Article

Current Biology

# Attentive Tracking of Sound Sources

## Highlights

- Humans track sound sources through feature space with a movable focus of attention
- Attentive tracking aids segregation of similar sound sources
- Tracking failures occur if sound sources pass nearby in feature space
- Tracking is robust to speech-like source discontinuities

## Authors

Kevin J.P. Woods, Josh H. McDermott

## Correspondence

kwoods@mit.edu

## In Brief

Hearing a sound source of interest amid other sources (the "cocktail party problem") is difficult when sources are similar and change over time, as in speech. Woods and McDermott show that humans segregate sources in such situations using attentive tracking—employing a moving locus of attention to follow a sound as it changes over time.

Woods and McDermott, 2015, Current Biology *25*, 1–9
August 31, 2015
http://dx.doi.org/10.1016/j.cub.2015.07.043

CellPress

Please cite this article in press as: Woods and McDermott, Attentive Tracking of Sound Sources, Current Biology (2015), http://dx.doi.org/10.1016/j.cub.2015.07.043

Current Biology
Article

# Attentive Tracking of Sound Sources

**Kevin J.P. Woods**[1,2,*] **and Josh H. McDermott**[1,2]
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Program in Speech and Hearing Bioscience and Technology, Harvard University, Cambridge, MA 02138, USA
*Correspondence: kwoods@mit.edu

http://dx.doi.org/10.1016/j.cub.2015.07.043

## SUMMARY

**Auditory scenes often contain concurrent sound sources, but listeners are typically interested in just one of these and must somehow select it for further processing. One challenge is that real-world sounds such as speech vary over time and as a consequence often cannot be separated or selected based on particular values of their features (e.g., high pitch). Here we show that human listeners can circumvent this challenge by tracking sounds with a movable focus of attention. We synthesized pairs of voices that changed in pitch and timbre over random, intertwined trajectories, lacking distinguishing features or linguistic information. Listeners were cued beforehand to attend to one of the voices. We measured their ability to extract this cued voice from the mixture by subsequently presenting the ending portion of one voice and asking whether it came from the cued voice. We found that listeners could perform this task but that performance was mediated by attention—listeners who performed best were also more sensitive to perturbations in the cued voice than in the uncued voice. Moreover, the task was impossible if the source trajectories did not maintain sufficient separation in feature space. The results suggest a locus of attention that can follow a sound's trajectory through a feature space, likely aiding selection and segregation amid similar distractors.**

## INTRODUCTION

The cocktail party problem is the challenge of hearing a source of interest given the mixture of sources that often enters the ears, as when following a conversation in a crowded restaurant. Figure 1A displays a spectrogram of one such scenario, in which two different speakers emit concurrent utterances. In such situations, sound energy produced by a particular source must be segregated from that of other sources and grouped together [1–4] into what is conventionally termed a "stream." The listener must select one (or perhaps more) of the streams for further processing [4–8]. The estimation of sound sources from mixtures is believed to rely on prior knowledge of the statistical regularities of natural sounds, such as common onset [9], harmonicity [10, 11], repetition [12], and similarity over time [1, 13–16], but less is known about the processes underlying attentional selection and their interaction with sound segregation [17–19].

Both segregation and selection could be aided by features of a target source that distinguish it from other sources, such as a unique pitch or location [20–23]. Studies of stream segregation have largely focused on cases such as this [1, 13–16, 24–30], in which competing sources are consistently separated in some representational space, giving them distinguishing features. However, real-world sources are not always separated in this way, as when we hear animals of the same species, machines of similar construction, or speakers of the same gender. An example of this latter case is shown in Figure 1. Speech results from a sound source (producing either a time-varying pitch or turbulent noise) that is filtered by the time-varying resonances of the vocal tract. Both source and filter are apparent in the frequency spectrum of brief segments of speech (Figure 1B). The regularly spaced peaks correspond to harmonics of the fundamental frequency (F0) that determines the pitch, whereas the peaks at coarser scales correspond to resonances of the vocal-tract configuration at that point in time, known as formants. Formants are one of the main determinants of phonemic structure (the vowel /oo/ in the example of Figure 1B). The fundamental frequency and first two formants are arguably the three most prominent features for human voices [31], but all three features vary substantially over time. Their trajectories for the two utterances in Figure 1A are plotted in Figure 1C; feature distributions across a set of utterances for each speaker are plotted in Figures 1D–1F. It is apparent that the voices largely overlap in all three features. This situation is the norm for speakers of the same gender: across the TIMIT database [32], 86.7% of randomly selected pairs of same-gender sentences (10,000 samples per gender) crossed each other at least once in all three features (26.7% if speakers were different genders, again with 10,000 samples). In these situations, faced with similar sources that cannot be separated on the basis of their features, how does the auditory system segregate and select sources of interest?

In this paper, we explore the possibility that attention might be used to track voices and other sound sources as they evolve over time, acting as a "pointer" by following a target as it moves through a feature space. By tracking a source's trajectory over time rather than relying on any consistent distinguishing features, attentive tracking could mediate segregation and selection when such features are not available. Although attentive tracking is well-established in the visual system [33–37], its existence in audition remains to be demonstrated and characterized.

