Perception & Psychophysics
1996, 58 (7), 1026–1036

A simulated "cocktail party" with up to three sound sources

WILLIAM A. YOST, RAYMOND H. DYE, JR., and STANLEY SHEFT
Loyola University, Chicago, Illinois

Listeners identified spoken words, letters, and numbers and the spatial location of these utterances in three listening conditions as a function of the number of simultaneously presented utterances. The three listening conditions were a normal listening condition, in which the sounds were presented over seven possible loudspeakers to a listener seated in a sound-deadened listening room; a one-headphone listening condition, in which a single microphone that was placed in the listening room delivered the sounds to a single headphone worn by the listener in a remote room; and a stationary KEMAR listening condition, in which binaural recordings from an acoustic manikin placed in the listening room were delivered to a listener in the remote room. The listeners were presented one, two, or three simultaneous utterances. The results show that utterance identification was better in the normal listening condition than in the one-headphone condition, with the KEMAR listening condition yielding intermediate levels of performance. However, the differences between listening in the normal and in the one-headphone conditions were much smaller when two, rather than three, utterances were presented at a time. Localization performance was good for both the normal and the KEMAR listening conditions and at chance for the one-headphone condition. The results suggest that binaural processing is probably more important for solving the "cocktail party" problem when there are more than two concurrent sound sources.

In 1953, Cherry wrote,

> How do we recognize what one person is saying when others are speaking at the same time (the "cocktail party problem")? On what logical basis could one design a machine ("filter") for carrying out such an operation? A few of the factors which give mental facility might be the following: (a) The voices come from different directions. (b) Lipreading, gestures, and the like. (c) Different speaking voices, mean pitches, mean speeds, male and female, and so forth. (d) Accents differing. (e) Transition-probabilities (subject matter, voice dynamics, syntax ...). (p. 925)

The *cocktail party effect* is an extensively cited auditory phenomenon (see Blauert, 1983) that has in recent years been reformulated as a problem of sound source determination (see Yost, 1992a) or sound source segregation (see Bregman, 1990). That is, how do we determine the sources of sound in multisource acoustic conditions? Cherry's quotation suggests several variables that might contribute to a solution to this problem. Over the years, several authors (see Yost, 1992a and 1992b, for a review) have added to Cherry's original list of possible solutions.

This research was supported by grants from the National Institute on Deafness and Other Communication Disorders and the Air Force Office of Scientific Research. We would like to thank our colleagues at the Parmly Hearing Institute for their comments on our work. Correspondence should be addressed to W. A. Yost, Parmly Hearing Institute, Loyola University, 6525 N. Sheridan Rd., Chicago, IL (e-mail: wyost@luc.edu).

Cherry felt that spatial separation was a major contributor to solving the cocktail party problem, yet spatial cues are just a subset of the possible cues that may be used to determine the sources of sound. Thus, spatial cues are neither necessary nor sufficient for sound source determination. A recent review of the literature (Yost, in press) suggests that spatial separation may not play a major role in sound source determination (i.e., listeners may be able to use many cues to solve the cocktail party problem). However, in that review it was pointed out that very few data have been collected in real-world listening situations, and most data relevant to the cocktail party problem have been obtained with only two competing sound sources. Finally, most work on the cocktail party effect has involved paradigms that would be considered "selective attention" tasks, in that the listener is to attend to one source in the presence of competing sources. In everyday listening, we often use "divided attention" (see Jones & Yee, 1993) to determine many concurrent sources in our acoustic world, until we select a source or sources of interest. That is, the cocktail party phenomenon reflects the ability to use spatial separation to segregate several sound sources. If there is a reason to process a single source, then one might selectively attend to that source while effectively ignoring the other sound sources. In the present study, listeners were asked to identify *all* active sound sources rather than to identify a particular source.

Therefore, the cocktail party phenomenon was investigated under somewhat real-world conditions by using up to three sound sources. Listeners were asked to identify words, letters, or numbers (utterance identification)

presented simultaneously over loudspeakers. They were also asked to indicate the location of each loudspeaker (utterance location) that presented the utterance. Since the loudspeakers were placed in the front horizontal plane at the level of the listener's pinnae and the utterances were low-pass filtered at 4000 Hz, sound source location would depend almost entirely on interaural differences of time and level (Wightman & Kistler, 1993). Performance was measured in three conditions: (1) *normal listening*—utterances were presented over loudspeakers in a sound-deadened room in which the listener was seated; (2) *one-headphone listening*—a single microphone at the position of the center of the listener's head sent the sounds to a single headphone to the listener seated in a remote sound-proof room (a condition in which binaural cues were removed); and (3) *KEMAR listening*—KEMAR (Knowles Electronic Manikin for Acoustical Research) was "seated" where the listener would have been and the left and right outputs from KEMAR's ears were led to the left and right channels of stereo headphones that were worn by a listener seated in a remote sound-proof room. Listeners in the normal listening condition were allowed to make head movements but were required to maintain a fixed distance from the loudspeakers (they could not lean). The KEMAR listening condition reintroduced some of the binaural cues that are available under normal free-field listening, but did not allow listeners to utilize information arising from head movements or from individualized head-related transfer functions.

METHOD

There were seven loudspeakers (Realistic-Minimus 2.5) in the sound-deadened room that were driven by Crown (Power Line Two) amplifiers (see Figure 1). The loudspeakers were equally spaced around the frontal hemisphere in 30° increments (from −90° azimuth on the left to +90° azimuth on the right) at a distance of 1.3 m from the listener, and at a height of 1.2 m (approximately the height of the head of the listener when seated). The room was 3.5 m long, 2.5 m wide, and 2.1 m high, constructed of sound-deadening office partitions and lined with sound-attenuating foam on all surfaces. Additional sound-attenuating foam was placed in the room at various locations (e.g., directly behind the listener) to equalize (as much as possible) the sound levels at the location of

