Selective attention in an overcrowded auditory scene: Implications for auditory-based brain-computer interface design

Ross K. Maddox
Department of Speech and Hearing Sciences, and Institute for Learning and Brain Sciences, University of Washington, Seattle, Washington 98195
rkmaddox@uw.edu

Willy Cheung
Institute for Learning and Brain Sciences, University of Washington, Seattle, Washington 98195
wllychng27@gmail.com

Adrian K. C. Lee[a]
Department of Speech and Hearing Sciences, and Institute for Learning and Brain Sciences, University of Washington, Seattle, Washington 98195
akclee@uw.edu

Abstract: Listeners are good at attending to one auditory stream in a crowded environment. However, is there an upper limit of streams present in an auditory scene at which this selective attention breaks down? Here, participants were asked to attend one stream of spoken letters amidst other letter streams. In half of the trials, an initial primer was played, cueing subjects to the sound configuration. Results indicate that performance increases with token repetitions. Priming provided a performance benefit, suggesting that stream selection, not formation, is the bottleneck associated with attention in an overcrowded scene. Results' implications for brain-computer interfaces are discussed.

PACS numbers: 43.66.Lj, 43.66.Mk [QJF]
Date Received: July 13, 2012 **Date Accepted:** September 17, 2012

1. Introduction

An auditory stream can be operationally defined as the percept of a group of successive or simultaneous sound elements appearing to emanate from a single source.[1] Many past studies have examined the acoustical properties that influence stream formation. For example, the number of streams perceived in an auditory scene depends on the similarity and proximity of sounds presented in sequence.[2] When evidence is ambiguous as to whether sound elements should be fused into one stream or not, the tendency for sound elements to split into different streams builds up with time exposure.[3]

Traditionally, these auditory streaming studies have concentrated on ambiguous stimuli, e.g., an alternating tone sequence,[4] leading to perceptual bi- or multistability.[5] However, real world stimuli are much more stable, e.g., voices and music, and stream formation is unambiguous due to the temporal coherence of the constituent spectrotemporal elements originating from the same source.[6] Listeners in general can direct their attention to these well-formed objects[7] but when there are many objects in the auditory scene, selective attention has also been shown to improve over time.[8]

[a] Author to whom correspondence should be addressed.

Can an auditory scene become overcrowded by so many streams that the ability to selectively attend breaks down? Does it take more time for a listener to selectively attend to one stream as an auditory scene gets more crowded? These questions are phrased not only to further our understanding of auditory object-based attention but also to maximize information that can be effectively conveyed in an auditory display. In recent years, auditory-based brain-computer interface (BCI) research has intensified, in hopes of providing completely locked-in syndrome patients a mode of communication based on classifying neural responses (usually measured with electroencephalography) resulting from the user's selective attention to predefined auditory stimuli.[9] Current methods usually use two streams;[10,11] however, the more sound streams an auditory scene contains, the more options can be presented in an auditory-based BCI. Critically, the effective transfer bit rate in an auditory-based BCI increases with the number of streams present in the scene while trading off with the user's ability to accurately attend to one auditory stream of choice.[11]

Here we designed two behavioral experiments to examine how well a listener can focus on one auditory stream in the presence of many. With the objective of maximizing BCI information transfer rate in mind, we presented listeners with several repeating letter streams (4, 6, 8, and 12 streams in the scene), with each stream having a unique virtual location and pitch to aid segregation. Each stream consisted of the same, unique repeated letter. In the first experiment, the letter (or token) start times were distributed uniformly within each 1 s repetition (i.e., once started, each stream was periodic). Therefore, the rate at which tokens were presented varied with the number of streams present in the scene. In the second experiment, the overall token rate was held constant. In both of these experiments, half of the trials were preceded by an additional primer in which all letters were played once sequentially (from the left side, lowest pitch to the right side, highest pitch), cueing the listener to the sound configuration. Results show that listeners can selectively attend to one stream quite accurately, even with up to 12 streams present in the scene. Furthermore, the auditory primer significantly improved behavioral performance. Together, these results provide us with crucial insights into how to optimize an auditory display for eventual BCI deployment.

