Published in final edited form as:

J Exp Psychol Hum Percept Perform. 2012 August ; 38(4): 998–1013. doi:10.1037/a0026789.

Feature Assignment in Perception of Auditory Figure

Melissa K. Gregg

Stony Brook University

Arthur G. Samuel

Stony Brook University

IKERBASQUE, Basque Foundation for Science, Bilbao 48011 Spain

Basque Center on Cognition Brain and Language, Donostia-San Sebastian 20009 Spain

Abstract

Because the environment often includes multiple sounds that overlap in time, listeners must segregate a sound of interest (the auditory figure) from other co-occurring sounds (the unattended auditory ground). We conducted a series of experiments to clarify the principles governing the extraction of auditory figures. We distinguish between auditory “objects” (relatively punctate events, such as a dog's bark) and auditory “streams” (sounds involving a pattern over time, such as a galloping rhythm). In Experiments 1 and 2, on each trial two sounds -- an object (a vowel) and a stream (a series of tones) -- were presented with one target feature that could be perceptually grouped with either source. In each block of these experiments, listeners were required to attend to one of the two sounds, and report its perceived category. Across several experimental manipulations, listeners were more likely to allocate the feature to an impoverished object if the result of the grouping was a good, identifiable object. Perception of objects was quite sensitive to feature variation (noise masking), whereas perception of streams was more robust to feature variation. In Experiment 3, the number of sound sources competing for the feature was increased to three. This produced a shift toward relying more on spatial cues than on the potential contribution of the feature to an object's perceptual quality. The results support a distinction between auditory objects and streams, and provide new information about the way that the auditory world is parsed.

Keywords

Auditory Figure; Auditory Scene Analysis; Auditory Perceptual Organization

Listeners are often faced with a substantial amount of sensory information from multiple sources overlapping in time and in space. Normal functioning in the world relies on the ability to successfully parse these complex auditory scenes. For example, in order to communicate successfully, a diner in a restaurant must segregate incoming speech from

Address correspondence to Arthur Samuel, Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69, 20009 Donostia, Spain. a.samuel@bcbl.eu.

Melissa K. Gregg: Department of Psychology, Stony Brook University; Arthur G. Samuel: (a) IKERBASQUE, (b) Basque Center on Cognition Brain and Language, and (c) Department of Psychology, Stony Brook University.

Melissa K. Gregg is now at the department of Psychology, University of Nevada, Las Vegas.

Publisher's Disclaimer: The following manuscript is the final accepted manuscript. It has not been subjected to the final copyediting, fact-checking, and proofreading required for formal publication. It is not the definitive, publisher-authenticated version. The American Psychological Association and its Council of Editors disclaim any responsibility or liabilities for errors or omissions of this manuscript version, any version derived from this manuscript by NIH, or other third parties. The published version is available at www.apa.org/pubs/journals/xhnp

other co-occurring sounds, such as background music, dishes clanging, and speech coming from a nearby table. Determining how human perceivers accomplish this feat is essential to a basic understanding of auditory perception.

The foundational research on this issue has been conducted within the framework of auditory scene analysis (ASA), most prominently by Bregman and his colleagues (see Bregman, 1990) for a review). Research from Bregman's lab has revealed that auditory components are perceptually grouped based on low-level Gestalt-based principles, such as similarity (Bregman & Campbell, 1971), temporal proximity (e.g., Dannenbring & Bregman, 1976), and spatial proximity (e.g., Steiger & Bregman, 1982), and by higher-level factors, such as prior knowledge and attention (Bregman, 1990). The result of perceptual grouping in the auditory modality has been called a "stream" by Bregman and his colleagues. "Stream" has been preferred over "sound" or "object" because an event may produce multiple sounds, e.g., the event of walking produces footsteps composed of several distinct sounds (Bregman, 1990).

ASA theory can account for many perceptual results, but there are a number of apparently problematic findings for the theory. For example, there are clear exceptions to ASA's principle of exclusive allocation: "The exclusive allocation principle says that a sensory element should not be used in more than one description at a time" (Bregman, 1990, p. 12) -- once a feature has been assigned to one sound source, it cannot be concurrently assigned to another, separate source. Though this is a core assumption of ASA theory, there are many cases in which an auditory feature does contribute to two objects at the same time (e.g., Fowler & Rosenblum, 1990; Rand, 1974).

A less appreciated problem with traditional ASA theory is that most of the research on auditory scene analysis has been conducted on simple, repeating auditory patterns (streams), rather than objects. ASA theory does not differentiate between streams and objects, but there may be important differences between the different types of sources (Remez, Rubin, Burns, Pardo, & Lang, 1994). The typical streams utilized in ASA experiments are perceptual groups for which identification depends on repetition (e.g., Bregman & Campbell, 1971), e.g., the identification of one or another rhythm. On the other hand, the identification of an auditory object, such as a cat's meow, does not depend on repetition, and thus principles derived from work that heavily focuses on streams may not be appropriate for auditory object recognition. While ASA research has provided rich information about how features are assigned to streams, the corresponding principles that guide how features are allocated to objects are not well specified. Thus, one major goal of this project is to examine potential differences in the perceptual organization of objects and streams.

The distinction between objects and streams will have some boundary cases, but this does not undercut the potential value of understanding how listeners perceptually organize objects versus streams. We should note that although Bregman has not distinguished between objects and streams, he has noted potential differences in the perceptual organization of sequential sounds (akin to streams) versus simultaneous sounds (akin to objects), and these have been confirmed empirically (see Shinn-Cunningham et al., 2007).

Although ASA theory has been most influential, there are other approaches that have produced interesting findings. Notably, the "ecological acoustics" approach is based on the fact that sound-emitting objects structure the sound reaching a listener's ear, and that this structure provides the information needed by the listener to recover the identity of the sound source (see Carello, Wagman, & Turvey, 2005, for an excellent discussion of this approach). For example, Kunkler-Peck and Turvey (2000) asked listeners to identify the shape of an unseen metal plate, based only on the sound that was made when the plate was struck, and

found that listeners performed well above chance. Similarly, listeners are relatively good at judging the length of wooden dowels, based only on the sounds made when they fall (Carello, Anderson, & Kunkler-Peck, 1998). The ecological acoustics literature has primarily focused on the idea that particular sound sources can be identified from the pattern of sound reaching the listener, even if the sound is degraded, as long as the relevant aspects of the pattern remain (e.g., sine-wave speech, Remez et al., 1994).

Some research that is closer to the ASA tradition has been more concerned with how the listener sorts acoustic cues when there are two potentially competing sounds in the environment. As it happens, these studies have often involved one stream and one object, even though no explicit distinction between objects and streams was made (beyond the ecological acoustics distinction between sounds with a determinable source and sounds with no determinable source). For example, Darwin and colleagues (e.g., Darwin, 1983; Darwin & Hukin, 1998; Darwin, Pattison, & Gardner, 1989) presented listeners with two sounds, a vowel (object) and a tone sequence (stream). The tone sequence was composed of tones that were the same frequency as the fourth harmonic of the vowel. When the vowel and the sequence were simultaneously presented, the sequence could capture the harmonic out of the vowel, changing the perceived category of the vowel from /e/ to /I/. Because listeners were instructed to attend to and identify the vowel, the perceptual contribution that the harmonic might have made to the stream is not known.

