Perception & Psychophysics
2007, 69 (1), 79-91

Cocktail party listening in a dynamic multitalker environment

Douglas S. Brungart and Brian D. Simpson
Air Force Research Laboratory, Wright-Patterson Air Force Base, Ohio

A priori information about the location of the target talker plays a critical role in *cocktail party* listening tasks, but little is known about the influence of imperfect spatial information in situations in which the listener has some knowledge about the location of the target speech but does not know its exact location prior to hearing the stimulus. In this study, spatial uncertainty was varied by adjusting the probability that the target talker in a multitalker stimulus would change locations at the end of each trial. The results show that listeners can adapt their strategies according to the statistical properties of a dynamic acoustic environment but that this adaptation is a relatively slow process that may require dozens of trials to complete.

One of the most difficult challenges faced by human listeners is the so-called *cocktail party problem* of attending to what one talker is saying when other talkers are speaking at the same time (Cherry, 1953). At the most basic level, most of the cues that listeners use to perform this cocktail party speech segregation task are based on monaural features of the competing speech signals, including such factors as the individual voice characteristics of the talkers (i.e., *F*0, vocal tract length, speaking style, etc.), the rhythmic and temporal cues related to the speech utterances themselves (onsets, offsets, and prosodic cues), and the listener's a priori knowledge about the constraints of the language and the context of the conversation. However, to the extent that these monaural cues can be used to segregate the competing voices into different perceptual streams, the ability to selectively focus attention on a single voice in a multitalker babble can be greatly enhanced by the binaural difference cues that occur when the competing talkers are located in different directions, relative to the listener (Hirsh, 1950). These binaural cues can also be exploited in real-world communication systems that use virtual source synthesis techniques to spatially separate the apparent locations of multiple speech channels and present them to the listener via stereo headphones (Crispien & Ehrenberg, 1995; Drullman & Bronkhorst, 2000; Nelson, Bolia, Ericson, & McKinley, 1999). However, the size of the overall benefit provided by these binaural cues depends, to some degree, on the amount of a priori knowledge the listener has about the locations of the target and interfering voices.

As a general rule, listeners are able to obtain substantially greater performance benefits from the spatial separation of the competing talkers when they know the location of the target talker in advance than when they do not know its location (Ericson, Brungart, & Simpson, 2004; Koehnke, Besing, Abouchacra, & Tran, 1998; Shinn-Cunningham & Ihlefeld, 2004). To this point, however, most of the studies in which the effect that a priori information about the target talker location has on multitalker speech perception has been examined have been limited to the two most extreme cases in which either (1) the location of the target talker remains fixed for all the trials in the experiment (the listener has perfect a priori information about the location of the target) or (2) the location of the target talker is varied randomly from trial to trial (the listener has no a priori information about the location of the target talker). The first case corresponds to the classical concept of *selective attention*, where the observer is asked to focus attention on a single source of information and to ignore any distracting inputs that might originate from other objects in the perceptual field (Broadbent, 1958; Cherry, 1953). Similarly, the second case corresponds to the classic concept of *divided attention*, where the observer is asked to spread the focus of attention across two or more sources and to respond to and process information that might originate from any one of them or even, in some cases, from more than one of them at the same time (Howard-Jones & Rosen, 1993; Moray, 1959; Spieth, Curtis, & Webster, 1954; Treisman, 1964; Yost, Dye, & Sheft, 1996).

In the real world, most listening situations fall somewhere between these two extremes. For example, multitalker listening tasks in command and control environments often require listeners to monitor a number of different communication channels for information that might originate from any of the active channels in the system. Some channels might be more likely to contain useful information than others, and the listener may have some access to some a priori information about how likely the target stimulus is to occur from any one particular location (Kidd, Arbo-

D. S. Brungart, douglas.brungart@wpafb.af.mil

gast, Mason, & Gallun, 2005), but all of the channels still must be monitored at all times to ensure that high-priority information originating from an unexpected source is not overlooked. This situation contrasts somewhat with real-world cocktail party listening environments, where listeners generally know which talker to listen for and can use verbal and nonverbal cues to guide their attention to the new target talker when a break in the conversation occurs. Even in these high-context conversational situations, however, there are instances in which a listener's attention can be drawn to highly relevant information originating from an unexpected source, such as an unexpected mentioning of his or her own name by a talker somewhere else in the room (Moray, 1959). Thus, it can be argued that in real-world situations, listeners rarely, if ever, have the luxury of focusing their attention exclusively on one talker, while ignoring the other speech signals in the environment.

