How many images are in an auditory scene?

Xuan Zhong[a] and William A. Yost
Department of Speech and Hearing Science, Arizona State University, Tempe, Arizona 85287, USA

(Received 22 September 2015; revised 24 January 2017; accepted 5 April 2017; published online 24 April 2017)

If an auditory scene consists of many spatially separated sound sources, how many sound sources can be processed by the auditory system? Experiment I determined how many speech sources could be localized simultaneously on the azimuth plane. Different words were played from multiple loudspeakers, and listeners reported the total number of sound sources and their individual locations. In experiment II the accuracy of localizing one speech source in a mixture of multiple speech sources was determined. An extra sound source was added to an existing set of sound sources, and the task was to localize that extra source. In experiment III the setup and task were the same as in experiment I, except that the sounds were tones. The results showed that the maximum number of sound sources that listeners could perceive was limited to approximately four spatially separated speech signals and three for tonal signals. The localization errors increased along with the increase of total number of sound sources. When four or more speech sources already existed, the accuracy in localizing an additional source was near chance. © *2017 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4981118]

[MAS] Pages: 2882–2892

I. INTRODUCTION

This paper deals with the question of how many sources producing simultaneous sounds in an auditory scene (Bregman, 1994; Cherry, 1953) can be identified and localized. The existing sound source localization literature has focused mainly on the localization or identification of a single sound source, in which case human listeners rely on localization cues such as interaural time and level differences as well as head-related transfer function (HRTF) cues (see Blauert, 1997).

There is not a large amount of literature measuring the ability of listeners to identify and/or localize multiple (more than two) simultaneously presented sounds, especially if the sounds are presented from different sources. It has been observed that when the total number of sound sources increases from one to two or three, the individual sources can still be individually localized (Blauert, 1997; Yost and Brown, 2013). The accuracy of sound source localization depends on the degree of temporal coincidence (Gardner, 1969) and, if the sounds are amplitude modulated on the rate of modulation (Yost and Brown, 2013). For even more sound sources, Blauert (1997) observed that summing localization would still occur as in the case of two loudspeakers, and the precision of the perception of an auditory event would depend on the degree of coherence. Lower coherence would lead to an increasingly diffuse auditory image, which could fill the entire perceived space. If the coherence exceeded 0.2, separate sound sources would be more easily localized. It was argued further that whatever cues are used for two-source localization would also apply to the case with more than two sound sources (Blauert, 1997), but the maximum number of perceived sound sources was not discussed or studied experimentally.

Yost *et al.* (1996) showed that three speech sounds were slightly less accurately identified and localized than two speech sounds. Spatially separating the sounds as opposed to having all sounds presented from one loudspeaker facilitated identification and localization performance. Santala and Pulkki (2011) investigated the influence of sound source distribution (up to 13 sound sources were used) on spatial sound perception using loudspeaker configurations restricted to the frontal horizontal plane. They found that, for simultaneous independent noises presented from loudspeakers with 15° spacing, up to three individual sound sources could be correctly perceived. Beyond five sound sources, the sources at the two ends of an array of loudspeakers were often not perceived, and the increasingly diffuse auditory image was likely to obscure the actual location of each individual sound source.

Several important recent studies examined sound source number judgments ("numerosity") in complex environments. Notably, Kawashima and Sato (2015) conducted experiments in which up to 13 sentences were played simultaneously from a maxima of six loudspeakers. In one experiment, up to six speech sounds were presented together from one loudspeaker. In this experiment, as the total number of sounds increased, listeners increasingly tended to underestimate the number of sounds. The perceptual limit was found to be between three and five, depending on the duration of the speech and gender of the talkers. In another experiment the sound sources (up to six) were spatially separated by 36° in different spatial arrangements. The spatial separation experiments were done both in a room and using HRTF simulations over headphones. There was a small improvement in the ability of the listeners to indicate the number of talkers when the sound sources were spatially

[a]Electronic mail: xuan.zhong@asu.edu

 0001-4966/2017/141(4)/2882/11/$30.00

separated as opposed to when the sounds came from the same loudspeaker, but the number of reported sources did not increase as the actual number of sources increased above three or four. Kawashima and Sato (2015) also measured reaction times in some of their numerosity experiments, and reaction times increased with increasing number of actual talkers. Kawashima and Sato (2015) asked listeners to indicate the number of talkers, rather than the sources of the sounds. Thus, sound source localization performance was not measured by Kawashima and Sato (2015), as it was in the present study.

More recently, Weller *et al.* (2016) simulated a complex room environment in an anechoic chamber with multiple talkers. Auditory "scenes" were constructed such that up to six talkers uttered 45-s portions of scripted monologues. The loudspeakers simulating the position of the talkers were distributed in azimuth and distance around the listener in the simulated room. Listeners with either normal hearing or hearing impairment were tested. Listeners were to count, locate, and identify the gender, of talkers. Normal-hearing listeners reported the actual number of talkers and their correct gender for up to four talkers, and performance on both tasks declined substantially for six talkers. Measuring sound source localization performance required a specialized analysis method to separate out localization errors associated with azimuth as opposed to distance locations. Estimated azimuth accuracy declined as a proportion of chance performance as the number of sources increased from one to six. Performance in all tasks was correlated with the degree of hearing loss, in that the greater the hearing loss the worse the performance.

The present study employed simple procedures to measure azimuth sound source localization performance for multiple simultaneous sounds. As was done by Kawashima and Sato (2015), but not Weller *et al.* (2016), the present study also compared numerosity performance when sounds were not spatially separated and when they were. The present study did not include any room or listening simulations, but measured performance in an actual acoustic space. While the present study and Weller *et al.* (2016) measured sound source localization accuracy and Kawashima and Sato (2015) did not, one of the procedures in the present study provides a simple, straight forward measure of azimuth sound source localization that is not potentially complicated by an interaction with distance as was the case in the Weller *et al.* (2016) study. The present study directly compares numerosity and sound source localization measures for words and for tones, suggesting how stimulus type might effect numerosity and sound source localization judgments, for simultaneous sounds presented from multiple sources. The present study used short duration words and tones rather than the longer duration sentences and sentence fragments used by Kawashima and Sato (2015) and Weller *et al.* (2016).

The study of the maximum number of perceived sound sources is also related to new techniques in spatial audio processing, especially audio rendering for virtual reality. A major application of such techniques is in interactive video games, in which the locations of virtual sound sources are generated and updated as the user explores a virtual space. A solution to this problem is to break the vibrating surfaces of virtual sources into small elements, and treat each element as an independent sound source. For example, Tsingos *et al.* (2004) developed software that could simulate 174 moving sound sources in real time. More recently, Moeck *et al.* (2007) suggested an algorithm that could process 1815 different sources. However, updating the locations of sound sources is computationally demanding, especially when the total number is comparatively large (for a brief review, see Verron *et al.*, 2010), and novel fast algorithms have been developed to reduce the computational load. If listeners are not aware of the locations of all, or at least a large portion of these virtual sources, the audio spatial rendering techniques may not be necessary in terms of human perception, even if they are computationally advantageous.