Recent work using stimuli of this type has measured the perception of both the object and the stream, with an emphasis on how the listener separates auditory figure and ground. Shinn-Cunningham, Lee, and Oxenham (2007) conducted trials in which either the vowel or the tone sequence was in perceptual figure (listeners attended to the vowel only or to the sequence only). As in Darwin's research, the sequence was composed of tones that were the same frequency as the fourth harmonic of the vowel /e/. If the feature were incorporated into the vowel, the category of the vowel would change from /I/ to /e/, and if the feature were incorporated into the sequence, its rhythm would change from "galloping" to "even." The feature could be presented at the same spatial location as the vowel (or tone sequence) to increase the probability of perceptual grouping; or, to reduce the likelihood of the feature grouping with one of the sources, the feature could be presented at a different spatial location.

Shinn-Cunningham et al. (2007) found that in the attend-tone block, listeners heard the feature as contributing to the tone sequence in two of the three spatial conditions: when it was presented at the same spatial location as the sequence and when it was presented in a spatial location that did not match either the sequence or the vowel. The feature did not contribute to the perception of the tone sequence when it was presented at the same spatial location as the vowel. This pattern shows that the spatial cues were not strong enough to prevent the feature and the stream from binding, except in the most extreme spatial configuration in which the feature was presented at the same spatial location as the object in the background. In this same condition in the attend-vowel block, the feature also did not group with the vowel -- the vowel was perceived as /I/. Because the feature did not contribute to either percept (the tone sequence or the vowel), it was as if the feature disappeared from the mixture, resulting in a case of feature non-allocation.

In fact, in the attend-vowel block, the feature never contributed strongly to the perception of the vowel. The asymmetry in the results of Shinn-Cunningham et al. (2007) suggests that this testing procedure may offer a method for examining potential differences between the perceptual organization of objects and streams. Because traditional ASA research has not distinguished between objects and streams, the generally accepted view of how auditory figure and ground are established may apply primarily to the special case of streams. In the

current study, we adapt Shinn-Cunningham et al.'s method to examine the differences in how features are assigned to objects and streams. The goal is to uncover general perceptual principles that determine how features are assigned to sounds in auditory figure.

A second major goal of this project was to pursue the curious case of a feature “disappearing” under certain competitive conditions (Shinn-Cunningham et al., 2007). It is surprising that when two sounds are competing for the same feature, an object (a vowel) in perceptual figure will not group with the feature despite perfect spatial alignment between that object and the feature (Shinn-Cunningham et al., 2007). This result contrasts with an extensive line of research demonstrating that an object in figure will group with a feature despite evidence that the feature belongs to another source (e.g., Cutting, 1976; Darwin & Hukin, 1998; Rand, 1974; Shinn-Cunningham & Wang, 2008). For example, demonstrations of duplex perception have shown that an object will group with a feature despite evidence based on spatial location that the feature does not belong to the object (Fowler & Rosenblum, 1990; Rand, 1974).

Given all of these prior demonstrations, why then is there a recent case (Shinn-Cunningham et al., 2007) in which an object in figure will not group with a feature even when the spatial cues are perfectly aligned? We suggest that perhaps an object will not group with the feature when it is already an identifiable object, i.e., when it has a clear categorical representation in memory, and incorporating the feature into the object would change the category of that object. This principle could help maintain perceptual stability in an unstable, constantly changing auditory world. A principle of this sort could make sense for objects, as they have relatively well-specified representations (e.g., what a cat's meow sounds like). This principle might not apply as naturally to streams, however, because streams are more “constructed” -- a stream is a sequence with patterned repetition that would not typically correspond to an entity stored in memory (e.g., no particular number of footsteps would presumably be represented, as this will vary widely from context to context).

There is support for our suggestion that the “goodness” of an object is important in perceptual organization: top-down processes have been claimed to guide perception of sensory input to the most likely, or best, interpretation (Davis & Johnsrude, 2007).

And, there is neurological evidence showing distinct brain areas for recognizable and unrecognizable sounds (Husain, McKinney, & Horwitz, 2006; Scott, Blank, Rosen, & Wise, 2000). Thus, we make the prediction that perceptual organization will differ with the “goodness” of an object. Specifically, we predict that features are more likely to be perceptually grouped with auditory objects when the resulting perception is a “good”, identifiable object. This distinction is clearly one of degree, not absolute, as there is variability in the goodness of auditory objects as well.

Experiment 1

In Experiment 1a, we replicate the conditions in which Shinn-Cunningham et al. (2007) observed the “disappearance” of a feature. In Experiment 1b, we then test whether an object in perceptual figure is significantly more likely to group with a feature if incorporating the feature produces a better, more identifiable object.

Experiment 1a: Replication of Shinn-Cunningham et al. (2007)

Method

Participants: Twenty four undergraduates from Stony Brook University (SBU) with no reported hearing deficiencies participated in this experiment. Their mean age was 22.7 years. All listeners received course credit as compensation for their participation.

Stimuli: Stimuli consisted of a vowel, a tone sequence, and a target feature that could contribute to the perception of each sound (see Figure 1a). If the feature is incorporated into the vowel, the extra energy near the vowel's first formant pushes the percept from /I/ (as in “pit”) to /ε/ (as in “pet”). The presence/absence of the feature also affects the rhythmic properties of the tone sequence. If the feature is perceived as part of the tone sequence, the sequence has an “even” rhythm; otherwise it is perceived as “galloping.”

A 60 ms vowel was synthesized at a sampling rate of 44.1 kHz with a source filter synthesis algorithm using Praat (Boersma & Weenik, 2007). The fundamental frequency of the vowel was 125 Hz. The first three formants were 490, 2100, and 2900 Hz, and their corresponding bandwidths were 90, 110, and 170 Hz. The fourth harmonic, 500 Hz, was filtered out of the vowel (see Figure 1b). This 500 Hz component was the target feature and was also used as each individual tone in the tone sequence. The amplitudes of the vowel and the tone were equated.

The vowel and the 500 Hz tone were combined to create 3-second long repeating sequences. There were 10 repetitions of a 300 ms sequence consisting of two 60 ms tones and a 60 ms vowel with a 40 ms silent interval after each component of the sequence (see Figure 1a). In the target present condition, the target feature was played at the same time as the vowel.

Spatial cues were manipulated by presenting stimuli over two speakers: one was positioned straight ahead of the listener, and one was 45° to the right of the listener. The amplitude of the stimuli was randomly roved over a 14 dB range to prevent listeners from using amplitude as a labeling cue; this manipulation overcame any amplitude cue that incorporating the feature with the vowel might have produced. All stimuli were presented at a comfortable listening level, never exceeding 80 dB.

Conditions: Four spatial configurations of the vowel and tones were used in the experimental conditions (see Figure 1c, bottom-left panel). There were three single-sound conditions for the vowel, and three for the tones (Figure 1c, right side), two of which were used as baselines for the experimental conditions (as in Shinn-Cunningham et al., 2007). In these two single-sound conditions, the vowel or the tone sequence was presented either with the feature at the same spatial location (target present), or without the target feature (target absent). These single-sound prototypes served as controls to confirm that the feature grouped with the tone sequence in the target present condition to produce the even rhythm percept, or with the vowel to produce the /ε/ percept. In the third (non-baseline) single-sound condition, each sound was presented with the feature at a different spatial location (target location doesn't match).

