynthesis.

***Procedure: Experiment 10 (schema learning with speech feature trajectories; mixture-probe format).*** Participants completed 100 trials in which they heard a mixture of two utterances followed by a probe utterance. On every other trial, the target utterance in the mixture was generated from a common schema (that varied across participants). For the remainder of the trials a random variant was drawn from each of the other sets with schema-based targets, resulting in statistically identical schema- and non–schema-based stimuli when pooled across participants. Following our online screening process (detailed above) (44), participants were given examples of utterance variants (two variants differing by transposition and/or time dilation, separated by a brief silence) and were told whether or not they were the same. This was intended to familiarize participants with examples of what would count as the same utterance during the experiment. Participants were then introduced to the mixture–probe task ("You will first hear two utterances played at the same time, then a single utterance. Did the single utterance appear in the mixture that came just before it?") and were provided six examples of trials along with the correct response to each; each example could be played repeatedly. A button at the bottom of the training screen progressed participants to the test section. Stimuli in the test section could be played only once.

***Procedure: Experiment 11 (schema-based distractors).*** This experiment was similar to experiment 10 except that alternate trials contained distractors from the same schema. That is, these alternate trials contained, only in the mixture, utterances from the same schema-based set, whereas the intended target was different across trials. For the non–schema-based trials a random variant was drawn from each of the other schema-based sets. Because experiment 11 had twice as many trials as experiment 10 (to ensure comparable exposure to the schema in each case), we doubled the number of schema-based variants by halving the time dilation and transposition step sizes between variants in a set. Task instructions were identical to those of experiment 10.

**Participants.** All experiments were approved by the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology. Sample sizes and demographic information for each experiment are shown in Table S1. Participants who reported hearing loss (4.1% across all experiments) were excluded from analysis. Additionally, in our main experiments using paradigm 2 (with the attentive tracking task; experiments 5–9) we excluded poorly performing participants (those with d′ <0.25 across experimental conditions) from analysis. In experiment 6, exclusion was based on performance in the first two blocks only because all subjects performed near chance in the third block. To maintain the balancing of the stimulus sets (in which all schema-based sets were used equally often so that the schema- and non–schema-based stimuli were identical when pooled across participants), we ran additional participants until each schema-based set was used an equal number of times subject to the performance inclusion criterion. These extra participants are not included in Table S1 (typically we needed to run about 30% extra participants whose performance was above the inclusion threshold in order to achieve the stimulus balancing). The exceptions to the balancing were experiments 6, 7, and C2. In experiments 6 and C2, balancing the stimulus sets was not relevant to the scientific conclusions. Experiment 7 was not balanced in order to maximize the number of participants. To compare methods of data collection, experiment S1 reports data from all participants who performed the task (attentive tracking) online and in the laboratory, without performance-based exclusions.

**Sample Sizes.** A power analysis performed on the data from experiment 1 indicated that 69 participants would be needed to reliably detect the effects of interest (β = 0.8, α = 0.05). This number of participants was targeted in most other experiments, but due to the nature of the screening process, the final participant counts varied somewhat across experiments. In experiments 2, 9, and 11, additional participants were run in anticipation of smaller effect sizes or to more strongly test the presence of a null result. To determine the sample size for experiment 7, a separate power analysis was performed on the pooled data from experiments 5 and 6, with bin sizes halved to seven trials per bin. This analysis indicated that 400 participants would be needed to resolve differences within individual bins (β = 0.8, α = 0.05).

**Statistics.** In paradigms 1 and 3, *t* tests were used to test for differences in performance between conditions. There was generally one such comparison per experiment, so no correction for multiple comparisons was employed. In paradigm 2, changes in performance over time were assessed with repeated-measures ANOVAs testing main effects of stimulus condition (schema- and non–schema-based) and for interactions between the effect of stimulus condition and trial number. Mauchly's test was used to test for violations of the sphericity assumption. When Mauchly's test was significant, the Greenhouse–Geisser correction was used. Data distributions were assumed to be normal and were evaluated as such by eye. Responses to the question assessing implicit learning in paradigm 2 were compared with a z test. For these data (shown in Fig. 3), error bars were computed via bootstrap because each participant gave only a single yes/no answer per query.