2. Experiment 1

2.1 Methods

Nineteen subjects (7 male, aged 19 to 45 yrs) took part in this experiment. All participants had pure-tone thresholds in both ears within 20 dB of normal-hearing thresholds at octave frequencies between 250 and 8000 Hz. All subjects gave informed consent to participate in the study as overseen by the University of Washington Institutional Review Board.

Stimuli consisted of streams of repeated spoken letters (randomly chosen, one-syllable letters from 1 female talker; average duration 431 ± 9 ms), monotonized using Praat software and processed with pseudo-anechoic head-related transfer functions (recorded from a KEMAR manikin at 1 m), as shown in Fig. 1. All stimuli were generated at a sampling rate of 24 414 Hz and sent to Tucker-Davis Technologies hardware for digital-to-analog conversion and attenuation, and then presented over in-ear headphones (Etymotic Research ER-2).

Listeners were instructed visually which letter stream to attend and to promptly press a button when an oddball "R" replaced a letter in the designated stream (responses made within 900 ms were counted as hits; all other presses were deemed false alarms). Each trial contained 16 repetitions of letters, with each letter repetition period fixed at 1 s regardless of the number of streams present (4, 6, 8, or 12). Thus each trial lasted 16 s. Two oddballs were placed in the designated stream and three oddballs were seeded randomly in other streams in each trial. Oddballs were more frequent in the non-target streams than they were in the target stream such that listeners that adopted a strategy of

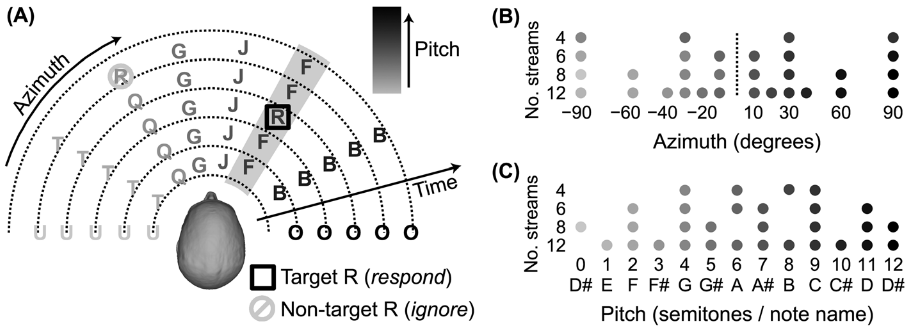

Fig. 1. Stimulus generation. (A) Repeating letter streams were presented, with each stream having a characteristic azimuth and pitch (shown in grayscale). The attended stream, "F," is highlighted. Where R replaces F is a target; where it replaces other letters is not. (B) The azimuths at which letter streams were presented for each stream count. Grayscale codes pitch as in (A). (C) The pitches at which letter streams were presented. Pitches were the first 4, 6, or 8 notes of a diatonic scale, or the chromatic scale in the case of 12 streams.

simply reporting every oddball heard would have high false alarm rates (they did not; see Sec. 2.2). In half the trials (randomly distributed), there was a primer in which each letter was played sequentially prior to the start of the trial in order to allow the listeners to learn that trial's sound configuration. To mimic the degree of energetic masking (caused by interference in the auditory periphery) and maximally reduce informational masking (i.e., other factors not explained by energetic masking), a control condition was included in which a target letter stream was still designated but all other streams were substituted with the letter "O" (12 stream condition; all target and non-target R's remained; designated stream was never O). A total of 60 trials per condition were tested and the testing procedure lasted less than 2 h (with breaks). Listeners received training prior to experimentation (training criterion of 8/10 targets in 5 trials before they could begin the experiment). Steady-state performance was calculated for each subject as the average hit percentage during the 5 s preceding the final second (final second was excluded since some subjects were reluctant to respond once the stimulus had ended). A two-way repeated measures analysis of variance (ANOVA) with factors of primer and number of streams and an interaction term was used for statistical testing of study hypotheses. In cases where the assumption of sphericity was deemed to be violated, Greenhouse–Geisser correction was used.