In this article, we present the results of experiments designed to examine how the dynamic properties of a multitalker environment influence performance in cocktail party listening tasks. In particular, the experiments were designed to evaluate multitalker speech perception as a function of the probability that a change in talker location would occur at the end of any given trial of the experiment. The results also provide insights about how listeners adapt to expected and unexpected changes in the location of a target talker that occur in multitalker listening situations. In order to provide the listeners with a means of identifying the target message without any advance knowledge about which talker would say the target message or where it would originate, we used a call-sign-based listening task (the coordinate response measure, or CRM) in which the target phrase was designated by the inclusion of a predetermined target call sign. In the next section, we will describe the design of the experiments in more detail.

METHOD

Experimental Design

In real-world listening environments, the ability to extract information from a target speech signal in a multitalker mixture can be influenced by an almost limitless number of possible factors. In this study, we focused our experimental design on three variables that were believed to be likely to influence performance in a dynamic cocktail party listening task: (1) the number of competing talkers and the spatial locations of those talkers, (2) the probability of a change in target talker location at the end of a trial, and (3) the manner in which the listening configuration changed when the target talker moved to a new location (which we refer to as *transition type*). The three variables will be described in more detail below.

Spatial configurations. Two factors that have a major influence on overall performance in cocktail party listening tasks are the number of competing talkers and the locations of those talkers, relative to the listener. A total of five different spatial configurations were tested in this experiment (see Figure 1). Two of the configurations involved two

