Perceptual limits in a simulated “Cocktail party”

Takayuki Kawashima¹ · Takao Sato²

Published online: 12 May 2015
© The Psychonomic Society, Inc. 2015

Abstract Numerosity judgments of simultaneous talkers were examined. Listeners were required to report the number of talkers heard when this number varied (1 to 13). Spatial location of talkers (1 or 6 locations), duration of talker voices (0.8 s, 5.0 s, and 15.0 s), and gender arrangement of talkers also were manipulated in four experiments. In all experiments, the proportion of correct numerosity judgments monotonically decreased as talker numbers increased. Perceptual limits, defined as talker numbers with proportion correct scores of 0.5, varied between 3 to 5 talkers, on average, depending on listening conditions, and were significantly higher for spatially separated talkers, for the longer voices, and for the mixed gender voices (Experiments 1, 2, and 3). In addition, Experiment 4 found that average numerosity response times increased monotonically over a range of one to four talkers. These results support the idea that, before counting talkers, listeners perceptually segregate talkers to render numerosity judgments. They also suggest that our functional auditory world for simultaneous voices may consist of, at most, three to five talkers depending on listening situations. In light of these results, possible causes for such perceptual limits are discussed.

Keywords Auditory organization · Numerosity perception · Response times · Counting

✉ Takayuki Kawashima
takayuki.kawashima@thu.ac.jp

¹ Faculty of Health and Medical Science, Teikyo-Heisei University, 2-51-4 Higashi-Ikebukuro, Toshima-ku, Tokyo 170-8445, Japan

² Graduate School of Humanities and Sociology, The University of Tokyo, Tokyo, Japan

When speakers cannot be seen by a listener, for example, because the speakers are situated on the other side of a wall, how many talkers can listeners perceive from this mix of voices? Numerosity perception of simultaneous objects has been studied mainly in the field of visual perception. A numerosity judgment of dots presented for a brief duration is quick and accurate up to four or five dots, although correct judgment requires more time if the number of presented dots increases because serial counting of the dots is required (Kaufman, Lord, Reese, & Volkman, 1949; Mandler & Shebo, 1982).

By contrast, a numerosity judgment of multiple simultaneous sounds has rarely been studied. We believe this subject to be of interest, not only because it extends our knowledge of numerical perception but also because listeners' performance can shed light on perceptual limits in so-called “cocktail party” situations (Cherry, 1953, p. 976). In everyday listening, we extract information about various separate sound sources from a compound signal comprising sounds with initially indistinct characteristics. This auditory function often is referred to as auditory scene analysis (Bregman, 1994). Because numerosity judgments require listeners to perceptually segregate each talker from a compound signal, the maximum number of talkers that listeners can judge correctly indexes a limit in the auditory scene organization.

In the present study, we conducted four experiments to investigate numerosity perception of simultaneous voices. In Experiment 1, we examined basic characteristics of numerosity perception for sounds presented at a single location. In Experiment 2, each voice was presented by one of six loudspeakers located in front of the listeners to examine the effect of spatial separation. Experiment 3 replicated previous experiments using a new control condition. In Experiment 4, we measured numerosity judgment response times (RTs) in order to analyze response strategies.

Experiment 1

Method

Participants

Ten university students (8 males and 2 females) participated in Experiment 1. The participants ranged in age from 19 to 21 years. All were native speakers of Japanese. Three of them had previously participated in another experiment on numerosity judgment (in Experiment 4). Their results were not qualitatively different from results of other participants. In questionnaires collected at the end of the experimental sessions, all participants reported that their hearing abilities were in the normal range.

Sample size was determined to be ten before data collection, and the data collection was stopped when the number of participants reached this value. The value was set to 10, because we estimated the size was large enough to reliably observe differences between conditions and because it was one of the multiples of two (for a counter balance of experimental condition orders). The estimation was based on results of both a pilot study and previous experiments that were similar to those in the present study (Kawashima, 2005; Experiment 4 in the present study). A similar procedure was adopted in Experiments 2, 3, and 4.

Apparatus and stimuli

Speech produced by ten males and ten females were taken from a corpus (Kobayashi, Itabashi, Hayami, & Takezawa, 1992). The corpus includes voices of approximately 60 talkers reading sentences selected from various Japanese novels and newspapers under the constraint that frequencies of occurrences of bi-syllabic combinations (and some tri-syllabic combinations) in the sentences were almost the same. The voices of 20 talkers were selected according to recording quality. Average fundamental frequencies of the male and female voices were 152 Hz and 288 Hz, respectively. Root mean square amplitude values of each voice were arranged to be equal during stimulus preparation.

Stimulus presentation was controlled by a personal computer (Apple MacBookPro). Signals were sent to a D/A converter (RME FireFace 400), and the voices were presented to participants through a single loudspeaker (Sound Device SD-0.6). The loudspeaker was positioned in front of the listener at a distance of 1.2 m and at a height of 1.2 m.

Measurements were conducted in a dark soundproofed room to prevent participants having prior knowledge of the number and position of the loudspeaker. The participants were tested individually. The experimental apparatus was concealed from the participants.

Procedure

To begin each trial, the participant pressed a button, which generated a 0.2-s warning beep (4-kHz tone burst). After a 0.4-s silent interval, stimulus voices were simultaneously presented and, after the presentation, the participant was required to report aloud the number of talkers. The participant's response was recorded by an experimenter who sat in a corner of the room. Onsets and offsets of the voices and the beeps were ramped (dampened) with a raised-cosine ramp (0.003 s).

To make it difficult for participants to base judgments on level-related cues, presentation levels of the stimuli were randomly changed between trials in the range of -18 dB to 0 dB at 1 dB steps (Kashino & Hirahara, 1996). In the extreme case involving sound level randomization, a talker's voice was presented at approximately 59 dB SPL. The sound levels were measured by a microphone located at the virtual center of a listener's head. At the start of the experiment, listeners were instructed that sound levels would change independently of the number of talkers and that sound level therefore would not be a reliable cue for the judgment. Participants were not told the exact range of the number of talkers, and they did not receive corrective feedback on each trial.

There were two voice durations: 0.8 s and 5.0 s. Trials in the two durations were separately blocked, and their order was counterbalanced between participants. In each block, the number of talkers was manipulated from 1 to 12 , for a total of 10 numbers: $1, 2, 3, 4, 5, 6, 7, 8, 10$, and 12 . Each trial was repeated 12 times, for a total of 120 trials in a block; trials were presented in random order. Ten practice trials preceded each duration condition. When the duration of a talker's voice (i.e., the length of a sound file) was not long enough for the condition (especially in the 5 -s condition), another sentence by the same talker was randomly selected, and the selected sentence was concatenated to the first voice to complete the duration.