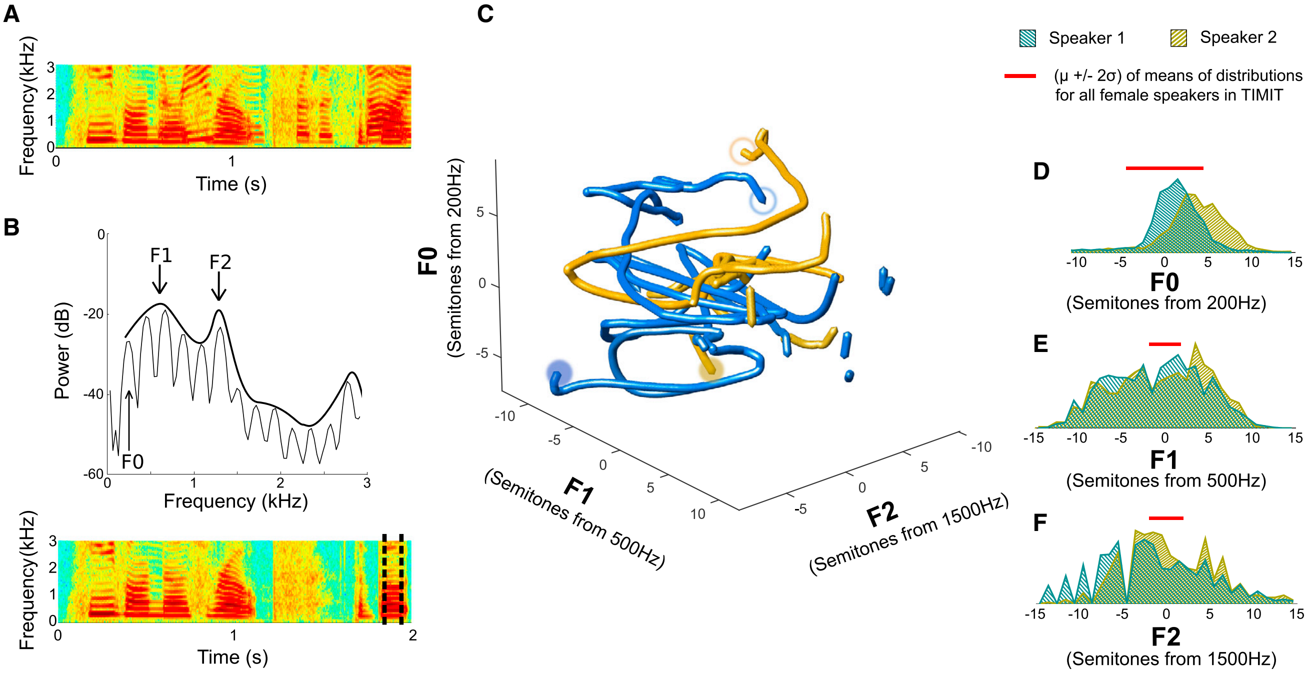

**Figure 1. Features in Natural Speech Vary over Time**

(A) Spectrogram of concurrent utterances by two female speakers.

(B) Example spectral structure of a single speaker. Top: power spectrum of a 100 ms segment of voiced speech excerpted from one of the utterances in (A). Resonances in the vocal tract produce formants—broad spectral peaks that determine vowel quality. Bottom: spectrogram of one of the utterances from (A). Dashed lines depict segment from which power spectrum in top panel was measured.

(C) Pitch and formant contours from the two utterances from (A), measured with PRAAT. The yellow line plots the trajectory for the utterance in (B). Open and closed circles denote the beginning and end of the trajectories, respectively.

(D–F) Marginal distributions of F0, F1, and F2 for all TIMIT utterances for these particular speakers. Red bars mark $\mu \pm 2\sigma$ of the means of such distributions for all 53 female speakers in TIMIT. Differences between the average features of speakers are small relative to the variability produced by a single speaker.

## RESULTS

Our approach was to ask listeners to distinguish sources (synthetic voices) that varied over time and overlapped in feature space such that they had no features that consistently distinguished them from each other. Our stimulus was intended as an abstraction of two concurrent speakers of the same gender, removing linguistic information so as to better isolate potential influences of attentive tracking. Each synthetic voice continuously varied in fundamental frequency (F0) and the first two formants (F1 and F2) over randomly generated trajectories (Figure 2A). The stimuli sounded like continuously modulated vowels.

On each trial, listeners first heard a “cue” (the starting portion of one synthetic voice) followed by a mixture of two synthetic voices. Listeners were then presented with a “probe” sound taken from the end of one of the voices and judged whether it belonged to the cued voice or not (Figure 2B; examples of stimuli can be heard at http://mcdermottlab.mit.edu/attentive_tracking/). Critically, the voice trajectories in each mixture were selected to cross each other in all feature dimensions, such that the voices could not be identified on the basis of any single feature. In addition, the distance between the cue and the two possible probes (the ends of the cued and uncued trajectories) in feature space was controlled to be the same, on average, such that the task could not be performed simply on the basis of the cue-probe distance. Rather, our task required the listener to be able to segregate the sources well enough to determine whether the cue and probe were part of the same source. This task could in principle be performed either by segregating and retaining the entirety of one or both sources in memory or by maintaining selective attention to the cued source as it changed over time (i.e., attentive tracking). We hypothesized that memory demands would limit the effectiveness of the first strategy and that listeners would instead rely on attentive tracking.

### Experiment 1: Stream Segregation without Distinguishing Features

A priori, it was unclear whether competing sources could stream correctly in the absence of distinguishing features, and so we began by testing whether listeners could perform our task. Performance was measured as sensitivity (d′) to whether the probe was drawn from the cued or uncued voice. Listeners performed much better than chance ($d' = 2.10$; $t(7) = 6.04$, $p < 0.001$), suggesting that the sources could be streamed correctly despite not having distinguishing features. However, listeners reported that the task was effortful and required attention to the cued voice. We thus used a second task to probe the focus of attention while subjects performed the streaming task.

Please cite this article in press as: Woods and McDermott, Attentive Tracking of Sound Sources, Current Biology (2015), http://dx.doi.org/10.1016/j.cub.2015.07.043

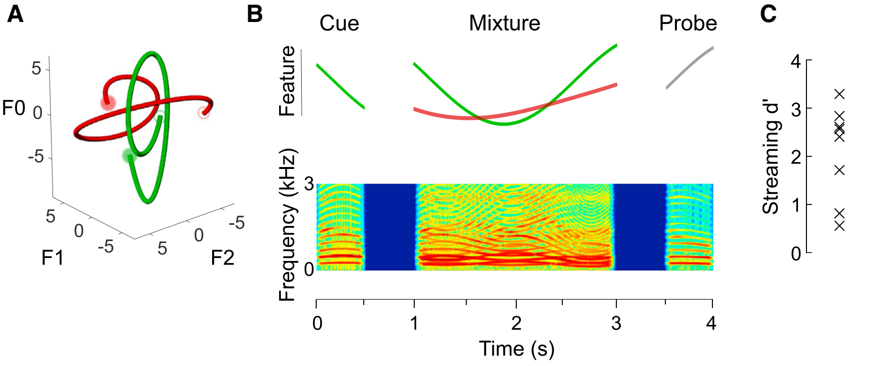

**Figure 2. Streaming Stimuli and Task**

(A) Representative stimulus trajectories from experiment 1 (stream-segregation task). Stimulus trajectories in all experiments crossed at least once in each feature dimension, such that the cued voice could not be selected on the basis of its average pitch or formant values. Here and elsewhere, open and closed circles denote the beginning and end of the trajectories, respectively.

(B) Listeners first heard a cue taken from the beginning portion of one voice, then a mixture of two voices, and finally a probe that could be taken from the end portion of either voice. Listeners had to decide whether the probe came from the cued voice. The graph depicts the stimulus variation along a single dimension for ease of visualization.

(C) Results of experiment 1 (stream-segregation task). Each marker plots the performance of an individual subject.

See also Figure S1 for block-by-block performance.

## Experiment 2A: Measuring the Distribution of Attention during Stream Segregation

Concurrent with the streaming task, we asked subjects to report a brief vibrato (i.e., pitch modulation) that could be presented in either source (Figure 3A). This vibrato appeared in half of all trials and occurred in either the cued or uncued voice equiprobably. We hypothesized that vibrato detection would be more accurate when the vibrato occurred in the focus of attention. After the end of each stimulus, the subject first reported whether or not the probe was from the cued source and then whether or not they heard vibrato anywhere in the stimulus, in either source. Subjects were not asked which source contained the vibrato, only whether they heard it or not.

