

Influence of Task-Relevant and Task-Irrelevant Feature Continuity on Selective Auditory Attention

ROSS K. MADDOX¹ AND BARBARA G. SHINN-CUNNINGHAM²

¹Hearing Research Center, Biomedical Engineering, 44 Cummington St, Boston, MA 02215, USA

²Center for Computational Neuroscience and Neural Technology, Biomedical Engineering, 677 Beacon St, Boston, MA 02215, USA

Received: 13 February 2011; Accepted: 19 October 2011; Online publication: 29 November 2011

ABSTRACT

Past studies have explored the relative strengths of auditory features in a selective attention task by pitting features against one another and asking listeners to report the words perceived in a given sentence. While these studies show that the continuity of competing features affects streaming, they did not address whether the influence of specific features is modulated by volitionally directed attention. Here, we explored whether the continuity of a task-irrelevant feature affects the ability to selectively report one of two competing speech streams when attention is specifically directed to a different feature. Sequences of simultaneous pairs of spoken digits were presented in which exactly one digit of each pair matched a primer phrase in pitch and exactly one digit of each pair matched the primer location. Within a trial, location and pitch were randomly paired; they either were consistent with each other from digit to digit or were switched (e.g., the sequence from the primer's location changed pitch across digits). In otherwise identical blocks, listeners were instructed to report digits matching the primer either in location or in pitch. Listeners were told to ignore the irrelevant feature, if possible, in order to perform well. Listener responses depended on task instructions, proving that top-down attention alters how a subject performs the task. Performance improved when the separation of the target and masker in the task-relevant feature increased. Importantly, the values of the task-irrelevant feature also influenced performance in some

cases. Specifically, when instructed to attend location, listeners performed worse as the separation between target and masker pitch increased, especially when the spatial separation between digits was small. These results indicate that task-relevant and task-irrelevant features are perceptually bound together: continuity of task-irrelevant features influences selective attention in an automatic, obligatory manner, consistent with the idea that auditory attention operates on objects.

Keywords: psychophysics, streaming, top-down, bottom-up, object-based attention

INTRODUCTION

Even when a cacophonous mixture of sound reaches our ears, we are typically very good at estimating what sound came from a particular physical sound source, forming a corresponding perceptual auditory object (e.g., see Darwin and Carlyon 1995; Shinn-Cunningham 2008). This ability relies on the fact that everyday sound sources are structured, containing low-level cues (e.g., interaural differences, harmonicity, intensity, etc.) that ordinarily work together to enable auditory scene analysis (e.g., see Bregman 1990; Darwin and Carlyon 1995). In typical settings, the resulting auditory objects have perceptual attributes (features) such as location, pitch, and loudness, derived from low-level acoustic cues, that vary slowly and continuously over time and perceptually link together sound from a particular source. As with visual objects, listeners can volitionally bias attention towards a particular auditory object in a scene by focusing on a desired auditory feature, such as pitch

Correspondence to: Barbara G. Shinn-Cunningham · Center for Computational Neuroscience and Neural Technology · Biomedical Engineering · 677 Beacon St, Boston, MA 02215, USA. Telephone: +617-353-5764; fax: 617-353-7755; email: shinn@cns.bu.edu

(e.g., Brungart and Simpson 2002; Rakerd et al. 2006; Shinn-Cunningham 2008; Helfer and Freyman 2009).

In vision, selective attention (the act of focusing on an object of interest in the presence of distracters in order to better analyze and understand that object) is known to be influenced not only by volitional, top-down attention but also by automatic, bottom-up attention driven by attributes of the visual scene (e.g., see Desimone and Duncan 1995). Many recent studies have examined selective auditory attention (e.g., Freyman et al. 1999; Eramudugolla et al. 2005; Kidd et al. 2005; Best et al. 2007; Brungart and Simpson 2007; Best et al. 2008; Ihlefeld and Shinn-Cunningham 2008a, b; Marrone et al. 2008); however, most of these studies ignored whether continuity of task-irrelevant auditory features affects selective attention. As a result, we know little about the degree to which auditory features that are irrelevant to a listener's goals bias selective auditory attention.

Past auditory studies that pitted auditory features against one another (e.g., in which two streams were presented with incongruent pitches and locations) show that, in the absence of top-down attention, all features contribute to the formation of an auditory stream over time (Darwin and Hukin 1999; Darwin and Hukin 2000a, b; Darwin et al. 2003; Ihlefeld and Shinn-Cunningham 2008a, b). Yet other studies hint that attention and object formation are not independent. For instance, the perceptual organization of tone sequences evolves over time in a manner that depends on how a listener directs attention (Carlyon et al. 2001; Macken et al. 2003; Cusack et al. 2004); however, it is not clear whether task-irrelevant features influence this evolution automatically. Thus, while we know that (1) listeners can bias selective attention by focusing on a specific auditory feature, (2) automatic formation of an object across time depends on the continuity of all of the features of a stimulus, and (3) top-down attention affects across-time object formation, it is not clear whether task-irrelevant features have an obligatory influence on perception, especially in processing complex signals like speech.

In vision, attention is "object based:" when an observer directs attention to one feature, all other features bound to the attended feature are perceptually enhanced. While most visual studies of attention consider static scenes, one study found that feature continuity automatically influences selective visual attention (Blaser et al. 2000). Specifically, when listeners were asked to attend to one of two overlapping, semitransparent images, each of which had features that changed smoothly through time, perception of slight feature discontinuities was enhanced in the attended image and suppressed in the unattended one. Here, in line with these past observations, we reasoned that if auditory attention is object based

(e.g., Alain and Arnott 2000; Cusack et al. 2000; Shinn-Cunningham 2008; Shinn-Cunningham and Best 2008), then feature continuity of a task-irrelevant feature should have an obligatory effect on the ability to volitionally focus selective auditory attention.

Listeners were presented with two consecutive digit pairs and asked to report, in different blocks, the digits matching either the location or pitch of a preceding primer phrase. Location and pitch sometimes worked in concert and sometimes conflicted across time (i.e., sometimes, words from the same location also had the same pitch, but sometimes they changed pitch). Importantly, listeners were aware that the continuity of the task-irrelevant feature could conflict with the task-relevant feature and therefore should be ignored, allowing us to test the idea that continuity of features, even ones irrelevant to performing a desired task, has an obligatory influence on selective attention, consistent with object-based auditory attention.