In the two-sound conditions, both the vowel and the tone sequence were presented without the target feature (control = no target) and with the target feature (target present condition). As can be seen in Figure 1c, for the target present condition, across the four versions, the location of the target matched one sound, both sounds, or neither sound. The spatial configurations depicted for each condition in Figure 1c were the only ones used in this experiment.

Procedure: There were three sequential phases: familiarization, training, and test. During the **familiarization** phase, the (single-sound baseline) target present and target absent versions of both the vowel and the tone sequence were presented while the correct label was displayed on the computer screen. For example, the vowel /I/ played while the label “i as in pit” was displayed. The vowel and the tone sequence were presented in separate blocks; within each block there was a one second interstimulus interval. There were 12 familiarization trials total (three repetitions of target present/absent versions of the vowel

and the tone sequence). Listeners were given the opportunity to repeat the familiarization phase before moving on with the experiment.

Listeners then completed a **training** phase in which they labeled each sound. Listeners responded (in this phase and in the test phase that followed) by pressing one of two buttons on a button board labeled “i” or “e” for the vowel training and “galloping” or “even” for the tone sequence training. The same labels were also displayed on the computer screen. Listeners received feedback during training: “Correct” or “Wrong” appeared after each response. The vowels and the tone sequences were presented in separate blocks, with a 500 ms inter-trial interval in each block. There were at least 24 training trials (6 repetitions of /I/, /e/, “even”, and “galloping”). Listeners repeated the training phase until they could achieve a 90% accuracy rate. Most listeners required one repetition of the vowel training block and no repetitions of the tone training block.

Next, listeners completed the **test** phase, which included an attend-tones block and an attend-vowel block. In the attend-tones block, listeners were instructed to listen to the tone sequences while ignoring any other sound that was presented. They were told to listen to the sequence on each trial and to label the sequence as “galloping” or as “even.” In the attend-vowel block, listeners were instructed to listen only to the vowel, while ignoring any other sounds. They were instructed to label the vowel they heard on each trial as “/I/ as in pit” or as “/e/ as in pet.”

In each block, the single- and two-sound conditions were randomly intermixed. There were 8 conditions within each block: There were 3 single-sound conditions: target absent, target present, and target present at a different location (see the right side of Figure 1c). There were 5 two-sound conditions: target absent (‘control’ on the top-left of Figure 1c), target present at the same location as the sound in figure, target present at the same location as the sound in ground, target present at the same location as both sounds, and target present at a location matching neither sound. There were 20 repetitions of each of the eight conditions, resulting in 160 trials per block. Brief rest breaks were provided half-way through each block and in between blocks. The next trial was launched 500 ms after each response, timing out after five seconds.

Order of blocks -- attend-tone or attend-vowel -- was counterbalanced across listeners. The experiment was conducted in a sound attenuated chamber and took about 40 minutes to complete.

Results and Discussion—As in the work by Shinn-Cunningham et al. (2007), an a priori criterion was set (here and in Experiments 2–3) excluding data from listeners who could not reliably distinguish between the single-sound target present and target absent prototypes: There were 8 conditions within each block: There were 3 single-sound conditions: target absent, target present, and target present at a different location (see the right side of Figure 1c). There were 5 two-sound conditions: target absent (‘control’ on the top-left of Figure 1c), target present at the same location as the sound in figure, target present at the same location as the sound in ground, target present at the same location as both sounds, and target present at a location matching neither sound. There were 20 repetitions of each of the eight conditions, resulting in 160 trials per block. Brief rest breaks were provided half-way through each block and in between blocks. The next trial was launched 500 ms after each response, timing out after five seconds. A d' score of at least 0.7 between the target present and target absent prototypes was required in order for listeners to be included in the data analysis. All listeners met the criterion with the tone stimuli. Data from 11 listeners were excluded because those listeners failed to meet the criterion with the vowel stimuli. While a subject loss of nearly half (11 out of 24) is high,

this rate is essentially the same as the attrition rate in Shinn-Cunningham et al; in their study, half of the subjects were excluded for failing to meet the vowel criterion. In the current study, we excluded data from one additional listener due to a failure to respond on 85% of the test trials. The following analyses were conducted on data from the remaining twelve listeners.

Accuracy (percent correct labeling) was computed for each listener in each condition. As in the work by Shinn-Cunningham et al. (2007), signal detection analysis was used to analyze the accuracy scores. This analysis provides a measure of perceptual distance between each stimulus and the single-object prototypes.

Two d' scores were calculated. The $d'_{\text{present:absent}}$ score measured the perceptual distance between the single-object target present and target-absent prototypes, i.e., the tone sequence or the vowel presented alone with (present) or without (absent) the 500 Hz tone. The $d'_{\text{stimulus:absent}}$ score measured the perceptual distance between a given stimulus, e.g., perception of the vowel in the Two-sound condition where the target matches the location of both sounds, and the single-object target absent prototype, e.g., perception of the vowel when presented alone without the 500 Hz tone. A δ score indicated whether the stimulus was perceived to be more like the target present or the target absent prototype, with: $\delta = d'_{\text{stimulus:absent}} / d'_{\text{present:absent}}$ (see Lee & Shinn-Cunningham, 2008a for more details). A δ value of < 0.5 indicates that a stimulus was perceived as more like the target absent version of the sound, while a $\delta > 0.5$ indicates that it was perceived as more like the target present version. As in Shinn-Cunningham et al., a normalization procedure was used to scale target-absent and target-present prototypes to 0 and 1, respectively.

The results of this decision theory analysis are depicted in Figure 2. The Single-sound (no match¹) condition (right-most point in both panels) replicated Shinn-Cunningham et al. (2007). Both the vowel (bottom panel; $\delta = 0.58$) and the tone sequence (top panel; $\delta = 0.81$) grouped with the target feature, despite being presented at separate spatial locations. It appears that listeners are tolerant of physical inconsistencies between features and objects or streams when there are no other sounds in the background that the feature could be attributed to.

The results of the Two-sound conditions also replicated Shinn-Cunningham et al. (2007). The critical result in that study was that the target feature did not contribute to either the vowel or the tone sequence when it was presented at the same spatial location as the vowel (referred to as the *critical condition* here and in Experiments 2–3). The results for the critical condition are circled in Figure 2. Consistent with Shinn-Cunningham et al., the target feature did not contribute to the perception of either the vowel or the tone sequence. As the top panel of Figure 2 shows, the feature grouped with the tone sequence at every spatial location except for the critical condition (mean $\delta = 0.41$). And, as shown in the bottom panel, the target feature also did not contribute to the vowel in the critical condition (mean $\delta = 0.41$). Because the target feature did not contribute to the percept of the tone sequence or the vowel in this condition, it is as if the feature disappeared from perception. This effect is of course not absolute, nor would one expect it to be. What is striking both here and in the original report by Shinn-Cunningham et al. is that a feature that is perceptually active in the Single-sound conditions does not get systematically allocated to either source when there are multiple possible attractors. Thus, just as it is theoretically problematic that features can get used by more than one source (violating the principle of exclusive allocation), it is important theoretically that features may get used by neither source. In the following experiments, we explore several factors that may control the allocation process.

¹Match refers to spatial location

Experiment 1b: The Feature Reappears

Shinn-Cunningham et al. (2007) and Experiment 1a have shown that when an object and a stream are competing for the same feature, there are certain spatial configurations that will result in a failure to allocate the feature to either source. Note that under these conditions one of the sounds in perceptual figure (the vowel) fails to group with the feature despite perfect spatial alignment of the two. Experiment 1b tests whether this non-allocation occurred because the object in perceptual figure was already an identifiable object without the feature. This experiment uses the same general procedures as Experiment 1a, but critically, in this version, the vowel was only an identifiable object when it grouped with the target feature.