Detection of vibrato was above chance overall (t(11) = 9.56, $p < 10^{-7}$). Performance also remained well above chance on the streaming task despite the concurrent vibrato task (t(11) = 6.60, $p < 10^{-5}$; Figure 3B). Additionally, there was no difference between streaming performance on trials with and without vibrato (t(11) = 0.45, p = 0.66), suggesting that the presence of vibrato did not interfere with subjects' ability to stream the voices in our task.

If listeners were tracking the cued voice with their attention, we might expect to see a bias in vibrato detection, with vibrato in the cued voice being more readily detected than vibrato in the uncued voice. We thus used hit rates for the two sets of trials to compute sensitivity to vibrato in the cued and uncued voices, using the false-alarm rate from the remaining trials without vibrato. Trials were included in this analysis only if the streaming task was performed correctly, to help ensure that the cued voice was in fact being tracked. Consistent with the notion that attention was directed to the cued voice, vibrato detection was better for the cued voice than for the uncued voice (t(11) = 3.25, $p < 0.01$; Figure 3C, right). Because the vibrato had the same distribution in feature space for both voices, the greater sensitivity when vibrato was in the cued voice suggests that the locus of attention was not constant over time and instead tracked the trajectory of the cued voice as it evolved.

If streaming performance in our task is mediated by attentive tracking, we might further expect subjects who are good at the streaming task to show greater attentional bias. We split our listeners into two equal-sized groups based on their streaming performance and examined attentional bias (the difference between vibrato sensitivity in the cued and uncued voices) separately for each group. The groups were defined by streaming performance on trials without vibrato, to avoid the possibility that the presence of vibrato might have differentially interfered with the streaming task (splitting subjects based on all-trial streaming performance would have resulted in the same groups).

The group that performed best on the streaming task showed a clear attentional bias toward the cued voice (t(5) = 5.35, $p < 0.005$), whereas the other (more poorly performing) group did not (t(5) = 0.89, p = 0.41) (Figure 3C, left). A two-factor ANOVA accordingly showed an interaction between streaming group and attentional bias ($F(1,10) = 7.78$, $p < 0.019$) (again using only correctly streamed trials). The two groups did not differ in their overall detection of vibrato (t(10) = 1.12, p = 0.29), indicating that they were not differentiated by more general factors that could affect performance (e.g., lack of engagement or fatigue). These results suggest that performance in the streaming task is linked to successful attentional selection of the cued voice via attentive tracking.

## Experiment 2B: Attentional Selection over Time

To test whether attentional selection was present throughout the cued source, we conducted a follow-up experiment in which stimuli were extended from 2 s to 3 s to provide more time points at which to probe vibrato detection. In addition, cue and probe durations were reduced from 500 to 250 ms to ensure that our general findings were robust to this parameter. The experiment was otherwise identical to experiment 2A. In particular, vibrato onset was uniformly distributed in time. We ran 12 new listeners in the experiment and screened for good streaming performance by rejecting those whose streaming performance fell below d′ = 1.5, as in experiment 2A (mean streaming performance for all 12 listeners tested was d′ = 1.31). As in experiment 2A, overall vibrato detection was not different between good and poor streamers (t(10) = 1.49, p = 0.17), but poor streamers did not show a significant vibrato detection advantage for the cued voice (t(5) = 1.56, p = 0.18).

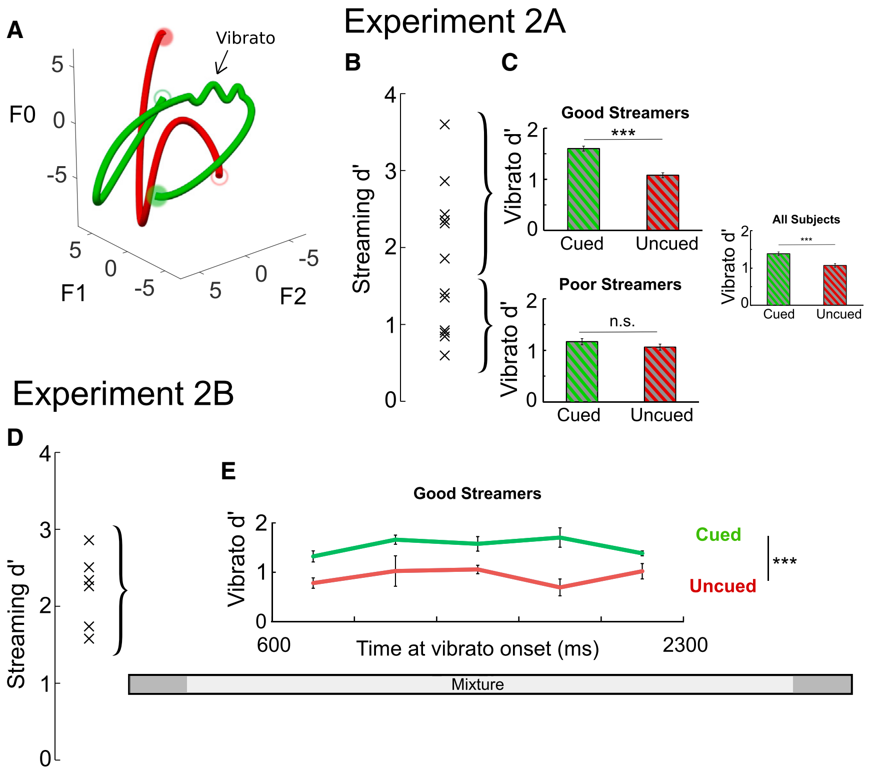

**Figure 3. Experiment 2: Vibrato Detection as a Measure of Attention during Streaming**

(A) Example stimulus trajectories. Either voice could contain vibrato (a brief pitch modulation, added in this example to the green trajectory). Listeners performed the stream-segregation task from experiment 1 but were additionally asked to detect vibrato in either stream. The trajectory shown is 2 s in duration (from experiment 2A); trajectories in experiment 2B were 3 s.

(B) Stream-segregation performance for the 12 participants in experiment 2A.

(C) Sensitivity to vibrato in the cued and uncued voices for subjects grouped by streaming performance (into two equal-sized groups; left) and pooled across groups (right). Includes only trials in which the stream-segregation task was performed correctly. Error bars here and elsewhere denote within-subject SEMs and thus do not reflect the variability in overall vibrato detection across subjects.

(D) Stream-segregation performance for the six best streamers in experiment 2B (3 s mixtures, 250 ms cue and probe, different group of listeners).