Gender of talkers in each trial was set to one of three conditions: only male, only female, and mixed (both female and male). In the mixed condition, the number of male and female talkers was same (in even talker number conditions) or differed by one (in odd talker number conditions). The three conditions appeared in a randomized order within a block.

This research was performed in accordance with ethical principles outlined in the Declaration of Helsinki. Written, informed consent was collected from each participant.

Data analysis

Within each number of talkers condition, the proportion of correct responses and average numerosity response were calculated for individual participants, in each of the duration conditions. Moreover, to obtain an index of perceptual sensitivity for each participant, we calculated the number of talkers

corresponding to proportion correct scores of 0.5 on a cumulative normal curve. We fitted the curve to the data by searching for the mean and SD of the cumulative normal curve that minimized the sum of squared errors.

Results

The proportion of correct responses and average numerosity response were calculated for individual participants in each condition. Because observed scores of both measures reflected minimal individual differences, values for each measure, averaged over listeners, are plotted in Fig. 1. Despite some differences due to duration, the two duration conditions had at least three results in common. First, listeners tended to report the number of talkers correctly only when this number of talkers was relatively small, i.e., the proportion of correct numerosity judgments declined abruptly around three or four talkers. Second, listeners consistently underestimated the number of talkers in conditions containing more than three or four talkers. Third, the participants' numerical reports gradually increased (with a slope of less than one) as the number of talkers increased.

To obtain an index of perceptual sensitivity for each participant, we calculated the numbers of talkers for participants as corresponding to proportion of 0.5 on a fitted cumulative normal curve as described above. The average values of the indices were 2.9 and 3.6 for 0.8 s and 5.0 s durations, respectively (Table 1). The results indicate that listeners could correctly judge the number of talkers in more than half of the trials only when the actual number of talkers was less than these limits (approximately 3). The proportion correct scores of 0.5 were adopted as criteria somewhat arbitrarily.

Listeners tended to perform better in the 5.0-s duration condition than in the 0.8-s condition. The difference of the average perceptual limits between the two conditions (i.e., 2.9 and 3.6) was statistically significant in a paired t test, $t(9) = 7.42$, $p < 0.001$, two-tailed,

Table 1 Mean perceptual limits and 95 % confidence intervals (CI) in each duration condition

Conditions	M	95 % CI
Experiment 1 ^a		
0.8	2.9	[2.6, 3.1]
5.0	3.6	[3.4, 3.8]
Experiment 2 ^a		
0.8	3.3	[3.1, 3.5]
5.0	4.5	[4.2, 4.9]
15.0	5.3	[4.8, 5.9]
Experiment 3		
L1N6	3.6	[3.0, 4.1]
L1N13	3.4	[3.0, 3.8]
L6N6	4.4	[3.8, 5.0]
Experiment 4 ^{a,b}		
0.8	3.2	[3.0, 3.5]
5.0	4.5	[4.0, 4.9]
15.0	5.0	[4.2, 5.8]

L1N6 = maximum six talkers were presented from one location; L1N13 = maximum thirteen talkers were presented from one location; L6N6 = maximum six talkers were presented from six locations.

^a Condition values are in seconds.

^b Results in voice-only conditions.

Cohen's d standardized with averaged standard deviations (SDs) was 2.25.

To evaluate possible effects of a talker's gender, trials involving single gender and mixed gender voices were analyzed separately to examine possible differences involving this variable. The single gender trials comprised only male trials and only female trials. One possible outcome was that listeners may hear voices in the mixed gender trials more readily than in the single-gender trials; the resulting analysis was consistent with this postulated outcome. The average perceptual limits in the 0.8 s were 2.7, 95 % confidence interval (CI) [2.5, 2.9], and 3.0 [2.7, 3.2] for single and mixed gender trials,

Fig. 1 Proportion of correct responses (left) and average numerosity responses (right) in Experiment 1. Error bars show 95 % CI. Horizontal line in the left panel shows 0.5 correct ratios and the diagonal line in the right panel shows a slope of 1.0 for reference

respectively. In the 5.0-s duration condition, limits were 3.3 [3.0, 3.6] and 3.9 [3.7, 4.2]. A main effect of the talker gender was significant in a two-way ANOVA (2 durations $\times$ 2 gender conditions), $F(1, 9) = 19.11$, $p = 0.001$, partial $\eta^2 = .68$; also the interaction was marginally significant, $F(1, 9) = 4.99$, $p = 0.05$. These results showed that listeners perceived the number of talkers more accurately in mixed gender trials than in same gender trials. There was no statistically significant difference between the limits for the male only and female only trials in a two-way ANOVA (2 durations $\times$ 2 gender conditions), $F(1, 9) = 0.23$, $p = 0.64$.

Discussion

The results of Experiment 1 indicate that the perceptual limits in listeners who experience a simulated cocktail party context are approximately three talkers (Table 1). Several explanations for the observed profile of average numerosity judgments are possible. One explanation holds that average responses represent the number of distinct talkers that listeners extract from compound signals and that the responses themselves reflect perceptual limits. Although this interpretation cannot be ruled out, we believe the average responses may be affected by factors other than the number of segregated sources, especially when the stimulus contains a relative large number of talkers. For example, it is certain that the intelligibility of each voice will decrease as the talker numbers increase (Bronkhorst, 2000). In this situation, listeners might bias their responses, because they know that in daily life it is more difficult to hear sounds when many other sounds are present.

Experiment 2

Experiment 2 extended Experiment 1 in two respects. First, the effect of spatial separation between talkers was examined. Spatial separation should enhance perception of each talker among simultaneous voices (Bronkhorst, 2000; Freyman, Helfer, McCall, & Clifton, 1999; Yost, Dye, & Sheft, 1996), and it also may facilitate enumeration of sources. Alternatively, it is possible that spatial separation may disturb listener judgments. Best, Gallun, Ihlefeld, & Shinn-Cunningham (2006) reported that listener performance in a divided listening task, where listeners reported keywords from two simultaneous talkers, became poorer with a wider separation between the talkers. They concluded that enlarging the spatial separation increases the cost associated with processing simultaneous talkers. Similarly, in the present study, increments in spatial separation could increase the cost of processing of multiple voices and might hinder listeners' numerosity judgments. In this study, we examined effects of location by comparing results of Experiment 2 with those of Experiment 1. The second difference in Experiment 2 involved stimulus

durations: the maximum duration of the voices was lengthened to 15.0 s to search for a possible saturation of an enhancing effect of duration.