Method

Participants: Sixteen SBU undergraduates with no reported hearing deficiencies served as listeners for course credit. Their mean age was 19 years. All listeners received course credit as compensation for their participation.

Stimuli: As in Experiment 1a, a tone sequence, a vowel, and a target feature were used. The tone sequence was perceived as rhythmically even with the target feature, and as galloping without the target feature. However, instead of a vowel that is perceived as /I/ without the feature and as /e/ with the feature, the vowel was perceived as a non-identifiable speech sound without the target feature and as /I/ with the feature. Figure 3 depicts this stimulus. The non-identifiable speech sound was synthesized using the same parameters as in Experiment 1a, but without the first formant (a critical formant for vowel identification). Thus, the sound had a fundamental frequency of 125 Hz and formants at 2100 and 2900 Hz with corresponding bandwidths of 110 and 170 Hz. As can be seen in Figure 3, the target in this experiment was the first formant of the vowel used in Experiment 1a, which peaked at 490 Hz (recall that the target in Experiment 1a was the 500 Hz harmonic of the vowel). The target was created by synthesizing a vowel with only the first formant (490 Hz with a bandwidth of 90 Hz) on a fundamental frequency of 125 Hz. As in Experiment 1a, instances of the target were used as the individual members of the tone sequence. That is, the tone sequence consisted of a number of repetitions of the shaded area shown in Figure 3. As such, it was not a pure tone sequence, but because the target was a narrow bandpass-filtered sound, the sequence was quite similar to tone sequences used in prior experiments.

As in Experiment 1a, the target feature was combined with the sounds to create 3-second long repeating sequences (see Figure 1a). There were 10 repetitions of a 300 ms sequence consisting of two 60 ms tones and a 60 ms non-identifiable speech sound, with 40 ms silent intervals after each component of the sequence. In the target present condition, the target feature was played at the same time as the non-identifiable speech sound. All other aspects of stimulus creation and presentation were as before.

Conditions: The conditions were the same as described in Experiment 1a (refer to Figure 1c).

Procedure: All aspects of the procedure were the same as described in Experiment 1a, except for the response button labels in the attend-vowel block. In this experiment, listeners labeled stimuli in this block as “not a vowel” or as “/I/ as in pit.”

Results and Discussion—All listeners met the a priori criterion with the tone stimuli. One listener did not meet this criterion with the vowel stimuli, one listener failed to respond on 75% of the test trials, and data from two listeners were lost due to experimenter error. The following analyses were conducted on the data from the remaining twelve listeners.

Attend-Vowel²: The results are presented in the bottom panel of Figure 4. In the Single sound condition (far right), the target feature was more likely to be perceptually grouped with the vowel than in Experiment 1a when the two sounds were presented at different spatial locations ($\delta = 0.90$ in this experiment vs. 0.63 in Experiment 1a, $p < .01$). Thus, the object was more likely to group with the target feature when doing so changed an unidentifiable sound into a vowel, even in the absence of competition.

In the Two-sound condition, the target feature “reappeared” perceptually. As can be seen in the circled marker in the bottom panel of Figure 4, the feature contributed to the perception of the vowel in the critical condition (mean $\delta = 0.89$) significantly more than in Experiment 1a (mean $\delta = 0.35$), $p < .01$. In fact, the feature contributed to the perception of the vowel in every Two-sound target present condition, except when the feature location only matched the tone sequence (left-most point). As we hypothesized, this result contrasts with the findings of Shinn-Cunningham et al. (2007) and Experiment 1a. We suggest that the feature “disappeared” in both of those cases because the feature was not necessary for perception of an identifiable object. When the feature is necessary to support perception of an identifiable object, listeners group the feature with the object even when the two are presented at different spatial locations.

Attend-Tones: In the Single-sound condition, when the target and tones were presented at different spatial locations (Figure 4, top panel, rightmost point), the tone sequence was less likely to group with the target feature than in Experiment 1a (mean $\delta = 0.37$ in this experiment vs. $\delta = 0.81$ in Experiment 1a; $p < .01$). It is tempting to conjecture that this reduction is the complement of the increased grouping we saw for the vowel, but there is no hint of a negative correlation ($r = .006$, n.s.). Though this condition was not a focus of the current study, a finding contrary to the results of Shinn-Cunningham et al. (2007) and Experiment 1a warrants a potential explanation. Recall that the tone sequence used in this experiment was made with the frequency components of the first formant of a vowel, rather than a 500 Hz pure tone. In all three experiments, two different percepts can occur when the tone sequence and the feature are presented at separate locations. The two can fuse into one rhythmic pattern of tones that repeats every 60 ms (as it did in Shinn-Cunningham et al. and in Experiment 1a). Or, the two can segregate into two different rhythms: a sequence consisting of two tones that repeat every 140 ms and a one-tone sequence that repeats every 240 ms. The additional frequency components used in the current experiment provided cues that made the segregated percept more likely to occur.

As in Experiment 1a, the target feature did not contribute to the perception of the tone sequence in the Two-sound critical condition (see the circled marker in the top panel of Figure 4). In fact, the feature contributed significantly less in this condition than it did in Experiment 1a, $\delta = 0.08$ vs. $\delta = 0.41$, $p < .01$. In the remaining Two-sound conditions, the target feature only contributed to the perception of the tone sequence when both the tone sequence and the feature were presented at the same spatial location (whether or not the competing object was also presented at this location).

Experiment 2: The Effect of Feature Variation on Perceptual Coherence

Experiment 1a confirmed that a feature can disappear from perception, and Experiment 1b revealed that the feature will reappear when the perceptual contribution of the feature results in an identifiable object. Experiment 2 examines the fate of the feature when it is altered relative to the sounds in figure and ground. Natural listening situations often involve many simultaneous sounds, and as a result the acoustic properties of each source are frequently

²As in Shinn-Cunningham et al. (2007), we make the assumption that the attended sound is in perceptual figure.

muddled together as they enter the listener's ear. One way to investigate how listeners resolve the mix of acoustic features and combine the correct features into auditory objects is to measure the effect of a modified feature on the perception of the sound in figure.

Thus, in Experiment 2 we test how modifying a feature affects perceptual coherence between the feature and an object, and between the feature and a stream. We tested two quite different types of modification. One type occurs in many natural listening situations -- noise masking. To contrast with noise masking, we also tested a modification that does not occur in natural listening situations -- mistuning of sound components. In both cases, the central issue is how modifying a feature affects perceptual coherence between the feature and an object or stream.

Method

Participants—Twelve SBU undergraduates (mean age = 19.2) with no reported hearing deficiencies participated in this experiment for course credit.

Stimuli—The target feature, tone sequence, and vowel stimuli from Experiment 1b were used in this experiment. The tone sequence was perceived as rhythmically even with the target feature, and as galloping without the target feature. The vowel was perceived as a non-identifiable speech sound without the target feature and as /I/ with the feature.

Stimuli for the noise manipulation were created by synthesizing a 60 ms segment of white noise using a preset function in Cool Edit Pro (version 2.1). Three different levels of noise were generated: noise 10dB lower than the amplitude of the target feature, noise equal in amplitude to the feature, and noise 10dB higher than the feature.

