JASA ARTICLE

Spatial cues can support auditory figure-ground segregation

Darrin K. Reed,[1] Maria Chait,[2] Brigitta Tóth,[3,a)] István Winkler,[3] and Barbara Shinn-Cunningham[4]

[1]*Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215, USA*

[2]*Ear Institute, University College London, London, United Kingdom*

[3]*Institute of Cognitive Neuroscience and Psychology, Center for Natural Sciences, Budapest, Hungary*

[4]*Carnegie Mellon Neuroscience Institute, Department of Biomedical Engineering, College of Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA*

ABSTRACT:

A study by Tóth, Kocsis, Háden, Szerafin, Shinn-Cunningham, and Winkler [Neuroimage **141**, 108 – 119 (2016)] reported that spatial cues (such as interaural differences or ITDs) that differentiate the perceived sound source directions of a target tone sequence (figure) from simultaneous distracting tones (background) did not improve the ability of participants to detect the target sequence. The present study aims to investigate more systematically whether spatially separating a complex auditory "figure" from the background auditory stream may enhance the detection of a target in a cluttered auditory scene. Results of the presented experiment suggest that the previous negative results arose because of the specific experimental conditions tested. Here the authors find that ITDs provide a clear benefit for detecting a target tone sequence amid a mixture of other simultaneous tone bursts.

 https://doi.org/10.1121/10.0001387

(Received 31 October 2019; revised 11 May 2020; accepted 20 May 2020; published online 9 June 2020)

[Editor: Adrian K. C. Lee] Pages: 3814–3818

I. INTRODUCTION

Spatial information is utilized in the processing of complex auditory scenes [for review see Bizley and Cohen (2013); Shinn-Cunningham *et al.* (2017)]. For stimuli consisting of brief tone bursts with randomly selected frequencies, detection of a target tone frequency that is repeated in successive bursts can be improved using spatial cues (Kidd *et al.*, 1994). Specifically, when the repeated target tone bursts are presented to one ear, detection of the target is improved when the interfering tones are presented to both ears instead of being presented only to the ear in which the target tones are presented. These results show that perceived differences in location promote the perceptual segregation of the target and masker. Spatial information can also aid in detecting patterns of tones when they are presented amid other randomly varying task-irrelevant tones. For instance, Kidd *et al.* (1998) demonstrated that an angular separation between a target tone pattern and random task-irrelevant tones improved the ability to detect the target. Because the interference from task-irrelevant tones likely results from increased stimulus uncertainty rather than energetic masking (i.e., masking of the representation of the target tones at the level of the cochlea), spatial separation could have provided a definitive cue allowing listeners to segregate and focus attention on the target from within the task-irrelevant tones.

The present study aims to investigate systematically whether spatially separating a complex auditory "figure" from the background auditory stream may enhance the detection of a target in a cluttered auditory scene.

The paradigm is based on a previous study (Tóth *et al.*, 2016) where listeners detected repeating target tones of inharmonic frequencies composed of a random set of pure tones within stimuli consisting of randomly varying tonal elements. The repeating pattern was perceived as a figure amid the randomly changing background (Teki *et al.*, 2011; Teki *et al.*, 2013). The aim of experiment 1 was to test whether a location difference between the frequency components assigned to the figure and the ground enhanced their perceptual separation (no difference, roughly 45° difference, or roughly 90° difference). Detection performance improved both as the number of pure tones making up each repeated complex increased (figure coherence; the figure contains either four or six tonal components), and as the number of repeated complexes increased (duration; the number of repeated chords was three, four, or five). Given that spatial information generally improves detection and pattern identification performance, the lack of benefit from spatial cues found in experiment 1 of Tóth *et al.* (2016) are surprising. Notably, these authors reported that a large spatial separation between a target and simultaneous masker actually interfered with target detection. They later reported correction (Tóth *et al.*, 2016) states that, due to a programming error in the code generating the stimuli, all lateralized events belonged to figure tones (the spatial location of the ground tones was not manipulated). The target, if it appeared, was either diotic or, on a small number of trials, lateralized. Importantly, there was no trial-by-trial feedback. This experimental structure may have biased the listeners to

a)Electronic mail: toth.brigitta@ttk.mta.hu

https://doi.org/10.1121/10.0001387

focus attention to sounds on the midline, which were always present, and to listen carefully to spectral rather than spatial cues to detect the target figure. Thus Tóth *et al*. (2016) did not convincingly test whether spatially separating the figure from the background auditory stream can enhance the detection of a target in a cluttered auditory scene. The goal of the current study was to test this issue in a systematic manner. Here we systematically vary the spatial separation between the figure and background to determine effects on perceptual segregation.