METHODS

Subjects

Six college-aged listeners were paid to participate in the experiment. All listeners were native speakers of American English, with normal auditory thresholds (confirmed by an audiometric screening). All subjects gave informed consent to participate in the study, as overseen by the Boston University Charles River Campus Institutional Review Board.

Basic task

Listeners were asked to identify a target sequence of two spoken digits in the presence of two competing interferer digits played simultaneously with the target. Target stimuli consisted of a sequence of two recorded digits whose location or pitch matched that of a preceding primer phrase. Each of the target digits was played with a simultaneous masking digit with a different location and pitch, as described further below (see Fig. 1).

The primer phrase ("Listen to this") was presented prior to the target and competing masker sequences. In half of the experimental blocks, listeners were instructed that the target sequence came from the same location as the primer. In these blocks, listeners had to direct spatial attention to the primer phrase to perform the task. In the other half of the blocks, listeners were instructed to report the digits that had the same pitch as the primer phrase (i.e., the listeners had to direct attention to the primer pitch to perform the task). Statistically speaking, the stimuli in *attend-location* trials and in *attend-pitch* trials were identical;

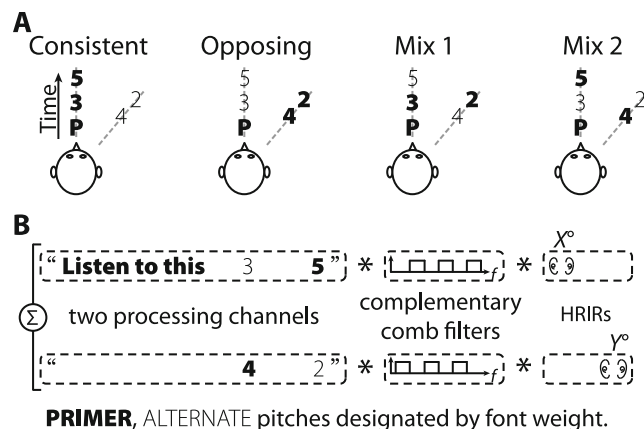


FIG. 1. **A** The four location–pitch conditions. Time is represented as distance from the head so that the primer phrase (*P*) comes first, followed by the two concurrent digit pairs (three and four form the first pair and five and two form the second). Pitch is denoted by font weight, with boldface type representing the primer pitch and light type representing the alternate pitch. **B** The method of stimulus generation. Each row is a processing channel that is comb filtered and convolved with an HRIR at a certain azimuth. Pitch is denoted by font weight as above. In this example, the correct response when attending location would be [3, 5], and when attending pitch would be [4, 5].

the only difference between blocks was the definition of which digits were target words and which were maskers.

Stimuli

Primer, target, and masker words were constructed to have distinct locations and pitches. Specifically, stimuli were constructed from recordings of one talker (male, a native speaker of American English). The primer was generated from ten independent tokens of the recorded phrase, “Listen to this.” Targets and maskers were constructed from ten recordings of the monosyllabic digits between zero and nine (excluding the disyllabic number “seven”), with 0 pronounced as “oh.”

The fundamental frequency of all tokens was near 100 Hz in the original recordings. Each token was used to generate a set of constant-pitch utterances with pitches between 75 and 134 Hz (100 Hz ± 5 semitones) using PRAAT software (www.praat.org). Pseudo-anechoic head-related impulse responses (HRIRs) were used to simulate spatial stimuli over headphones. The HRIRs were recorded from a Knowles Electronic Manikin for Acoustics Research, positioned in the center of an echoic classroom at azimuths ranging from -90° to $+90^\circ$. These HRIRs were then time windowed to remove the reflected energy, leaving only the direct sound (Shinn-Cunningham et al. 2005).

On each trial, two locations (*X* and *Y*) and two pitches (*A* and *B*) were chosen. The primer phrase

was always set to have one of these locations (*X*) and one of these pitches (*A*). In each simultaneous digit pair, one digit had location *X* and one location *Y*; likewise, one had pitch *A* and one pitch *B*. The values of pitch and location were chosen randomly and independently with the constraints that, across the trials in the block, (1) there were equal numbers of trials in which $|X - Y| = 10^\circ, 40^\circ,$ and 90° in azimuth and (2) there were equal numbers of trials in which $|A - B| = 5, 250,$ and 500 cents in pitch (where one cent, abbreviated ¢ , corresponds to a frequency ratio of $2^{1/1,200}$, or 1/100 of a musical semitone).

The individual digit tokens were of similar durations, ranging from 396 to 648 ms (the primer phrase was 1,143 ms long). The onsets of the digits in a simultaneous digit pair were time aligned, with the shorter duration digit zero padded at the end to make them equal duration. The primer, initial digit pair, and subsequent digit pair were temporally concatenated without any intervening silence, producing continuous streams of speech in which the digit-to-digit separation was 549 ms, on average.

The same scheme was used to generate the stimuli in all blocks. With this scheme, when listeners were asked to attend to the location of the primer, the correct, target digits were the two sequential digits that had location *X*; given that the pitches of the target digits were chosen independently of location, the pitches of each of the target digits could either match (*A*) or not match (*B*) the primer, as well as match or not match the pitch of the other target digit. Listeners were instructed to try to ignore pitch in order to select the target digits from the sound mixture when the target was defined by the primer location. Similarly, in an *attend-pitch* block, the target digits were the sequential digits with pitch *A* whose locations could either match or not match the primer; listeners were instructed to try to ignore space to perform well on these trials.