Method

Participants

Twelve university students participated in Experiment 2 (3 males and 9 females). The participants ranged in age from 18 to 22 years. All were native speakers of Japanese. None of them had participated in other experiments on numerosity judgment before, and none participated in Experiment 1. Their hearing abilities were reported in the normal range using the same procedure as in Experiment 1.

Procedure

Apparatus and stimuli were the same as in Experiment 1. The procedure was basically the same as that in Experiment 1, with the following exceptions. The maximum number of talkers was six, and six loudspeakers (Sound Device SD-0.6) were used for presentation of voices. Positions of the six loudspeakers were fixed on a semicircle and separated by 36 degrees; the listener sat at the center of the semicircle. Spatial separation between talkers in a trial was manipulated in narrow and wide conditions. This was accomplished by activating loudspeakers in different configurations. In the narrow condition, separations between the activated loudspeakers were kept as narrow as possible. For example, to present four talkers' voices in the narrow condition, only the four loudspeakers directly in front of the listener were used. In contrast, in the wide condition, loudspeakers that were located to each side of the listener (those directly facing the listener's ears), along with speakers at relatively center positions, were used. To clarify, we represent the six loudspeaker positions with a series of six letters. Thus, an example of the narrow conditions with four talkers was configured as: IAAAAI (A: activated speaker, I: inactivated speaker); an example of the wide conditions was configured as: AAIIAA. For the six talker condition, the spatial positions of talkers were the same between the wide and the narrow conditions, because all six loudspeakers were always activated. Within a trial block, the number of times each loudspeaker was used was kept equal for the left and right side of each listener. In the one talker condition, one of the six loudspeakers was randomly selected for the stimulus presentation.

Three duration conditions (0.8, 5.0, and 15.0 s) were used in Experiment 2 as blocked variables. The order of duration blocks was counterbalanced over participants. A block contained 72 trials, because each of the six numbers of talkers (1, 2, 3, 4, 5, and 6) was repeated 12 times in randomized order.

Only in the one talker condition, speech-like-noises (SLNs) were presented simultaneously with a voice to evaluate possible influence of acoustic interference between simultaneous sounds. The SLN was a type of Gaussian white noise whose power spectrum was filtered to be the long-term average spectrum of each talker and was amplitude-modulated with an envelope of a stimulus voice (Bronkhorst & Plomp, 1992; Brungart, Simpson, Ericson, & Scott, 2001). A main advantage of the SLNs was that, although they are not voices, on average their acoustic energy is equal to that of a human voice. The SLNs enabled us to manipulate signal-to-noise ratios of voices without using voices (they also were used in Experiment 4). By presenting five SLNs with a single voice, we evaluated the amount of possible acoustic interference that would occur to the single voice by five other voices. Each of the five SLNs had a different spectral shape based on a talker's long-term speech spectrum. Six talkers used in the one talker condition (i.e., the five talkers supplying the spectra for the SLNs plus a single target talker) were selected randomly in each trial in the same manner as in the six talker condition. Both the voice of a talker and the SLN were presented at approximately 60 dB SPL. Data were analyzed in the same way as in Experiment 1.

Results

As in Experiment 1, average numerosity response and proportion of correct responses are reported, because these two measures were qualitatively similar for each participant (Fig. 2). Results were similar to those obtained in Experiment 1: Average responses were almost equal to actual talker numbers when a relatively small number of talkers were presented. Although numerosity responses gradually increased as the number of talkers increased, they underestimated the actual numbers. The effect of voice duration was similar to that in Experiment 1. Mean perceptual limits were 3.3, 4.5, and 5.3 for 0.8 s, 5.0 s, and 15.0 s durations, respectively. This effect of voice duration was significant in a one-way ANOVA, $F(2,$

22) = 63.3, $p < 0.001$, $\eta^2 = 0.85$. All pair-wise differences between the means were significant in post-hoc comparisons. This indicates that the effect of voice duration was not completely saturated between 5.0 and 15.0 s. In the present report, results of post-hoc comparisons are based on Tukey's HSDs with 0.05 significance level.

One purpose of Experiment 2 was to examine the effect of spatial separation between talkers by comparing these results with those of Experiment 1. Results of comparable conditions (i.e., those with same duration and same number of talkers) in the two experiments are plotted in Fig. 3 (replotted from Figs. 1 and 2). Mean perceptual limits were 2.9 and 3.3 for 0.8 s and were 3.6 and 4.5 for 5.0 s. Although the differences were relatively small, a two-way mixed design ANOVA (2 experiments $\times$ 2 durations) on perceptual limits showed a significant effect of interaction, $F(1, 20) = 6.68$, $p = 0.02$, partial $\eta^2 = .05$. Simple main effects of experiment in each duration also were significant, $F(1, 20) = 9.22$, $p = 0.006$ for 0.8 s and $F(1, 20) = 19.95$, $p < 0.001$ for 5.0 s, respectively. These results support the hypothesis that spatial separation between talkers enhances numerosity judgment.

To explore the effect of spatial separation, differences in the two degrees of spatial separation between talkers, narrow and wide, were analyzed for the data in Experiment 2 (only for 2- to 6-talker conditions). Perceptual limits of each participant were calculated in both conditions then analyzed in a two-way ANOVA (2 spatial conditions $\times$ 3 durations). A main effect of spatial condition was significant, $F(1, 11) = 6.81$, $p = 0.02$, partial $\eta^2 = 0.38$, and an interaction with duration was not significant, $F(2, 22) = 0.04$, $p > 0.1$. These analyses suggest that the spatial separation contributes to a facilitation of numerosity judgments.

In the one talker condition, five SLNs were presented with a single voice: averages of proportion correct scores over listeners in these conditions were: 0.96 for 0.8 s, 0.96 for 5.0 s, and 0.98 for 15.0 s, respectively (upper leftmost symbols in Fig. 2). These are rather high accuracy measures, suggesting that acoustic interference on a single voice by other five voices

Fig. 2 Proportion of correct responses and average numerosity response in Experiment 2. Error bars show 95 % CI

Fig. 3 Proportion of correct responses and average response in Experiments 1 and 2. Dotted lines were plotted referenced to the left axis, and solid lines were plotted referenced to the right axis. Left and right panels show the results for 0.8 s and 5.0 s, respectively. Error bars show 95 % CI

is not so intense as to render the single voice totally unavailable in the numerosity judgment.