II. METHODS

The stimuli used in the study were based on the design of Tóth *et al*. (2016). Signals were composed of successive inharmonic tone complexes, referred to as "chords." The pure tones comprising the chords were selected from a 179–7246 Hz frequency range with uniform logarithmical spacing in steps of 0.5 semitones. Each trial consisted of a sequence of 40 randomly generated chords of 50 ms duration. Chords were temporally adjacent and included 10 ms raised-cosine onset/offset ramps to reduce spectral splatter (spread of spectral content at abrupt onsets and offsets). It contrasts with the design of Tóth *et al*. (2016), in which each chord consisted of 9–21 tones. All chords here contained a fixed number of ten tones. In half of the stimuli, four of the ten tones in a chord were repeated over either three or five chords. These repeated tones collectively formed a "figure" amid the background of random chords. The onset of these figure chords randomly occurred between the 15th and 20th chord (750–1000 ms after the stimulus onset) as in Teki *et al*. (2011) and Teki *et al*. (2013). In the other half of the stimuli, all ten tones were randomly selected from chord to chord, forming background-only (control) trials. Throughout the study, participants were instructed to indicate whether or not they detected the presence of a figure by pressing one of two response buttons at the end of each trial.

The percept of lateralization in the present study refers to the perceived difference in lateral direction of a set of tonal elements (either the figure or a comparable number of background tones) relative to the diotic background tones. Lateralization was achieved using only interaural time differences (ITD). This insured that any improvement in detection of the figure was not simply a result of an increase in the power of the figure tones relative to the background tones in a given ear but was instead a result of perceived spatial differences. Schematic representations of the three interaural conditions tested in the present study (diotic, lateral-burst, and lateral-stream) are shown in Fig. 1. Note that for each of these conditions, we included trials that contained a figure (top row of Fig. 1) and control (background-only) trials (bottom row of Fig. 1). For simplicity, Fig. 1 only depicts configurations with the ITDs promoting the stimuli to be lateralized to the right of midline; however, in the actual experiment, both right- and left-lateralized stimuli appeared with equal probability.

In the diotic condition (whether the trial included a figure or was a control trial), all tones were presented diotically, with at an ITD of 0 μs (target and background were both perceived at the midline). In lateral-burst figure trials, the figure tones were presented with an ITD of either +685 μs or −685 μs (roughly at $\pm 90^{\circ}$ perceived angle), whereas all background tones were presented diotically. Thus, in these trials, the background was at the midline, while partway through the trial, a lateralized target figure appeared at an angle to the left or to the right. We hypothesized that the brief lateral event is salient and draws attention exogenously both to the direction and to the time at which the target will appear if it is present. This exogenous draw of attention may make it easier for the listener to detect figure tones.

The study included the lateral-stream condition as well, which was tested by an alternative explanation for the interference of lateralization observed in the Tóth *et al*. (2016) study. We hypothesized that the contrast between the irregularity of the background and the regularity of the figure may results in a salient transition, which might aid in detecting the figure amid the background. In the lateral-burst control trials, a subset of the background tones (equal in number, duration, and temporal probability to the figure tones) was presented with an ITD of either +685 μs or −685 μs while all other tones were presented diotically. In other words, in these trials, the whole background was presented at the midline at the start of the trial, while a portion of the background switched to a lateral position partway through the trial. In the lateral-stream conditions, for the entire duration of the trial, six of the ten tones in each background chord were presented diotically for the entire duration of the trial, while the remaining four tones were presented at an ITD of either +685 μs or −685 μs (lateral-stream control trials). In the lateral-stream figure trials, partway through the trial, the figure tones were added with an ITD that matched the ITD of the lateralized part of the ongoing background.