(E) Sensitivity to vibrato versus temporal position of vibrato onset (equal-sized bins of uniformly distributed onset times) in the cued and uncued voices for the six best streamers in experiment 2B. Only trials in which the stream-segregation task was performed correctly are included. The gray bar below depicts the time course of the mixture; regions matching the cue and probe are in dark gray.

Figure 3E shows vibrato detection over time in the cued and uncued voices for the six good streamers from experiment 2B. Notably, the attentional bias seen in good streamers did not change significantly over the course of the stimulus. A two-factor within-subject ANOVA on vibrato detection for the good streamers showed a main effect of whether vibrato was in the cued or uncued voice ($F(1,5) = 25.7$, $p < 0.005$) but no effect of the time at which vibrato occurred ($F(4,20) = 1.40$, $p < 0.27$) and no interaction ($F(4,20) = 0.68$, $p < 0.62$). These data suggest that successful streaming entails attentional selection throughout the duration of the stimulus, providing further evidence for attentive tracking.

## Experiment 3: Effect of Speech-like Discontinuities

Natural speech consists not only of voiced sounds produced using the vocal folds (e.g., vowels), but also unvoiced sounds (e.g., certain consonants) and pauses. In contrast, the stimuli in experiments 1 and 2 were continuously voiced. Given that attentive tracking would presumably fail if discontinuities between voiced segments were sufficiently long, we sought to determine whether tracking could remain effective for sources with discontinuities like those found in speech. We had subjects perform the streaming task as before, but with half of all stimuli containing discontinuities intended to mimic those found in speech. To create speech-like discontinuities, we took our usual stimulus trajectories and zeroed-out segments by drawing in alternation from the distributions of voiced and unvoiced segment durations in the TIMIT corpus (estimated using STRAIGHT [38]; Figure 4A and 4B). Streaming of stimuli with speech-like discontinuities was not different from continuous stimuli ($t(4) = 0.78$, $p = 0.48$; Figure 4C), indicating that attentive tracking is robust to such discontinuities and could play a role in the perception of natural speech.

## Experiment 4: Effect of Source Proximity

What causes streaming errors? If attention aids streaming by providing a moving pointer to the cued voice, then streaming errors could arise if the focus of attention occasionally switches onto the uncued source by accident. Such switches might be more likely if the two competing sources briefly take similar feature values at the same time, potentially because the resolution of attention might be limited and thus prone to switching onto the wrong source when it passes close by. To examine the effect of source proximity, we made stimuli where the two sources' closest pass in feature space (Figure 5A) was parametrically varied over eight steps. Subjects performed the same streaming task as in experiments 1–3.

Performance in the condition with the lowest minimum distance (0.5 semitones) was not different than chance ($t(11) = 0.39$, $p = 0.70$; Figure 5B). As the minimum distance between sources was increased, performance increased to a mean d′ of 2.1 in the highest distance condition of 7.5 semitones ($F(7,77) = 27.4$, $p < 10^{-7}$). Performance was thus tightly constrained by whether the source trajectories passed close to each other. The average distance separating trajectories also increased as minimum distance increased (bin 1, $\mu = 6.7$ semitones, SD = 2.1; bin 8, $\mu = 9.9$ semitones, SD = 0.8) but was less predictive of performance: the correlation between minimum distance and performance, partialling out average distance, was $r = 0.33$ ($p < 10^{-8}$), while the correlation between

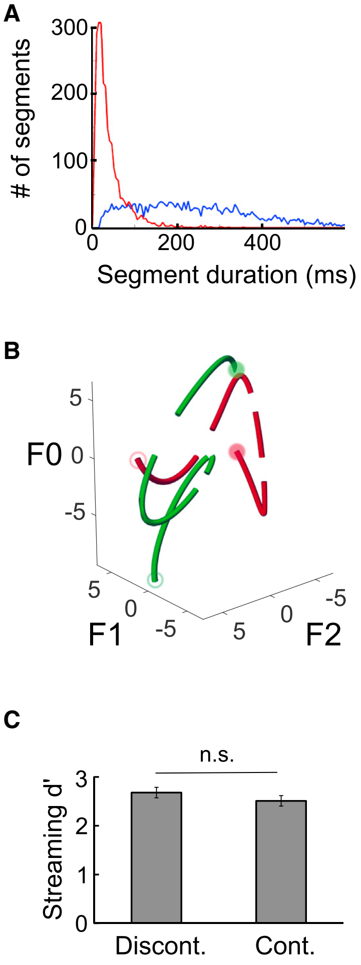

**Figure 4. Experiment 3: Speech-like Discontinuities**

(A) Histograms of the durations of discontinuities (red) and voiced segments (blue) in the stimuli.

(B) Example stimulus trajectories from experiment 3, containing speech-like discontinuities.

(C) Stream-segregation performance for discontinuous and continuous sources.

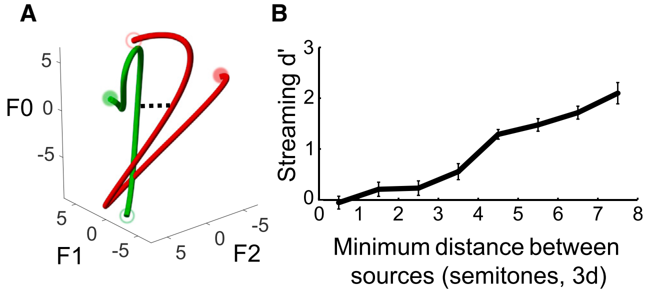

**Figure 5. Experiment 4: Source Proximity**

(A) Example stimulus trajectories; dashed line indicates the sources' closest pass in feature space.

(B) Stream-segregation performance as a function of this minimum distance between sources.

average distance and performance, partialling out minimum distance, was *r* = 0.13 (p = 0.02).

Although the results could reflect the resolution of attention, performance might also be limited by the ability to discriminate the two voices when they take on similar feature values (presumably necessary to maintain attention to one voice rather than the other). While our data cannot directly distinguish these alternatives, it is interesting to compare the results with those of prior segregation experiments using static stimuli such as double vowels [39, 40], in which performance plateaus once sources differ in F0 by more than 1 semitone [41]. In our paradigm, a difference of less than 1 semitone in F0, F1, and F2 (condition 1) yielded chance performance, and performance improved continuously as the source distance was increased well beyond a semitone. This result is thus consistent with the possibility that performance was partly limited by attention-specific resolution limits, though this is difficult to prove using our current paradigm. Regardless of the cause, the effects of minimum distance place a pronounced limit on stream segregation.