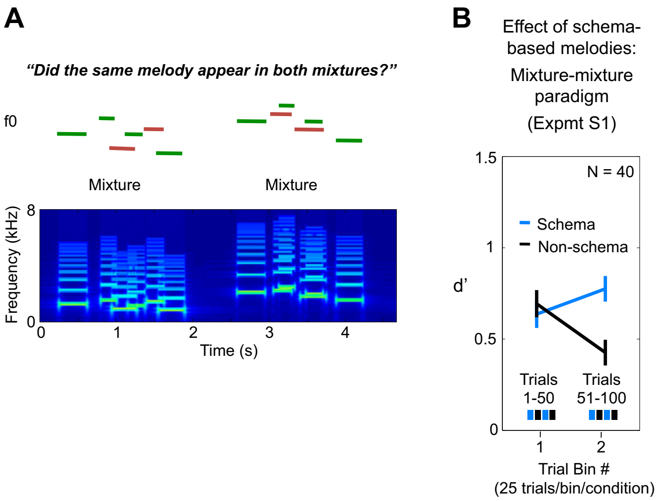

**Fig. S1.** Schema learning occurs without isolated exposure to the schema. (*A*) Schematic of trial structure (*Upper*) and spectrogram of an example stimulus (*Lower*). A target melody (green line segments) was presented concurrently with two distractor notes (red line segments), followed by a probe melody (green line segments) also presented with two distractor notes. The experiment was otherwise identical to experiment 1: Listeners judged whether the probe melody matched the target melody. The probe melody was transposed up or down in pitch by a random amount, and on every other trial the target melody was generated from a common schema. (*B*) Results of experiment S1 (*n* = 160). Error bars denote SEM. Performance is overall lower than in experiments 1–4 (in which the probe was presented in isolation), presumably because there are twice as many opportunities to make grouping errors.

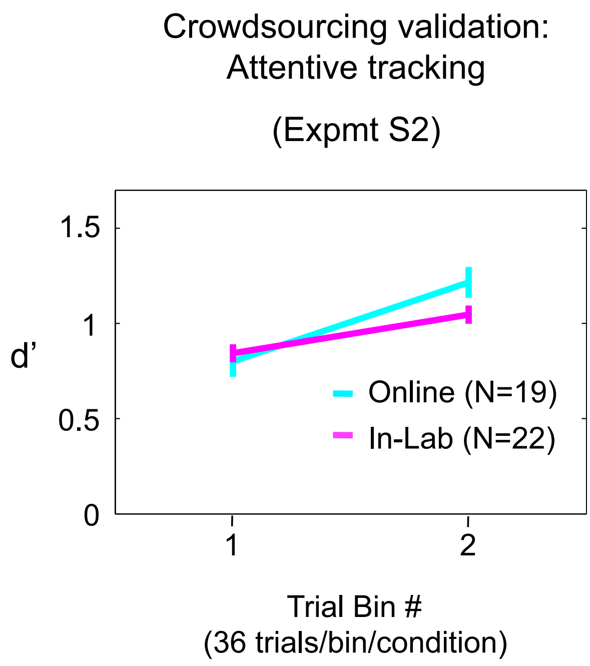

**Fig. S2.** Validation of online performance. A priori it seemed possible that learning effects might be rapid. To resolve changes in performance from one trial to the next, we ran large numbers of participants in relatively brief web-based experiments. Crowdsourcing is widespread in psychology (64–66) but is rarely used for auditory research due to concerns about sound quality. We have developed (and use here) methods to better control online sound presentation by screening out participants who are not wearing headphones (44), but we nonetheless wanted to validate the online version of the attentive tracking task that we used to explore schema learning. We ran 22 participants in the laboratory and then used the same stimulus set to run participants online ($n = 19$ after our screening protocol). The results show that online participants performed as well as those in the laboratory, with no significant difference in tracking performance in either the first half or the second half of the experiment [respectively: $t(39) = 0.13$, $P = 0.90$; $t(39) = 0.43$, $P = 0.67$], and with no significant interaction between time and testing method ($F(1,39) = 1.36$, $P = 0.25$), indicating that changes in performance over the course of the experiment were similar in the two participant pools. Given that our task is effortful and requires focused attention (46), this result suggests that online participants can be as motivated and engaged as in-laboratory participants given appropriate screening procedures (44).

**Table S1. Participant demography**

| Experiment | *N* | *N* female | Mean age, y | Excluded (d′ <0.25) |
|---|---|---|---|---|
| Experiment 1 | 160 | 79 | 37.4 | N/A |
| Experiment 2 | 192 | 102 | 38.1 | N/A |
| Experiment 3 | 88 | 30 | 36 | N/A |
| Experiment 4 | 68 | 33 | 36.6 | N/A |
| Experiment 5 | 86 | 43 | 33.9 | 75 |
| Experiment 6 | 133 | 54 | 33.2 | 86 |
| Experiment 7 | 183 | 82 | 33.0 | 180 |
| Experiment 8 | 86 | 31 | 35.7 | 57 |
| Experiment 9 | 172 | 86 | 37.4 | 185 |
| Control experiment 1 | 45 | 16 | 36.6 | 66 |
| Control experiment 2 | 51 | 26 | 37.7 | 12 |
| Experiment 10 | 93 | 53 | 39.2 | N/A |
| Experiment 11 | 202 | 103 | 38.0 | N/A |
| Experiment S1 | 40 | 17 | 38.3 | N/A |
| Experiment S2 | 41 | 16 | 30.2 | N/A |

N/A, not applicable.