The present experiments focused on the role spatial separation plays in the perception of the multiple sound sources. In experiment I, multiple speech signals (one word country names) were played from up to 12 differently located loudspeakers at the same time. The total number of sound sources was randomized between one and eight. The loudspeaker locations of each talker were also randomized within a trial. The task for the listeners was (1) report the total number of loudspeakers presenting the words, and (2) report the individual locations of all the loudspeakers. In experiment II, a number of different voices were played from the same number of fixed loudspeaker locations three times in succession (three intervals). In the second interval, an additional voice at a new loudspeaker location was added to the existing sound sources. The task was to report the location of the added sound source in the second interval. Experiment III largely resembled experiment I, except that the stimuli were tones instead of speech. One reason for comparing speech and tonal stimuli is that these two types of sounds appear to represent the ends of a continuum of sound source localization accuracy with broadband stimuli (e.g., noises, Yost and Zhong, 2014; speech, Grantham *et al.*, 2007) having the highest accuracy and narrowband stimuli, such as tones, the worst (Yost and Zhong, 2014; Stevens and Newman, 1934). Experiment IV served as a control experiment to determine the extent to which performance in experiments I and III was due to spatial separation of the sounds, and not just to their spectral/temporal differences.

II. GENERAL METHODS

A. Instrumentation

All three experiments were conducted in a reflection-reduced room at Arizona State University (see Zhong and Yost, 2013, for a full description of this room). The dimensions of the room were $15' \times 12' \times 10'$ (length $\times$ width $\times$ height). The broadband reverberation time (RT60) was 97 ms (Yost *et al.*, 2015). Sounds were played from three 12-channel digital-to-analog converters (Echo Gina 12, Santa Barbara, CA) running at 44.1 kHz per channel. The generated signals were amplified with AudioSource AMP 1200 amplifiers (AudioSource, Portland, OR) before they were fed to the loudspeakers (Boston Acoustics Soundware 100, Woburn, MA). Twelve of the 24 loudspeakers on the azimuth plane, with even angular spacing of

30° (same as the 30° separation used in the study by Yost and Zhong, 2014) were arranged in the horizontal plane in a ring centered on the listener with a radius of 5 ft. The vertical level of the loudspeakers was the same as that of the listener's pinnae. The loudspeakers were numbered from 0° based on their locations, as shown in Fig. 1. In all tests, listeners indicated the location of loudspeakers by reporting these 12 numbers. Listeners' heads were not fixed. However, listeners were asked to keep their heads in place as much as possible in all experiments. The position of their heads was monitored during the experiment.

B. Speech materials

The voices of six female and six male American English talkers were recorded. Each sound was the recording of a single word: names of one-word countries. A total of 24 words were recorded for each talker. The words were Belgium, Britain, Burma, China, Congo, Cuba, Haiti, Japan, Korea, Libya, Mali, Mexico, Nauru, Norway, Oman, Peru, Russia, Sudan, Syria, Togo, Tonga, Turkey, Yemen, and Zambia. After the words were recorded, the levels were normalized. A computer algorithm detected the onset of each word spoken and time-aligned the recorded words. The exact length of each recorded word spoken had some variation, but the variation should influence all test conditions in the same way. Silence was added to the end of each recorded word so that all sound presentations were 1-s long. In the experiments, when a given loudspeaker was playing sound, ten words were picked randomly from the same talker's recordings and combined to form a 10-s stimulus, which streamed through a particular sound card channel corresponding to the given loudspeaker.

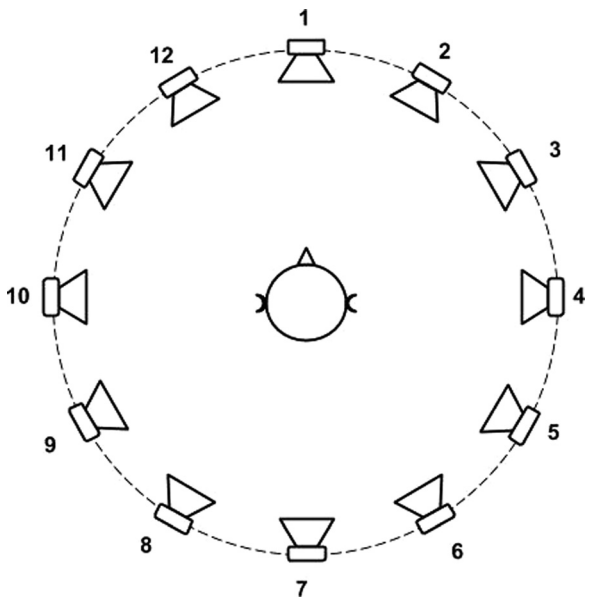

FIG. 1. Test setup for the localization tasks. Loudspeakers were numbered 1–12 with 30-degree spacing on the horizontal plane at the level of the listeners' ears. A circular loudspeaker array on the horizontal plane was used. As a result, the shorter angular path from the reported location to the actual location was used in error calculation. For instance, the angle between loudspeaker #12 and #1 was calculated as 30° (not 330°) since these two loudspeakers are separated by just one loudspeaker position.

C. Subjects

Eight normal-hearing listeners voluntarily participated in experiments I–III. All listeners had normal hearing, i.e., they had hearing thresholds not exceeding 20 dB hearing loss (HL) across the octave frequencies between 250 Hz and 8 kHz as measured by a Maico (Eden Prairie, MN) (MA 53) audiometer. Six additional listeners were tested in experiment IV. All procedures used in all of the experiments of this study were approved by the Arizona State University Institutional Review Board (IRB).

III. EXPERIMENT I: LOCATING MULTIPLE SPEECH SOURCES

A. Tasks

In experiment I, a number of loudspeakers in the azimuth plane played speech sounds at the same time. On each trial, the total number (1–8) of loudspeakers presenting sound was randomly distributed across the 12-loudspeaker array (see Fig. 1). The sound level from each loudspeaker was 65 dBA as measured at the center of the listening position. Once the total number was decided randomly, the individual locations of each active loudspeaker (12 loudspeakers) were also randomly assigned. Speech recordings were played from the loudspeakers at the same time. The total duration of sounds was fixed at 10 s for each trial. The silence time between any two consecutive repetitions was 400 ms. During this process, the same talker was assigned to the same loudspeaker, but the country name was randomized. An example trial for two loudspeakers would be loudspeaker three (as shown in Fig. 1) playing "Britain, China, Haiti, Japan,…" uttered by a particular male voice, while loudspeaker eight played "Korea, Libya, Mali, Mexico,…" uttered by a particular female voice. The task was to identify (1) the total number of loudspeakers that were playing sounds, and (2) the individual locations of all loudspeakers presenting sounds. Identifying the location of the sound sources was continued until the total number of reported sound sources was reached. For example, if listeners reported four sound sources, they were to indicate the location of each of the four sources by typing in four different numbers associated with the loudspeaker locations as shown in Fig. 1. Listeners were instructed to guess if necessary.