## Experiment 5: Stream Segregation of Sources Varying in Just One Feature

The effects of proximity in experiment 4 raise the possibility that streaming could be additionally limited by the number of features in which sources vary. If the cued and uncued voices vary in only a single dimension, then they will necessarily pass through each other if constrained to similar ranges. In principle, listeners could utilize the smoothness of source trajectories to correctly stream through situations where two sources briefly coincide in their features. However, the poor performance at close proximities in experiment 4 suggests that this might not be the case, as does the observation by Bregman and others that crossing frequency-modulated sweeps are heard to "bounce" [1, 42].

We compared performance in our streaming task for stimuli varying in one or three dimensions (Figure 6A). In both conditions, stimuli always crossed at least once in every dimension along which they varied, but for the three-dimensional stimuli, these crossings did not occur at the same point in time (as in the preceding experiments). Replicating the results of the preceding experiments, stimuli varying in three dimensions yielded performance much better than chance (t(9) = 3.13, p < 0.01). In contrast, when stimuli varied in only one dimension, performance was not different than chance (t(9) = 1.35, p = 0.21) and was different than performance with three dimensions (t(9) = 2.42, p < 0.05; Figure 6B). These results suggest that multiple features allow accurate streaming where single features cannot (see also [42]), possibly because multiple feature dimensions make it less likely that sources will attain similar values in all features at once. These results also suggest that successful segregation of time-varying voices depends on the joint representation of multiple features rather than any single feature alone.

## DISCUSSION

Auditory scenes often contain multiple similar sound sources, complicating the processes of segregation and selection crucial to hearing out a source of interest. We designed a task to

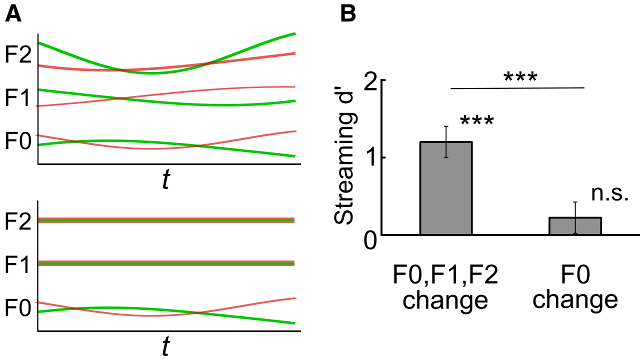

**Figure 6. Experiment 5: Sources Varying in Just One Feature**

(A) Example feature trajectories in the two conditions of experiment 5, in which sources could vary over time in either three dimensions (F0, F1, and F2) or one (F0).

(B) Stream-segregation performance for sources changing in F0, F1, and F2 and sources changing only in F0.

measure stream segregation under such conditions and to probe whether it would depend on a listener's ability to track sound sources with their attention. In a first experiment, we found that competing sources can in fact stream without distinguishing features. In a second experiment, we found that successful streaming was associated with better detection of perturbations in the cued voice than in the uncued voice. This detection bias was present throughout the time course of the stimulus. Because perturbations in the cued and uncued voices had the same distribution in feature space, it is unclear how to explain this detection bias without supposing a locus of attention that followed the cued voice as it evolved over time. In three additional experiments, we found that streaming was robust to speech-like discontinuities in the sources but that it broke down when the source trajectories passed close to each other and that it was nearly impossible if they varied in only a single feature dimension. The results thus suggest that attentive tracking could contribute to the segregation of natural speech and suggest some of the factors that could limit real-world segregation performance.

Our experiments differ from most prior studies in presenting sources without distinguishing features that could otherwise guide streaming and selection (e.g., "A is always higher than B") [1, 13–16, 24–30]. If sources do not have distinguishing features, streaming and selection must instead rely on the source trajectories, for instance on their continuity (e.g., "A(t) is closer to A(t − 1) than B(t − 1)"). It may be the case that streaming and attention usually rely on source trajectories in this way under a wide range of conditions, but because most studies use stimuli with distinguishing features, a simpler explanation has usually been available.

## The Role of Attention in Stream Segregation

The extent to which attention affects streaming is a topic of ongoing debate [24–29, 43, 44]. Some studies have argued that stream segregation can occur for unattended sources [25, 28]; other streaming phenomena are known to be sensitive to attention [26, 27, 44]. For example, if presented with sequences of alternating high and low tones, listeners can guide stream segregation and choose to hear the stimulus as one or two streams [27]. Our results suggest that attention can guide stream segregation by tracking the target source with a moving locus of attention, causing it to be grouped over time. Evidence for attentive tracking came from the finding that good streamers showed an attentional bias toward the cued voice while poor streamers were equally good at detecting perturbations to the cued and uncued voices. However, since even the poor streamers streamed well above chance, it would appear that attentional bias to the cued voice is helpful but not completely necessary for streaming in our task. One possibility is that listeners have some ability to maintain both source trajectories (i.e., to stream) even when they are unable to fully select the cued voice.

We also found that good streamers were no worse than poor streamers at detecting vibrato in the uncued voice, i.e., that the attentional bias came from enhanced vibrato detection in the cued voice. One explanation is that the good streamers enhanced the representation of the cued voice without suppressing the uncued voice. However, it is also possible that the good streamers suppressed the uncued voice while also being better overall at vibrato detection, with the two effects offsetting to produce similar performance to the poor streamers for the uncued voice. More work will be needed to disentangle these possibilities.

Even with the aid of attentive tracking, we found that streaming failed when source trajectories coincided. The continuity of source trajectories could in principle have been used to correctly stream them even when they passed close to one another, but this predictability is evidently not exploited by the auditory system. Similar results are present in work by others. For example, Bregman demonstrated that concurrent ascending and descending melodies are heard to "bounce" off each other rather than pass through each other [1], and Culling and Darwin found that bouncing could be eliminated if the two streams were given different timbres [42]. We have informally observed bouncing to be robust to the trajectories' angle of intersection and to discontinuities inserted at the point of intersection (up to several hundred milliseconds in duration). Our results suggest that this reflects a potentially general effect of source proximity which persists even under conditions of attentive tracking.

## Relation to Visual Attentive Tracking

Although to our knowledge the present study provides the first unambiguous evidence for attentive tracking in audition, analogous phenomena have been studied in vision for decades. For example, many studies have presented visual displays in which several identical items move along independent spatial trajectories [33, 35–37]. Attentively tracking one or more target items maintains awareness of their trajectories, allowing the target to be identified when the items on the display stop moving. The properties of visual attentive tracking are relatively well established and include constraints due to speed [45, 46], object similarity [47], crowding [48], and capacity limits [37]. One avenue of future work will be to investigate whether auditory and visual attentive tracking exhibit functional parallels, potentially reflecting shared mechanisms.