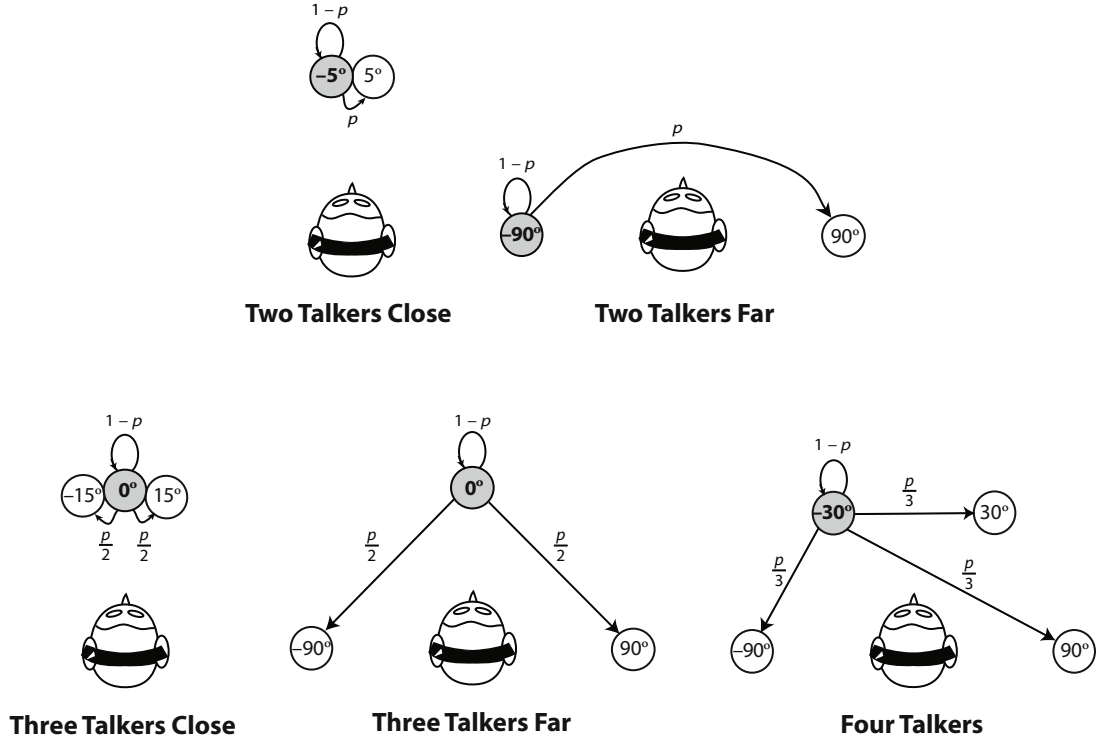

Figure 1. The five different spatial configurations that were used in the experiment: two-talker-close, two-talker-far, three-talker-close, three-talker-far, and four-talker. The arrows in the figure illustrate the transition probability model that governed changes in the location of the target talker in each condition. At the end of each trial, the target talker (indicated by the angular location in bold type) remained in the same location with a probability of $1 - p$ (where p was the transition probability for that condition) and changed to one of the other locations with a probability $p/(n - 1)$ (where n was the number of competing talker locations in that condition). See the text for further details.

competing talkers: a *two-talker-close* configuration, in which the talkers were located at ±5º in azimuth, and a *two-talker-far* configuration, in which the talkers were located at ±90º in azimuth. Another two configurations involved three competing talkers: a *three-talker-close* configuration, with talkers at −15º, 0º, and +15º in azimuth, and a *three-talker-far* configuration, with talkers at −90º, 0º, and +90º in azimuth. The final configuration was a *four-talker* configuration, with competing talkers located at ±90º and ±30º in azimuth.

Transition probability. Within each of the spatial configurations shown in Figure 1, the dynamic aspects of the multitalker listening task were varied by manipulating the probability p that the target phrase would move to a new location at the end of any given trial. This transition probability scheme is illustrated in the form of a Markov model by the arrows in each panel of Figure 1. In each case, the current location of the target talker is indicated by the shaded circle, and the arrow looping back to that location indicates the probability of the target's remaining in the same location on the next trial of the experiment ($1 - p$). The arrows drawn from the target location to the other locations indicate the probability of the target talker's moving to that new location on the next trial of the experiment. On any given transition, the target was assumed to be equally likely to move to any of the other talker locations in the configuration, and the total probability of a transition's occurring was set to the transition probability value p. Thus, the probability that the target would move to any particular location on any given trial with n competing talkers was equal to $p/(n - 1)$. In the two-talker and four-talker configurations, these p values were 0 (no transitions), 1/8, 1/4, 1/2, and 1 (transitions on every trial). In the three-talker configurations, these p values were 0, 1/6, 1/3, 2/3, and 1.

Transition type. Three different types of transitions were also examined in the experiment. The first type of transition was designed to approximate a somewhat natural real-world cocktail party situation in which talkers remain in fixed positions over the course of the conversation but the source of the most pertinent target information moves from talker to talker in an irregular pattern. Thus, a Type I transition block consisted of a set of trials on which the *talkers* remained in fixed positions but the *target phrase* moved to a different talker at a different location whenever a transition occurred.

The second type of transition was designed to simulate a situation in which the listener knows *who* the target talker is but does not know *where* that talker is located. Thus, a Type II transition block consisted of a set of trials on which the target phrase was always spoken by the same talker (who was randomly selected at the start of the block) but the location of that talker changed whenever a transition occurred. Also note that the locations of the other competing talkers were randomly changed whenever a Type II transition occurred.