B. Procedures

The listeners were first instructed as to the purpose and procedure of the experiment. Then they practiced for 5–10 min. Each run consisted of eight trials, with the total number of sound sources being 1–8 in randomized sequence. Each listener was tested for 20 runs, i.e., a total of 160 trials. Each experiment took 1–1.5 h. After the listeners finished 25%, 50%, and 75% of the experiment, they were offered a short rest. During the practice sessions, the listeners were given the correct answers to familiarize them with the experimental setting and the protocol. In the formal sessions, no feedback was given to avoid practice effects.

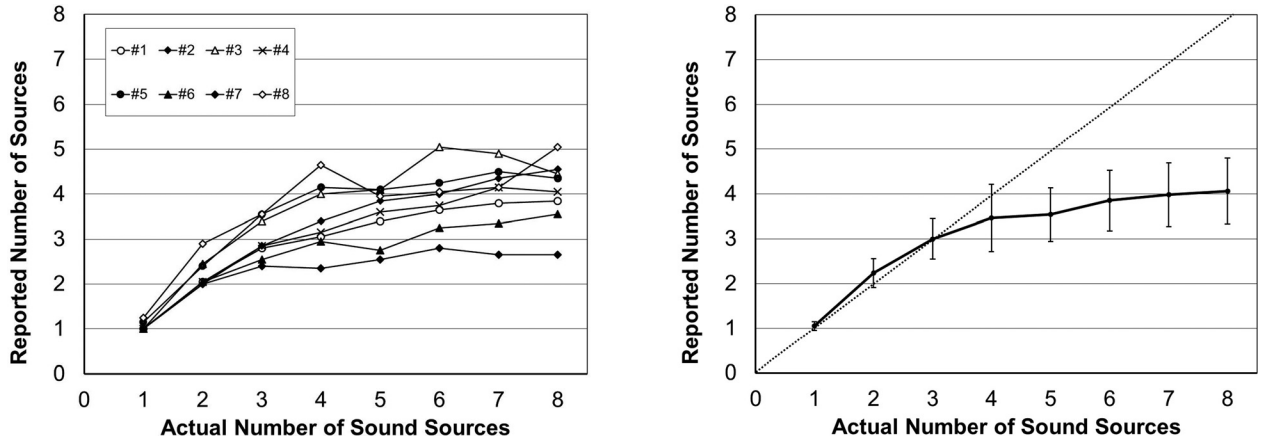

FIG. 2. (Left) Individual results of all eight listeners in experiment I (speech) showing the relationship between the reported and actual total number of sound sources. (Right) Mean and plus/minus one standard deviation across the eight listeners. The dotted diagonal line represents correct (ideal) responses. Vertical lines are +/− one standard deviation.

C. Results

The listener's ability to correctly report (1) the total number of sound sources, and (2) the locations of all sound sources was analyzed. For reporting the total number of sound sources, individual performance levels are shown in the left panel of Fig. 2 and mean performance on the right in terms of reported number of sound sources as a function of the actual number of sound sources. If the perceived number of sources was the same as the actual number of sources, the data should follow the dotted diagonal line. For the cases of one, two, and three sound sources, the mean reported total number of sources were 1.1, 2.2, and 3.0, respectively, which were within one standard deviation compared to the actual number of sources in all cases. For the case of four sources, the mean perceived total number of sources was 3.5, which is smaller than the actual total number (four), but still within one standard deviation of ideal performance. In cases of five to eight sources, the mean reported total number of sources was smaller than the actual number, and was never within one standard deviation of ideal performance. Moreover, the reported total number of sources was rarely more than four even when eight sources presented sound. On rare occasions with four or fewer actual sources, a listener reported more sources than there were actual sources. The variation in the number of reported sources for any set of actual sources for any individual listener was small (i.e., approximately the same range as the between listener range, e.g., one or two different reported sources for one actual source and two to four different reported sources for anyone listener when there were eight actual sources). As a consequence, the median number of individual listener responses was very similar to the mean number of responses as shown in the left panel of Fig. 2.

Figure 3 shows individual listener localization accuracy results (as proportion of correct location responses) on the left panel and the mean of these individual results and standard deviations on the right panel. A "hit" indicated a loudspeaker number that was one of the actual loudspeakers presenting a word. For the proportion of correct location responses, hits were summed and divided by the actual number of sound sources (e.g., if five actual loudspeakers presented words, and a listener indicated that there were three

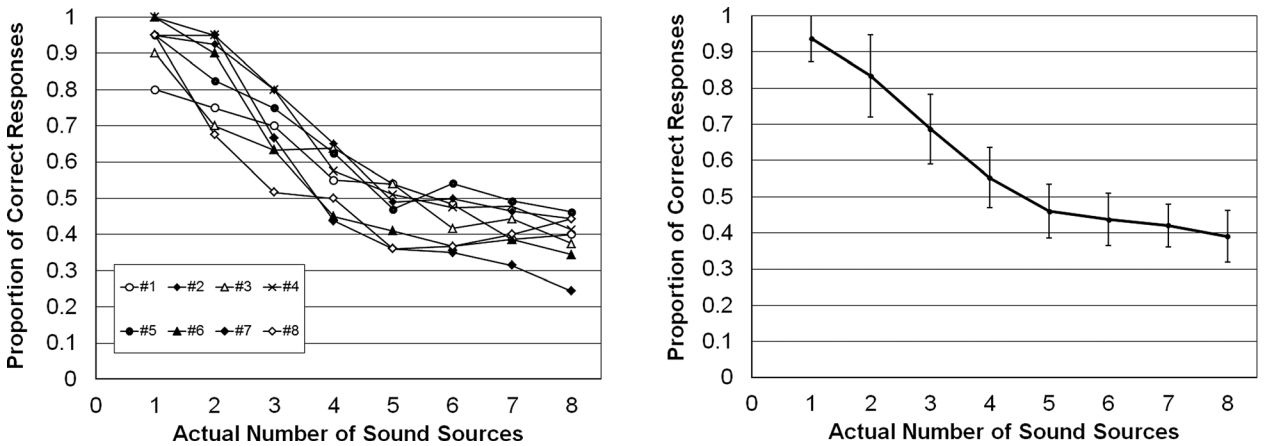

FIG. 3. (Left): Individual results (eight listeners) in experiment I (speech) showing the relationship between proportion of correct location responses (Hits) and the actual number of sound sources. (Right) Mean and plus/minus one standard deviation across the eight listeners.