Attentive tracking in vision has most often been studied using targets that move through space, and auditory attentive tracking might also occur under such circumstances [49]. We chose to examine attentive tracking of a sound source's acoustic features rather than its location in physical space because of the potential relevance of these features to the cocktail party problem (Figure 1) and the challenges of rendering realistic spatial motion for complex sounds. However, visual attentive tracking is also not limited to tracking though physical space. In a study closely analogous to ours, Blaser et al. [34] asked subjects to track one of two spatially overlapping gratings that changed smoothly in three feature dimensions (orientation, frequency, and hue). The authors found that the gratings could be tracked through feature space despite the absence of any consistent distinguishing feature by which they could be individuated. Our results indicate that both visual and auditory objects can be tracked in this way.

Our study examined the role of attentive tracking in segregating and selecting similar concurrent sources, but attentive tracking could be advantageous under other conditions as well. For example, if a single speaker is talking over a noisy background, attentively tracking the target could potentially improve the extraction of its detail even if the speaker is unlikely to be confused with another sound source. Attentive tracking in such conditions could be another fruitful topic for future studies.

## EXPERIMENTAL PROCEDURES

### Stimuli and Design

Source trajectories were 2 s in duration with the exception of experiment 2B (3 s). These durations were long enough to demonstrate tracking, yet short enough to yield a large number of trials (320 trials over a 1 hr session). Although natural speech utterances frequently exceed this duration, 2 and 3 s are well within the range of typical spoken English sentences. The cue consisted of the initial 500 ms of the cued voice, and the probe was the last 500 ms of the cued or uncued voice (with the exception of experiment 2B, with 250 ms cue/probe). Cue and probe durations were chosen to be long enough to clearly identify the voice from which they came, yet short enough that the streaming task could not be performed with a simple comparison of the cue and probe. Stimuli were generated by Klatt synthesis [50], in which the instantaneous values for F1 and F2 formed the poles of two cascaded filters intended to simulate vocal-tract resonances. Stimuli were synthesized at a sample rate of 8 kHz with 16-bit resolution. 100 ms linear ramps were applied to the onset and offset of cues, probes, and mixtures. For facilitation of segregation, the cued voice began 50 ms before the uncued voice in the mixtures.

Our design relied on the use of pairs of source trajectories that crossed each other in each feature dimension, such that the cued voice could not be selected by attending to any particular value of any feature. This was achieved by generating many trajectories that smoothly changed directions over time, and selecting pairs of trajectories that crossed at least once in each dimension during the middle portion of the mixture (i.e., excluding the regions corresponding to the cue and probe). The trajectory of each feature of each source stimulus was generated (independently, so that features did not covary) by sampling an excerpt of Gaussian noise (500 Hz sampling rate) and filtering it between 0.05 and 0.6 Hz (by setting the amplitudes of frequencies outside this range to zero in the frequency domain). The chosen band limits resulted in trajectories that were not monotonic and that could change directions as many as three times over a 2 s duration, increasing the likelihood that pairs of trajectories would cross. Pilot results indicated that streaming performance was somewhat worse for faster trajectories, though well above chance.

Trajectories for each feature were scaled and centered to cover a physiologically appropriate range, spanning 100–300 Hz for F0, 300–700 Hz for F1, and 800–2,200 Hz for F2. Feature means and SDs (expressed in semitones from the mean) were as follows: F0, $\mu$ = 206.2 Hz, SD = 3.9 semitones; F1, $\mu$ = 436.0 Hz, SD = 3.3 semitones; F2, $\mu$ = 1306.8 Hz, SD = 4.1 semitones. Importantly, the distributions of distances from the cue to the correct probe and from the cue to the incorrect probe were similar (cue-probe distance, $\mu$ = 7.41 semitones, SD = 3.05 semitones; cue-foil distance, $\mu$ = 8.56 semitones, SD = 3.20 semitones), a side effect of the fact that the trajectories were generated from filtered noise. To ensure that the slight difference in cue-probe and cue-foil distance did not influence task performance, we reanalyzed experiments 1, 2A, 3, and 5 using a subset of trials (139/160) in which the average cue-probe and cue-foil distances were equated (by throwing out the trials with the largest difference between these distances until the means of the two distances were nearly equal; 7.94 and 7.93 semitones, with SDs of 2.84 and 2.85, respectively). The results of these reanalyses did not differ qualitatively from those with the full set of stimuli (all statistical tests yielded the same outcomes in both sets of analyses, and all results graphs appeared nearly identical).

A fixed set of trajectories was used in experiments 1, 2A, 3, and 5. These trajectories were selected to not pass closer than 5.5 semitones from each other (Euclidean distance in the three-dimensional feature space of F0 × F1 × F2; achieved by rejecting trajectory pairs that did not meet this criterion).

The vibrato of experiment 2 was achieved by randomly selecting a 200 ms segment of either voice's F0 trajectory and adding to it a 200 ms excerpt of a 10 Hz sinusoid 0.5 semitones in amplitude (with an initial phase of zero, such that no discontinuities were introduced). In experiment 2A, vibrato could begin anywhere from 600 to 1,300 ms from the beginning of the mixture. In experiment 2B, vibrato could begin anywhere from 600 to 2,300 ms from the beginning of the mixture (drawn from a uniform distribution in both cases).

The speech-like discontinuities of experiment 3 were created by drawing in alternation from distributions of durations of voiced and unvoiced segments in the TIMIT corpus (estimated using STRAIGHT [38]; see Figure 4A for the resulting distributions of durations) and using the resulting sequence of segment durations to gate voicing in the Klatt synthesis procedure. Voicing intensity in each segment was Hanning-windowed with ramps of duration equal to one-quarter of that segment, in order to avoid artificial-sounding onsets and offsets.

In experiment 4, stimuli were generated by the same process as in the other experiments, except that noise was filtered between 0.05 and 0.3 Hz (instead of 0.6 Hz), slowing trajectories so that they crossed just once or twice in each dimension (trajectories in experiments 1–3 and 5 could cross up to three times per dimension). The slowing served to reduce the number of close passes for each trajectory pair, such that there was one closest pass whose distance could be used to assign the trajectory to a condition. Trajectory pairs whose minimum distance fell within designated bin limits were then selected for each experimental condition in order to parameterize the minimum distance between sources. The mean minimum distance for stimuli in the first bin was approximately half a semitone, and in the last bin, 7.5 semitones (Euclidean distance in three-dimensional feature space, equivalent to 0.375 and 4.30 semitones, respectively, in each of the three feature dimensions). Bin limits (in three-dimensional semitones) were 0–1, 1–2, 2–3, 3–4, 4–5, 5–6, 6–7, and 7–8. The mean minimum distances of stimuli in these bins were (also in three-dimensional semitones) 0.65, 1.52, 2.49, 3.50, 4.45, 5.48, 6.46, and 7.45. Source trajectories were additionally constrained such that cue-probe and cue-foil distances had similar means and SDs within each condition. In each condition, we generated 150 stimuli and then removed stimuli until the difference between the cue-probe and cue-foil distances had a mean near zero. Then, from the remaining set of stimuli, subsets of 40 were drawn at random until a subset was found in which cue-probe and cue-foil distances had similar SDs as well as similar means (mean cue-probe and cue-foil distances across conditions were 8.7 and 8.5 semitones, respectively, with mean within-condition SDs of 3.4 and 3.1 semitones). It was also the case that cue-probe and cue-foil distances were similar across conditions (cue-probe distances, *F*(7,312) = 1.42, p = 0.20; cue-foil distances, *F*(7,312) = 1.06, p = 0.39). The average distance separating trajectories increased somewhat with minimum distance (bin 1, $\mu$ = 6.7 semitones, SD = 2.1; bin 8, $\mu$ = 9.9, SD = 0.8) but was less predictive of performance than minimum distance (see the Results).