Ten individuals (seven male, three female) participated in the current study with age ranging from 18 to 34 yr (mean 25.2). All had normal hearing sensitivity, with standard audiometric air-conductive thresholds of 20 dB hearing level (HL) or less for pure tones from 0.125 to 8 kHz. Participants provided written informed consent as approved by the Boston University Institutional Review Board. Stimuli were presented through Etymotic ER-1 insert earphones in an acoustically treated booth. The overall stimulus presentation level was 70 dB sound pressure level (root-mean-squared). All stimuli were randomly generated on a trial-by-trial basis using MATLAB.

Prior to beginning the experiment, all participants completed a 15-min training session to familiarize them with the task and temporal structure of the trials. A series of five stimulus blocks with trial-by-trial feedback were delivered to participants during training. The first three training blocks contained trials with a figure/control duration of 12, 8, or 5 chords (presented in order of decreasing duration) using the diotic configuration. The fourth and fifth blocks demonstrated the lateral-burst and the lateral-stream interaural

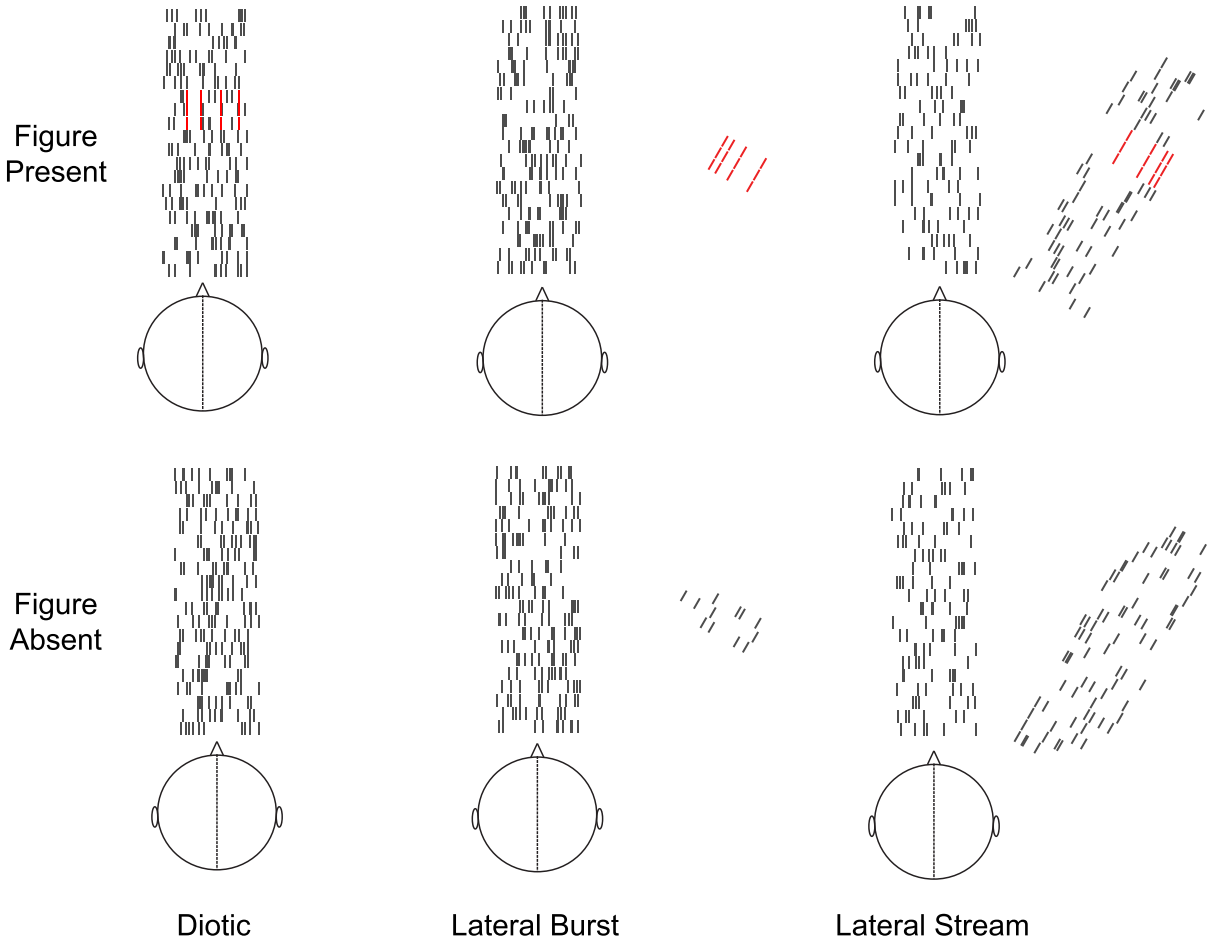

FIG. 1. (Color online) Schematic illustration of the figure and control (figure absent) trials for each of the three interaural conditions. The trial types are shown in separate rows and interaural conditions are show in separate columns. Black vertical lines depict random tones (background) while red dotted lines represent repeating target patterns (figure). The perceived position of the stimulus is shown relative to the listener's head. In the diotic target and background were both perceived at the midline. In lateral-burst figure trials, the background was perceived at the midline (figure partway through the trial) appeared at an angle to the left or to the right. In the lateral-burst control trials, at the start of the trial, the whole background was presented at the midline, while a portion of the background switched to a lateral position partway through the trial.

conditions, respectively, using figure/control duration of five chords. Participants were given the opportunity to repeat any of the blocks until they felt comfortable with the task. Although listeners did not have to meet some strict criteria in the training phase before beginning the experiment, they typically exceeded 70% accuracy in all three of the interaural configurations for the five-chord figure duration.

Following the practice session, participants took part in a test session consisting of 48 trials for each of the 12 trial types: both figure and control trials for each of the three interaural conditions, all tested with figure durations of three and five chords (2 figure presence $\times$ 3 interaural condition $\times$ 2 figure durations $=$ 12 trial types). For the lateral-burst and lateral-stream conditions, the 48 trials were divided equally between left and right lateral presentations. The grand total of 576 trials (48 trials for each of the 12 conditions) were broken into 16 blocks of 36 trials each. Block delivery was self-paced: Between blocks, participants could take a break. Within a block we randomly intermixed different interaural conditions (diotic, lateral burst, and lateral stream) that had the same target duration (either 3 or 5). All participants were first presented with the (easier) five-chord duration trials (the first eight blocks). Because the focus of this study was to investigate differences related to the interaural conditions, this potential ordering effect of the duration parameter does not confound our intended research question while allowing for a shorter training session. Listeners were provided both trial-by-trial visual feedback (green cues after correct responses and red cues following incorrect responses) and a performance summary after each block. Participants typically took approximately 40 min to complete the test session.

III. RESULTS

The accuracy of discriminating figure and background-only trials was quantified using d′ (Green and Swets, 1966). A linear statistical model relating performance across blocks to the serial order of the block showed no significant effect [$F(1,478) = 0.18$, $p = 0.68$]. A similar linear model relating performance to the figure lateral position, i.e., left lateralized versus right-lateralized stimuli, also showed no

https://doi.org/10.1121/10.0001387

significant effect [$F(1,78) = 0.17$, $p = 0.67$]. Therefore, trials were collapsed across stimulus blocks and the lateralization direction, separately for the two figure durations, for the three lateralization conditions, and for figure and control trials.

It is worth noting that the distribution of residuals from a linear model fitted to the data violated the assumption of normally distributed data (Shapiro-Wilk; $W = 0.99$, $p = 0.03$). To assess the impact of this assumption violation, a variety of non-parametric tests were conducted, including Friedman's test and pairwise comparisons using percentile bootstrap methods (Wilcox, 2011). Qualitatively, the results of this more complex analysis lead to conclusions similar to those from the methods described above. Therefore, for the sake of simplicity, results of the more conventional methods are reported here.

Figure 2 shows group performance for the three interaural conditions and two different figure/chord durations. Supplemental Fig. 1 depicts the individual performance for each condition and stimulus type.[1] A two-way repeated measures analysis of variance (ANOVA) was performed on the collapsed dataset with figure duration and interaural condition as within-subjects factors. Listeners were much better at detecting figures with longer duration, confirmed by the statistically significant main effect of duration [$F(1,18) = 92$, $p < 0.001$]. A statistically significant main effect of interaural configuration was also observed [$F(2,18) = 35$, $p < 0.001$]. The interaction between figure/chord duration and interaural condition was not statistically significant.