Single- and mixed-gender trials were analyzed separately to examine the effect of talker gender as in Experiment 1. Again, data showed that listeners perceived the number of talkers more accurately in the mixed-gender trials than in the same gender trials. The mean perceptual limits were 3.2 and 3.5 in the single and the mixed-gender trials, respectively, in the 0.8-s condition. Corresponding limits in the 5.0-s condition had values of 4.4 and 4.5; in the 15.0-s condition, they had values of 5.1 and 6.0. A main effect of the talker gender was significant in a two-way ANOVA (3 durations $\times$ 2 gender conditions), $F(1, 11) = 5.78$, $p = 0.03$, partial $\eta^2 = 0.34$; the interaction of these variables was not statistically significant, $F(2, 22) = 1.43$, $p = 0.26$. Nor was there a statistically significant difference on perceptual limits between male only and female only conditions in a two-way ANOVA (3 durations $\times$ 2 gender conditions), $F(1, 11) = 1.37$, $p = 0.27$.

Discussion

Perceptual limits increased beyond 5.0 s to 15.0 s. The improvement suggests that listeners utilized a cognitive function that integrates information over relatively long time intervals.

Results of Experiments 1 and 2 indicate that the spatial separation between talkers enhanced listeners' numerosity perception. However, it also is possible that this resulted from a procedural difference between the experiments, specifically the difference in the maximum number of talkers tested. Listeners may have assigned the same number, i.e., four, for the maximum number of talkers presented in each experiment, which may have led to smaller numerosity responses in Experiment 1. This possible scaling strategy is consistent with the data plotted in Figs. 1 and 2, in that the average responses were around four in the maximum number of talkers in both experiments.

Experiment 3 had two purposes: First, the effect of spatial separation was examined again within a single experiment. Second, the effect of the maximum number of talkers in a block was examined.

Experiment 3

Method

Participants

Twelve university students participated in Experiment 3 (1 male and 11 females). The participants ranged in age from 19 to 20 years. All were native speakers of Japanese. Four of them had participated in other experiments on numerosity judgment before (one in Experiment 1 and three in Experiment 2). Their results were not qualitatively different from results of other participants. Their hearing abilities were reported to be in the normal range by the same procedure as in the previous experiments.

Procedure

The procedure was almost the same as in Experiment 2, with the following exceptions. All stimuli were presented through headphones (Sennheiser HDA200) instead of loudspeakers, and locations of stimuli were manipulated using Head Related Transfer Functions (HRTFs). All HRTFs used in the current study were measured to sound sources at 0 degrees elevation. Two sets of HRTFs were used, one measured at Tohoku University (2001) with a human subject and another measured at Nagaoka University of Technology (2002) with an artificial head and torso. Four of the 12 participants were tested with the former set of HRTFs, and 8 of them were tested

with the latter set. The differences between the two HRTF sets did not affect the main conclusions of the experiment.

Sound level was established by setting root mean square amplitude values of single voices to be constant, as in the previous experiments, and the values were determined to produce sounds of 59 dB SPL after the voices were processed by HRTFs corresponding to 0 degrees azimuth and elevation (in a maximum case of sound level randomization).

The duration of voices was 5.0 s throughout the experiment. There were three experimental conditions. The first was a control condition, with all voices presented at a location in front of the listener, and the number of talkers varying from 1 to 6: 1, 2, 3, 4, 5, and 6 (condition L1N6). In the second condition (L1N13), all voices were presented at one location as in L1N6, and the number of talkers varied from 1 to 13, excluding 10 and 12. In the third condition (L6N6), voices were presented at six locations, -90 , -50 , -15 , $+15$, $+50$, and $+90$ degrees azimuth (where negative signs indicate locations to the listener's left side), and the number of talkers was manipulated as in the first condition. Unequal spatial separations between talkers were introduced because the two sets of HRTFs did not contain functions allowing equal separations. The three conditions were blocked, and the order of the blocks was counterbalanced under the constraint that L1N6 and L6N6 were always measured in succession (the order of the two conditions was also counterbalanced). Details of spatial separation conditions in L6N6 (i.e., narrow and wide) were determined according to the same rule used in Experiment 2. Each number of talkers was repeated 12 times; therefore, a block contained 72 trials in L1N6 and L6N6, and 132 trials in L1N13. Data were analyzed in the same way as in Experiment 1.

Results

Average numerosity responses and proportions of correct responses in the three experimental conditions are plotted in Fig. 4. One goal of Experiment 3 was to examine the role of the maximum number of talkers on numerosity judgments. If participants scaled their answers depending on the maximum number of talkers, the responses to voices actually containing the same number of talkers would differ for L1N6 and L1N13 (i.e., larger in the former condition). Perceptual limits were estimated as in the previous experiments, and they were 3.5, 3.4, and 4.4 in L1N6, L1N13, and L6N6 conditions, respectively. A one-way ANOVA on the limits revealed a significant main effect of experimental conditions, $F(2, 22) = 16.3$, $p < .001$, $\eta^2 = 0.59$; however, the difference between L1N6 and L1N13 was not significant in multiple comparisons. In addition, a two-way ANOVA was conducted on average numerosity responses (3 experimental conditions $\times 6$ numbers), and a significant interaction was observed, $F(10, 110) = 10.3$, $p < 0.001$. The interaction seemed to result from differences

between L6N6 and the two other conditions and, contrary to the prediction described above, no significant differences between L1N6 and L1N13 for any number of talkers emerged in follow-up multiple comparisons. Furthermore, there were no significant order effects (i.e., whether a listener participated in L1N6 and L6N6 first or L1N13 first did not affect the average responses and ratios of correct responses). Results of Experiment 3 were not consistent with the hypothesis that the maximum number of talkers in a block affects listener numerosity judgments.

The effect of spatial separation was examined in Experiment 3. As previously described, the average perceptual limits were 3.5 and 4.4 in L1N6 and L6N6 conditions, respectively. Multiple comparisons following the one-way ANOVA described above revealed that the difference was significant. In addition, the effect of extent of spatial separation between talkers (i.e., narrow vs. wide) was evaluated in L6N6. Perceptual limits were 4.3 and 4.5 on average for narrow and wide conditions, respectively. Although the difference was relatively small, a paired t test revealed that the difference was significant, $t(11) = 2.6$, $p = 0.03$, two-tailed, Cohen's d standardized with averaged SDs was 0.28. The small but statistically significant difference between narrow and wide separation was consistent with the hypothesis that spatial separation of talkers can enhance numerosity judgments.