### Procedure

Each experiment contained 320 trials run in eight blocks of 40 trials each. Conditions were randomly ordered across an experiment. Listeners were encouraged to take short breaks between blocks. Feedback was provided on each trial in the streaming task. No feedback was given for the vibrato-detection task (experiments 1A and 2B). Total testing time for each experiment was approximately 45 min (55 min for experiment 2B). Performance tended to improve over the first three blocks and then stabilize (see Figure S1).

Stimuli in experiments 1, 2A, 3, and 5 were generated from a fixed set of 80 voice trajectory pairs. Each pair was used twice in each condition, once with each of the two possible assignments of cued and uncued voice, giving 160 trials per condition. These same trajectory pairs were also used in the other conditions (with vibrato added in experiment 2A, with discontinuities added in experiment 3, and with F1 and F2 change removed in experiment 5; experiment 1 included another condition that is not analyzed here). Thus, stimuli for different conditions across these experiments were the same apart from the experimental manipulation. New voice trajectory pairs were generated for experiment 2B (same procedure, yielding 80 pairs of 3 s trajectories) and experiment 4 (40 pairs in each of eight bins yielding 320 unique stimuli, with the cued voice randomly chosen on each trial).

### Participants

All experiments were approved by the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology. Eight subjects (two female, mean age of 26.4 years) participated in experiment 1. Twelve subjects (seven female, mean age of 25.3 years) participated in experiment 2A. Twelve subjects (five female, mean age of 25.6 years) participated in experiment 2B. Five subjects (two female, mean age of 21.8 years) participated in experiment 3. Twenty subjects (11 female, mean age of 25.2 years) participated in experiment 4. Eight of these 20 subjects were excluded from experiment 4 due to overall d′ scores below 0.1 (mean across conditions). For the 12 subjects included in analysis, seven were female, with a mean age of 25.0 years. Ten subjects (six female, mean age of 27.2 years) participated in experiment 5. Three subjects who participated in experiment 2A subsequently participated in experiment 3, and one subsequently participated in experiment 5. Two of these repeat subjects were categorized as good streamers in experiment 2A.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes one figure and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2015.07.043.

## ACKNOWLEDGMENTS

We thank Michael Cohen, Nai Ding, and Jason Fischer for comments on the manuscript. This work was supported by a McDonnell Scholar award and an NSF CAREER award to J.H.M.

Received: April 6, 2015
Revised: June 17, 2015
Accepted: July 15, 2015
Published: August 13, 2015

## REFERENCES

1. Bregman, A.S. (1990). Auditory Scene Analysis: The Perceptual Organization of Sound (The MIT Press).
2. Darwin, C.J., and Carlyon, R.P. (1995). In Auditory Grouping. The Handbook of Perception and Cognition, *Volume 6*, B.C.J. Moore, ed. (Academic Press).
3. McDermott, J.H. (2009). The cocktail party problem. Curr. Biol. *19*, R1024–R1027.
4. Carlile, S. (2014). Active listening: speech intelligibility in noisy environments. Acoust. Aust. *42*, 90–96.
5. Alain, C., and Arnott, S.R. (2000). Selectively attending to auditory objects. Front. Biosci. *5*, D202–D212.
6. Fritz, J.B., Elhilali, M., David, S.V., and Shamma, S.A. (2007). Auditory attention–focusing the searchlight on sound. Curr. Opin. Neurobiol. *17*, 437–455.
7. Shinn-Cunningham, B.G. (2008). Object-based auditory and visual attention. Trends Cogn. Sci. *12*, 182–186.
8. Lee, A.K.C., Larson, E., Maddox, R.K., and Shinn-Cunningham, B.G. (2014). Using neuroimaging to understand the cortical mechanisms of auditory selective attention. Hear. Res. *307*, 111–120.
9. Darwin, C.J. (1984). Perceiving vowels in the presence of another sound: constraints on formant perception. J. Acoust. Soc. Am. *76*, 1636–1647.
10. Moore, B.C.J., Glasberg, B.R., and Peters, R.W. (1986). Thresholds for hearing mistuned partials as separate tones in harmonic complexes. J. Acoust. Soc. Am. *80*, 479–483.
11. Micheyl, C., and Oxenham, A.J. (2010). Pitch, harmonicity and concurrent sound segregation: psychoacoustical and neurophysiological findings. Hear. Res. *266*, 36–51.
12. McDermott, J.H., Wrobleski, D., and Oxenham, A.J. (2011). Recovering sound sources from embedded repetition. Proc. Natl. Acad. Sci. USA *108*, 1188–1193.
13. Winkler, I., Denham, S.L., and Nelken, I. (2009). Modeling the auditory scene: predictive regularity representations and perceptual objects. Trends Cogn. Sci. *13*, 532–540.
14. Shamma, S.A., Elhilali, M., and Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. Trends Neurosci. *34*, 114–123.
15. Moore, B.C.J., and Gockel, H.E. (2012). Properties of auditory stream formation. Philos. Trans. R. Soc. Lond. B Biol. Sci. *367*, 919–931.
16. Bendixen, A. (2014). Predictability effects in auditory scene analysis: a review. Front. Neurosci. *8*, 60.
17. Ding, N., and Simon, J.Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. USA *109*, 11854–11859.
18. Mesgarani, N., and Chang, E.F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. Nature *485*, 233–236.
19. Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. Neuron *77*, 980–991.
20. Darwin, C.J., Brungart, D.S., and Simpson, B.D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. J. Acoust. Soc. Am. *114*, 2913–2922.
21. Carlyon, R.P. (2004). How the brain separates sounds. Trends Cogn. Sci. *8*, 465–471.
22. Kidd, G., Jr., Arbogast, T.L., Mason, C.R., and Gallun, F.J. (2005). The advantage of knowing where to listen. J. Acoust. Soc. Am. *118*, 3804–3815.
23. Bregman, A.S., Liao, C., and Levitan, R. (1990). Auditory grouping based on fundamental frequency and formant peak frequency. Can. J. Psychol. *44*, 400–413.
24. Snyder, J.S., Alain, C., and Picton, T.W. (2006). Effects of attention on neuroelectric correlates of auditory stream segregation. J. Cogn. Neurosci. *18*, 1–13.
25. Sussman, E.S., Horváth, J., Winkler, I., and Orr, M. (2007). The role of attention in the formation of auditory streams. Percept. Psychophys. *69*, 136–152.
26. Carlyon, R.P., Cusack, R., Foxton, J.M., and Robertson, I.H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. J. Exp. Psychol. Hum. Percept. Perform. *27*, 115–127.
27. van Noorden, L.P.A.S. (1975). Temporal Coherence in the Perception of Tone Sequences (Institute for Perceptual Research).
28. Macken, W.J., Tremblay, S., Houghton, R.J., Nicholls, A.P., and Jones, D.M. (2003). Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. J. Exp. Psychol. Hum. Percept. Perform. *29*, 43–51.
29. Gutschalk, A., Rupp, A., and Dykstra, A.R. (2015). Interaction of streaming and attention in human auditory cortex. PLoS ONE *10*, e0118962.