Since there were two pairs of digits, there were four possible location–pitch conditions, which we classified based on the relationship between the primer pitch and the pitches of the two subsequent digits from the primer location *X*. In *consistent* trials, both digits at the primer location (*X*) also had the primer's pitch (*A*) (see Fig. 1A). In *opposing* trials, both digits at the primer location (*X*) had the alternate pitch (*B*). In *mix type 1* trials, the first digit at the primer location (*X*) had the primer's pitch (*A*), while the second digit at location *X* had the alternate pitch (*B*). Finally, in *mix type 2* trials, the first digit at the primer location (*X*) had the alternate pitch (*B*) and the second digit from location *X* had the primer pitch (*A*). Each of these four conditions was equally likely to occur on any given trial, with the constraint that each of the four types of trials was presented in exactly one

quarter of the trials in a given block. As discussed below, the primary motivation for including *mix* trials was to ensure that, within a block, the relationship between the task-relevant and task-irrelevant features was completely unpredictable, so that listeners would try to ignore the task-irrelevant feature to perform the task. Most analysis focused on *consistent* and *opposing* trials, where stimulus and instruction effects were easier to interpret.

Given the scheme described above, each trial consisted of (1) a primer followed by two consecutive digits from location *X*, played simultaneously with (2) a silence followed by two consecutive digits from location *Y*. To reduce the effects of energetic masking, these two spatial streams were filtered by complementary comb filters that alternated in the frequency domain between values of 0 and 1, every 1/3 octave. One comb filter passed one half of the frequency spectrum; the other passed the other half. The cutoff frequencies of the comb filters were shifted randomly from trial to trial over a 1/3-octave range, with the shift chosen from a uniform distribution on a log frequency scale. Although half of each spectrum of each digit was removed, all digits were still easily identified from the remaining acoustic energy. See Figure 1B for a diagram illustrating how the stimuli were generated.

Procedure

Subjects were seated in a sound-treated booth (Industrial Acoustics Company, Winchester, UK). Stimuli were generated in MATLAB and presented using TDT hardware (Tucker Davis Technologies, Alachua, FL) over ER-1 in-ear headphones (Etymotic Research, Elk Grove Village, IL).

As described above, listeners were presented with stimuli that were statistically identical in all blocks; only the instructions differed between blocks. In every other block (blocks 1, 3, 5, etc.), listeners were instructed to report the two digits whose location matched that of the primer. In the other half of the blocks (blocks 2, 4, 6, etc.), listeners were told to report the digits whose pitch matched that of the primer. There were 12 blocks in total (six pitch blocks and six location blocks). The first two blocks were training blocks whose results were not included in the data analysis. The remaining ten blocks (five attend-location blocks, five attend-pitch blocks) were analyzed to produce the results reported here.

Each block consisted of 288 trials (eight trials of each of three angles, three pitch separations, and four conditions) with a required break halfway through each block and as many optional breaks as desired. On average, blocks lasted less than half an hour. Because there were nine possible digits at each time

point and there were two digits to report, the probability of reporting both target digits correctly by randomly guessing between all potential responses was 1.2%. However, if the listeners always heard at least one of the two digits from each simultaneous pair, then the probability of reporting both target digits by randomly selecting between the competing digits was one in four, or 25%.

Listeners entered their responses using a numerical keypad GUI on a computer screen in the sound-treated booth. The subject was required to answer on each trial, guessing if they were uncertain of the response. Feedback was given after each trial indicating whether the listener responded correctly (to both digits) or not. Trials were self-paced, initiated 500 ms after the response to the previous trial was entered.

RESULTS

Responses were classified into one of four types. Responses were *correct* when both target digits were correctly reported. In *wrong-feature* errors, listeners reported the digits that would have been correct if the instructions had been to attend to the other feature (e.g., if a subject reported digits that had the same pitch as the primer when instructed to report digits matching the primer's location). If the reported sequence of digits was present in the stimulus in the reported order, but the response was neither *correct* nor *wrong-feature*, a response was classified as a *confusion* error (e.g., if the subject reported the first target digit and the second masker digit). Finally, when at least one of the reported digits was not present in the corresponding target-masker pair (e.g., the first reported digit was "three" when the first target digit was "two" and the first masker was "one"), the response was scored as a *guess* error.

These different categories of responses were analyzed separately for *attend-location* and *attend-pitch* trials, for each of the four types of trials (*consistent*, *opposing*, *mix type 1*, and *mix type 2*). Within these eight categories (2 instructions \times 4 trial types), results were then further broken down based on the size of the spatial separation and pitch separation of the competing target/masker digits in the stimulus. Specifically, results were combined across trials in which the angular separations of the competing target and masker digits were equal (taking on values of 10°, 40°, or 90°), ignoring the absolute locations of target and masker, and across trials in which the pitch separations of target and masker digits were equal (5¢, 250¢, or 500¢), ignoring the absolute pitches. (Note that this analysis may collapse results over trials differing in difficulty; for instance, the *attend-location*

task is likely to be easier when the target and masker are at 0° and 10°, respectively, than when they are at 80° and 90°.) For each of the resulting nine spatial-pitch combinations (3 spatial separations × 3 pitch separations) in each of the eight trial types, we calculated the percentage of *correct*, *wrong-feature*, *confusion*, and *guess* responses for each subject. For brevity and because the response patterns were very similar across subjects, we report only the across-subject averages.

While the experimental design we employed included *consistent*, *opposing*, *mix type 1*, and *mix type 2* trials, the primary motivation for including the two types of mix trials was to ensure that the first digit pair in a trial carried no information about the second. In the *mix* trials, it is difficult to tease apart any contributions of feature continuity that are obligatory and automatic from those that depend on directing top-down attention, the primary focus of our experiment. Therefore, the only planned statistical analyses we performed were two multi-way ANOVAs, one exploring results of *consistent* trials and one exploring results of *opposing* trials.

The first subsection below provides a qualitative overview of the results as a function of instruction and condition to help orient the reader; in this section we also report the result of one post hoc statistical test of the *mix* trial results. The remaining two sections focus on results in the *consistent* condition and *opposing* condition, including results of our planned statistical comparisons. The trends described in section A, while not all tested statistically, are supported by the statistical tests reaching significance that are described below.