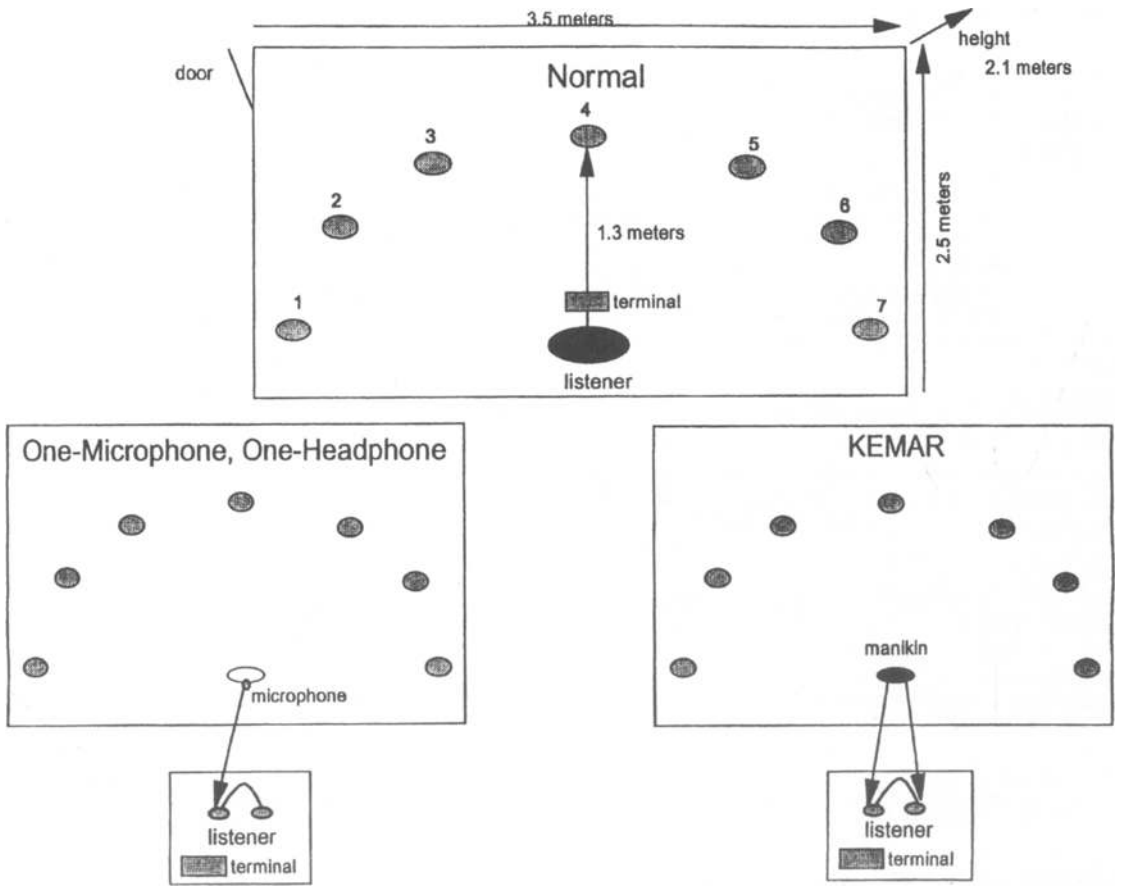

Figure 1. A drawing showing the spatial layout of the sound-deadened listening room (top), and the one-headphone and KEMAR listening conditions (bottom). In the listening room, seven loudspeakers are located in a semicircle in front of the listener. In the one-headphone listening condition, a single omnidirectional microphone placed at the location of the listener delivers the sound to a single headphone of the listener seated in a remote sound-proof room. In the KEMAR listening condition, the binaural recordings from a stationary KEMAR are delivered to the stereo headphones of the listener seated in the remote room. The numbers 1–7 were used by the listeners to indicate the spatial location of each of the seven loudspeakers.

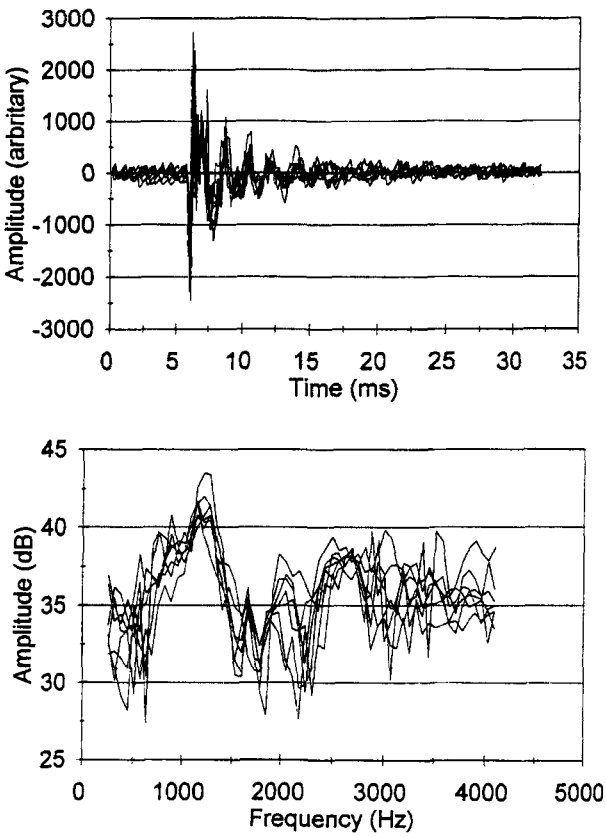

Figure 2. On top are the superimposed average time waveforms from the seven loudspeakers excited with a 122-μsec click. The measurements were made at the location of the listener, and the waveforms represent the average of 10 clicks. On the bottom are the superimposed average amplitude spectra of the clicks recorded from each of the seven loudspeakers. The spectra represent the average amplitude in each adjacent 25-Hz spectral region from 300 to 4096 Hz.

the listener's head that were produced by each of the seven loudspeakers. The loudspeaker outputs were matched in dB(A) output, and the seven loudspeakers were chosen to be within 2 dB of each other in the spectral region of 300–4000 Hz (the approximate bandwidth of the stimuli).