30. Demany, L., Erviti, M., and Semal, C. (2015). Auditory attention is divisible: segregated tone streams can be tracked simultaneously. J. Exp. Psychol. Hum. Percept. Perform. *41*, 356–363.

31. Latinus, M., and Belin, P. (2011). Human voice perception. Curr. Biol. *21*, R143–R145.

32. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., and Pallett, D.S. (1993). TIMIT acoustic-phonetic continuous speech corpus (Linguistic Data Consortium). https://catalog.ldc.upenn.edu/LDC93S1.

33. Pylyshyn, Z.W., and Storm, R.W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. Spat. Vis. *3*, 179–197.

34. Blaser, E., Pylyshyn, Z.W., and Holcombe, A.O. (2000). Tracking an object through feature space. Nature *408*, 196–199.

35. Intriligator, J., and Cavanagh, P. (2001). The spatial resolution of visual attention. Cognit. Psychol. *43*, 171–216.

36. Scholl, B.J., Pylyshyn, Z.W., and Feldman, J. (2001). What is a visual object? Evidence from target merging in multiple object tracking. Cognition *80*, 159–177.

37. Alvarez, G.A., and Franconeri, S.L. (2007). How many objects can you track? Evidence for a resource-limited attentive tracking mechanism. J. Vis. *7*, 14.1–10.

38. Kawahara, H., and Morise, M. (2011). TANDEM-STRAIGHT: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. Sadhana *36*, 713–722.

39. Culling, J.F., and Darwin, C.J. (1993). Perceptual separation of simultaneous vowels: within and across-formant grouping by F0. J. Acoust. Soc. Am. *93*, 3454–3467.

40. Assmann, P.F., and Summerfield, Q. (1990). Modeling the perception of concurrent vowels: vowels with different fundamental frequencies. J. Acoust. Soc. Am. *88*, 680–697.

41. Meddis, R., and Hewitt, M.J. (1992). Modeling the identification of concurrent vowels with different fundamental frequencies. J. Acoust. Soc. Am. *91*, 233–245.

42. Culling, J.F., and Darwin, C.J. (1993). The role of timbre in the segregation of simultaneous voices with intersecting F0 contours. Percept. Psychophys. *54*, 303–309.

43. Alain, C., and Izenberg, A. (2003). Effects of attentional load on auditory scene analysis. J. Cogn. Neurosci. *15*, 1063–1073.

44. Cusack, R., Deeks, J., Aikman, G., and Carlyon, R.P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. J. Exp. Psychol. Hum. Percept. Perform. *30*, 643–656.

45. Verstraten, F.A.J., Cavanagh, P., and Labianca, A.T. (2000). Limits of attentive tracking reveal temporal properties of attention. Vision Res. *40*, 3651–3664.

46. Howard, C.J., and Holcombe, A.O. (2008). Tracking the changing features of multiple objects: progressively poorer perceptual precision and progressively greater perceptual lag. Vision Res. *48*, 1164–1180.

47. Makovski, T., and Jiang, Y.V. (2009). Feature binding in attentive tracking of distinct objects. Vis. Cogn. *17*, 180–194.

48. Shim, W.M., Alvarez, G.A., and Jiang, Y.V. (2008). Spatial separation between targets constrains maintenance of attention on multiple objects. Psychon. Bull. Rev. *15*, 390–397.

49. Huddleston, W.E., Lewis, J.W., Phinney, R.E., Jr., and DeYoe, E.A. (2008). Auditory and visual attention-based apparent motion share functional parallels. Percept. Psychophys. *70*, 1207–1216.

50. Klatt, D.H. (1980). Software for a cascade/parallel formant synthesizer. J. Acoust. Soc. Am. *67*, 971–995.

Current Biology

Supplemental Information

# Attentive Tracking of Sound Sources

Kevin J.P. Woods and Josh H. McDermott

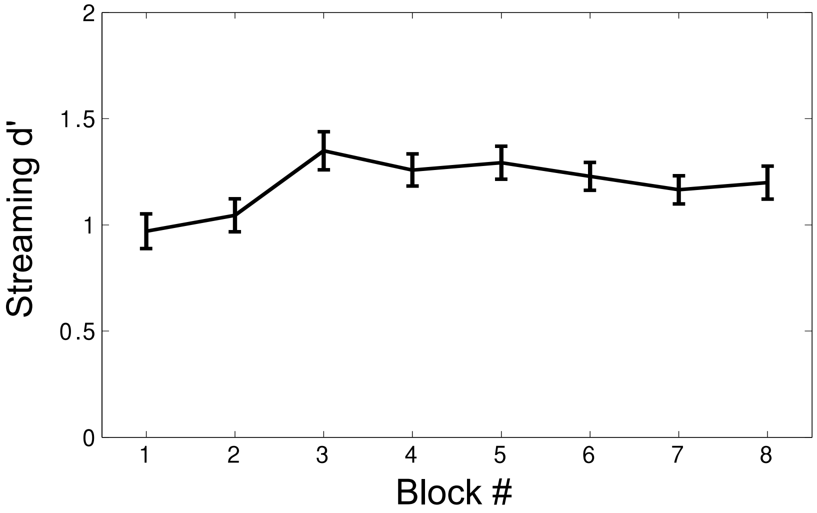

**Figure S1.** Block-by-block Performance, Related to Figure 2
Data is pooled across Experiments 1, 2A, 2B, and 4 (n=44); Experiments 3 (discontinuities) and 5 (single-feature only) were omitted from this analysis because they had conditions interleaved which did not have the usual streaming stimuli. Overall performance is relatively low because 12 of the subjects were from Experiment 4, which had many conditions with low proximity. Error bars show within-subject SEMs.