TABLE I. Statistics for proportion of correct localizations in Fig. 3. N = number of presented sounds (1–8); combin(12,N) = combination of 12 sources taken N at a time (number of total possible source combinations); n = number of reported sounds (1–4) with n:Fig. 2 being n from Fig. 2; P_N(Hits):Fig. 3 = the proportion of hits from Fig. 3; n_{Hits} = number of correct localization responses measured out of the n responses, P_n(Hits) = proportion of Hits based on n reported responses; TOT_c = combination of N things taken n at a time (i.e., number of total situations in which all n responses out of N possibilities were hits); and P(Cor):Guess = guessing proportion, i.e., getting all n responses correct = TOT_c/N by chance. Measures of localization performance using the procedure of experiment I are only possible if N is considerably less than 12. Consequently, the current experiment uses 8 as the upper limit.

N	combin(12,N)	n:Fig. 2	P_N(Hits):Fig. 3	n_{Hits}	P_n(Hits)	TOT_c = combin(N,n)	P(Cor):Guess
1	12	1	0.93	0.93	0.93	1	0.0833
2	66	2	0.84	1.68	0.84	1	0.0152
3	220	3	0.70	2.10	0.70	1	0.0045
4	495	3	0.55	2.20	0.73	4	0.0081
5	792	4	0.47	2.35	0.59	5	0.0063
6	924	4	0.45	2.70	0.68	15	0.0162
7	792	4	0.42	2.94	0.74	35	0.0442
8	495	4	0.40	3.20	0.80	70	0.1414

loudspeakers presenting words, and two out of the three location responses were correct, proportion of correct locations would have been 2/5 or 0.4).[1]

When the number of actual sources was one, the proportion of correct location responses was on average 0.93. For the measure of proportion of correct locations described above, the proportion decreased as the actual number of sources increased from one to five. As the number of actual sources increased from five to eight, the proportion of correct locations slightly decreased.

Determining a measure of performance regarding judgments of the location of multiple sources is challenging because of the procedure used (e.g., reporting all locations at one time) and because listeners for most conditions reported fewer sources than actual sources. While we believe the method used for indicating location accuracy describes the overall location performance of the listeners, other measures and analyses could be performed. Table I describes a few of the other ways in which localization performance might have been measured. In Table I, N represents the actual number of presented sound sources (N = 1–8) and n represents the reported number of sound sources (n could vary between 1 and 8, but was primarily between 1 and 4; see Fig. 2). The second column in Table I represents total number of combinations of 12 possible actual sources taken N at a time for each N [combin(12,N)]. The third column (n:Fig. 2) is the number (rounded to the nearest integer) of reported sound sources for each N actual sound sources taken from Fig. 2. The fourth column indicates the proportion of correct locations (hits) for each N taken from Fig. 3 [P_N(Hits):Fig. 3]. Another way to have calculated proportion of correct locations would have been to divide the sum of the hits by n (reported number of sources) rather than N (actual number of sources), which is what was done for Fig. 3. The fifth column represents the number of hits when only the reported number of sources was used (n_{Hits}). The sixth column shows the proportion of correct locations when the proportion is based on the number of reported (n) sources [$P_n(\text{Hits}) = n_{\text{Hits}}/n$] and not the number of actual (N) sources (the number in first column). A comparison between the calculations of columns four and six indicates that while the percent of "correct" responses declines monotonically to 40% as the number of actual sources (N) increases from one to eight (one of the reasons we used this measure in Fig. 3), percent correct decreases non-monotonically as the number of reported (n) sources increases from one to four. Additionally, a measure of percent correct based on n would indicate the same or better performance than that based on N. The last two columns in Table I attempt to provide some information related to chance performance in judging the location of the sound sources in experiment I. We do so in terms of the listener being perfect in guessing the locations of all of the reported (n) sound sources. When there are N actual sources and the listener reports there are n sources, the number of n responses that are all correct (all hits, TOT_c) is the combination of N things taken n at a time [so TOT_c = combin(N,n)], which is displayed in the seventh column as a function of N. The eighth or last column is the proportion of correct responses if all n sources were correctly located and the listener guessed [P(Cor):Guess=Tot_c/combin(12,N)]. The last column denotes chance performance based on listeners getting all hits (being totally correct) for each value of n and N shown in Table I, and these values of n for each N represent how the listeners responded on average. The chance performance decreases non-monotonically with increases in either N or n. Estimates of chance performance can vary by a lot based on the assumptions one makes (in Table I, the estimate of guessing varies by a factor of 31.4 from N = 3 to N = 8). As a result, it is difficult to estimate how a measure like percent correct relates to chance performance. Estimates like those in Table I clearly suggest that there is higher probability of being correct when there are eight actual sources than for any other number of sources (1–8). Thus, it is probably the case that no single measure of an estimate of how well listeners judged sound source location will fully describe such performance, given the procedure used in experiment I.

IV. EXPERIMENT II: LOCATING AN ADDED SPEECH SOURCE

It was difficult to determine the localization accuracy of individual sound sources based on experiment I alone. In the literature, it has been documented that localization of a new or "extra" sound source was increasingly less accurate when the number of existing sound sources, or distractors, grew

from zero to two (Langendijk *et al.*, 2001). Based on this study and experiment I, as the number of distractors further increased, the localization accuracy of an individual sound source could be expected to further decrease.

The following experiment requires subjects to determine the location of an additional sound source among a set of existing sound sources, which allows for a clearer estimate of sound source localization accuracy than was possible using the procedure of experiment I as explained in Table I. When localization accuracy as measured by rms (root-mean-square) error decreases to chance level, it is reasonable to assume that the maximum number of perceived sound sources has been reached. It is unlikely that one can accurately localize five or six sources if one can only indicate that there are four sources. However, numerosity experiments alone cannot tell one what the sound source localization accuracy may be when the number of sources is less than the numerosity estimate. As a result, experiment II is designed to make the current study more complete.

A. Tasks

Experiment II expands on the second part of experiment I in which the perception of sound source location was measured. In experiment II a number of loudspeakers played sounds in each of three intervals. In the first and third intervals, the total number and locations of loudspeakers that played sounds remained the same. During the second interval, an additional source also played a different sound from a different talker. The task of the listener was to indicate the location of the additional sound source in the second interval (the listener was told that a new word at a new location would be presented in the second interval). During this process, the same talker produced the base speech sounds in the three intervals, but the country names were randomized. For example, in the first interval, loudspeakers four and ten might have played "Britain" and "China," respectively, for the same talker. In the second interval, the same loudspeakers (four and ten) might have played "Haiti" and "Korea," respectively, by the same talker used for interval one, while a new talker at a new loudspeaker location, e.g., number two, played "Libya." Then in the final (third) interval, loudspeakers four and ten played "Mali" and "Cuba" again from the same talker as was used for the base conditions in the previous two intervals. As previously mentioned, the extra word in the second interval always came from a loudspeaker that did not present a base word (i.e., in the example the extra word would not be presented from loudspeakers four or ten). The silence time in between the consecutive intervals was 400 ms. The correct answer was number two in the example above, which is the location of the added source in the second interval. The overall sound levels from each loudspeaker were 65 dBA as measured at the center of the listening position. The sound levels of the second and the third intervals randomly varied by +/− 2 dB compared to the first interval. A new talker was chosen at random for each three-interval trial.