Figure 2 displays the time waveform averaged over 10 presentations and smoothed spectra of a train of ten 122-μsec transients (presented at the rate of 5 per second) delivered over each of the seven loudspeakers and recorded (ElectroVoice 112 omnidirectional microphone) through a sound-level meter (IVIE, IE-30-A) placed at the position of a listener's head. The spectra were obtained with a 4,096-point FFT in which each frequency was represented and then smoothed by plotting the average level for each consecutive 25-Hz band. Although there were some acoustical differences among the seven loudspeakers located at the seven different positions in the room, the differences were relatively small. The acoustic controls, the measurements (see Figure 2), and the nature of the results made it unlikely that the findings were contaminated by idiosyncratic acoustic properties of the room or particular loudspeakers. The room was reflective (the most intense first echo was approximately 25 dB down from the source), but not to the extent that any one loudspeaker produced utterances that were more identifiable than those produced by the other loudspeakers.

The speech materials were 42 NU-6 words, the 26 letters of the alphabet, and the numbers 1–9 spoken by seven male talkers. The talkers were in the sound-deadened room, and they spoke the words into a microphone (ElectroVoice 112 omnidirectional microphone) that fed the signals directly to a low-pass filter (4000-Hz cutoff frequency). A MASSCOMP computer sampled the signals at a rate of 8192 Hz with a 16-bit A/D converter and then stored the stimulus files. The final sets of words obtained from all talkers were determined by a panel of three judges to be equally intelligible.

During each test condition, each loudspeaker presented an utterance of a unique male talker (e.g., Loudspeaker 1 would always present the utterances of Male Talker 2, etc.). The loudspeaker location of each "talker" was fixed for one listener, but changed randomly from listener to listener. Utterances were presented *one at a time*, *two at a time*, or *three at a time*, from one, two, or three randomly chosen loudspeakers. These three conditions were presented in the order listed above, and the listeners were never told for any condition how many sources would be presented. The loudspeakers and the words were chosen randomly, the only constraint being that a particular utterance could be spoken by only one talker at a time. In each of these conditions (see Table 1), there were three lists. (1) 42 NU-6 words (e.g., 6 words for each of the seven talkers) were presented the first time without any prior exposure of the listener to the list (W1*y*); (2) the listener then had 10 min to study the words, and following this study period, the same NU-6 words were presented again but in a different random order (this repetition allowed an estimate of learning) (W2*y*); and (3) the 26 letters and 9 numbers ("let"; see Table 1, note) were presented (letters and numbers were assumed to be highly overlearned utterances that would minimize any learning effects).

There were three listening conditions as described in the introduction: (1) normal listening, (2) one-headphone listening, and (3) KEMAR listening. KEMAR was fitted with Zwislocki couplers, and the recordings were made with an ER-11 microphone/preamplifier system. An ElectroVoice 112 omnidirectional mi-

Table 1
The Conditions and Groups of the Experiment

Group	Listening Condition	No. of Utterances	Gender M	Gender F	Age Range
1	normal	1 at a time	3	2	19–23
2	normal	2 at a time	3	2	22–24
3	normal	3 at a time	2	3	21–22
4	one headphone	1 at a time	2	3	19–41
5	one headphone	2 at a time	1	4	19–22
6	one headphone	3 at a time	2	3	18–20
7	KEMAR	1 at a time	2	3	19–25
8	KEMAR	2 at a time	2	3	17–32
9	KEMAR	3 at a time	1	4	20–25
Follow-up	all three	3 at a time	2	5	22–45

Note—Tasks comprised (1) utterance identification, with NU-6 words the first time (W1*y*), NU-6 words the second time (W2*y*), letters and numbers (let) (*y* represents one of the three levels of scoring), and (2) utterance location, with loudspeaker localization (loc) three times (loc1, loc2, loc3). Words were scored as follows: words as originally entered (W*x*1), level-2 scoring (W*x*2), level-3 scoring (W*x*3). *x* represents one of the three levels of the utterance identification task. Groups 1–9 performed all tasks in the following order: W1*y*, loc1, W2*y*, loc2, let, loc3; and all utterances were analyzed using all three word scoring measures (W*x*1 ,W*x*2, W*x*3). Thus the data for each group were recorded as: W11, W12, W13, loc1, W21, W22, W23, loc2, let, loc3. In the follow-up group, the order of listening conditions was the following: normal (data not counted), normal, one headphone, then KEMAR; and within each listening condition, the listeners performed identification and then localization for the condition in which the utterances were presented three at a time.

crophone and a Sennheiser Model H05 earphone were used in the one-headphone listening condition. In the normal listening condition, the listeners were free to move their heads, but they could not move from the chair, lean forward, or lean to the side, since a motion detector would then cause the trial to be aborted. In the KEMAR conditions, the manikin remained stationary.

The first time through each list for each listening condition and number of utterances, the listeners were asked to enter into the computer all of the words, letters, or numbers that they heard (utterance identification). They could listen to each utterance or group of utterances as often as they wanted, and the numbers of times that they listened were recorded. After all of the utterances were presented, they were repeated in the same order, and the listeners indicated from which loudspeakers (1–7, see Figure 1) utterances were heard (utterance localization). For each utterance or group of utterances, the words, letters, or numbers that each listener had identified (whether or not they were actually presented) were displayed sequentially on the terminal and the listener indicated the number of the loudspeaker (1–7) from which the utterance was delivered. That is, the locations of all utterances were not determined—only those of the words, letters, and numbers that a listener listed during the utterance identification part of the experiment. During the utterance localization task, listeners could also listen to each utterance or group of utterances as often as they wanted, and the numbers of times that they listened were recorded.

There were nine groups of listeners in the experiment (see Table 1); three listening conditions by the three different numbers of simultaneous utterances (one, two, or three). Each group had 5 listeners. The listeners were recruited from the introductory psychology course at Loyola University, Chicago. Besides receiving credit for the course, listeners were paid; each of them could earn as much as $25 for 1 h of participation. The amount of payment per correct utterance was scaled according to the anticipated difficulty of the task. There was a criterion number of correct responses that a listener had to achieve before being allowed to earn money and before the data were used in the study. The average amount earned in each group ranged from $18.50 to $20.25, so it is unlikely that any major differences in the data are due to differences in motivation across groups of listeners.