Overview of responses

Figure 2 plots the across-subject means of each kind of response in both *attend-location* and *attend-pitch*

conditions (Fig. 2A, B, respectively), collapsing across the size of the location separation and pitch separation between target and masker digits. In general, guessing errors were rare (less than 6% of all responses; see top white section of each individual bar in Fig. 2). In addition, the rate of *guess* errors was nearly constant, independent of whether listeners were attending to location or to pitch and independent of what kind of trial was presented (compare white bars in Fig. 2A to those in B as well as within each sub-figure). There was a slight tendency for listeners to give correct answers more often for the second digit than the first digit (average of 70.5% vs. 73.5%, across all four trial types); however, this difference was not significant ($p=0.06$) and sufficiently small that it is henceforth ignored.

Overall, listeners were much more likely to report the correct answer in the *attend-location* than in the *attend-pitch* trials (in all cases, the black portions of the bars in Fig. 2A are bigger than the black portions of the corresponding bars in Fig. 2B). Consistent with this, both *wrong-feature* errors and *confusion* errors were more likely when listeners were instructed to attend pitch than when instructed to attend location (the light and dark gray portions of each of the bars in Fig. 2B are larger than the corresponding bar portions in Fig. 2A). Indeed, in the *attend-pitch* trials, the likelihood of a *correct* response was smaller than the likelihood of a *wrong-feature* response in both types of *mix* trials (in the two rightmost bars of Fig. 2A, the black sections are smaller than the corresponding light gray sections; a *post hoc* paired *t* test found this difference to be statistically significant: $p<0.001$). The likelihood of *wrong-feature* responses in these *attend-pitch*, *mix* trials was also greater than the likelihood of *confusion* errors (in the two rightmost bars of Fig. 2A, the light gray sections are larger than the corresponding dark gray sections). This pattern of results is notable in

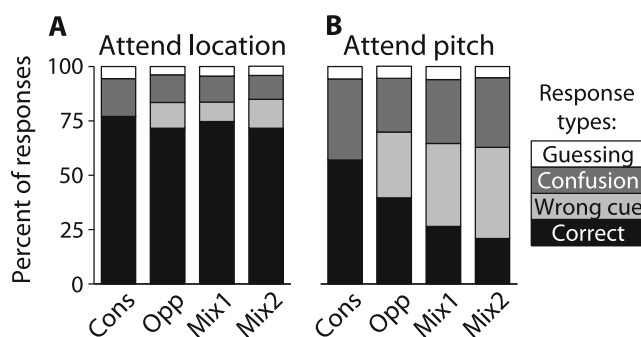


FIG. 2. Summary response data collapsed across angle and pitch separation when attending location (A) and pitch (B). Each column is one of the four possible conditions (in the same order as they appear in Fig. 1A), and each color-coded portion of the bar corresponds to a specific response type. There were four response types: *correct*, in which both digits were correctly reported; *wrong-feature*, in which

the subject reported the digits that would have been correct under the alternate attention instructions; *confusion*, which is any error that is not a *wrong-feature* error, but no digits which were not present in the stimulus were reported; and *guessing*, in which the subject reported at least one number that was not present in the stimulus.

that if the listeners were unable to attend to the correct pitch and were simply selecting digits randomly from the mixture, the number of *correct* responses and *wrong-feature* errors should be equal, and one half the number of *confusion* errors.

When breaking down responses by location and pitch separation, a few general trends were observable. In both the *attend-location* and *attend-pitch* blocks, the probability of a *correct* response tended to increase as the task-relevant feature became stronger, and the probability of a *wrong-feature* response decreased. *Confusion* errors, in which all of the reported digits were present in the stimulus but were a combination of target and competing digits, tended to decrease as the strength of the task-relevant feature separation increased, but this effect was larger in *attend-location* trials than in *attend-pitch* trials. *Guessing* errors did not depend on the strength of either feature. Below, we explore these trends by considering how listener instructions as well as the location and pitch separation of the competing target and masker digits affected performance for *consistent* and *opposing* trials, described in the next two sections.

Effect of instructions and feature separation in consistent trials

One key question is whether listeners change how they respond depending on the instructions they receive (rather than being influenced exclusively by automatic, obligatory processes). The presence of any such effect is easy to evaluate in *consistent* trials, where

the correct answer is the same whether instructions were to attend location or to attend pitch: if performance in the *consistent* trials differs with task instruction, then listeners must be adopting different listening strategies based on instructions since the stimuli are identical in all blocks.

In the *consistent* trials, listeners had a higher probability of reporting the correct answer when they were instructed to attend location than when they were instructed to attend pitch. This result is shown in Figure 3, which plots the probability of a *correct* response for *consistent* trials from the *attend-location* blocks (solid lines) and *attend-pitch* blocks (dashed lines), both as a function of the angle separation (Fig. 3A) and as a function of the pitch separation (Fig. 3B, which replots the same data shown in A). In every case, *correct* responses were about 20–25% more likely for *attend-location* trials than for the corresponding *attend-pitch* trials (solid lines are above dashed lines in all panels). In addition, the likelihood of responding correctly increased as the size of the pitch or location separations increased (in both Fig. 3A, B, the probability of a *correct* response tends to increase from the left panels to the right panels; similarly, within each panel, there is a tendency for performance to increase from left to right).

The subject-specific *correct* response percentages were analyzed with a multi-way repeated-measures ANOVA, using main factors of instruction, location separation, and pitch separation. Greenhouse–Geisser correction was used to correct for non-sphericity, where a *p* value of 0.05 or less was considered significant. All three main