B. Procedures

The listeners were first instructed about the purpose and procedures of the experiment. Then they practiced for 5–10 min to get familiar with the procedures. On each trial, a simple MATLAB interface played the sounds and then asked the listeners to type the number of the loudspeaker of the added sound source in the second interval. Each run was comprised of eight trials, with the total number of base sources equaling 1–8. The sequence of trials within a run was randomized. Each listener was tested for 20 runs, i.e., a total of 160 trials. Each experiment took 0.5–1 h.

C. Results

In experiment II, only the location of the single added source was reported, so rms error in degrees could be computed (see Yost and Zhong, 2014). Since a circular loudspeaker array on the horizontal plane was used, the shorter angular path from the reported location to the actual location was used when calculating errors (Fig. 1), e.g., the angle between loudspeaker #12 and #1 was calculated as 30° (not 330°) since these two loudspeakers are separated by just one loudspeaker position. Listeners could indicate any of the 12 loudspeaker locations as that containing the additional sound source in interval 2, but sound was only presented for the additional loudspeaker in interval 2 from those loudspeakers not used to produce the other sounds in intervals 1–3. For the example provided in Sec. IV A, the additional loudspeaker could have been any loudspeaker except #4 and #10. That is, in any interval in which multiple sounds were played simultaneously, any loudspeaker in the 12 locations was allowed to play at most 1 sound. Listeners sometimes reported a loudspeaker that had presented the sound in all three intervals as that presenting the additional sound in interval two, especially when there were several loudspeakers presenting sounds. Chance performance in terms of rms error [see Eq. (4) in Rakerd and Hartmann, 1986] was estimated using a 10 000 replication Monte Carlo method (as examples, see Grantham *et al.*, 2007, or Yost and Zhong, 2014). The Monte Carlo method used the exact same procedure for calculating rms error as described in Yost and Zhong (2014) and reported for the results shown in Fig. 4. The average rms error by this simulation was 104.8°, which is used as the estimate of chance performance in experiment II.

The data in Fig. 4 show an overall upward trend of rms error as the number of existing sound sources increased, meaning that as the number of sources increased, it was more and more difficult for listeners to accurately locate the additional source in interval two. In cases of four or more existing sound sources, listeners' performance was lower than or within one standard deviation of chance performance. In the current experiment, only the location of the additional source was recorded and analyzed. Locations of sound sources in intervals one and three were not recorded. Hence, although there may be coincidence of responses with non-target sources, such responses were not reported.

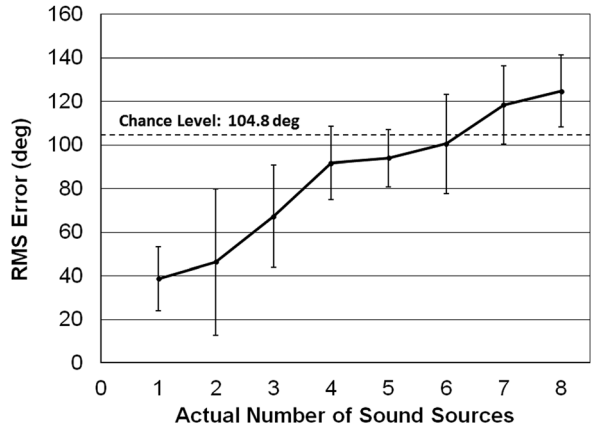

FIG. 4. Mean and plus/minus one standard deviation (over six listeners) in experiment II: added sound source (speech), showing the relationship between localization rms error and the actual number of sound sources (chance: 104.8 deg).

V. EXPERIMENT III: LOCATING MULTIPLE TONAL SOURCES

A. Tasks

In experiment III the test setup and tasks were the same as in experiment I except that the sounds were tones instead of talkers. The frequencies were 313 Hz, 419 Hz, 541 Hz, 733 Hz, 863 Hz, 1019 Hz, 1277 Hz, 1511 Hz, 1993 Hz, 2633 Hz, 3457 Hz, and 5051 Hz. All of the frequencies were prime numbers, i.e., no combination of them constituted a harmonic series. Each tone was 1-s long, with a 50-ms cosine-squared rise-fall time. In each trial, the same randomized set of sounds was played ten times at the same randomized set of loudspeakers. The ten generated tones were played repeatedly and the repetition was controlled by computer program. The silence time between each consecutive tone was 100 ms. The total duration of sounds was fixed at 11 s for each trial. The task for the listener was to first identify the total number of loudspeakers that were playing sound. Listeners were then asked to indicate the individual loudspeaker location of each of the tonal sounds they identified in their first answer. This was the same method as in experiment 1.

B. Procedures

The listeners were told of the purpose and procedures of the experiment. Then they practiced for 5–10 min. In each trial, a simple MATLAB interface asked them to enter the total number of sources, and then the locations of all the loudspeakers presenting sounds. Each run comprised eight trials, with the total number of sources one to eight in a randomized sequence. Each listener was tested for 20 runs, i.e., a total of 160 trials. Each experiment took 1–1.5 h.

C. Results

The results are shown in Figs. 5 and 6 using the same format and calculations as for Figs. 2 and 3. The individual and mean performance of all listeners in reporting the total number of sound sources is shown in Fig. 5. As in the case of multiple speech sources, as the total number of tonal sounds increased, the reported number increased. This number grew at a rate less than unity, and more than three sound sources were never reported.

The individual and mean proportion of correct location responses (computed as in experiment I) are shown in Fig. 6 in the left (individual data) and right panels (mean and standard deviation), respectively, as in Fig. 3. When the number of existing sources was one, the proportion of correct location responses was on average 0.68. As for the speech words (Fig. 3) all measures of proportion of correct location for the tones (Fig. 6) decreased as the actual number of sources increased from one to four. See Table I for additional information that may be useful in evaluating the measures of proportion of correct locations used in experiment II.