2.2 Results

Figure 2(A) shows the across-subject mean [± 1 standard error of the mean (SEM)] of the steady-state performance (bars) and performance on the first repetition of each trial (white dots). Without a primer the steady-state performances were 93.8 ± 1.8%, 85.4 ± 3.2%, 73.2 ± 3.7%, and 60.4 ± 4.4% for 4, 6, 8, and 12 streams, respectively (mean ± SEM). With a primer, they were 95.9 ± 1.3%, 89.1 ± 2.9%, 76.3 ± 3.8%, and 61.2 ± 4.0%. On the control condition, steady-state performance was 94.3 ± 1.6%. Without a primer the first-repetition performances were 61.8 ± 6.4%, 48.7 ± 6.8%, 42.1 ± 7.7%, and 38.2 ± 7.7% for 4, 6, 8, and 12 streams, respectively. With a primer, they were 90.8 ± 3.9%, 75.0 ± 5.7%, 68.4 ± 7.4%, and 34.2 ± 6.4%. On the control condition, first-repetition performance was 68.4 ± 6.3%. There was a significant effect of the number of streams present in the scene on steady-state performance [$F(3,54) = 64.7$, $p < 0.001$] and on first-repetition performance [$F(3,54) = 22.8$, $p < 0.001$]. The existence effect of the auditory primer was not significant for steady-state [$F(1,18) = 4.17$, $p = 0.056$] but was for first-repetition performance [$F(1,18) = 18.3$, $p < 0.001$]. There was a significant interaction between primer and the number of streams for first-repetition performance [$F(3,54) = 6.68$, $p = 0.001$]. Steady-state performance in the

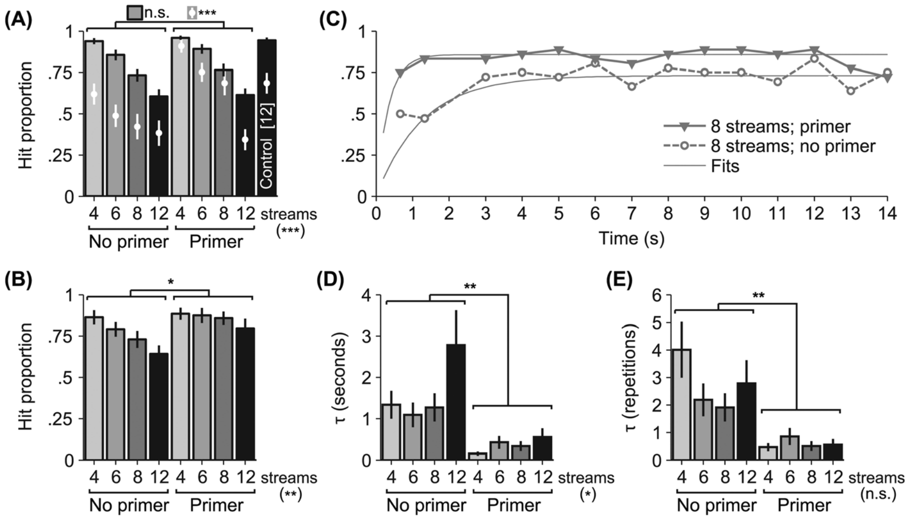

Fig. 2. Behavioral results. (A) The steady-state performance (bars) and first-repetition performance (white dots) for the first experiment (cross-subject mean ± SEM). (B) Steady-state performance for Experiment 2. (C) Example data (markers) from Experiment 2 and their exponential fits (solid, thin curves). Triangles with a solid line show performance on eight streams with primer, and circles with a dashed line show eight streams without a primer. (D) Time constant in terms of absolute time (seconds) for each condition (cross-subject mean ± SEM). (E) Time constants in terms of repetitions plotted as in (D).

control condition was near ceiling level and on par with the easiest condition (four streams, with primer). There were an average of 0.31 false alarms per trial.

2.3 Discussion

Results from Experiment 1 suggest that listeners can selectively attend to the target stream quite easily; even with 12 streams present, performance was better than 60%. Providing listeners with an auditory primer significantly improved their performance at the beginning of the trial but not at the end. Even though the control condition had similar levels of energetic masking as in the 12-stream condition, listeners achieved the same hit rate as in the 4-stream condition. This suggests that the difficulty associated with listening in an overcrowded environment is not due to energetic masking but is primarily due to informational masking, or errors of selection. Listeners also reached steady-state performance very quickly (after 1 or 2 repetitions). However, it is unclear whether this buildup is a function of time or repetition number because the repetition period was fixed across all conditions. Furthermore, we wanted to disentangle task difficulty due to the number of streams present in the scene from the token rate. Thus in the following experiment we varied the repetition period by fixing the token rate instead.