The third type of transition was designed to simulate a situation that might occur in a command and control task in which a listener is required to monitor multiple channels of radio traffic but has no way of knowing which talkers will be speaking on which channels or which channel will contain the most relevant information at any given time. There is reason to believe that operators who work in such environments could greatly benefit from the use of a *spatial intercom system* that uses virtual audio display technology to cause the different speech channels of the system to appear to originate from different locations, relative to the listener's head. The Type III condition was designed to replicate the level of performance that might occur with this type of spatial intercom system. Thus, the Type III transition block consisted of a set of trials on which the different competing talkers were randomly assigned different locations on *every trial* but the location of the target phrase changed only when a transition occurred. The only restriction on the randomization of the talker locations was that the identity of the target talker had to change every time a transition occurred (i.e., the target phase could not move to the same talker at a different location; it always had to move to both a different talker *and* a different location whenever a transition occurred). Thus, on one trial in the Type III condition, a listener might hear Talker 1's voice at −90º (the target location), Talker 2's voice at 0º, and Talker 3's voice at +90º. On the next trial, all of the talkers would move locations even if no transition had occurred, so the target phrase would stay at −90º but would now be spoken by Talker 2, Talker 3 would now be heard at 0º, and Talker 1 would now be heard at +90º. In practical terms, the main contrast between the Type III condition and the other two conditions was that the talker locations and the target talker identity changed on *every trial*, whereas the talker locations and target talker identity in the other two conditions changed only when there was a change in the location of the target talker (a transition).

Figure 2 provides illustrated examples of each of these three types of transitions in the four-talker listening configuration. The leftmost panel of the figure illustrates the situation just prior to a target talker transition from Talker B at Location 2 (30º to the listener's left) to a different target talker at Location 3 (30º to the listener's right). The second panel shows the situation after a Type I (talkers fixed) transition: In these transitions, the four talkers (A–D) remain fixed at their respective locations, and the target phrase moves to a different talker with a different fixed location (C in this case). The third panel shows the situation after a Type II (target talker fixed) transition: In these transitions, the target talker (B) moves to the new target location, and the other talker locations change randomly after each transition. The final panel shows the situation after a Type III (random talker location) transition: In this case, the talker locations change randomly after every trial, with the restriction that the target phrase always changes to a different talker whenever a transition occurs.

Subjects

A total of 9 experienced paid volunteer listeners with normal hearing participated in the experiments. All but 2 were right handed. Seven of them participated in all of the experimental conditions. One participated only in the two-talker and three-talker spatial configurations, and the last one participated only in the four-talker spatial configuration.

Apparatus and Stimulus Generation

Speech materials. The experiment was based on the CRM, a call-sign-based intelligibility test that has been used in a number of previous listening experiments involving two or more simultaneous talkers (Brungart, 2001a, 2001b). The CRM phrases were taken from the publicly available CRM speech corpus for multitalker communications research (Bolia, Nelson, Ericson, & Simpson, 2000), which contains phrases of the form "Ready (call sign) go to (color) (number) now," spoken by four male and four female talkers with all possible combinations of eight call signs ("arrow," "baron," "Charlie," "eagle," "hopper," "laker," "Ringo," and "tiger"), four colors ("blue," "green," "red," and "white"), and eight numbers (1–8). Thus, an example phrase from the CRM corpus would be "Ready baron go to blue six now."

In this experiment, the stimulus presented to the listener always consisted of a mixture of two, three, or four phrases that were randomly selected from phrases spoken by male talkers in the CRM corpus: a target phrase, which was randomly selected from the phrases containing the call sign *baron*, and one, two, or three competing phrases, which were randomly selected from the phrases with a different call sign, color, and number than the target phrase. Note that all of the target and competing phrases were equalized to have the same overall RMS power, which was roughly equivalent to 75 dB SPL in the headphone-presented stimuli used in this experiment. In order to balance for differences in the intelligibility of the different talkers, only two of the male talkers in the corpus were used in the two-talker condition, and only three of the male talkers were used in the three-talker condition. Note that all of the phrases in the corpus were synchronized so that the onset of the word "ready" occurred at the same time in all of the competing sentences. No effort was made to synchronize the onset of the color and number keywords within the individual competing phrases.