An important question related to experiments I (speech) and III (tones) was whether listeners were reporting the total number of sounds or the total number of spatially separated sound sources. It is known that human listeners are able to segregate sounds based on speech features such as fundamental frequency (F0), even if they are from exactly the same location. Experiment IV was conducted to determine the extent to which data from experiments I and III reflect an effect based on spatial separation of the sound sources or the spectral-temporal differences of the sounds. The logic of the control experiment was to present multiple sounds from a

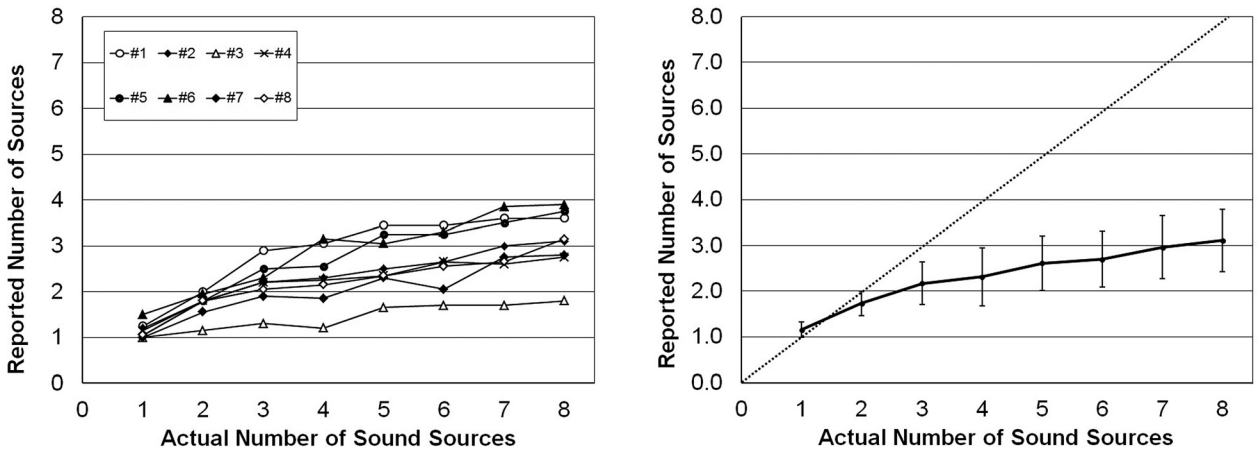

FIG. 5. Same format and calculations as in Fig. 2, but for the tonal data (experiment III). The dotted diagonal line represents correct responses.

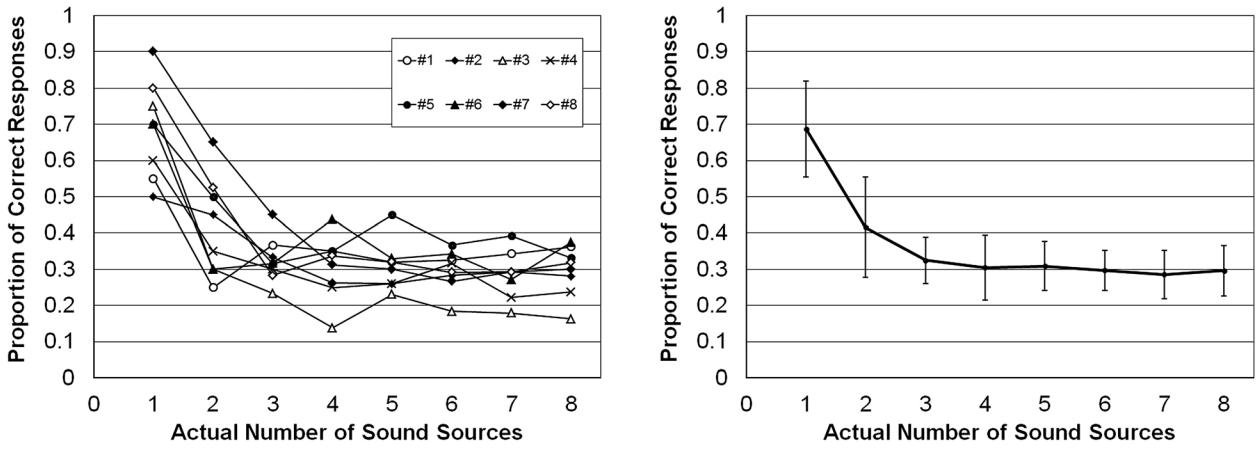

FIG. 6. Same format and calculations as in Fig. 3, but for the tonal data (experiment III).

single loudspeaker or from spatially separated loudspeakers as was done in experiments I and III.

VI. EXPERIMENT IV: ROLE OF SPATIAL SEPARATION

The previous experiments have shown that there is a limit to the number of separate sound sources that can be located by human listeners. However, it is also well known that human listeners can separate different voices that are played from a single sound source (Kawashima and Sato, 2015). As a result, an interesting question is which part of the limits reported in previous experiments is due to the limit of auditory perception, and which part is due to spatial processing. Experiment IV investigated the effect of spatial separation of sounds.

A. Task and procedure

Experiment IV resembled experiments I (speech) and III (tones) as closely as possible. In experiment IV in the nonspatial condition, the sounds (same speech words and tones used in experiments I and III) were mixed and presented from one loudspeaker randomly chosen from trial to trial from the 12 locations shown in Fig. 1. The spatial condition of experiment IV was the same as that used in experiments I and III (i.e., the sounds were spatially separated). Only a total of five actual loudspeakers was used in experiment IV, but the same pause between stimulus presentations and number of stimulus presentations were used in experiment IV as in experiments I and III. In all conditions of experiment IV, listeners were asked to indicate how many sounds (rather than sound sources) they could detect.

B. Subjects

Six listeners (four females and two males between 21 and 33 years of age) who reported normal hearing participated in experiment IV. No listener in experiment IV participated in any of the other experiments.

C. Results

As can be seen in Fig. 7 (mean data plotted as they were in Figs. 2 and 5), listeners estimated fewer sounds when all the sounds came from one loudspeaker as opposed to coming from different loudspeakers. Listeners could determine more speech sounds than tones. Thus, spatial separation did aid in determining the number of sounds.

The benefit of spatial separation of sound sources for speech words was larger than for steady tones. The results in Fig. 7 also showed that when multiple sounds were played from the same location, the number of perceived sounds was more than one. This clearly demonstrated that at least a part of the ability to detect multiple sound sources is based on an ability unrelated to spatial processing. That is, the ability to separate (segregate) sounds is also probably based on pitch, timbre, and/or speech features (e.g., temporal modulation).

VII. DISCUSSION

A. Perceptual limits of auditory scene analysis

All experiments showed that the ability of human listeners to identify the number of sound sources and to localize these sound sources decreases with increasing total number of sources. About four speech, or about three tonal, spatially separated sound sources presented simultaneously can be identified in an auditory scene. In experiments I and III, during each second of a 10-s trial, different words or tones were played from the same set of loudspeakers, i.e., the listener had ten "looks" at the locations of the loudspeakers. It was highly unlikely that more repetitions would have improved performance, but fewer may have made performance worse. Ten presentations were used because the task is difficult and this paradigm will make it possible to test other variables, such as rotating the sounds around the loudspeaker array.

When the actual number of speech sound sources was small (between one and three), the listeners tended to report the total number of sources correctly. When the total number of actual sources was four or more, the average reported number of sources plateaued around four. At the same time listeners produced larger errors in localization accuracy.