Sixty-one students participated in the study, and the data from 52 listeners were included in this paper. That is, 9 listeners failed to perform at the criterion level of performance during one of the three presentation conditions in the normal listening environment (2 failed for the two-at-a-time condition, and 7 failed for the three-at-a-time condition). Forty-five listeners were used in the nine groups described above, and 7 in a follow-up experiment. In the follow-up experiment, 7 (unpaid) listeners ran in all three listening conditions with letters and numbers as utterances ("let" and "loc"; see Table 1). They listened only to utterances pre. ted three at a time, and the listening conditions were presented in the following order: two runs in the normal listening condition (the data for the first run in the normal listening condition were not used), in the one-headphone condition, and finally in the KEMAR condition. The follow-up experiment allowed us to have a within-groups comparison so that we could determine whether certain effects in the main experiment might have been due to the between-groups design.

All utterances were bandpass filtered between 300 and 4000 Hz (they were played out at an 8192-Hz rate through a 16-bit, 8-channel, buffered D/A converter on the MASSCOMP computer). The sounds were normalized to the same RMS level, and when more than one word, letter, or number was presented at a time, the utterances were aligned at their temporal midpoints. Since the difference in duration of the utterances was maximally 128 msec, the maximum onset (offset) separation was 64 msec. The level of each utterance was measured at the location of the listener's head for each loudspeaker. These levels ranged over the duration of the utterances and across utterances from approximately 66 to 74 dB(A) (from the slow meter reading on the sound-pressure level meter). The speech was mixed with the same broadband white noise presented continuously to all seven loudspeakers at an overall level of 55 dB(A) (measured individually for each loudspeaker). The background noise served to mask any extraneous sounds.

The data for the letters and numbers were scored only for total percent correct. For the words, there were three levels of scoring analysis: (1) Level 1 (W*x*1), scored directly as the listener responded; (2) Level 2 (W*x*2), in which corrections were made for spelling and homonyms (e.g., *dear* for *deer*); and (3) Level 3 (W*x*3), in which additional corrections were made for words or near words that might have been the combination of the words presented (e.g., *mop* and *fall*, yielding *mall*), for close homonyms (e.g., *drain* for *rain*), and suffixes added by the listener that were not in the spoken word (e.g., *homes* for *home*).

RESULTS

Figures 3–5 show the average percent correct scores, *P*(C), for the conditions of listening to one (Figure 3), two (Figure 4), or three utterances at a time (Figure 5). The condition labels are those defined in Table 1. That is, the data in each figure are shown for the identification and localization tasks for the words (both sets of presentations) and for the letters and numbers. Performance decreased as the number of competing utterances increased from one at a time to three at a time. In general, performance was best under the normal listening conditions, poorest in the one-headphone condition, and intermediate in the KEMAR condition. For each condition, performance increased as the listener had more experience and/or familiarity with the utterances. For some conditions and listeners, there were fairly large differences in performance as a function of how the NU-6 words were scored (for the three levels of scoring, see Table 1). The letter and number conditions ("let") were the easiest to score (i.e., they required the least amount of interpretation of the listener's responses), and they always yielded the best performance. Standard errors of the mean in the identification tasks ranged from 3% for utterances presented one at a time, to 7% for utterances presented two at a time, to 13% for utterances presented three at a time. Variability decreased as listeners obtained more experience with the utterances, with the smallest variability being that for the "let" condition.

Localization performance (loc1, loc2, loc3) was very good in the normal condition, fairly accurate with KEMAR, and at or near chance (chance was 1/7, or 14.3%) in the one-headphone condition. In the one-headphone condition, some listeners performed below chance in that they recognized some of the voices and assigned them a consistent loudspeaker number, but most of the time this assignment was incorrect. This appeared to have happened for 11 listeners and for a few talkers (three or fewer) for each of these listeners. Such a strategy caused 1 listener to score above chance, because she correctly recognized one of the talkers and assigned him to the correct loud-

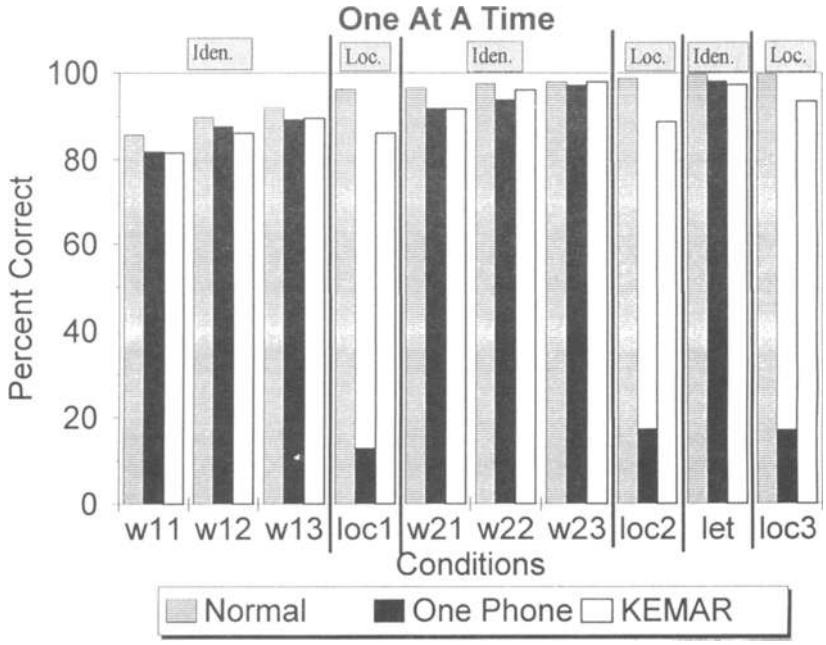

Figure 3. Average percent-correct performance in utterance identification (Iden. = identification task) and utterance localization (Loc. = localization task) for each listening condition when the utterances were presented one at a time and for the three ways of scoring word identification. Table 1 should be consulted for a description of the W*xy* conditions. In general, they represent the first ($x = 1$) or second ($x = 2$) time the listener was presented the words and the three levels of scoring (y). Standard error of the mean was 3% across all identification and across all localization conditions.