Spatial processing. The spatial configurations used in the experiment were implemented over headphones with head-related transfer functions (HRTFs; Wightman & Kistler, 1989b) designed to repro-

duce the spatial audio cues that would normally occur in anechoic space for sound sources at each of the talker locations shown in Figure 1. Virtual acoustic simulations of this type have been shown to produce multitalker listening performance that is comparable to the free field without the need for an anechoic chamber or a complicated apparatus for generating the spatially separated speech signals (Best, 2004; Brungart, Ericson, & Simpson, 2002; Brungart & Simpson, 2002; Brungart & Simpson, 2003; Crispien & Ehrenberg, 1995; Drullman & Bronkhorst, 2000; Hawley, Litovsky, & Colburn, 1999; Hawley, Litovsky, & Culling, 2004; Shinn-Cunningham, Schickler, Kopco, & Litovsky, 2001). In this experiment, the HRTF processing was accomplished with a digital signal processor (TDT RP2) that convolved each competing speech signal with two different HRTFs, one for the left ear and one for the right ear. This HRTF processing was done at a 50-KHz sampling rate with 128-point HRTF filters (provided by TDT) that were measured in the ears of a live female listener (Subject S.D.L.) at the University of Wisconsin by Wightman and Kistler (1989a).[1] The HRTF-processed signals from each competing talker were then digitally summed and presented to the left and right ears of the listener over stereo headphones (Sennheiser HD-520). Note that no headtracking was used in this study, so the apparent locations of the virtual sources remained fixed in location, relative to the listener's head, independently of any head movements made during the experiment.

Experimental Procedure

The data collection was divided into blocks of trials, with each trial consisting of a single presentation of a set of phrases from the CRM corpus (lasting roughly 2 sec) and each block consisting of either 40 trials (in the two-talker conditions) or 60 trials (in the three- and four-talker conditions). The data were collected with the listeners seated at a control computer located in one of three quiet sound-treated rooms. The listeners were instructed to listen for the target phrase containing the call sign "baron" and to use the mouse to select the color–number combination contained in that target phrase from an array of colored digits displayed on the CRT of the control computer. Once that selection had been made, the experimental program determined whether a transition should occur prior to the next stimulus presentation, reassigned the locations of the target and competing talkers accordingly, and initiated the next trial in the sequence. The listeners were not provided with any information about the spatial configuration, transition type, or transition probability prior to the start of each block of trials, and no correct answer feedback was given at the end of each trial of the experiment. Note that the listeners were given as long as they wished to make their responses and that the stimulus in the next trial always started approximately 1 sec after the mouse click that completed the response in the previous trial. Although total trial time was not measured in this experiment, previous results in similar experiments have shown that each trial in the self-paced CRM task takes an average of roughly 5.24 sec to complete, with a standard deviation of 0.8 sec.

Performance in the experiment was expected to vary with both the identity and the location of the target talker, so care was taken to make sure that these two factors were balanced across the different conditions tested for each listener in the experiment. In the Type I and Type III transition conditions, in which the target talker identity changed randomly throughout each block, one block of trials was collected with each possible initial target location, with a random distribution of initial talker locations in each block (see the top panel of Figure 3). This resulted in two blocks of trials in each two-talker