Perceptual limits in single- and mixed-gender trials were analyzed. The mean perceptual limits were 3.4 and 4.2 for the single and the mixed gender trials, respectively, in L1N6. They were 3.2 and 3.8 in L1N13 and 4.1 and 4.9 in L6N6. A main effect of the talker gender was significant in a two-way ANOVA (3 experimental conditions $\times 2$ gender conditions), $F(1, 11) = 37.94$, $p < 0.001$, partial $\eta^2 = 0.78$, and the interaction of these two variables was not statistically significant, $F(2, 22) = 0.71$, $p = 0.50$. There was no statistically significant difference on perceptual limits between male-only and female-only conditions in a two-way ANOVA (3 experimental conditions $\times 2$ gender conditions), $F(1, 11) = 1.3$, $p = 0.28$.

Discussion

Results in Experiment 3 indicated that the maximum number of talkers in a block does not affect numerosity judgments. This finding is consistent with our conjecture that spatial separation caused the differences between Experiments 1 and 2. Spatial separation should enhance perception of each talker among simultaneous voices (Bronkhorst, 2000; Freyman et al., 1999; Yost et al., 1996). This enhancing effect suggests that perceptual separation of simultaneous voices enhances numerosity judgments (Yost et al., 1996).

Fig. 4 Proportion of correct responses and average responses in Experiment 3. L1N6 = maximum six talkers were presented from one location; L1N13 = maximum thirteen talkers were presented from one location; L6N6 = maximum six talkers were presented from six locations. Error bars show 95 % CI

Experiment 4

In Experiment 4, response times (RTs) in the numerosity judgment were measured to examine listeners' response strategies. RT has been commonly used in the study of numerosity perception in vision and has provided evidence for subitizing and counting in making judgments (Kaufman et al., 1949; Mandler & Shebo, 1982). On the other hand, RTs have rarely been used in the study of auditory numerosity judgments, with the exception, perhaps, of Repp (2007), who measured RTs in enumeration of rapidly presented tone sequences. In the numerosity judgment of simultaneous sounds, there are at least two logical possibilities for the listeners' strategies: (a) numerosity judgments rely on the counting of perceptually segregated sources; and (b) numerosity judgments do not rely on counting perceptually segregated sources, but instead are based on impressions of the sound properties unrelated to segregated sources, such as timbre. We think that in the former case, RTs for the numerosity judgment should increase as the number of talkers increases, although this cannot be predicted readily from the latter case. Therefore, we hypothesized that RT latency would increase as a function of number of talkers.

In addition, we examined the impact of signal-to-noise ratios of multiple voices on RT latency in the numerosity judgments. As the number of talkers increases, the signal-to-noise ratio of each voice decreases. In principle, lower signal-to-noise ratio should render it more difficult for listeners to differentiate voices. This effect of signal-to-noise ratios on RT latency was examined with conditions where voices were simultaneously presented with SLNs.

Method

Participants

Fourteen university students participated in Experiment 4 (3 males and 11 females). They ranged from 19 to 21 years. All

were native speakers of Japanese. None had participated in other experiments on numerosity judgment before. Their hearing abilities were reported in the normal range using the same procedure as in Experiment 1.

Sample size was determined to be 12 before data collection as in the previous experiments. The data collection was stopped once when the number of participants reached 12. One of the first 12 participants, however, showed an exceptional performance: Her correct ratios were less than 0.5 in one- and two-talker conditions. Results of this participant were excluded, and data of two extra participants were successively collected to accomplish a counterbalance of experimental condition orders. Because the data of the two participants seem to be unexceptional, the data of one participant who participated former were included in data analysis and those of the latter participant were not included. Consequently, results of the 12 participants were reported.

Procedure

The procedure was basically the same as that in Experiment 2, with the following exceptions. In each trial, in order to respond, participants had to press a button on a keyboard that was placed across their knees when they judged talker number. The button press was required first in order to keep listeners' bodily movements the same regardless of their answers. They were explicitly told to respond as soon as they had an answer. RT was defined as a time interval between an onset of the stimulus voices and the button press. When the button was pressed during presentation of voices, the voices were turned off.

As in Experiment 2, Experiment 4 contained six talker number conditions (1 to 6 talkers). In addition, four conditions were included in a block. In all of four conditions, voices were presented simultaneously with SLNs to manipulate signal-to-noise ratios of the voices. For the first of four conditions, one voice was presented with two SLNs. For the second condition,

two voices were presented with one SLN, respectively. In these two conditions as well as in the three-talker (voice only) conditions, the signal-to-noise ratios of each voice are approximately same, because the number of sound sources (i.e., voices and SLNs) is always constant (i.e., three). Therefore, we were able to evaluate listener performances as a function of the number of talkers under a constant signal-to-noise ratio. In third and fourth conditions, one SLN was presented with one voice and one SLN was presented with three voices, respectively. We included these conditions to evaluate the effect of a decrease of signal-to-noise ratio on RTs.

These four conditions were mixed with six voice only conditions in one block, resulting in a total of 10 sound source conditions. These 10 conditions were repeated 9 to 16 times, creating a block of 120 trials for each duration condition. Twenty practice trials preceded each duration condition.

Data analysis

In Experiment 4, RTs were measured as one of the dependent variables, in addition to perceptual limits and average numerosity response. One trial in 5.0-s condition of a participant was not included in analysis, because the participant had started the next trial before she answered.

Results

Perceptual limits

Results in voice-only conditions are plotted in Fig. 5. Mean perceptual limits in the voice-only conditions were 3.5, 4.5, and 5.0 for 0.8 s, 5.0 s, and 15.0 s, respectively (Table 1). A main effect in a one-way ANOVA was significant, $F(2, 22) = 26.5$, $p < 0.001$, $\eta^2 = 0.70$, and all the differences were significant in multiple comparisons. The results showed that listener performances were similar between Experiments 2 and 4 despite some differences in experimental settings (Fig. 2).

The effect of extent of spatial separation between talkers (i.e., narrow vs. wide) was evaluated. Perceptual limits were

3.2 and 3.3 on average for narrow and wide conditions in the 0.8 s condition, respectively. They were 4.5 and 4.5 in the 5.0-s condition and 4.9 and 5.2 in the 15.0-s condition. Although the differences between the average values were relatively small, a two-way ANOVA (3 durations $\times$ 2 spatial conditions) on the average limits revealed a significant main effect of spatial separation, $F(1, 11) = 9.81$, $p = 0.01$, partial $\eta^2 = 0.47$, and an interaction with the durations was not significant, $F(2, 22) = 1.29$, $p = 0.30$. These analyses indicate that the spatial separation facilitated listener numerosity judgments.