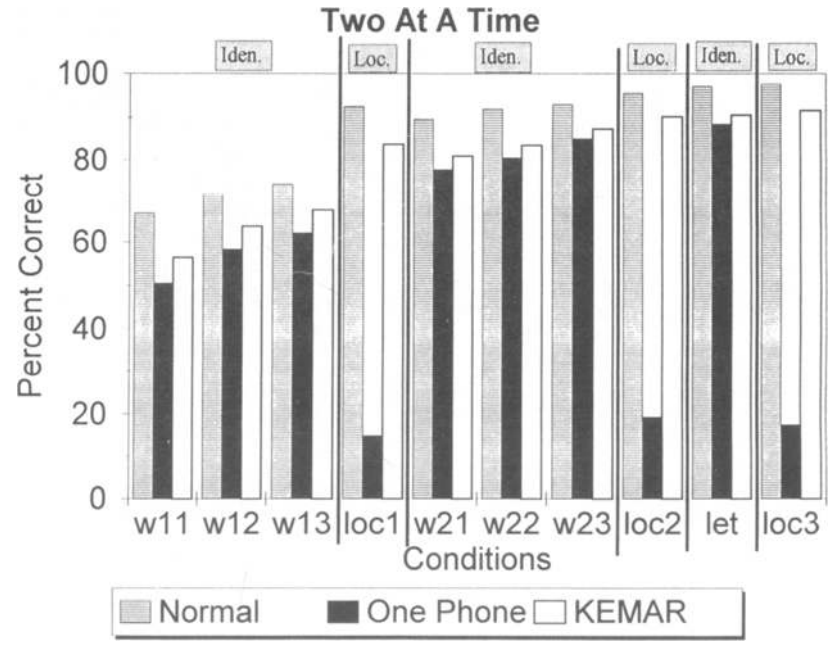

Figure 4. Same as Figure 3, except for utterances presented two at a time. Standard error of the mean was 7% across all identification and 5% across all localization conditions.

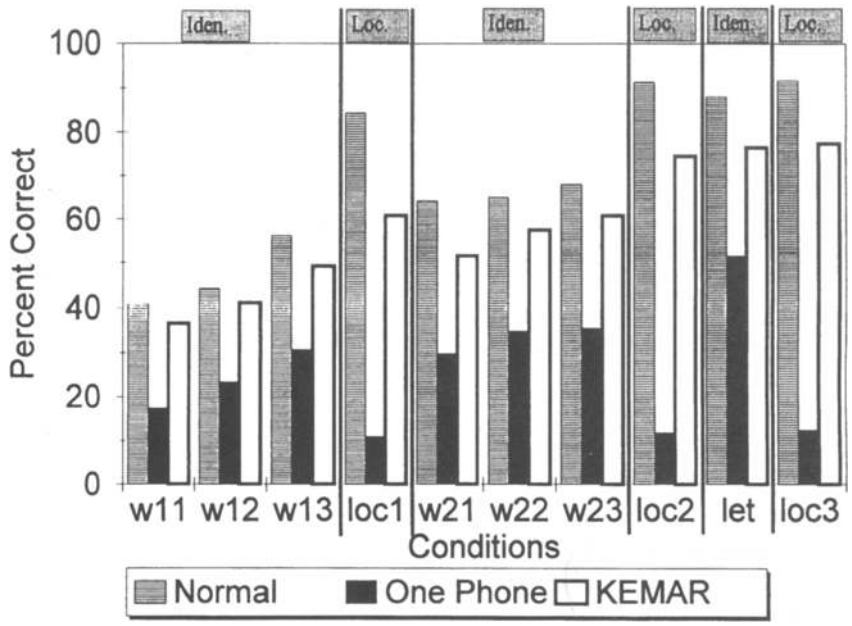

Figure 5. Same as Figure 3, except for utterances presented three at a time. Standard error of the mean was 13% across all identification and 11% across all localization conditions.

speaker. Recall, that the percent-correct scores for localization were only tabulated on the basis of the utterances that the listener listed during the identification part of the experiment. Standard errors of the mean in the localization tasks for normal listening and KEMAR listening ranged from 3% for utterances presented one at a time, to 5% for utterances presented two at a time, to 11% for utterances presented three at a time. Variability in the localization task decreased as the listener gained more experience with the utterances (as it did in the identification task), with the smallest variability found in the "let" conditions. Variability in localization performance in the one-headphone conditions depended on whether or not the listeners were able to identify one or more of the talkers, as explained above. The standard error of the mean for localization performance in all one-headphone conditions was 8%.

The data were subjected to two three-way analyses of variance (ANOVAs; listening condition × number of utterances × practice), one for identification and one for localization. For both ANOVAs, all three main effects were significant at the .01 level, indicating that the listening condition [$F(2,108) = 39.68$ for identification and $F(2,108) = 2,839.8$ for localization], number of utterances [$F(2,108) = 331.97$ for identification and $F(2,108) = 53.18$ for localization], and amount of experience [$F(2,108) = 120.51$ for identification and $F(2,108) = 16.54$ for localization] produced significant changes in both identification and localization performance. For the identification task there were also significant interactions (at the .01 level of significance) between listening condition and number of utterances [$F(8,108) = 14.42$] and between listening condition and practice [$F(8,108) = 12.59$]. Both interactions were most likely due to the fact there was very little change in identification performance across the three listening conditions or the three levels of scoring for the condition in which the utterances were presented one at a time. For the localization task there was a significant interaction (at the .01 level of significance) between listening condition and number of utterances [$F(8,108) = 9.05$], most likely because localization performance for the one-headphone listening condition changed very little as a function of the number of utterances.

Planned comparisons showed no significant difference at the .01 level in letter and number identification performance ("let" condition) between the conditions in which the utterances were presented two at a time and when they were presented one at a time [$F(1,108) = 5.64$, which is significant at the .05 level]. There were significant differences at the .01 level between the conditions in which the utterances were presented three at a time and either one at a time [$F(1,108) = 105.82$] or two at a time [$F(1,108) = 34.12$]. Because of this, and because the "let" condition produced the best performance and was easiest to score, much of the discussion is based on the "let" conditions for utterances presented three at a time.