3. Experiment 2

3.1 Methods

Nine subjects (2 male, aged 19 to 30 yrs) who took part in Experiment 1 returned to participate in this experiment. All stimulus generation and delivery methods were identical as described in Experiment 1 except that (1) the overall token rate was set at 12 tokens/s for all conditions and (2) each trial was 15 s in duration. Because repetitions were faster, letter tokens were truncated to 250 ms; letters that became unintelligible were excluded (mean duration before truncation, 401 ± 8 ms).

In Experiment 1 we reported the first-repetition performance, but Experiment 2 allowed us to analyze the time course of the responses. To do so, a rising exponential was fit to each subject's performance for each condition. Using the steady-state value as the asymptote A, a fit P of the form

$$P = A(1 - \mathrm{e}^{-t/\tau}), \tag{1}$$

with time constant τ as the free parameter was estimated by computing the log likelihood curve and choosing the τ that resulted in the maximal value. Accordingly, the τ parameter was larger for longer buildup times. An ANOVA was performed on the values of τ with the same factors as in Experiment 1. Example fits can be seen compared to corresponding raw data in Fig. 2.

3.2 Results

Figure 2(B) shows the across-subject mean (±SEM) of the steady-state hit performance. Without a primer the steady-state performances were 86.3 ± 4.3%, 79.1 ± 4.4%, 73.1 ± 5.2%, and 64.3 ± 5.2% for 4, 6, 8, and 12 streams, respectively (mean ± SEM). With a primer they were 88.5 ± 3.7%, 87.4 ± 4.5%, 85.8 ± 4.0%, and 79.7 ± 5.8%. As in Experiment 1, there was a significant effect of the number of streams present in the auditory scene [$F(1.44,11.5) = 9.46$, $p = 0.006$] but here the presence of the auditory primer [$F(1,18) = 9.561$, $p = 0.015$] also had an effect on steady-state performance. Figures 2(D) and 2(E) show the across-subject mean (±SEM) of τ. The time-constant τ can be expressed in terms of time in seconds [Fig. 2(D)], or in terms of token repetitions [Fig. 2(E)]. When expressed in terms of absolute time, there was a significant effect on the number of streams [$F(3,54) = 3.57$, $p = 0.029$] and of the primer [$F(1,18) = 13.7$, $p = 0.006$]. However, when expressed in terms of repetition number, only the effect of the primer was significant [$F(1,18) = 18.8$, $p = 0.003$]; the time constant no longer depended on the number of streams [$F(3,54) = 1.21$, $p = 0.328$]. No interaction terms were significant in any of the tests. There were an average of 0.48 false alarms per trial.

To compare across experiments [Figs. 2(A) and 2(B)], we ran a repeated measures ANOVA with the same factors as before but with an added experiment factor. Steady-state performance depended on the primer [$F(1,8) = 17.5$, $p = 0.003$] and number of streams [$F(3,24) = 30.1$, $p < 0.001$] but not on the experiment. However, there was a significant interaction term of experiment and number of streams [$F(3,24) = 7.43$, $p = 0.001$], which reflects generally lower, compressed performance in the second experiment.

3.3 Discussion

Results from Experiment 2 suggest that buildup appears to be a function of repetition number, and not time. As in Experiment 1, cueing the subjects with the sound configuration in the primer conditions provided great improvement in their performance across all streaming conditions. Performance at the beginning of a trial was dictated by the number of repetitions of the letters the listener had heard, as opposed to the time in seconds, since the trial began.

4. General discussion

The results from these two experiments agree with our intuition—listeners generally found it harder to selectively attend to one designated stream when there were more streams present in the auditory scene. However, performance in the most crowded scene (12 streams in total) was still surprisingly good. Comparing the results of these two experiments, we found that the faster the tokens were presented, the harder it was for listeners to perform the task [e.g., compare the 4-stream results in Experiment 1 with a repetition rate of 3 repetitions/s and in Experiment 2 with a repetition rate of 1 repetition/s in Figs. 2(A) and 2(B)]. This indicates that the token rate also influences the cognitive load of the listeners.