Stimuli for the mistuning manipulation were created by shifting the frequency of the target feature both upward (+) and downward (–) using Cool Edit Pro. Previous research has demonstrated, with single-object conditions, that mistuning a feature by 8% is sufficient to remove the feature from an object, i.e., to cause the feature to no longer contribute to the identity of the object (see Darwin & Gardner, 1986); pilot testing in our lab confirmed that 8% was an appropriate minimum level of mistuning to use. Thus, the pitch of the target feature was shifted by $\pm 8\%$, $\pm 16\%$, and $\pm 32\%$ in this experiment.

Conditions—The conditions from Experiment 1b were modified to include the noise and the mistuning manipulation. As can be seen in Figure 5, both the Single- and Two-sound conditions were modified to include three levels of noise masking of the target feature: low (10dB lower than the feature), medium (same level as the feature), and high (10dB higher than the feature). Three levels were also used for the mistuning manipulation: low (8%), medium (16%), and high (32%). The conditions depicted for the noise manipulation in Figure 5 are analogous to the three levels of mistuning used for the mistuning manipulation. These conditions test how altering the target feature affects the probability of the feature binding with the object or stream. The two-sound target absent condition from Experiment 1b was not included in this experiment; this condition was unnecessary to measure an effect of noise and mistuning on perceptual coherence. The Single-sound target present/absent prototype conditions from Experiment 1b (with no noise or mistuning manipulation) were used as before, to allow a comparison of how each modified condition was perceived relative to these Single-sound prototypes.

There were 20 conditions per sound, as follows: There was one *Single-Sound* Target Absent Prototype and one *Single-Sound* Target Present Prototype. As before, the prototypes were not directly included in any data analyses because they were used to calculate the δ scores for the other conditions. There were two Single-Sound cases and four Two-Sound cases,

yielding six conditions that were orthogonally combined with the three levels of noise or mistuning manipulation (low, medium, and high). The result was 3 *Single-Sound* Feature Matches (feature matches location of sound), 3 *Single-Sound* Feature Doesn't Match (feature does not match location of sound), 3 *Two-Sound* Feature Matches Both, 3 *Two-Sound* Feature Matches Tones, 3 *Two-Sound* Feature Matches Vowel, and 3 *Two-Sound* Feature Matches Neither.

Procedure—Each manipulation (noise and mistuning) and each block (attend-vowel and attend-tones) included 20 repetitions of the 20 conditions. The result was 1600 trials: 400 trials in the attend-tones noise manipulation block, 400 in the attend-vowel noise manipulation block, 400 in the attend-tones mistuning manipulation block, and 400 in the attend-vowel mistuning manipulation block. Each listener completed 400 trials a day for four consecutive days. On each day, listeners completed one of the manipulations (noise or mistuning) and one of the blocks (attend-tones or attend-vowel). A balanced Latin square was used to determine the order of manipulation and block for each participant. Rest breaks were given after every 100 trials. All other procedures were the same as described in Experiment 1b.

Results and Discussion

Data from one listener did not meet the a priori criterion for the attend-vowel block and was excluded from the data analyses. Data from an additional listener was excluded because the listener failed to respond to any two-object condition. The following analyses were conducted on data from the remaining ten listeners.

How does the level of variation affect perceptual coherence?—The effect of feature variation on perceptual grouping is reflected in the δ scores. Recall that high δ scores ($\delta > .5$) indicate that the feature and the object/stream have been perceptually grouped; low δ scores ($\delta < .5$) indicate that the feature does not contribute to the perception of a sound. The feature was more likely to contribute to the perception of the object/stream at low and medium levels (mean δ scores were .51 and .47) of mistuning than at the high level of mistuning (mean $\delta = .34$), both p 's $< .05$. Similarly, the feature was more likely to contribute to the perception of the object/stream at the low (mean $\delta = .58$) and medium (mean $\delta = .50$) levels of noise than the high (mean $\delta = .26$) level of noise, both p 's $< .01$.

ANOVAs were conducted with Sound (tone sequence or vowel), Condition (the four Two-sound and two Single-sound cases), and Manipulation Level (low, medium, or high) as factors on the δ scores from both the mistuning and the noise manipulations. Both manipulations produced a significant effect of Level: mistuning, $F(2, 18) = 32.44$, $p < .01$, $\eta^2 = .78$, and noise, $F(2, 18) = 21.45$, $p < .01$, $\eta^2 = .70$.

The finding that both the object and the stream grouped with the modified feature (at low and moderate levels) suggests that listeners are flexible in interpreting evidence that two auditory components may belong together. In this experiment, listeners made use of incomplete (feature masked by noise) and inconsistent (feature with a different pitch) information to accomplish coherent perception. However, with more severe mismatching, incorporation of the feature was significantly reduced.

How does the type of variation affect perceptual coherence?—Figures 6a and 6b depict the results. Recall that the *Single-Sound* Target Absent Prototype and the *Single-Sound* Target Present Prototype were not directly included in the data analyses because they were used to calculate the δ scores for the other conditions, leaving the five conditions shown in these figures. The diamond markers in the figures represent the baseline δ scores

from Experiment 1b, the “x” markers represent the δ scores from the noise manipulation, and the circle markers represent the scores from the mistuning manipulation. As can be seen in the figures, the answer to the question of which type of variation is tolerated better depends on whether the sound in figure was the object (the vowel) or the stream (the tone sequence).

Figure 6a shows that the natural (i.e., noise) and unnatural (i.e., mistuning) manipulations had similar effects on the stream. Perception of the stream was not affected very much by either manipulation, particularly in the critical condition (see circled markers in figure). However, analysis of the tone sequence did yield a significant Experiment $\times$ Condition interaction (noise: $F(4, 36) = 4.30, p < .01, \eta^2 = .32$; mistuning $F(4, 36) = 11.83, p < .01, \eta^2 = .57$). The noise and the mistuning manipulation each significantly decreased the feature's contribution to the perception of the tone sequence relative to the baseline. The noise manipulation resulted in the feature contributing to the tone sequence less than in Experiment 1b in the *Single-Sound* Feature Doesn't Match, *Two-Sound* Feature Matches Tones, and *Two-Sound* Feature Matches Neither conditions, $p's < .05$. The mistuning manipulation decreased the feature's contribution to the perception of the tone sequence relative to Experiment 1b in the *Single-Sound* Feature Doesn't Match, *Two-Sound* Feature Matches Tones, *Two-Sound* Feature Matches Neither, and the *Two-Sound* Feature Matches Both conditions, $p's < .05$.

The δ scores from the condition of particular interest in this experiment (the critical condition) were similar to the δ scores from the corresponding condition in Experiment 1b. Recall that in Experiment 1b, the feature did not contribute to the perception of the tone sequence in the critical condition. Given this lack of incorporation to start with, it makes sense that the feature would continue to be of no use to the stream when the feature was modified by noise and by mistuning.

When an object (the vowel) was in perceptual figure, the natural manipulation decreased the feature's contribution to the perception of the object more than the unnatural manipulation (see Figure 6b). There was a significant Experiment $\times$ Condition interaction for both vowel manipulations (noise: $F(4, 36) = 36.94, p < .01, \eta^2 = .80$; mistuning: $F(4, 36) = 35.74, p < .01, \eta^2 = .80$). As can be seen in Figure 6b, both manipulations made the feature less likely to contribute to the perception of the vowel. The mistuning manipulation made the feature significantly less likely to contribute to the perception of the vowel in 4 of the 5 conditions: *Single-Sound* Feature Doesn't Match, *Two-Sound* Feature Matches Both, *Two-Sound* Feature Matches Neither, and *Two-Sound* Feature Matches Vowel (critical condition), $p's < .01$. Note, however, that although the mistuning manipulation decreased the feature's contribution to the perception of the object in the critical condition, the feature was still usually perceived as part of the object (mean $\delta = 0.69$).