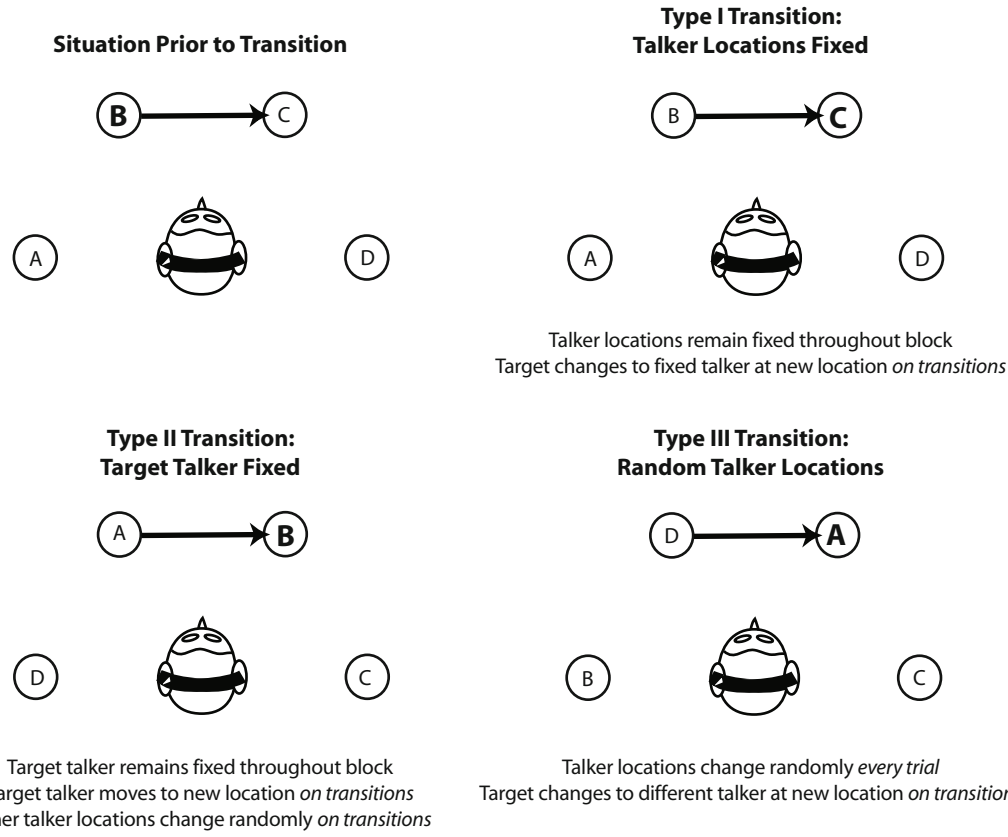

Figure 2. Examples of the three types of target transitions tested in the four-talker configuration of the experiment. The four letters A, B, C, and D represent the four different talkers, with the bold text indicating the current location of the target talker. See the text for further details.

condition, three in each three-talker configuration, and four in each four-talker configuration. In the Type II transition conditions, the target location changed randomly within the block, but the target talker identity did not. Thus, in those conditions, one block of trials was collected with each possible target talker identity, with a random distribution of target talker locations (middle panel of Figure 3). Again, this resulted in two blocks of trials in each two-talker configuration, three in each three-talker configuration, and four in each four-talker configuration. There was also one special case that occurred when the transition probability was equal to zero in the Type I and Type II blocks. In these conditions, neither the target talker nor the target identity ever changed over the course of a block of trials. Thus, it was necessary to test all possible combinations of initial target talker and initial target location in these conditions (bottom panel of Figure 3). This resulted in 4 blocks of trials in the two-talker conditions, 9 blocks in the three-talker conditions, and 16 blocks in the four-talker conditions. However, because there was effectively no difference between the Type I and the Type II transitions in the zero transition probability conditions, the same blocks of data were used to analyze both conditions. Across all the spatial configurations and transition probabilities, this set of initial conditions resulted in a total of 60 blocks of trials in the two-talker conditions, 96 blocks of trials in the three-talker conditions, and 68 blocks of trials in the four-talker conditions for each subject. The data were collected in 20- to 30-min sessions over a period of approximately 3 months. Each of the 8 subjects participated in the three-talker condition first, then in the two-talker condition, and finally in the four-talker condition. This resulted in a grand total of 12,240 trials for each of the 9 listeners who participated in the experiment.

RESULTS AND DISCUSSION

ANOVAs were conducted separately on the data from the two-talker, three-talker, and four-talker spatial configurations of the experiment. In the two-talker and three-talker cases, three-factor ANOVAs were conducted, with the factors of spatial separation (closely spaced or widely spaced talkers), transition type (I, II, or III), and transition probability (the five values of p used in that configuration). In the four-talker case, only one spatial configuration was used, so a two-factor ANOVA was conducted on the factors of transition type (I, II, or III) and transition probability. In each case, the percentages of correct color *and* number identifications in each condition were calculated separately for each listener, normalized with an arcsine transform, and subjected to a within-subjects repeated measures ANOVA. The results of these analyses show that at the $\alpha = .05$ level, the main effect of spatial configuration was significant in both the two-talker and the three-talker conditions [$F(1,7) = 8.56$ and 69.0, respectively]; the main effect of transition probability was significant only in the three-talker and four-talker conditions [$F(4,28) = 36.91$ and 17.92, respectively]; and the main effect of transition type was significant in the two-talker, three-talker, and four-talker conditions [$F(2,14) =$ 18.52, 24.21, and 9.77, respectively]. There was a marginally significant interaction between spatial configuration and transition probability in the two-talker condition [$F(4,28) = 2.78, p = .046$], but no other interactions were significant. Thus, we will disregard the interactions and will focus our discussion on the three main effects, each of which will be discussed in detail below.