Past auditory streaming studies suggest that buildup of streaming typically takes several seconds[12] but these experiments used ambiguous stimuli that would lead to bi- or multi-stable percepts.[5] In our experiments, all repeated letter tokens can be unambiguously grouped into objects since they all have distinct onset/offset times, pitches, and locations. Furthermore, there was not a substantial buildup in performance when auditory primers were used; i.e., the buildup was nearly instantaneous.

Therefore, we argue that the buildup of performance in the no-primer conditions in our experiments may be related to object selection rather than to stream formation.

Previous studies focused on selective attention with two or more streams have depended on successfully reporting a message masked by other speech.[13,14] Here, the content of the stream is predictable (with oddball detection used as a proxy for successful selective attention), making that an additional useful feature. This difference stems from the study's BCI-centric design.

Results from these experiments also provide us with insight into the tradeoffs associated with the display design used in future auditory-based BCI. One potential design for an auditory BCI involves presenting a number of auditory "options" (streams) to a listener and decoding which stream he or she is attending (e.g., using two competing streams[10]). If such a scheme uses repeating tokens, a faster repetition rate can increase the information transfer rate because the buildup in performance is associated with the repetition number, and not time. However, a faster repetition rate can also make the task harder, and the reduction in users' accuracy in performing the task will also lower the BCI effective information transfer rate. Providing the user with the sound configuration significantly improves his or her performance but this is at a one-time cost of playing the auditory primer to the users. These observations may provide a reasonable starting point for optimizing the effective information transfer rate in future auditory-based BCI designs.

Acknowledgments

This work was supported by Grant Nos. R00 DC010196 for A.K.C.L. and T32 DC005361 for R.K.M. from NIH/NIDCD. We thank Dr. Eric Larson and Dr. Jonathan Brumberg for comments on an earlier version of this manuscript, as well as Sara Chai and Mihwa Kim for data collection.

References and links

[1] B. C. J. Moore and H. E. Gockel, "Properties of auditory stream formation," Philos. Trans. R. Soc. London, Ser. B **367**(1591), 919–931 (2012).

[2] A. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA, 1990), 773 pp.

[3] S. Anstis and S. Saida, "Adaptation to auditory streaming of frequency-modulated tones," J. Exp. Psychol. Hum. Percept. Perform. **11**(3), 257–271 (1985).

[4] L. Van Noorden, "Temporal coherence in the perception of tone sequences," Doctoral dissertation, Eindhoven University of Technology, Eindhoven, the Netherlands, 1975.

[5] S. L. Denham, A. Bendixen, R. Mill, D. Tóth, T. Wennekers, M. Coath, T. Bőhm, O. Szalardy, and I. Winkler, "Characterising switching behaviour in perceptual multi-stability," J. Neurosci. Methods **210**, 79–92 (2012).

[6] S. A. Shamma, M. Elhilali, and C. Micheyl, "Temporal coherence and attention in auditory scene analysis," Trends Neurosci. **34** (3), 114–123 (2011).

[7] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," Trends Cogn. Sci. **12**(5), 182–186 (2008).

[8] V. Best, E. J. Ozmeral, N. Kopco, and B. G. Shinn-Cunningham, "Object continuity enhances selective auditory attention," Proc. Natl. Acad. Sci. U.S.A. **105**(35), 13174–13178 (2008).

[9] M. Schreuder, T. Rost, and M. Tangermann, "Listen, you are writing! Speeding up online spelling with a dynamic auditory BCI," Front. Neurosci. **5**, 112 (2011).

[10] N. J. Hill and B. Schölkopf, "An online brain–computer interface based on shifting attention to concurrent streams of auditory stimuli," J. Neural Eng. **9**(2), 026011 (2012).

[11] M. A. Lopez-Gordo, E. Fernandez, S. Romero, F. Pelayo, and A. Prieto, "An auditory brain–computer interface evoked by natural speech," J. Neural Eng. **9**(3), 036013 (2012).

[12] C. Micheyl, B. Tian, R. Carlyon, and J. Rauschecker, "Perceptual organization of tone sequences in the auditory cortex of awake macaques," Neuron **48**, 139–148 (2005).

[13] A. Treisman, "Verbal cues, language, and meaning in selective attention," Am. J. Psychol. **77**, 206–219 (1964).

[14] J. C. Webster and L. N. Solomon, "Effects of response complexity upon listening to competing messages," J. Acoust. Soc. Am. **27**, 1199–1203 (1955).