The noise manipulation nearly eliminated any incorporation of the feature into the vowel. Planned comparisons revealed that the noise decreased the feature's contribution to the vowel in all five conditions, $p's < .01$. As can be seen in Figure 6b, the noise manipulation resulted in the feature barely contributing to perception of the vowel at all (overall mean $\delta = .37$). This was particularly apparent in the critical condition (mean $\delta = .37$, a large drop from its powerful influence in Experiment 1b).

In summary, the answer to the question of how feature variation affects perceptual coherence is: It depends. When the stream was in figure, perception was not strongly affected by either manipulation. In contrast, when the vowel was in perceptual figure, perception of the identity of the object was changed by the natural manipulation, but only minimally affected by the unnatural manipulation.

The results of Experiment 2 suggest a potentially important distinction between streams and objects: streams appear to be more robust to feature variation than objects. Perhaps this is the case because the repeated rhythm of streams provides a predictable context that makes the incorporation of “good enough” features into the perception of streams more likely. Objects appear to be less tolerant of noise variation in particular. This might reflect the nature of objects. The identification of objects does not typically depend on repetition, so listeners have less time and fewer chances to correctly allocate features to objects. As a result, listeners may be more selective in allocating features to objects. This suggestion is of course based on a very limited sample: the stream and object used in the present investigation, though as we discuss in the General Discussion there are several findings that support the object-stream distinction. We believe that it would be very useful to test other types of objects and streams to determine how generalizable this pattern is.

Experiment 3: Multiple Sounds Competing for the Same Feature

Experiment 1 demonstrated that auditory features are more likely to be grouped with a sound if doing so produces a more identifiable object, when there is competition between the object and a stream. Because natural listening situations often involve even more complex scenarios, Experiment 3 tests whether the grouping principle that operated in Experiment 1 holds when there are three uses for the same feature, rather than two. Of course, in a busy real-world environment there may be many more, but testing a three-sound environment is a useful step beyond the one- or two-sound cases that have been the typical focus of study. As we shall see, even an increase to three sources produces interesting developments.

Method

Participants—Twenty-six SBU undergraduates (mean age = 20.2 years) served as listeners in this experiment. No listener reported any hearing deficiency. All listeners received course credit for their participation.

Stimuli—The target feature from Experiment 1b was used. Recall that this feature was the first formant of the vowel /I/, which peaked at 490 Hz. Three additional stimuli were used: a tone sequence and two vowels. The tone sequence and one of the vowels were the same as in Experiment 1b. The tone sequence is perceived as “even” if it groups with the target feature and as “galloping” if it does not. The vowel, called Vowel 1 in this experiment, creates the percept of /I/ if it groups with the feature and a non-identifiable speech-like sound (“not a vowel”) if it does not group with the feature. The second vowel was the same vowel used in Experiment 1a. This vowel, called Vowel 2, creates an /e/ percept if it groups with the feature and /I/ if it does not group with the feature. The feature used in Experiment 1a shifts the percept of this vowel from /I/ to /e/ because it adds energy to the first formant. While the feature used in Experiment 3 is different from the one used in Experiment 1a, the category of the vowel that this feature creates is the same as Experiment 1a. Vowel 2 was used in this experiment to examine the effects of perceptual competition in denser auditory scenes when one of the sounds is already an identifiable object without the feature (recall that Vowel 1 is only an identifiable object when it groups with the feature).

Conditions—Figure 7 depicts the conditions that were used. As can be seen in the figure, on a Single-sound trial, one of the three sounds was presented alone, either with or without the target feature. On target present trials, the feature was either presented at the same spatial location as the sound (to promote grouping) or at one of three different spatial locations (to discourage grouping). All three sounds were presented simultaneously in the Three-sound conditions; the three sounds and the target feature were presented at various combinations of spatial locations: The target feature was presented at the same spatial location as the

attended sound, at a spatial location matching one of the other sounds, or at a spatial location that did not match any of the three sounds (see below).

Procedure—There were three blocks. In each block, the task was to attend to one of the sounds: the tone sequence, Vowel 1, or Vowel 2. The attended sound was always presented at a speaker located directly in front of the listener. In the attend-tone block, the task was to label the tone sequence as “even” or as “galloping” by pressing the corresponding key on a button board. In the attend-Vowel 1 block, the task was to label the vowel as “i as in pit” or as “not a vowel.” In the attend Vowel 2 block, the task was to label the vowel as “e as in pet” or as “i as in pit.” Sounds could be presented at four possible spatial locations: the center (where the attended object/stream was always presented), 45° to the right, 90° to the right, and 90° to the left.

There were seven conditions in each block: *Single-sound* target present, *Single-sound* target absent, *Single-sound* target location does not match, *Three-sound* target location matches sound in figure, *Three-sound* target location matches one sound in ground, *Three-sound* target location matches the other sound in ground, and *Three-sound* target location matches no sound. As before, the first two *Single-sound* conditions were used to compute a baseline for the other conditions. There were 24 repetitions of each condition in each block, except for the two conditions in which the target location matched one of the sounds in ground. This condition was split so that there were 12 repetitions for each sound, e.g., in the Attend-Vowel 1 block, there were 12 repetitions of the feature being presented at the same location as Vowel 2 and 12 repetitions of the feature being presented at the same location as the Tones. There were 144 trials per block with order of blocks counterbalanced across listeners. Listeners received a rest break after each block. All other procedures were as before.

Results and Discussion

Thirteen listeners failed to meet the a priori criterion for Vowel 2 and one listener failed to meet the criterion for Vowel 1. The following analyses are reported for the remaining twelve listeners.

Attend-Tones—As can be seen in the top of Figure 8, the feature only contributed to the perception of the tone sequence when the two were presented at the same spatial location (see circled marker in figure). This pattern produced a significant effect of Condition, $F(4, 44) = 89.16$, $p < .01$, $\eta^2 = .89$. Planned comparisons confirmed that the δ scores in the Three-sound condition in which the target feature matched the location of the tones were significantly higher than the δ scores in all other conditions, p 's $< .01$. This finding is quite different from the results of Experiment 1b. Recall that in Experiment 1b the target feature contributed to the tone sequence when the feature matched the location of the tones and when it matched the location of both sounds -- the tones in figure and the vowel in ground. Thus, it appears that when three sounds are competing for the same feature, assignment of the feature to the tone sequence is more selective than when only two sounds are competing for the feature: As competition increases, there is a shift towards more of a winner-takes-all assignment process, with shared location taking on a much more potent role.

Attend-Vowel 1—Similarly, the target feature only contributed to the perception of Vowel 1 when the two were presented at the same spatial location (see circled marker in middle of Figure 8). This trend contributed to a significant effect of Condition, $F(4, 44) = 9.53$, $p < .01$, $\eta^2 = .46$. Planned comparisons confirmed that the δ scores in the Three-sound condition in which the target feature location matched Vowel 1 were significantly higher than the δ scores in all other conditions, p 's $< .05$. Recall that in Experiment 1b, the feature contributed to the perception of the vowel in five of the six conditions. As with the tone sequence, the

feature only contributed to the object in figure when there was evidence based on spatial location that the feature belonged only to the object. Again, winner takes all, and spatial coincidence is powerful.