To assess differences in performance between the three interaural conditions for the two figure duration, a *post hoc* analysis was conducted using Bonferroni adjusted alpha levels of 0.00833 (0.05/6) per test. Results from paired t-tests indicate that listeners were better at detecting the figure in the lateral-burst condition as compared to the diotic condition for both figure durations three [$t(9) = -6.58$, $p < 0.001$] and five [$t(9) = -6.41$, $p < 0.001$]. Participants were also better at detecting the figure in the lateral-stream condition as compared to the diotic condition for figure duration five [$t(9) = -3.97$, $p = 0.003$]. With respect to the two binaural conditions, detection performance was better in the lateral-burst than in the lateral-stream conditions for figure duration three [$t(9) = 4.43$, $p = 0.002$].

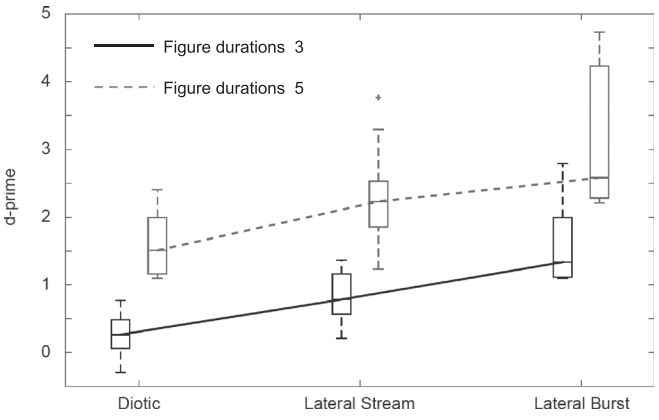

FIG. 2. Group-averaged ($N = 10$) d′ values (standard error of mean represented by bars) are shown across the three interaural conditions and two different figure durations. The figure chord durations marked by the white color for duration 3 and the dark grey color for duration 5. The three interaural conditions are shown in the three separate columns.

IV. DISCUSSION

Results demonstrate a clear benefit of spatial information on figure detection ability. Although this result is consistent with expectations derived from prior studies using similar stimuli (Kidd *et al.*, 1994; Kidd *et al.*, 1998), it conflicts with the findings of Tóth *et al.* (2016). We argue that this discrepancy is most likely driven by differences in details of the experimental procedures and methods, ultimately concluding that binaural cues do indeed facilitate figure detection.

In the study of Tóth *et al.* (2016), whenever a lateralized event was present, it was always a figure. In many ways, this makes it even more surprising that in this study, participants in that study performed worse in detecting the figure in the lateral-burst condition than in the diotic condition. However, all stimulus types were intermingled in the original design, many of which were very challenging diotic trials. Especially, given that no feedback was provided, we believe that the results of Tóth *et al.* were a consequence of the mixed design in which participants may have simply focused their attention to the midline to facilitate detection on the majority of trials. A strategy of focusing attention to the midline would tend to suppress lateralized sounds, and thus would make it more difficult to detect less salient lateralized figure objects. Indeed, percent-correct results (not shown in the original manuscript, which only reported sensitivity) support this interpretation. Specifically, performance was poor in the lateral-burst-figure condition because of a high incidence of missed (lateralized) targets. This supports the view that focused attention to midline overcame what one might think was a salient, sudden appearance of a lateralized target. Consistent with the idea that listeners were highly focused on midline, performance in the diotic conditions was particularly good (e.g., as compared to similar conditions tested in Teki *et al.*, 2013).

In addition to testing the effect of separating the figure from the background auditory stream, the current study included the lateral-stream condition, which was motivated by an alternative explanation for the interference of lateralization observed in the Tóth *et al.* (2016) study. We hypothesized that the contrast between the irregularity of the background and the regularity of the figure may result in a salient transition, which might aid in detecting the figure amid the background. If this hypothesis were correct, then performance in the lateral-stream configuration should have exceeded performance in the lateral-burst configuration, since the change in the ongoing lateral stream would provide a clear cue. Instead, the opposite was found: listeners performed best in the lateral-burst configuration, and were on average better in the lateral-burst configuration than the