In the data analysis for Figs. 3 and 6 it is not possible to calculate a rms error [e.g., when the correct response was #1, #4, #6, and #9, and the listener's response was #1, #4, and #7, it was impossible to tell whether #6 or #9 was missed, and which sound (word or tone) was incorrectly localized at

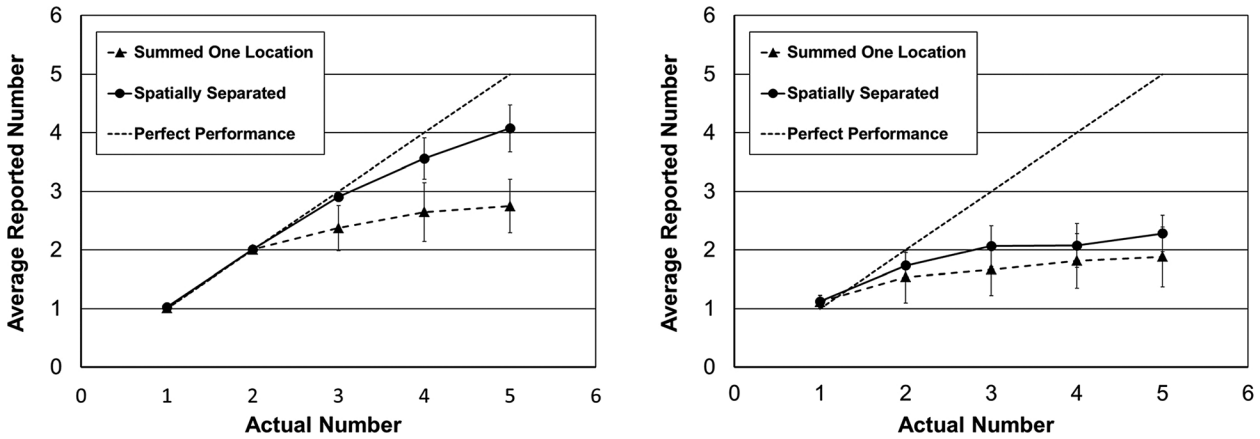

FIG. 7. (Left) The relationship between the mean and plus/minus one standard deviation of the reported number and the actual number of speech sounds (six listeners, speech stimuli) for experiment IV. Solid line and circles for spatially separated sound sources, and dashed line and triangle for all words co-located at the same loudspeaker. The dotted diagonal line represents perfect performance. (Right) Same relationships as shown in the (left) panel but for the tonal stimuli.

#7]. Experiment II was designed to validate the findings of experiment I by incrementing the number of sound sources, and by asking the listener to report only the location of that added source. In this case an rms error could be computed. The rms error increased as the total number of actual sound sources grew, and stayed within one standard deviation of chance level when the total number of speech sources was four or more.

The result of experiment II also agreed with the conclusion of experiment I that the processing of sound sources changed very little when there were more than four sound sources. We interpret that this result reflects the localization error of each sound source among a group of sources. Previously Langendijk *et al.* (2001) showed that the localization performance of a target sound source was degraded when the number of distracting simultaneous sounds increased from zero to two. In one aspect, the current study confirmed the tendency observed by Langendijk *et al.* (2001). In another aspect, our study used distractors that were played prior to the target (as well as at the same time as the target), which was not the case in the study by Langendijk *et al.* (2001). In addition, we used a larger number of distractors. Overall, experiments I and II produced converging evidence on sound source localization accuracy of several spatially separated sources producing simultaneous sound.

In experiment II, only the location of the additional source was reported by the listeners in a trial, whereas in other experiments both the total number and locations were required in each trial. As noted in Sec. III, listeners sometimes reported a loudspeaker that had presented the sound in all intervals as that presenting the additional sound only in interval two. As a result, the errors that the listener made can potentially be categorized into coincident and non-coincident types to show how existing sound sources distract the judgment of additional sound source locations. Meanwhile, a slight difference between experiments I and II was that subjects had to distinguish the target source in the second interval from the sources in the other two intervals in experiment II, whereas in experiment I they only needed to locate all sound sources. The effect of those differences is a good topic for future studies. The duration of stimuli in experiment II was shorter compared to the other experiments, which was unavoidable due to differences in experiment designs. In future studies, the effect of repetitions of stimuli on localization errors will also be discussed.

Experiment III expanded the findings of experiment I to include tonal signals. It was found that when multiple independent tones were played from multiple loudspeakers ($N > 3$), a large portion of the sources was missed, and the total number plateaued around three. The results of experiment III compared to experiment I are consistent with the literature cited in the Introduction suggesting that wideband sounds, such as speech, are more accurately localized than tones (i.e., the average proportion of correct responses for one speech source was 0.94 and 0.68 for one tonal source). Experiment IV indicates that spatially separating sound sources increases the ability to determine the number of sound sources and their locations as compared to when the sources are not spatially separated.

The current study along with the several studies described in the Introduction (Yost *et al.*, 1996; Santala and Pulkki, 2011; Kawashima and Sato, 2015; Weller *et al.*, 2016) all reached a similar conclusion that the number of sounds that can be correctly identified (numerosity) is limited to between three and four. It is also the case that sound source localization accuracy is very poor, if not near chance performance, when the number of simultaneous and spatially separated sound sources exceeds three to four. This basic outcome occurs for different sounds [words, sentences, sentence fragments (of different durations), noises, and tones], different perceptual judgments (number of sounds, number of sound sources, talker gender, azimuth sound source localization, and distance sound source localization), different psychophysical measures (percent correct identification, sound source localization accuracy, and reaction time), different simple and complex listening environments (sound fields, simulated rooms, and HRTF headphone simulations), and different listeners (normal-hearing and hearing impairment). The consistent outcome from these divergent experimental approaches suggests that the findings about the perceptual limitation of sound

source localization are robust. In several cases small, but significant, differences have been found based on different stimuli (e.g., words versus tones as measured in the present study), different speech durations (Kawashima and Sato, 2015), different listeners (normal-hearing versus hearing impaired, Weller *et al.*, 2016), and whether sounds are produced by the same sound source or spatially separated sound sources (the present study; Kawashima and Sato, 2015).

The primary aim of the present study was to investigate the role of spatial separation and sound source localization in numerosity judgments. The results imply that spatially separating sound sources does have an impact on numerosity for speech and tones. That is, spatial separation of sound sources is one of many stimulus variables that allow different sounds to be perceived when there are multiple simultaneously presented sounds. Data from the sound source localization measures of experiment I and III and the rms measures of sound source localization accuracy of experiment II indicate that sound source localization accuracy decreases as the number of sources producing simultaneous sounds increases from one, and is near chance for five or so sound sources. This is consistent with the results of Yost and Brown (2013) who showed a significant increase in sound source localization rms error between localizing one sound source and two. The end result is that the number of spatially separated sources producing simultaneous sound than can be accurately localized is limited to three of four. This limit has important consequences for segregating sound sources in an auditory scene or cocktail party context, e.g., spatial release from masking, if such masking release depends on sound source localization, would probably be limited to four or fewer maskers.