Figure 6 shows identification performance for the "let" conditions for the three listening conditions and different numbers of utterances. This figure shows clearly that there was only a small change in performance when the utterances were presented two at a time as the listening condition changed from the normal to the one-headphone

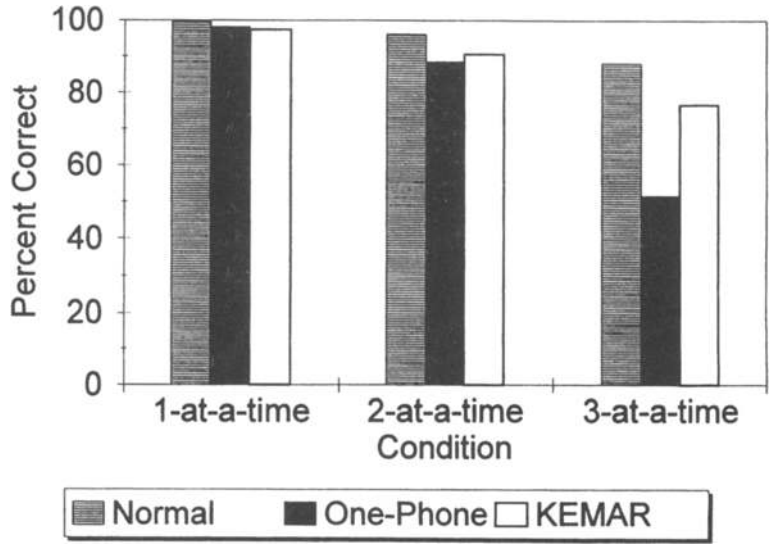

Figure 6. Average percent-correct letter and number identification comparing the three listening conditions for each of the three numbers of utterances. Standard error of the mean was 8.5% across all conditions.

condition. There was a much larger change when the utterances were presented three-at-a-time across conditions, and listening in the KEMAR condition to utterances presented three at a time produced performance intermediate to that obtained in the two other listening conditions.

Figure 7 shows the identification results for the "let" condition, when three letters or numbers were presented simultaneously, as a function of the separation between the loudspeakers. As the separation between the loudspeakers increased, performance improved for the normal and KEMAR listening conditions, but not for the

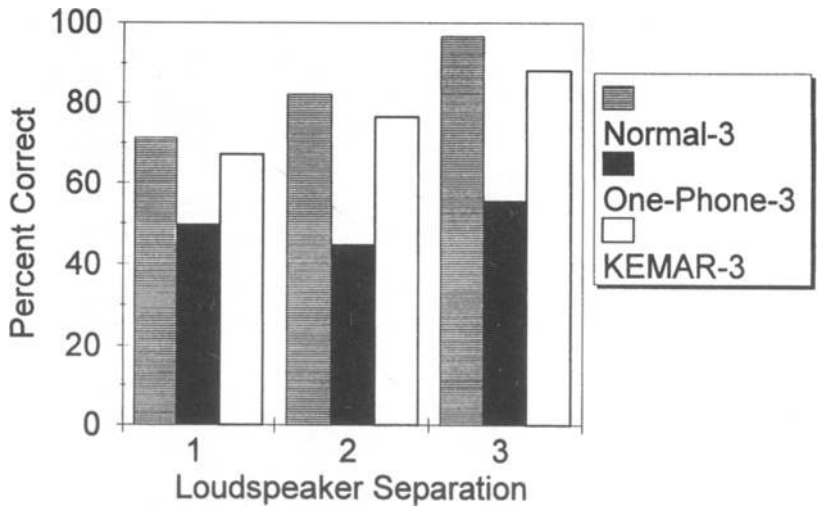

Figure 7. Average percent-correct letter and number identification as a function of the loudspeaker separation and listening condition when the utterances were presented three at a time. In this condition, the letters and numbers could come from loudspeakers separated by one (e.g., Loudspeakers 4, 5, 6), two (e.g., Loudspeakers 1, 3, 5), or three (Loudspeakers 1, 4, 7) loudspeakers.

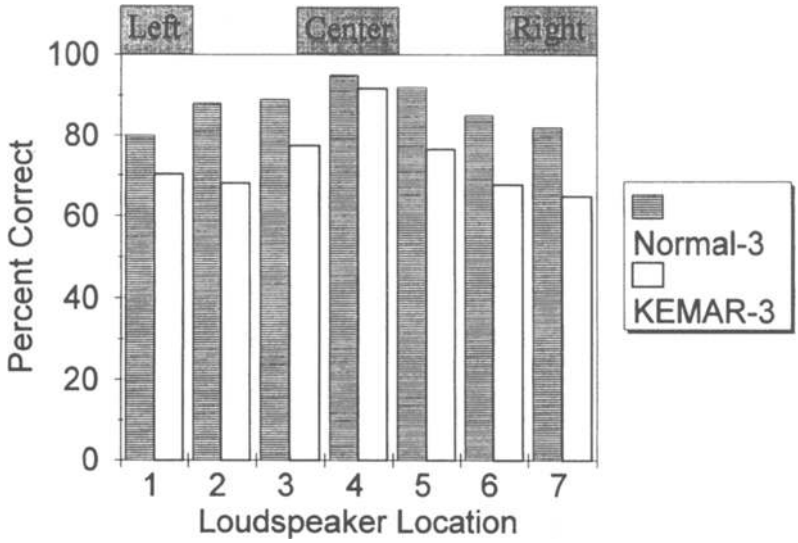

Figure 8. Average percent-correct localization accuracy for the letters and numbers ("let" condition) as a function of which loudspeaker delivered the utterance in the normal and KEMAR listening conditions when the utterances were presented three at a time.

one-headphone condition. Figure 8 shows the localization accuracy for the "let" condition as a function of loudspeaker location. Data for the one-headphone listening condition are not shown, since performance in this condition was essentially at chance. For both listening conditions that are shown, listeners were more accurate in determining the correct location of sounds coming from the loudspeakers at −30°, 0°, and +30° than at −90°, −60°, +60°, and +90°. Since the listeners (and KEMAR) were oriented toward the speaker at 0°, this result is consistent with the finding that localization acuity is best for sounds in front as compared with those at