lateral-stream configuration. Thus, the current results suggest that the brief lateral event draws attention exogenously both to the direction and to the time at which the target will appear if it is present. This exogenous draw of attention appears to make it easier for the listener to decide whether that event is a repeated figure or a random pattern of background tones (consistent with Kidd *et al.*, 1998). In the lateral-stream condition, listeners also do better than in the diotic condition, but not as well as in the lateral-burst condition. Presumably, when there is a lateral stream at the start of a trial, listeners know to focus their attention in that direction and monitor the lateral stream for the appearance of the target; that is, they know where to direct attention, but still require spectral cues to determine when the target appears, if it is present. Because the number of background tones in the lateral stream is smaller than the number of background tones at the midline in the diotic condition, figure detection is enhanced relative to the diotic condition. This explanation accounts for the beneficial effects of spatial cues while also explaining performance improvements in the lateral-stream configuration over the diotic configuration. Put another way, the mechanisms contributing to target detection differ across configurations. In the diotic and lateral-burst conditions, listeners must monitor a random background for the emergence of a repeated target that occurs at some unknown time; this regularity is more prominent when listeners need to monitor only a subset of background tones (lateral-burst configuration) than when they must monitor a denser background (diotic configuration). In contrast, in the lateral-burst condition, attention is drawn exogenously in both time and location to a small set of tones that are either repeated (target present) or unstructured (control), removing all temporal uncertainty.

An alternative but arguably less likely explanation of the improved performance in the lateralized burst condition relative to the diotic condition would suggest that listeners were using the change in the number of tones comprising the chord in the midline as a task cue. A reduced number of midline tones would then prompt listeners to shift their attention to one of the two lateralized bursts that they then evaluate for the presence of a target. Again, once attention shifts to the side, they can evaluate the stimulus for the presence of the figure.

An interesting question for future work is to determine how an acoustic feature such as ITD influences brain responses (e.g., as measured with M/EEG) when detecting figure objects amid random background tones.

ACKNOWLEDGMENT

This work was supported by National Institute of Deafness and Communication Disorders R01 DC013825 to B.S.C., Erasmus Mundus Auditory Cognitive Neuroscience grant (to B.T.), the National Research Development and Innovation Office (123790 projects) awarded to T.B., and by the Hungarian Academy of Sciences [Magyar Tudományos Akadémia (MTA)], the János Bolyai personal grant awarded to T.B. The authors declare no conflict of interest.

[1]See supplementary material at https://doi.org/10.1121/10.0001387 for individual d′ values across the three interaural conditions and two different figure durations.

Bizley, J. K., and Cohen, Y. E. (**2013**). "The what, where and how of auditory-object perception," Nat. Rev. Neurosci. **14**, 693–707.

Green, D. M., and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics* (Wiley, New York), Vol. 1.

Kidd, G., Jr., Mason, C. R., Deliwala, P. S., Woods, W. S., and Colburn, H. S. (**1994**). "Reducing informational masking by sound segregation," J. Acoust. Soc. Am. **95**, 3475–3480.

Kidd, G., Jr., Mason, C. R., Rohtla, T. L., and Deliwala, P. S. (**1998**). "Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns," J. Acoust. Soc. Am. **104**, 422–431.

Shinn-Cunningham, B., Best, V., and Lee, A. K. C. (**2017**). "Auditory object formation and selection," in *Auditory System Cocktail Party* (Springer, New York), pp. 7–40.

Teki, S., Chait, M., Kumar, S., Shamma, S., and Griffiths, T. D. (**2013**). "Segregation of complex acoustic scenes based on temporal coherence," Elife **2**, e00699.

Teki, S., Chait, M., Kumar, S., von Kriegstein, K., and Griffiths, T. D. (**2011**). "Brain bases for auditory stimulus-driven figure-ground segregation," J. Neurosci. **31**, 164–171.

Tóth, B., Kocsis, Z., Háden, G. P., Szerafin, Á., Shinn-Cunningham, B. G., and Winkler, I. (**2016**). "EEG signatures accompanying auditory figure-ground segregation," Neuroimage **141**, 108–119. Correction in Neuroimage 2017 Dec 9.

Wilcox, R. (**2011**). *Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction* (CRC, Boca Raton, FL).