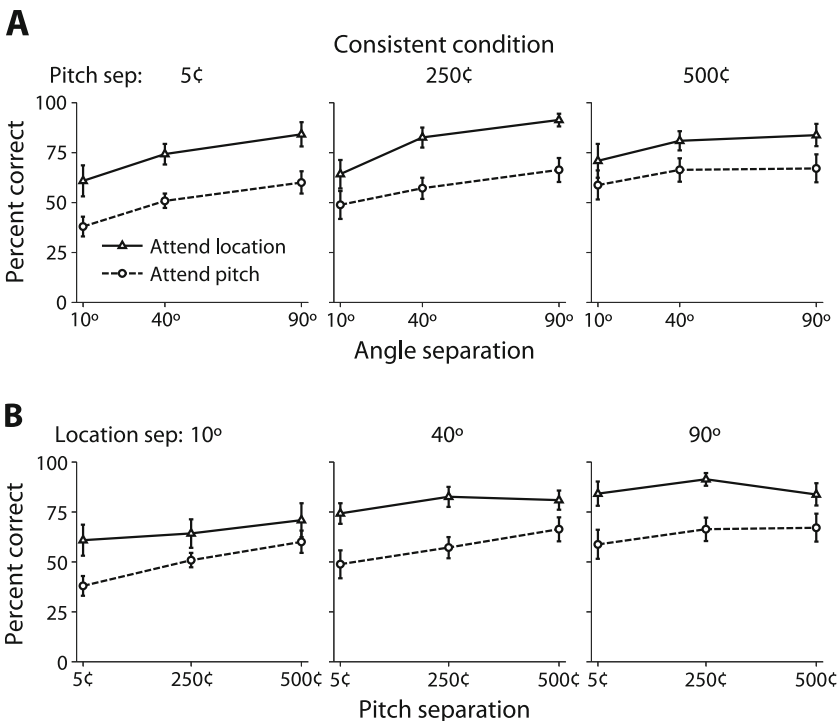


FIG. 3. The effect of directing attention in the *consistent* condition. There was a higher percentage of *correct* responses when attending location (solid lines, triangular markers) than when attending pitch (dashed lines, circular markers), despite the fact that for both sets of instructions, both location and pitch could have helped reinforce the correct digits. The x-axes of the upper panel (A) represent angle separation, with each of the three plots corresponding to one of three pitch separations, increasing from left to right. The lower panel (B) shows the same data with the x-axes now representing pitch separation and each of the three plots corresponding to one of the three location separations. *Correct* responses in the *attend-location* blocks increased with spatial separation, but were unaffected by pitch separation. In the *attend-pitch* trials, the percentage of *correct* responses increased with both the task-relevant feature and the task-irrelevant feature. Error bars represent ± 1 SEM.

factors were statistically significant ($p < 0.001$ for instruction, $p = 0.001$ for location, $p = 0.016$ for pitch); however, no interactions were significant. These results show that listeners changed how they performed the task based on what feature they were instructed to attend. Moreover, the fact that both location separation and pitch separation were statistically significant supports the idea that performance improved as any feature separation increased, whether or not the increase was in a feature that was task relevant.

Effects of instructions and feature separation in opposing trials

Another way to test whether instructions alter how listeners respond to identical sensory inputs is to compare, for the same stimuli, how often listeners reported digits that matched a particular feature of the primer in *opposing* trials, where continuity of location conflicts with continuity of pitch. Figure 4A shows the percentage of times that listeners reported the pair of digits that matched the primer location,

while Figure 4B plots the percentage of responses in which listeners reported the pair of digits that matched the primer pitch, directly contrasting results when attention is directed to location (solid lines) versus when it is directed to pitch (dashed lines). Results are shown as a function of the separation of the feature of the reported digits that matched the primer (angle in Fig. 4A and pitch in B). Each individual panel shows results for one of the three separations of the feature of the reported digits that did not match the primer.

Just as in the results from the *consistent* trials, there is a very strong effect of instructions: in all cases, listeners were more likely to report the digits that matched a particular primer phrase feature when they were instructed to attend to that feature than when instructed to attend to the other feature. Specifically, listeners were more likely to respond with digits that matched the primer location when instructed to attend location than when instructed to attend pitch (in Fig. 4A, solid lines always fall above the dashed lines). Similarly, listeners were more likely to respond

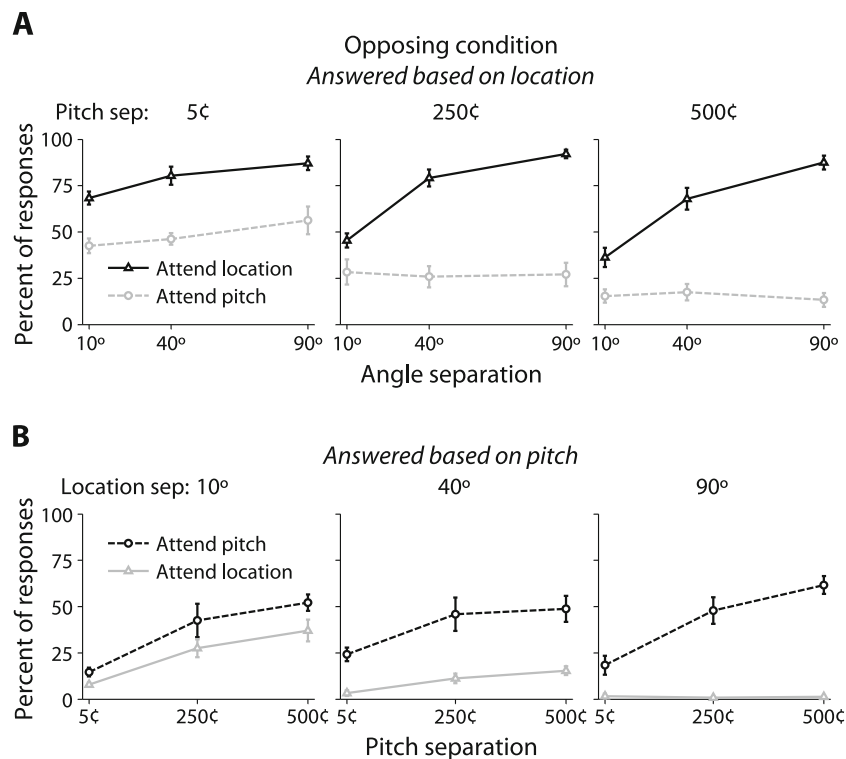


FIG. 4. The effect of directing attention in the *opposing* condition. **A** The percentage of trials for each configuration in which the listeners based their responses on location. As in Figure 3, trials from attend-location blocks are shown as *solid lines with triangular markers*, and trials from attend-pitch blocks are shown as *dashed lines with circular markers*. Using the same color codes from Figures 2 and 3, the *correct* responses (answered based on location in an *attend-location* block) are shown in black and wrong-feature responses (answered based on location in an *attend-pitch* block) are shown in light gray. Angle separation is on the x-axis and each of the

three plots in the row corresponds to one of the three pitch separations, increasing from *left to right*. **B** The same as (A), but for responses based on pitch. Accordingly, the x-axes now represent separation with each of the three plots corresponding to an angle separation. The line styles to indicate attend-pitch and attend-location trials are the same; however, the colors have been reversed to indicate that responding based on pitch is correct (*black*) in attend-pitch trials, and constitutes a *wrong-feature* response (*light gray*) in attend-location trials. Error bars are ± 1 SEM; where not visible, they are small enough to be hidden by the markers.