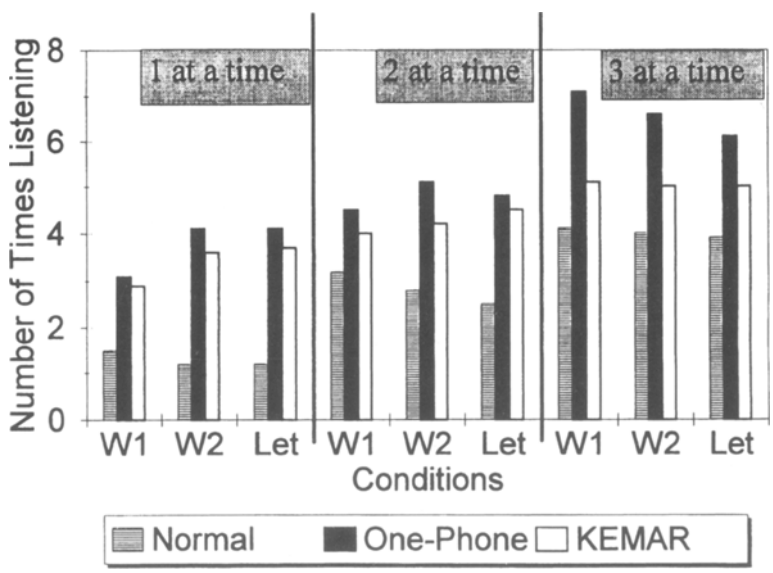

Figure 9. The average number of times the listeners elected to listen to each utterance or group of utterances as a function of the first two times listening to the words (W1 and W2) and listening to the letters and numbers ("let") and for the three listening conditions and number of utterances.

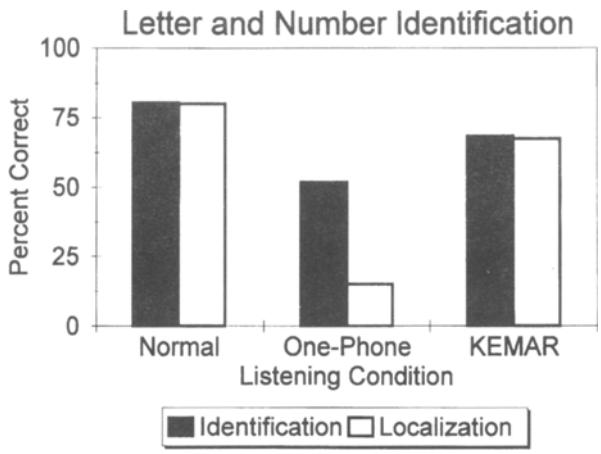

Figure 10. The average percent-correct identification and localization scores for a group of 7 listeners in the follow-up study who listened to the letters and numbers presented three at a time in all three listening conditions. Standard error of the mean was 7.5% across all identification and 5% across all localization conditions.

the side (Mills, 1958). The results of the utterance identification task showed no apparent effect of speaker location on performance.

For 3 of the 45 listeners, identification performance was slightly better for one talker (the same talker for all 6 of these listeners) than for the other six talkers. For these 3 listeners, this talker yielded performance that was on average 6.7% better than that for the other talkers. These 3 listeners were each in different groups. Although no formal acoustical analysis was done, this talker appeared to have the lowest fundamental voicing frequency. The data from Figures 7 and 8 also indicate that the differences in performance as a function of which loudspeakers presented the utterance were due primarily to their spatial location rather than to any unusual acoustical feature of one or more loudspeaker (e.g., local spectral differences among loudspeakers that can be seen in Figure 2).

Figure 9 shows the average number of times the listener chose to listen to each utterance in the identification task. The listeners required the largest number of repetitions in the one-headphone condition and when there were three simultaneous utterances. For utterances presented three at a time, only one listener ever listened fewer than three times, and she did so only twice. That is, for utterances presented three at a time, listeners almost always chose to have the utterances presented at least three times before they were satisfied that they had "heard" all of the utterances. During the localization task for the normal and KEMAR listening conditions, the listeners chose on average 3.3 fewer repetitions of the utterances than they did during the identification task. In the one-headphone listening condition, the listeners on average chose only 2.1 repetitions of the utterances. Many listeners reported that there was no spatial information in the one-headphone conditions, and they adopted a guessing strategy that required very few repetitions of the utterances.

Figure 10 shows the average results for the 7 listeners in the follow-up study in which only letters and numbers were presented and the number of simultaneous utterances was always three. Both identification and localization performance were best in the normal listening condition, poorest in the one-headphone condition, and intermediate for KEMAR listening. The trends in these data are the same as those obtained in the main experiment, which used a between-groups design.

DISCUSSION

The results show that listeners can identify on the average more than 90% of letters and numbers delivered two at a time and over 80% of the letters and numbers delivered three at a time. They are less accurate in identifying utterances that are unfamiliar, but they are clearly able to perform well above chance. That is, three simultaneously presented sound sources can be identified with a relatively high degree of accuracy. However, identifying the utterances in these tasks was not easy. This claim is supported by the data of Figure 9 indicating that listeners had to listen often to each utterance to be sure they "heard" all utterances, by the fact that 9 out of 61 could not perform the task even at a low level of accuracy (and their data were therefore not used in the study), by the verbal reports of the listeners, and by the relatively large variability in the data. The listeners reported that it was easier to localize the source of the utterance in the normal and KEMAR listening conditions than it was to identify the utterance. The listeners also needed fewer repetitions of the utterances for localization than for identification, and there was less variability in the data for localization than for identification.

There are very few published results with which to compare the data for utterances presented three at a time, but the results for utterances presented two at at time appear to be consistent with the literature (see Yost, in press). However, very few data have been collected in a divided-attention task in listening conditions intended to simulate real-world listening.

A major aim of this study was to investigate the role of binaural hearing in listeners' ability to attend to multiple sound sources. The results suggest that there is not a large advantage provided by spatial listening when there are two sound sources. This finding is somewhat consistent with the existing literature (see Yost, in press), in which for selective attention tasks involving identification of one of two sound sources there is a 2–5 dB advantage when the information is presented dichotically in some fashion. Thus, it appears likely that the other cues that differentiate one talker from another (e.g., fundamental frequency, vibrato, prosody, etc.) allow listeners to identify words accurately, so that providing binaural cues is of little additional benefit (see Shackleton & Meddis, 1992, and Yost, 1992a).