Perceptual limits in single and mixed gender trials were analyzed. In this analysis, the data of one participant were excluded because proportions of correct responses of this individual were unusually high even within the five- and the six-talker conditions in the 15.0-s mixed-gender trials (0.75 and 1.0, respectively); in fact, the estimated perceptual limit for this participant departed from the group mean by more than 3 SDs. The average perceptual limits were 3.0 and 3.3 in the single- and the mixed-gender trials, respectively, in the 0.8-s condition. These limits were 4.2 and 4.6 in the 5.0-s condition and 4.6 and 4.8 in the 15.0-s condition. Statistical analyses of these data revealed a significant main effect of talker gender in a two-way ANOVA (3 durations $\times$ 2 gender conditions), $F(1, 10) = 6.37$, $p = 0.03$, partial $\eta^2 = 0.39$; however, the interaction of these variables was not statistically significant, $F(2, 20) = 0.14$, $p = 0.87$. There was no statistically significant difference of perceptual limits between male-only and female-only conditions in a two-way ANOVA (3 durations $\times$ 2 gender conditions), $F(1, 10) = 2.99$, $p = 0.11$.

Reaction times

Voice only conditions To evaluate outlying values, RT latencies greater than and less than 3 SDs from the mean RTs were counted in each duration condition. However, we did not exclude these trials in the analyses, because the relevant percentages were low. That is, percentages of these trials ranged from 0.3 % to 1.9 %, depending on individual participants (the average was 1.0 % with SD of 0.49). Median RT values were

Fig. 5 Proportion of correct responses and average numerosity response in Experiment 4 (voice-only conditions). Error bars show 95 % CI

calculated for each participant and then averaged over participants in each condition. There were no trials where participants responded before the start of stimulus voices. Average RTs in voice only conditions were plotted in Fig. 6.

In all duration conditions, RTs increased monotonically as the number of talkers increased from one to four talkers: subsequently, slopes showing increases asymptote around four talkers. Two-way ANOVA (3 durations $\times$ 6 numbers of talkers) revealed a significant interaction, $F(10, 110) = 26.6$, $p < 0.001$, partial $\eta^2 = 0.7$, and simple main effects of talker numbers also were significant, $F(5, 55) = 19.8$ for 0.8 s, $F(5, 55) = 70.0$ for 5.0 s, and $F(5, 55) = 62.5$ for 15.0 s, respectively, all $ps < 0.001$. Multiple comparisons showed that average RTs differed significantly between two and four talkers in all duration conditions, although the corresponding differences between four and six talkers (for 0.8 s and 5.0 s) and between five and six talkers (for 15.0 s) were not significant. These results indicated that listeners tended to take more time to respond as the number of talkers increased, at least when less than five talkers were simultaneously presented.

Voices with speech-like-noises conditions The conditions involving the voice plus speech-like-noises (SLNs) were included in Experiment 4 with the goal to evaluate the effects of signal-to-noise ratios of individual talkers on RTs. As the number of talkers increased, acoustic signal-to-noise ratios of individual talkers decreased; consequently, the increase of RTs observed in the voice-only conditions may be explained by the decrease in the signal-to-noise ratios. Average RTs in the trials where voices were presented with SLNs were calculated in the same way as in voice-only trials (Fig. 7, left panel). In Fig. 7, the three rows below the abscissa

show the number of talkers (i.e., voices), the number of SLNs, and the summation of both numbers in each column, respectively. An increment in this sum indicates a decrease of the signal-to-noise ratios of voices.

The SLNs were used to create approximately the same signal-to-noise ratios in the one-, two-, and three-talker conditions. The RTs in these conditions were plotted as the line-connected three symbols at the right side of the left panel in Fig. 7 (for # talkers 1, 2, and 3). In all three duration conditions, RTs increased with increases in the number of talkers. In a two-way ANOVA on RTs (3 durations $\times$ 3 numbers), an interaction of these variables was significant, $F(4, 44) = 14.3$, $p < 0.001$. Simple main effects of the talker numbers were significant for all three duration conditions, $F(2, 22) = 9.91$ for 0.8 s, $F(2, 22) = 59.2$ for 5.0 s, and $F(2, 22) = 50.6$ for 15.0 s, respectively, all $ps < 0.001$. The results indicate that RTs increase as the number of talkers increases, although the signal-to-noise ratios of the voices were expected to be fairly constant. The right panel in Fig. 7 shows the proportion of correct responses and the average numerosity response of voices in the SLNs conditions. The average numerosity response indicates that the increase of RTs was roughly accompanied with increases in listeners' numerical reports.

The SLNs were also used to evaluate the effects of decreases of signal-to-noise ratios on RTs. In the left panel of Fig. 7, three line-connected pairs of symbols in each duration condition (left side in the panel) show average RTs in one, two, and three talker conditions, respectively. In these pairs, the left symbol of each pair shows RTs when only voices were presented (replotted from Fig. 6), whereas the right symbol of plotted pairs shows RTs when the voices were presented with one SLN. In other words, the signal-to-noise ratios in the right symbol conditions were always lower compared with the paired conditions, although the numbers of talkers or voices were held constant. A three-way repeated measures ANOVA on RTs (3 durations $\times$ 2 SLN presentation levels $\times$ 3 numbers of talkers) revealed a significant interaction between number of voices and SLN presentation, $F(2, 22) = 4.01$, $p = 0.03$. However, the three-way interaction was not significant, $F(4, 44) = 0.24$, $p = 0.91$. Simple main effects of SLN presentation were significant in one- and two-talker conditions, $F(1, 11) = 24.7$, and $F(1, 11) = 18.9$, respectively, both $ps < 0.002$, although the SLN presentation effect was not significant in a three-talker condition, $F(1, 11) = 1.34$, $p = 0.27$. These results indicate that RTs increased when a SLN was presented with voices in conditions with one and two talkers. In the right panel of Fig. 7, the average numerosity response indicates that the RT increases caused by the addition of a SLN to voices were not accompanied by increases in listeners' numerical reports (line-connected symbols in the corresponding one- and two-talker conditions).