with digits that matched the primer pitch when instructed to attend pitch than when instructed to attend location (in Fig. 4B, dashed lines fall above the solid lines). These results confirm that the listeners changed how they performed the task based on what feature they were supposed to attend (i.e., listeners did not base judgments on some obligatory across-time grouping, but used feature-selective attention in performing the task, as instructed).

As noted previously, the likelihood of a *correct* response (solid lines in Fig. 4A and dashed lines in B) generally increased with the feature separation in the feature that the listeners were instructed to attend. In the *opposing* trials, this is seen in the fact that the *correct* responses are monotonically increasing in all panels of Figure 4 (in Fig. 4A, solid black lines are positively sloped; in Fig. 4B, dashed lines are positively sloped). In addition, when listeners were told to attend to the primer location (solid lines), the likelihood of responding correctly decreased as the task-irrelevant pitch separation increased, particularly when the task-relevant spatial separation was small (in Fig. 4A, the slope of the solid lines increases from the leftmost to the rightmost panels, primarily because the left endpoints of these lines, corresponding to the smallest spatial separation, drops lower as one looks from left to right panels). In other words, when listeners tried to direct attention to primer location, there was a clear influence of the to-be-ignored feature of pitch on how well they performed.

In contrast, when listeners were told to attend to the primer pitch, the likelihood of responding correctly was not strongly affected by the competing, to-be-ignored feature of location (in Fig. 4B, the dashed lines in each panel are very similar). Of course, unlike when listeners were instructed to attend location, performance was relatively poor in conflicting trials in which listeners were asked to attend pitch, reaching only about 60% correct for the largest pitch separation of 500¢; thus, this lack of an effect may be due to a floor effect, rather than providing evidence that location cues were not influencing performance. This interpretation is further supported by the relatively large percentage of *wrong-feature* responses in the *attend-pitch*, *opposing* results (see below).

In both the *attend-location* and the *attend-pitch* blocks, the likelihood of making a *wrong-feature* response decreased as the size of the task-relevant feature separation increased. For instance, the percentage of the time that listeners reported digits that matched the primer location when they were instructed to pay attention to the primer pitch decreased as the size of the pitch separation of target and masker digits increased (in Fig. 4A, the dashed

gray line is highest in the left panel and lowest in the right panel). Similarly, the percentage of the time that listeners reported digits that matched the pitch of the primer when they were instructed to pay attention to the primer location decreased as the size of the location separation of target and masker digits increased (in Fig. 4B, the solid gray line is highest in the left panel and lowest in the right panel).

In the *attend-location* trials, the percentage of the time that listeners incorrectly reported digits matching the primer pitch depended on an interaction between the size of the location separation and the size of the pitch separation. Specifically, when the location separation was small, the percentage of the time that listeners responded with digits that matched the primer pitch increased as the pitch separation increased (in the leftmost panel of Fig. 4B, the solid line is positively sloped). As the location separation increased, the likelihood of making a *wrong-feature* response in the *attend-location* trials decreased, to the point that they almost never occurred when the target and masker digits were separated by 90° (the solid line is near zero for all pitch separations in the rightmost panel of Fig. 4B).

In the *attend-pitch* trials, the percentage of the time that listeners incorrectly reported digits matching the primer location was relatively large, reaching as high as 50% (see dashed line in left panel of Fig. 4A). These *wrong-feature* errors increased weakly with location separation for the smallest pitch separation (in the leftmost panel of Fig. 4A, the dashed line has a small positive slope). However, no dependence on location separation was evident when the pitch separation was 250¢ or 500¢ (dashed lines are essentially flat in the middle and right panels of Fig. 4A).

Because the number of *correct* responses and *wrong-feature* responses are not independent, we performed repeated-measures, multi-way ANOVAs (with non-sphericity correction, as described above) only on the *wrong-feature* responses from *opposing* trials, separately for *attend-pitch* and *attend-location* instructions. Each ANOVA analyzed main factors of location separation and pitch separation, as well as their interaction. In the *attend-pitch* blocks, the percentage of trials in which listeners incorrectly based their answers on location (Fig. 4A, dashed gray lines) was significantly influenced by pitch separation ($p=0.017$); however, neither the location separation nor the location–pitch interaction term was significant ($p=0.089$ and 0.11 , respectively). In the *attend-location* blocks, the percentage of trials in which listeners incorrectly based their answers on pitch (Fig. 4B, solid gray lines) depended jointly on the location separation and the pitch separation (the location–pitch interaction term was significant; $p=0.004$). In addition, the main effect of pitch

separation was significant in its own right ($p < 0.001$), while the location separation was not ($p = 0.088$). Thus, consistent with the descriptions above, when listeners were instructed to attend location in the *opposing* trials, responses were influenced by the size of the task-irrelevant pitch separation, as well as the size of the task-relevant location separation. In the *opposing* trials in which listeners were instructed to attend pitch, responses were influenced directly by the task-relevant pitch separation, but the task-irrelevant location separation had no statistically significant influence.

DISCUSSION

Performance is not limited by audibility or intelligibility

In all conditions and for all location and pitch separations, there was a very low rate of *guessing* errors; instead, listeners nearly always reported digits that were present in the sound mixture (see Fig. 2). The fact that listeners rarely made *guessing* errors shows that they nearly always perceived a sequence of digits from the competing words making up the sound mixture. Rather than being constrained by limited intelligibility or audibility of the competing digits, in this experiment, performance was determined by how effectively listeners could direct top-down attention to select the correct target digits from the sound mixture.