However, when the sound field becomes more complex with three concurrent sounds, spatial cues appear to play a greater role in listeners' ability to identify the three

sources. This conclusion is based on both the comparison of identification performance under the three listening conditions (Figures 3–5 and the two ANOVAs) and on some of the other analyses. For instance, identification performance improved as the distance between the loudspeakers increased (Figure 7), but only in the normal and KEMAR listening conditions. This is consistent with the notion that spatial separation is a useful cue when the utterances were presented three at a time. Clearly there were spatial cues available when the utterances were presented both two at a time and three at a time in the normal and KEMAR conditions, since the percent correct for localization was high in each of these conditions. In addition, the changes in localization performance with speaker location are consistent with the known fact that localization acuity is greater (see Figure 8) at zero azimuth and decreases as the sound source moves toward the side of the head (Mills, 1958). It is also clear that essentially no localization cues were available in the one-headphone listening condition.

Localization performance was poorer for KEMAR listening than for normal listening. There are a number of possible reasons for this low level of performance. KEMAR does not preserve the head-related transfer function (HRTF) of the individual listener. With the loudspeakers placed in the azimuthal plane in front of the listeners at an elevation equal to the height of the pinnae and with low-pass filtering of the utterances, the interaural differences of time and level would be the dominant cues for localization (see Wightman & Kistler, 1993), and one might expect KEMAR to preserve these cues quite well. Since nonindividualized HRTFs support good azimuthal localization of virtual sources once front–back reversals are resolved (Wenzel, Arruda, Kistler, & Wightman, 1993), one would expect the ability of subjects in the KEMAR listening condition to localize sources to be nearly as good as in the normal listening condition if performance was dependent solely on static spatial cues. All listeners complained that listening in the KEMAR condition was frustrating because head movements by the listener did not allow them to "face" a source. This frustration was especially high in the follow-up group of listeners who listened in all three conditions. They reported that being able to move their heads in the normal listening condition in order to "face" a potential sound source was helpful, and they felt disadvantaged by not being able to use head movements in the KEMAR listening condition. However, the brevity of the utterances meant that listeners could not actually move their heads to "face" a source *while* an utterance was being presented. They could have remembered where a source might have been and then faced in that direction when the utterance presentation was repeated. Thus, the poor localization performance in the KEMAR listening condition relative to that in the normal listening condition may be a result of the listeners' inability to use head movements appropriately. However, the data of Figures 8 and 9 suggest that the listeners did not tend to move their heads in the normal listening localization task to the extent that was allowed. First, they did not ask for the utterances to be repeated for localization as often as they did for identification. Often, the listeners estimated the location of the utterance without asking to have the utterances repeated. Second, the errors in judging the location of the loudspeakers were greater for loudspeakers off to the side. If the listeners had turned to the loudspeakers that they believed had delivered the utterance, these differences in localization errors across spatial locations should have disappeared. Perhaps identification and localization performance would have improved in the KEMAR listening condition had listeners been allowed to change the orientation of KEMAR during and between trials in a manner that simulated head movements that subjects made when participating in the normal listening condition. However, performance declined more for localization than for identification between the normal and KEMAR listening conditions, suggesting that head movements play a larger role in localization than in identification.

Finally, since the utterances were low-pass filtered, the role of high frequencies in the cocktail party effect could not be determined. The speech sounds used in this study contained high-frequency information, and these high frequencies are known to aid in localization, especially in the vertical direction. Thus, the conclusions of this study are limited to the processing of low frequencies, and, therefore, primarily to the role of interaural differences of time and level.

CONCLUSIONS

Spatial hearing does seem to play a role in divided attention tasks that characterize the cocktail party problem. This was especially true when three words were presented simultaneously. Coupling a listener's head movements to the position of KEMAR relative to the sound sources might improve the listener's performance in the KEMAR condition. Thus, in answer to Cherry's (1953) question, "On what logical basis could one design a machine ('filter') for carrying out such an operation (solving the cocktail party problem)?" spatial hearing does appear to provide one such logical basis, especially when there are more than two sound sources.

REFERENCES

BLAUERT, J. (1983). *Spatial hearing*. Cambridge, MA: MIT Press.

BREGMAN, A. S. (1990). *Auditory scene analysis*. Cambridge, MA: MIT Press.

CHERRY, C. (1953). Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America*, **25**, 975-981.

JONES, M., & YEE, W. (1993). Attending to auditory events: The role of temporal organization. In S. McAdams & E. Bigand (Eds.), *Thinking in sound* (pp. 69-106). Oxford: Oxford University Press, Clarendon Press.

MILLS, A. W. (1958). On the minimum audible angle. *Journal of the Acoustical Society of America*, **30**, 237-243.

SHACKLETON, T. M., & MEDDIS, R. (1992). The role of interaural time difference and fundamental frequency difference in identification of concurrent vowel pairs. *Journal of the Acoustical Society of America*, **91**, 3579-3581.

WENZEL, E. M., ARRUDA, M., KISTLER, D. J., & WIGHTMAN, F. L. (1993). Localization using nonindividualized head-related transfer functions. *Journal of the Acoustical Society of America*, **94**, 111-123.

WIGHTMAN, F. L., & KISTLER, D. J. (1993). Sound localization. In W. A. Yost, A. N. Popper, & R. R. Fay (Eds.), *Auditory psychophysics* (pp. 155-193). New York: Springer-Verlag.

YOST, W. A. (1992a). Auditory image perception and analysis. *Hearing Research*, **56**, 8-19.

YOST, W. A. (1992b). Auditory perception and sound source determination. *Current Directions in Psychological Sciences*, **1**(6), 15-19.

YOST, W. A. (in press). The cocktail party problem: 40 years later. In R. H. Gilkey & T. R. Anderson (Eds.), *Binaural and spatial hearing*. Mahwah, NJ: Erlbaum.

(Manuscript received March 2, 1995;
revision accepted for publication December 11, 1995.)