B. Implications for three-dimensional (3-D) audio rendering techniques

The results of the current study have several implications for the development of spatial audio rendering techniques. First, simulating a large number of individual sound sources may not be useful in some contexts if humans are to make perceptual judgments about the multiple sound sources. Tsingos *et al.* (2004) created a complex auditory scene with 174 sound sources, and claimed that their audio rendering technique did not affect sound source localization error. However, the behavioral experiment in their pilot test involved exploring a virtual 3-D space that combined interactive video and audio. As a result, their experiments were only suggestive of how well a single sound source location could be virtualized when visual/auditory interactions were allowed, not how good their algorithm was at creating complex auditory scenes or how realistic the scenes could be.

Second, the current results show that more efficient spatial audio compression is possible. When a spatial clustering technique was used (Tsingos *et al.*, 2004), the single-source localization error of human listeners at different angles was used to decide the lobe width of the clusters. Within the same cluster, virtual sound sources share the same set of transfer function, e.g., HRTF, to the two ears. Several previous studies (such as those reviewed in Verron *et al.*, 2010) unnecessarily separated the spatial sphere into a large number of elements or clusters, which added to the computational load of the processors. The current experiments demonstrate that the localization error of a new source will grow with the number of existing sound sources. So fewer and broader clusters can potentially be used without influencing the perception of the auditory scene.

Finally, the findings in the current study are also meaningful in the design of machine perception algorithms (Zhong *et al.*, 2016). In the designs of machine hearing systems based on human auditory system models, computational limitations may affect both systems.

VIII. CONCLUSIONS

In summary, the maximum number of sound sources that listeners perceived on the horizontal plane was around four for independent speech signals and three for tonal signals of different frequencies. The localization errors increased along with the increase of total number of sound sources. When four or more speech sources were present, listeners were at chance level in localizing an additional sound source. These data along with other data in the literature (Yost *et al.*, 1996; Santala and Pulkki, 2011; Kawashima and Sato, 2015; Weller *et al.*, 2016) suggest that an auditory scene appears to contain only a few separated sound sources that are perceptually different, implying there are probably only a few voices at a cocktail party that can be located. The virtual-audio representation of more than three or four sound sources will probably have little effect on human perception of the location of these virtual sound sources.

ACKNOWLEDGMENTS

This research was supported by a grant from the Air Force Office of Scientific Research (Grant No. FA9550-12-1-0312) awarded to W.A.Y. Most of the experiments were part of X.Z.'s Ph.D. dissertation completed in the Department of Speech and Hearing Science at Arizona State University (Zhong, 2015). X.Z. is grateful to his dissertation committee, Dr. William Yost, Dr. Yi Zhou, and Michael Dorman, for their advice and support. The assistance of the Laboratory Coordinator, Anbar Najam, and Dr. Torben Pastore are also recognized.

[1]In a very few cases listeners reported more sound sources than were actually presented (e.g., two were presented and a listener reported three). In these few cases the listener indicated the sound source locations for the number of reported locations (as was done for all other conditions), and as a consequence the proportion of correct location responses (when based on dividing by the number of actual source presentations) can be 1.0. In cases in which the number of reported sources is less than the actual number of presented sources, the proportion of correct locations can never be 1.0.

Blauert, J. (**1997**). *Spatial Hearing*, 2nd ed. (MIT, Cambridge, MA), pp. 1–494.

Bregman, A. S. (**1994**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT, Cambridge, MA), pp. 1–792.

Cherry, C. (**1953**). "Some experiments on the recognition of speech with one and with two ears," J. Acoust. Soc. Am. **25**, 975–981.

Gardner, M. B. (**1969**). "Image fusion, broadening, and displacement in sound location," J. Acoust. Soc. Am. **46**, 339–349.

Grantham, D. W., Ashmead, D. H., Ricketts, T. A., Labadie, R. F., and Haynes, D. S. (**2007**). "Horizontal-plane localization of noise and speech signals by postlingually deafened adults fitted with bilateral cochlear implants," Ear Hear. **28**, 524–541.

Kawashima, T., and Sato, T. (**2015**). "Perceptual limits in a simulated 'cocktail party,' " Atten. Percept. Psychophys. **77**(6), 2108–2120.

Langendijk, E. H., Kistler, D. J., and Wightman, F. L. (**2001**). "Sound localization in the presence of one or two distracters," J. Acoust. Soc. Am. **109**, 2123–2134.

Moeck, T., Bonneel, N., Tsingos, N., Drettakis, G., Viaud-Delmon, I., and Alloza, D. (**2007**). "Progressive perceptual audio rendering of complex scenes," in *Proceedings of the 2007 Symposium on Interactive 3D Graphics and Games*, pp. 189–196.

Rakerd, B., and Hartmann, W. M. (**1986**). "Localization of sound in rooms, III. Onset and duration effects," J. Acoust. Soc. Am. **80**, 1695–1706.

Santala, O., and Pulkki, V. (**2011**). "Directional perception of distributed sound sources," J. Acoust. Soc. Am. **129**, 1522–1530.

Stevens, S. S., and Newman, E. B. (**1934**). "The localization of pure tones," Proc. Natl. Acad. Sci. U.S.A. **20**, 593–596.

Tsingos, N., Gallo, E., and Drettakis, G. (**2004**). "Perceptual audio rendering of complex virtual environments," ACM Trans. Graphic. **23**, 249–258.

Verron, C., Aramaki, M., Kronland-Martinet, R., and Pallone, G. (**2010**). "A 3-D immersive synthesizer for environmental sounds," IEEE/ACM Trans. Audio, Speech, Lang. Process. **18**, 1550–1561.

Weller, T., Best, V., Buchholz, J. M., and Young, T. (**2016**). "A method for assessing auditory spatial analysis in reverberant multitalker environments," J. Am. Acad. Audiol. **27**, 601–611.

Yost, W. A., and Brown, C. A. (**2013**). "Localizing the sources of two independent noises: Role of time varying amplitude differences," J. Acoust. Soc. Am. **133**, 2301–2313.

Yost, W. A., Dye, R. H., and Sheft, S. (**1996**). "A simulated 'cocktail party' with up to three sound sources," Percept. Psychophys. **58**, 1026–1036.

Yost, W. A., and Zhong, X. (**2014**). "Sound source localization identification accuracy: Bandwidth dependencies," J. Acoust. Soc. Am. **136**, 2737–2746.

Yost, W. A., Zhong, X., and Najam, A. (**2015**). "Judging sound rotation when listeners and sounds rotate: Sound source localization is a multisystem process," J. Acoust. Soc. Am. **138**, 3293–3310.

Zhong, X. (**2015**). "Dynamic spatial hearing by human and robot listeners," Doctoral dissertation, Arizona State University.

Zhong, X., Sun, L., and Yost, W. (**2016**). "Active binaural localization of multiple sound sources," Rob. Auton. Syst. **85**, 83–92.

Zhong, X., and Yost, W. A. (**2013**). "Relationship between postural stability and spatial hearing," J. Am. Acad. Audiol. **24**, 782–788.