Listeners adjust how they selectively attend based on task goals

Some past experiments have explored how the continuity of different auditory features influences what listeners perceive as coming from one auditory stream (Darwin and Hukin 1999; Darwin and Hukin 2000a, b; Darwin et al. 2003). Such studies prove that all features in a stimulus contribute to linking together words across time and that the influence of each specific feature depends on the strength of the continuity of that feature. However, these studies did not manipulate which feature listeners were instructed to attend. As such, these studies implicitly treat auditory features as having some obligatory, fixed influence on across-time grouping, rather than testing whether top-down attention can alter how listeners selectively listen in a sound mixture.

Here, listeners changed how they performed a selective attention task based on what feature they were told to attend. In *consistent* trials, target digits shared both the same location and the same pitch as the primer, and the correct response was the same, independent of instructions. However, performance was markedly better when attention was directed to location than when it was directed to pitch (see

Fig. 3). In addition, when location and pitch were pitted against one another (in *opposing* trials), the percentage of the time listeners reported digits that matched a particular feature was always greater when the listeners were told to attend to that feature than when they were told to attend to the opposing feature (see Fig. 4).

These results demonstrate that listeners change how they selectively listen to a sound mixture depending on which feature they are explicitly instructed to attend. Such fluid weighting of different acoustic features is likely to play an important role in everyday listening. For instance, if two competing talkers with similar-sounding voices come from different directions, a listener can focus on the known location of the talker of interest; in contrast, if two talkers are standing very close together, a listener can focus instead on the timbre of the desired talker's voice.

Continuity of a task-irrelevant feature influences performance

Although listeners altered how they performed our task based on what feature was task relevant, continuity of the task-irrelevant feature also influenced performance. The influence of the task-irrelevant feature was strongest when the task-relevant feature was weakest; specifically, when the task-relevant feature was weaker, the likelihood of responding with digits matching the task-irrelevant feature was higher.

As discussed below, probably because of the specific choices of stimulus parameters used here, location was relatively strong compared with pitch in the current study. This helps explain why in the *attend-pitch* trials, the size of the task-irrelevant location separation had little effect on performance: the task-irrelevant location feature had a quite large effect even for the smallest spatial separation. Specifically, for all tested location separations, *wrong-feature* responses were very likely to occur when listeners were instructed to attend pitch. Indeed, when listeners were instructed to attend pitch, *wrong-feature* responses, where listeners responded with digits that matched the primer location, were more likely than *correct* responses in both types of *mix* trials (see Fig. 2B). In addition, for the smallest pitch separation tested (5¢), the likelihood of a *wrong-feature* response in *opposing* trials was roughly 50% for all three spatial separations, much greater than the likelihood of a *correct* response (about 20%; see Fig. 4). In other words, because pitch separation was relatively weak compared with the spatial separations used here, the task-irrelevant location separation had a strong influence on performance in all *attend-pitch* trials, independent of the size of the angular separation of target and masker digits.

In *attend-location* trials, performance was better overall, presumably because the tested spatial separations were relatively salient. For *attend-location* trials, the likelihood of a *wrong-feature* response was always much smaller than the likelihood of a *correct* response, never exceeding about 40%. However, even though the task-irrelevant pitch feature was relatively weak compared with the task-relevant location feature, performance was affected by the size of the pitch separation in *attend-location* trials. For instance, in the *opposing* trials in which listeners were instructed to attend location but the task-relevant location separation was only 10°, the probability of answering correctly with digits matching the primer location decreased as the pitch separation increased (see the leftmost points of the solid lines in Fig. 4A). Similarly, in the *opposing* trials in which listeners were instructed to attend location, the likelihood of listeners reporting digits that matched the task-irrelevant primer pitch increased as the separation of the target and masker pitches increased for the small 10° spatial separation (in the leftmost panel of Fig. 4B, *wrong-feature* responses, shown by the solid lines, increase with pitch separation). If listeners had been able to completely ignore the relatively weak pitch feature when it was task irrelevant, there would be no such dependence on pitch separation in the *attend-location, opposing* trial responses. Similarly, if listeners were simply unable to use a relatively weak spatial separation, but could ignore the task-irrelevant pitch, there would be a large number of *wrong-feature* errors and of *confusion* errors in *attend-location* trials when the separation was only 10°; however, the percentages of these specific types of errors would be independent of the pitch separation. Instead, in the *attend-location* trials that had a small spatial separation, we found that the likelihood of *wrong-feature* errors increased as the pitch separation increased. Thus, even when a feature should be ignored to perform the task as instructed, its continuity has an obligatory influence on performance.

Previous studies have argued that auditory attention is object based (e.g., Alain and Arnott 2000; Shinn-Cunningham 2008). While not hard proof, the current results are consistent with this idea. Specifically, these results suggest that continuity of a task-irrelevant auditory feature influences performance, an effect that suggests that whatever word is in the attentional foreground is perceptually linked to subsequent words that share its features, even when these features are known to be task irrelevant.

Location and pitch have qualitatively similar effects

For the specific values of location and pitch that we tested in this study, location was a stronger cue than pitch. For instance, overall, performance was much

better when listeners directed attention to pitch than location (see Fig. 2). In *consistent* trials, performance was also markedly better when listeners directed attention to location than pitch (see Fig. 3). Furthermore, in *attend-pitch mix* trials, *wrong-cue* responses, where listeners reported digits that matched the primer location, were more likely than *correct* responses (see Fig. 2B).

Although performance was better in *attend-location* trials than in *attend-pitch* trials, the choice of the location and pitch separations of target and masker digits used here was arbitrary, not “matched” in any systematic way. Moreover, in both *attend-location* and *attend-pitch* trials, performance improved as the task-relevant feature separation increased. Together, these observations suggest that using smaller location separations and/or larger pitch separations would decrease the influence of location and increase the influence of pitch on performance. Thus, despite the fact that for the values we tested here, location had a stronger influence on performance than pitch, there is no reason to think that the way these two different features influence selective auditory attention is fundamentally different. Indeed, we find that location and pitch play similar dual roles: each can act as feature to which listeners can direct top-down attention and as a bottom-up cue to object continuity that influences the ability to selectively hear out a desired source from a sound mixture.