Attend-Vowel 2—As can be seen in the bottom of Figure 8, in the perception of Vowel 2 all of the δ scores fell below the 0.5 cutoff. There were no differences in the δ scores across the Conditions, $F(4, 44) = 1.09$, $p > .05$. This finding converges nicely with prior experiments (done with two sources): The target feature never contributed to this vowel in Shinn-Cunningham et al.'s (2007) study with two objects and only contributed to one of six conditions in Experiment 1a. Observing the same outcome here, with three sources, confirms the central conclusion of Experiment 1: A critical factor in integrating features with objects is the degree to which such integration is needed to establish a “good” object. This vowel was already an identifiable object (the vowel /I/) without the feature, and thus the feature was not assigned to it.

These clear results of Experiment 3 confirm and extend the findings of Experiments 1a, 1b, and 2. The contrast between Vowel 2 and Vowel 1 confirms the pattern established in Experiments 1a and 1b: Because Vowel 2 was already a good auditory object, the target feature was not integrated with it, whereas such integration did occur for Vowel 1 (which was not a good auditory object in the absence of the feature). The winner-takes-all pattern found for the Tones and for Vowel 1 extends the prior results, demonstrating that spatial coincidence plays a greater role when there are more potential competitors in the auditory environment.

General Discussion

The central new finding of this project was that a feature is more likely to contribute to an object in figure if the result of the perceptual grouping is a “good,” identifiable object. This perceptual principle may explain one way in which listeners maintain perceptual constancy of auditory objects in a constantly changing auditory world. This principle did not apply to streams in perceptual figure. Feature assignment to a stream was guided by lower-level acoustic and spatial factors, such as those described by Bregman and colleagues (Bregman, 1990).

The finding that a feature is more likely to contribute to the perception of an object when the result is a good, identifiable percept applied across a variety of two-sound contexts. In Experiment 1b, the feature contributed to the perception of the object when the two were presented at different spatial locations, even when a second sound (the stream) that could group with the feature was also presented at that location. In Experiment 2, the feature contributed to the object in perceptual figure even when the pitch of the feature was mistuned relative to the object. This occurred when the object and the feature were presented at the same spatial location, even if the stream was also presented at that location.

The increased probability of perceptual grouping of an object and feature to produce an identifiable object was apparent in the two-sound scenes in this study. But, when the auditory scene was more crowded, there was a pronounced shift toward “winner takes all” with spatial coincidence playing a major role. When three sounds were presented in a scene, the feature contributed to the perception of the impoverished object only when the feature and the object were presented at the same spatial location (and no other sound was also presented at that location). The same finding occurred with the stream: the feature only contributed to the perception of the stream when the stream and feature were presented at the same location. Thus, the priorities in auditory scene analysis shift when there is increased competition for a feature: listeners primarily rely on spatial cues if there are

multiple “takers” for one feature, even if the result is a poor, non-identifiable object in figure or ground.

Objects vs.Streams—One of the goals of this project was to examine differences in the perceptual organization of objects and streams (an underappreciated distinction in the ASA literature). We observed several important differences. The results of all three experiments demonstrate that objects are more sensitive than streams to variation in the input that affects how good they are. This tendency reflects the quite different representations of objects and streams. Objects presumably have well-specified representations (e.g., what a cat's meow sounds like), whereas streams are patterns that may not always correspond to a specific entity in memory.

The key result of Experiment 1a, and of Shinn-Cunningham et al. (2007), was that a feature disappeared from perception. The principles of ASA theory (Bregman, 1990) cannot naturally explain this “ungrouping” of a feature and object or of a feature and a stream. ASA theory asserts that energy trading holds between objects in a scene, but there are conditions under which the components in a scene do not split the potential energy available (see Shinn-Cunningham et al., 2007). In particular, a feature can contribute to two sounds at the same time (e.g., Fowler & Rosenblum, 1990; Rand, 1974), and a feature can completely disappear from perception of a scene; both of these observations are problematic for the energy trading view.

Given the inadequacy of ASA theory to account for our findings (and others), it seems necessary to either modify ASA theory or to look beyond it for an explanation. An augmented ASA theory could potentially explain the reappearance of the feature in Experiment 1b. The low-level grouping principles of ASA would predict that the feature could contribute to the perception of either the tone sequence or the vowel. The competition between the two sounds could result in listeners relying on higher-level processes to accomplish segregation, rather than low-level processes. However, these higher-level processes would have to include a mechanism sensitive to the structure of the sound in figure in relation to the sounds in ground. No such mechanism is currently included in ASA theory. A revised version in which the properties of the sounds in figure and in ground affect perceptual coherence would provide a more comprehensive picture of how auditory objects are extracted from perceptual ground. Based on the present data, any such model would have to factor in the “goodness” of an object, what kind of overlapping sounds are present, how many other objects are in the background, and the relationship between the grouping cues available for the object and the stream.

Some of the studies by Shinn-Cunningham and her colleagues can be viewed as providing amendments to ASA theory. For example, Lee and Shinn-Cunningham (2008b) built on Bregman's distinction between simultaneously-organized sounds (e.g., the vowel) and sequentially-organized ones (e.g. the tone sequence). They replicated the central result of Shinn-Cunningham et al. (2007), and showed that when reverberation is added the spatial cues play a somewhat reduced role. In two other studies (Lee, Babcock, & Shinn-Cunningham, 2008; Schwartz & Shinn-Cunningham), these researchers used a (simultaneous) tone complex rather than a vowel as a competing sound with the tone sequence, and found results that were more consistent with energy trading than in the other studies or in the current study. However, these two studies also used a richer harmonic target sound, one which changed the perceived pitch of the tone complex when the two were integrated. Thus, there were enough differences to make a direct comparison difficult. Lee and Shinn-Cunningham (2008b) suggest that spatial cues are more important for sequentially-organized sounds than for simultaneously-organized ones. The shift to spatial cue dominance in Experiment 3 for both the vowels (simultaneously organized, in their

terms) and the tones demonstrates the importance of looking at denser auditory scenes in trying to determine general principles of auditory organization.

An alternative to augmenting ASA theory is to consider other views of scene analysis. The literature on ecological acoustics (e.g., Carello, Wagner, & Turvey, 2005; Gaver, 1993; Gaver, Rosenblum & Robart, 2007) may offer a natural basis for the distinction we have drawn between objects and streams³. As we have noted, this literature has established that listeners can determine the size and shape of objects from the sounds that they produce (Carello, Anderson, & Peck, 1998; Kunkler-Peck & Turvey, 2000). From this perspective, one important guiding principle for perceptual organization is that features are integrated when they provide information about the source of the sound, such as the shape. As Carello, Wagner, and Turvey (2005) note, when this type of information is not available, listeners may resort to the low-level rules established by Bregman (1990); ecological processes do not apply to the non-ecological sounds (e.g., tone sequences) that are often presented in psychoacoustic experiments. Thus, the differences between objects and streams could reflect differences in how a sound with a determinable source (e.g. the vowel), and a sound with no determinable source (e.g., the stream), are perceptually organized, rather than differences that are due to object versus stream per se. This account is somewhat different than the explanation we have offered for our results, but it is not incompatible. We have argued that there are important differences in the perceptual organization of objects and streams. Whether these differences are due to differences in memory representations, determinable sources, or some other factor is an empirical issue, with potentially interesting theoretical implications.