Fig. 6 Average response times in voice only conditions in Experiment 4. Error bars show 95 % CI

Fig. 7 Results in the noise presented conditions in Experiment 4. Left panel shows RTs and right panel shows correct ratios and average responses. In the right panel, dotted lines were plotted referenced to the

left axis, and solid lines were plotted referenced to the right axis. Error bars show 95 % CI. The horizontal line in the right panel indicates the referent of 0.5 for proportion correct

Discussion

RTs increased in accordance with an increase in the number of talkers as expected. This finding is consistent with the interpretation that listeners give numeric answers by counting perceptually segregated sources and suggests that listeners do not respond using sound impressions other than segregated information, such as timbre, when less than approximately four talkers are presented.

The results obtained using SLNs indicate that an increase of RTs in tandem with the number of talkers does not require a decrease of signal-to-noise ratios of voices (Fig. 7). This finding is consistent with the above conjecture that the observed differences of RTs in the voice-only conditions are caused by counting. That is, this finding suggests that factors other than a decrease of signal-to-noise ratios of voices are required to explain the increase of RTs in response with increments in number of talkers. We think a strong candidate for these factors is the process of counting of voices.

Furthermore, it is clear from Fig. 7 that listeners' RTs to the same numbers of voices increase as a result of the simple addition of an SLN to the voices. Moreover, the slopes of line-connected pairs in the one and two talker conditions are clearly different: They are steeper in the two-talker condition (left side of Fig. 7, left panel). This difference suggests that in the voice only conditions, adding one talker to two talkers creates greater difficulties in distinguishing among talkers than adding one talker to one talker. In fact, this tendency was observed in the 5.0-s and 15.0-s voice-only conditions (Fig. 6).

General discussion

In four experiments the results reveal that it is difficult for us to hear and distinguish more than three to five voices reliably from simultaneous voices. Furthermore, the performance

changes depending on listening conditions. Performance improved for spatially separated talkers, for mixed gender voices, and for longer voices. The effects of spatial separation of sound sources indicate that a numerosity judgment depends on auditory organization. In addition, RTs for these judgments slowed as the number of talkers increased from one to four talkers, which is consistent with an interpretation that listeners' reported numbers are based on a process of counting perceptually segregated individual sources. These findings indicate that our limits in auditory organization of multiple voices are approximately three to five talkers.

In everyday listening, we extract information about each sound source from a compound signal composed of sources whose characteristics are not initially distinct. Computationally, this poses the auditory system an ill-defined problem of restoring auditory scenes (Bregman, 1994; Yost, 2008). The data in the present study provide new references for computational models of auditory scene analysis (Stern, 2005). For example, models that incorporate functions comparable to those used by humans should be able to restore almost all talkers from the four talker voices of 5.0-s duration (Fig. 2).

At least two unresolved issues about auditory numerosity judgments arise in the present study. First, it is not clear what is segregated and counted. Second, it is not clear what makes numerosity judgments difficult when there are more than three to five talkers. There are several candidates for what is segregated: These may involve acoustic properties of voices, such as pitch, timbre, and spatial locations. It is widely believed that the perceptual segregation of a sound source is accomplished by a grouping of appropriate frequency components in a compound signal (Bregman, 1994; Darwin & Carlyon, 1995; Yost, 2008). These properties may be segregated from a compound signal through auditory grouping process. Although we were not able to narrow the number of segregated candidates, it is possible that spatial locations might have played a role, as

indicated by Santala and Plukki (2011) who reported that it is possible for listeners to identify the locations of up to three or five different noises that are presented simultaneously from different locations. This suggests that listeners in the present study might have been able to segregate the locations of several talkers. However, sound source location is known to be a relatively weak cue in auditory grouping relative to other features, such as a common fundamental frequency (Darwin & Carlyon, 1995; Turgeon, Bregman, & Roberts, 2005). In the present experiment, perceptual limits tended to be higher in mixed-gender trials than in same-gender trials. It is known that different gender voices have different average fundamental frequencies. Therefore, this result might suggest that the pitch of voices was segregated from a compound signal during numerosity judgments. It seems plausible that segregated characteristics, such as the locations of talkers, and the pitch of voices are related to enumeration.

Next, consider what makes numerosity judgments difficult when there are more than three to five talkers. We entertain three possible answers to explain observed perceptual limits. They are not mutually exclusive. Respectively, they involve the effects of energetic and informational masking, the capacity of attention/short-term memory, and a perceptual indexing mechanism (Pylyshyn & Storm, 1988; Pylyshyn, 2001).

First, as the number of talkers increased, the signal-to-noise ratios of each voice decreased. Therefore, energetic masking between voices (i.e., interference by frequency overlap) should make segregation of voices difficult, because with more voices each signal becomes less salient. In this way, a decline in performance might be explained by energetic masking. However, the role of energetic masking in numerosity judgments seems to be limited. On average, the listeners' responses tended to *increase* gradually with decreased signal-to-noise ratios in all four experiments; this is contrary to what is expected if energetic masking strongly influences the perception. Also, informational masking should make detection of a target stimulus difficult when similar stimuli (distracters) are presented simultaneously with the target, even when these distracters levy minor energetic masking effects on the target (Watson, Kelly, & Wroton, 1976). Informational masking might impact the accuracy of numerosity judgments by disturbing the detection of voices (Brungart et al., 2001).

Second, limits on perception of voices also may be explained by capacity of attention (Kahneman, 1973) and/or short-term memory. Both segregation and counting depend on attention. It has been reported that the formation and maintenance of an auditory stream are influenced by attention (Carlyon, Cusack, Foxton, & Robertson, 2001). In addition, internalized counting itself may require allocations of attention to each target (Piazza, Mechelli, Price, & Butterworth, 2006; Vuokko, Niemivirta, & Helenius, 2013). Thus, limits of attention may result in reduced accuracy of numerosity

judgments of simultaneous voices. In addition, memory capacity (Cowan et al., 2005) might affect numerosity judgments, because segregated voices must be stored during serial counting. However, memory storage seems not to be a major factor causing the observed limits at least in the 0.8-s and 5.0-s conditions. If storage capacity had caused the perceptual limits, increasing the duration of voices would not have enhanced the perceptual limits.