In vision studies, spatial attention is often contrasted with attention to non-spatial features, such as color or shape, as if the mechanisms governing spatial and non-spatial attention are fundamentally different. Such thinking is likely driven by the fact that object location is encoded automatically in the retina, with sources from different directions exciting different receptors; in contrast, other visual features must be computed more centrally. In audition, location must be computed from the signals reaching the left and right ears; as such, location information is similar to other features, such as pitch, that are derived from neural computations. Consistent with this view, the current results suggest that spatial and non-spatial features influence selective auditory attention similarly, both through bottom-up object continuity, and through top-down attentional modulation.

Conclusions

In natural listening conditions, any number of across-time characteristics can be used to distinguish and select between different sound sources. Normally, various cues and features such as location and pitch work in concert to define what sounds are linked perceptually into an ongoing auditory object to which a listener can attend. Here, we show that listeners can

bias selective attention by focusing on either location or pitch in an ongoing sound mixture. However, even though listeners can voluntarily modulate the influence of different auditory features on source selection, continuity of a task-irrelevant feature affects performance. In the most extreme cases, a strong task-irrelevant cue can overpower a weak task-relevant one (as when listeners try to attend to a small pitch difference, but end up reporting words that share the same location across time). Results are consistent with the view that auditory attention is object based, selecting auditory streams whose across-time grouping is influenced by auditory feature continuity of both task-irrelevant as well as task-relevant features; however, the influence of specific auditory features is modulated by volitional, top-down attention.

ACKNOWLEDGMENTS

We would like to thank Rafael Alvarez-Berrios and Timothy Streeter for their help with stimulus generation and data collection. Funding for this research was provided by the Office of Naval Research and the National Institutes of Health.

REFERENCES

- ALAIN C, ARNOTT SR (2000) Selectively attending to auditory objects. *Front Biosci* 5:D202–D212
- BEST V, OZMERAL EJ ET AL (2007) Visually-guided attention enhances target identification in a complex auditory scene. *J Assoc Res Otolaryngol* 8(2):294–304
- BEST V, OZMERAL EJ ET AL (2008) Object continuity enhances selective auditory attention. *Proc Nat Acad Sci* 105(35):13174–13178
- BLASER E, PYLYSHYN ZW ET AL (2000) Tracking an object through feature space. *Nature* 408(6809):196–199
- BREGMAN AS (1990) *Auditory scene analysis: the perceptual organization of sound*. MIT Press, Cambridge
- BRUNGART DS, SIMPSON BD (2002) The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. *J Acoust Soc Am* 112(2):664–676
- BRUNGART DS, SIMPSON BD (2007) Effect of target-masker similarity on across-ear interference in a dichotic cocktail-party listening task. *J Acoust Soc Am* 122(3):1724
- CARLYON RP, CUSACK R ET AL (2001) Effects of attention and unilateral neglect on auditory stream segregation. *J Exp Psychol Hum Percept Perform* 27(1):115–127
- CUSACK R, CARLYON RP ET AL (2000) Neglect between but not within auditory objects. *J Cogn Neurosci* 12(6):1056–1065
- CUSACK R, DEEKS J ET AL (2004) Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *J Exp Psychol Hum Percept Perform* 30(4):643–656
- DARWIN CJ, CARLYON RP (1995) Auditory grouping. In: Moore BCJ (ed) *Hearing*. Academic, San Diego, pp 387–424
- DARWIN CJ, HUKIN RW (1999) Auditory objects of attention: the role of interaural time differences. *J Exp Psychol Hum Percept Perform* 25(3):617–629
- DARWIN CJ, HUKIN RW (2000A) Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *J Acoust Soc Am* 107(2):970–977
- DARWIN CJ, HUKIN RW (2000B) Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention. *J Acoust Soc Am* 108(1):335–342
- DARWIN CJ, BRUNGART DS ET AL (2003) Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *J Acoust Soc Am* 114(5):2913–2922
- DESIMONE R, DUNCAN J (1995) Neural mechanisms of selective visual attention. *Annu Rev Neurosci* 18:193–222
- ERAMUDUGOLLA R, IRVINE DR ET AL (2005) Directed attention eliminates ‘change deafness’ in complex auditory scenes. *Curr Biol* 15(12):1108–1113
- FREYMAN RL, HELFER KS ET AL (1999) The role of perceived spatial separation in the unmasking of speech. *J Acoust Soc Am* 106(6):3578–3588
- HELFER KS, FREYMAN RL (2009) Lexical and indexical cues in masking by competing speech. *J Acoust Soc Am* 125(1):447–456
- IHLEFELD A, SHINN-CUNNINGHAM B (2008A) Disentangling the effects of spatial cues on selection and formation of auditory objects. *J Acoust Soc Am* 124(4):2224–2235
- IHLEFELD A, SHINN-CUNNINGHAM B (2008B) Spatial release from energetic and informational masking in a selective speech identification task. *J Acoust Soc Am* 123(6):4369–4379
- KIDD G JR, ARBOGAST TL ET AL (2005) The advantage of knowing where to listen. *J Acoust Soc Am* 118(6):3804–3815
- MACKEN WJ, TREMBLAY S ET AL (2003) Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. *J Exp Psychol Hum Percept Perform* 29(1):43–51
- MARRONE N, MASON CR ET AL (2008) Tuning in the spatial dimension: evidence from a masked speech identification task. *J Acoust Soc Am* 124(2):1146–1158
- RAKERD B, AARONSON NL ET AL (2006) Release from speech-on-speech masking by adding a delayed masker at a different location. *J Acoust Soc Am* 119(3):1597–1605
- SHINN-CUNNINGHAM BG (2008) Object-based auditory and visual attention. *Trends Cogn Sci* 12(5):182–186
- SHINN-CUNNINGHAM BG, BEST V (2008) Selective attention in normal and impaired hearing. *Trends Amplif* 12(4):283–299
- SHINN-CUNNINGHAM BG, KOPCO N ET AL (2005) Localizing nearby sound sources in a classroom: binaural room impulse responses. *J Acoust Soc Am* 117(5):3100–3115