Whatever the basis turns out to be for the difference between the organizational properties of streams and objects, the results of the current study suggest that this is an important distinction. It is notable that a distinction of this general form emerges from quite different ways of thinking about perceptual organization: Objects versus streams, simultaneous versus sequential organization, and source-based versus non-source-based all reflect the need to make a distinction. By examining the characteristics of objects and streams in figure, and factors that affect the likelihood of perceptual grouping, the three experiments in the current study have provided a more complete picture of how auditory scene analysis is accomplished.

Acknowledgments

This work was supported by NIH grants R01-51663 and R01-059787. We thank Donna Kat for the many hours she devoted to programming our experiments and for providing the technical assistance that was so valuable to the completion of this project. We also thank two undergraduate research assistants, Carolyn Creary and Kavita Patel, for their assistance with data collection, and three anonymous reviewers for their constructive suggestions. This work was submitted as partial fulfillment of Melissa K. Gregg's doctoral dissertation. This work also was presented at the 2009 meeting of the Psychonomic Society.

References

- Boersma, P.; Weenik, D. Praat: Doing phonetics by computer (Version 4.6.01 [Computer program]. 2007. Retrieved May 16, 2007, from <http://www.praat.org/>
- Bregman, AS. Auditory scene analysis: The perceptual organization of sound. MIT Press; Cambridge, MA: 1990.
- Bregman AS, Campbell J. Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*. 1971; 89:244–249. [PubMed: 5567132]

³We thank an anonymous reviewer for suggesting this alternative

- Carello C, Anderson KL, Peck A. Perception of object length by sound. *Psychological Science*. 1998; 9:211–214.
- Carello, C.; Wagman, JB.; Turvey, MT. Acoustic specification of object properties. In: Anderson, J.; Anderson, B., editors. *Moving image theory: Ecological considerations*. Southern Illinois University Press; Carbondale, IL: 2005. p. 79-104.
- Cutting JE. Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. *Psychological Review*. 1976; 83:114–140. [PubMed: 769016]
- Dannenbring GL, Bregman AS. Stream segregation and the illusion of overlap. *Journal of Experimental Psychology: Human Perception and Performance*. 1976; 2:544–555. [PubMed: 1011004]
- Darwin, CJ. Auditory processing and speech perception. In: Bouma, H.; Bouwhuis, DG., editors. *Attention and performance X*. Hillsdale, NJ; Lawrence Erlbaum Associates: 1983.
- Darwin CJ, Hukin RW. Perceptual segregation of a harmonic from a vowel by interaural time difference in conjunction with mistuning and onset asynchrony. *Journal of the Acoustical Society of America*. 1998; 103:1080–1084. [PubMed: 9479762]
- Darwin CJ, Pattison H, Gardner RB. Vowel quality changes produced by surrounding tone sequences. *Perception & Psychophysics*. 1989; 45:333–342. [PubMed: 2710634]
- Davis MH, Johnsrude I, S. Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*. 2007; 229:132–147. [PubMed: 17317056]
- Deutsch D. Octave generalization and tune recognition. *Perception & Psychophysics*. 1972; 11:411–412.
- Fowler CA, Rosenblum LD. Duplex perception: A comparison of monosyllables and slamming doors. *Journal of Experimental Psychology: Human Perception and Performance*. 1990; 16:742–754. [PubMed: 2148589]
- Gregg MK, Samuel AG. Change deafness and the organizational properties of sounds. *Journal of Experimental Psychology: Human Perception and Performance*. 2008; 34:974–991. [PubMed: 18665739]
- Gregg MK, Samuel AG. The importance of semantics in auditory representations. *Attention, Perception, & Psychophysics*. 2009; 71:607–619.
- Husain FT, McKinney CM, Horwitz B. Frontal cortex functional connectivity changes during sound categorization. *Neuroreport*. 2006; 17:617–621. [PubMed: 16603922]
- Kunkler-Peck A, Turvey MT. Hearing shape. *Journal of Experimental Psychology: Human Perception and Performance*. 2000; 1:279–294. [PubMed: 10696618]
- Lee AKC, Babcock S, Shinn-Cunningham BG. Measuring the perceived content of auditory objects using a matching paradigm. *Journal of the Association for Research in Otolaryngology*. 2008; 9:388–397. [PubMed: 18473138]
- Lee AKC, Shinn-Cunningham BG. Effects of frequency disparities on trading of an ambiguous tone between two competing auditory object. *Journal of the Acoustical Society of America*. 2008a; 123:4340–4351. [PubMed: 18537385]
- Lee AKC, Shinn-Cunningham BG. Effects of reverberant spatial cues on attention-dependent object formation. *Journal of the Association for Research in Otolaryngology*. 2008b; 9:150–160. [PubMed: 18214613]
- Remez RE, Rubin PE, Berns SM, Pardo JS, Lang JM. On the perceptual organization of speech. *Psychological Review*. 1994; 101:129–156. [PubMed: 8121955]
- Scott SK, Blank CC, Rosen S, Wise RJS. Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*. 2000; 123:2400–2406. [PubMed: 11099443]
- Steiger H, Bregman AS. Competition among auditory streaming, dichotic fusion, and diotic fusion. *Perception & Psychophysics*. 1982; 32:153–162. [PubMed: 7145585]
- Shinn-Cunningham BG, Lee AKC, Oxenham AJ. A sound element gets lost in perceptual competition. *Proceedings of the National Academy of Sciences*. 2007; 104:12223–12227.
- Shinn-Cunningham BG, Wang D. Influences of auditory object formation on phonemic restoration. *Journal of the Acoustical Society of America*. 2008; 123:295–301. [PubMed: 18177159]

- Van Noorden, LPAS. Unpublished doctoral dissertation. Eindhoven University of Technology; 1975. Temporal coherence in the perception of tone sequences.
- Whalen DH, Liberman AM. Speech perception takes precedence over nonspeech perception. *Science*. 1987; 237:169–171. [PubMed: 3603014]

Figure 1.

Stimuli and conditions that were used in Experiment 1. (A) Stimuli in the two sound condition consisted of a 3 second long repeating sequence of tones, a vowel, and a target feature that can group with either object. (B) Spectrum of the synthetic vowel that was used in Experiment 1a. If the feature is perceived as part of the vowel, the percept is /e/, otherwise the percept is /I/. (C) Spatial configuration of the stimuli in both the Single- and Two-Sound conditions.

Figure 2.
Experiment 1a results. The diamond markers represent δ scores.

Figure 3. Spectrum of the synthetic vowel that was used in Experiment 1b. If the feature is perceived as part of the vowel, the percept is /I/, otherwise the percept is a non-identifiable speech sound.

Figure 4.
Experiment 1b results. The diamond markers represent δ scores.

Figure 5.
Noise masking conditions used in Experiment 2.

6A

6B

Figure 6.

(A). Comparison of results from Attend-Tones block of Experiment 1b and Experiment 2 (collapsed across the low, medium, and high levels of noise and mistuning). (B) Comparison of results from Attend-Vowel block of Experiment 1b and Experiment 2 (collapsed across the low, medium, and high levels of mistuning). The markers represent d scores.

Figure 7.
Conditions that were used in Experiment 3.

Figure 8.
Experiment 3 results. The diamond markers represent δ scores.