The third explanation for perceptual limits of numerosity judgments involves perceptual indexing, or namely a bottom-up process that tags certain salient features of a scene for later processing (Pylyshyn & Storm, 1988). Listeners must be able to maintain a distinction between segregated voices (extracted from a compound signal), because if they cannot, they might count a talker more than once. In this respect, the numerosity judgment task resembles a multiple visual object tracking task, in which observers visually track a subset of objects (e.g., crosses in a display), all of which move around independently on a screen for a certain duration (Pylyshyn & Storm, 1988). Multiple acoustic streams or moving visual objects must be individuated in both tasks. In the latter, it has been reported that observers can track up to four or five moving objects. It is noteworthy that these limits roughly correspond to the perceptual limits found in the present study for numerosity judgments of simultaneous voices. The performance in visual tracking often is considered to reflect a property of a visual indexing mechanism that individuates four or five individual objects in a visual scene (Pylyshyn & Storm, 1988; Pylyshyn, 2001). An auditory version of the indexing mechanism might be worth considering as a possible bottom-up mechanism that explains the perceptual limits in numerosity judgments of voices.

Efficiency in numerosity processing for a relatively large number of objects in vision often has been related to competence for mathematics (Halberda, Mazzocco, & Feigenson, 2008). In the present study, the average numerosity response gradually increased along with the talker numbers (Fig. 1), implying that listeners are sensitive to the differences in number of talkers beyond the perceptual limits. It would be of interest to investigate the cause for the gradual increase and to examine the relationship between the discrimination sensitivity and competence in mathematics of listeners.

Massaro (1976) reported that accuracy of numerosity judgment for a series of tone bursts was worse when the tones were presented separately at different spatial locations. By contrast, the present study found that spatial separation enhances listener performance in numerosity judgment. These different effects of spatial separation may reflect bottlenecks of these two types of numerosity judgment are different.

Human voices may be special for the human auditory system (Lieberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967); therefore, the present findings may not be observed for other types of natural sounds. Studies of numerosity

judgment using natural sounds other than human voices may be useful for further understanding of the auditory cognition.

Acknowledgments The authors thank Dr. Tsuyoshi Kuroda and two anonymous reviewers for their helpful comments on the manuscript.

References

- Best, V., Gallun, F. J., Ihlefeld, A., & Shinn-Cunningham, B. G. (2006). The influence of spatial separation on divided listening. *Journal of the Acoustical Society of America*, 120, 1506–1516.
- Bregman, A. (1994). *Auditory scene analysis*. Cambridge, MA: MIT Press.
- Bronkhorst, A. W. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acustica*, 86, 117–128.
- Bronkhorst, A. W., & Plomp, R. (1992). Effect of multiple speech-like maskers on binaural speech recognition in normal and impaired hearing. *Journal of the Acoustical Society of America*, 92, 3132–3139.
- Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *Journal of the Acoustical Society of America*, 110, 2527–2538.
- Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, 27, 115–127.
- Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 25, 975–979.
- Cowan, N., Elliott, E., Scott Saults, J., Morey, C., Mattox, S., Hismjatullina, A., & Conway, A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, 51, 42–100.
- Darwin, C., & Carlyon, R. (1995). Auditory grouping. In B. C. J. Moore (Ed.), *Hearing* (pp. 387–424). San Diego, CA: Academic Press.
- Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *Journal of the Acoustical Society of America*, 106, 3578–3588.
- Halberda, J., Mazocco, M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, 455, 665–668.
- Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.
- Kashino, M., & Hirahara, T. (1996). One, two, many-Judging the number of concurrent talkers [Abstract]. *Journal of the Acoustical Society of America*, 99, 2596. doi:10.1121/1.415287
- Kawashima, T. (2005). A psychophysical study of auditory scene analysis. (Doctoral thesis, University of Tokyo, Tokyo, Japan. Written in Japanese).
- Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkman, J. (1949). The discrimination of visual number. *American Journal of Psychology*, 62, 498–525. doi:10.2307/1418556
- Kobayashi, T., Itabashi, S., Hayami, S., & Takezawa, J. (1992). ASJ continuous speech corpus for research. *Journal of the Acoustical Society of Japan*, 48, 888–893.
- Lieberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74, 431–461.
- Mandler, G., & Shebo, J. (1982). Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General*, 111, 1–22. doi:10.1037/0096-3445.111.1.1
- Massaro, D. W. (1976). Perceiving and counting sounds. *Journal of Experimental Psychology: Human Perception & Performance*, 2, 337–346.
- Nagaoka University of Technology (2002). Head related transfer functions measured and presented online (Nov. 2005). Original internet home page containing HRTF databases closed and moved to http://www.nagaoka-ct.ac.jp/ee/lab_syano/index_e.html
- Piazza, M., Mechelli, A., Price, C., & Butterworth, B. (2006). Exact and approximate judgments of visual and auditory numerosity: An fMRI study. *Brain Research*, 1106, 177–188. doi:10.1016/j.brainres.2006.05.104
- Pylyshyn, Z. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition*, 80, 127–1589.
- Pylyshyn, Z., & Storm, R. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, 3, 179–197.
- Repp, B. H. (2007). Perceiving the numerosity of rapidly occurring auditory events in metrical and nonmetrical contexts. *Perception & Psychophysics*, 69, 529–543.
- Santala, O., & Pulkki, V. (2011). Directional perception of distributed sound sources. *Journal of the Acoustical Society of America*, 129, 1522–1530.
- Stern, R. (2005). Signal separation motivated by human auditory perception: Applications to automatic speech recognition. In P. Divenyi (Ed.), *Speech separation by humans and machines* (pp. 135–154). Dordrecht, Netherlands: Kluwer Academic.
- Turgeon, M., Bregman, A., & Roberts, B. (2005). Rhythmic masking release: Effects of asynchrony, temporal overlap, harmonic relations, and source separation on cross-spectral grouping. *Journal of Experimental Psychology: Human Perception and Performance*, 31, 939–953.
- Tohoku University (2001). Head related transfer functions measured and presented online by Tohoku University Research Institute of Electrical Communication. <http://www.ais.riec.tohoku.ac.jp/index.html>
- Vuokko, E., Niemivirta, M., & Helenius, P. (2013). Cortical activation patterns during subitizing and counting. *Brain Research*, 1497, 40–52.
- Watson, C., Kelly, W., & Wroton, H. (1976). Factors in the discrimination of tonal patterns. II. Selective attention and learning under various levels of stimulus uncertainty. *Journal of the Acoustical Society of America*, 60, 1176–1185.
- Yost, W. (2008). Perceiving sound sources. In W. Yost, A. Popper, & R. Fay (Eds.), *Auditory perception of sound sources* (pp. 1–12). New York, NY: Springer.
- Yost, W., Dye, R. H., & Sheft, S. (1996). A simulated “cocktail party” with up to three sound sources. *Perception & Psychophysics*, 58, 1026–1036.