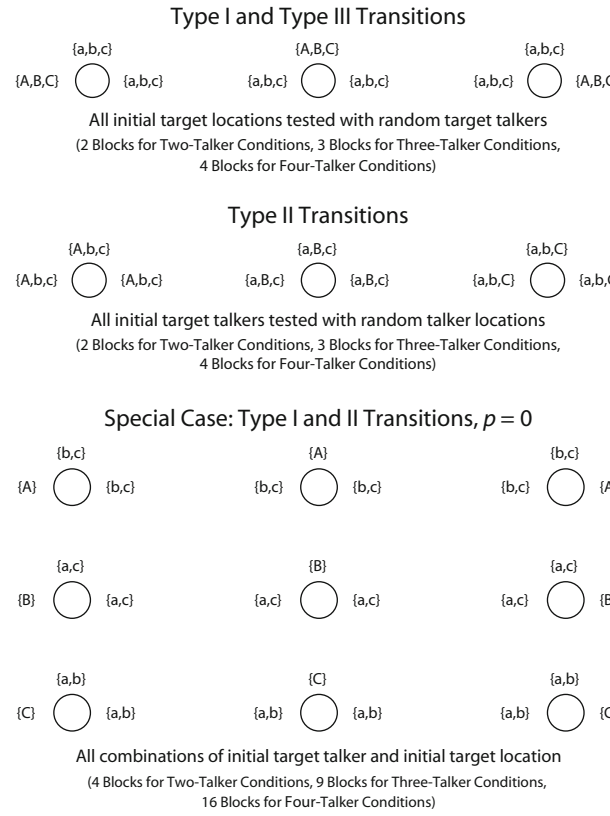

Figure 3. Pictorial representation of the initial conditions used in each block of trials in the three-talker configurations of the experiment. The letters within the braces represent the possible talkers (a, b, or c) that might occur at that location at the start of a block of trials. The capitalized letters represent the possible locations and identities of the initial target talker.

Spatial Configuration

Figure 4 shows the overall percentages of correct color and number identifications in the CRM task as a function of the location of the target talker for each of the five spatial configurations tested in the experiment. From these results, we can make the following observations.

Overall performance decreased as the number of competing talkers increased. Averaged across all the conditions, performance dropped from about 92% correct responses in the two-talker conditions to 72% correct responses in the three-talker conditions and to 62% correct responses in the four-talker conditions. A post hoc analysis (Fisher LSD) of the individual scores of the 7 subjects who participated in all three conditions revealed that these conditions were significantly different from one another at the $p < .05$ level.

Overall performance was better in the widely spaced spatial configurations than in the closely spaced spatial configurations. Averaged across all the target talker locations, overall performance in the two-talker-far configuration was about 3 percentage points better than that in the two-talker-close configuration. Performance in the three-talker-far configuration was about 11 percentage

points better than that in the three-talker-close configuration. Thus, not surprisingly, it seems that listeners do better overall when listening to competing talkers who are spaced far apart than when listening to competing talkers who are close together.

Overall performance was better when the target phrase originated from the leftmost or rightmost talker locations than when it originated from the central target talker locations. On average, performance in the three-talker and four-talker conditions was roughly 20 percentage points better when the target originated from the leftmost or rightmost target talker positions than when it originated from the more medial positions tested. A post hoc analysis (Fisher LSD) confirmed that this difference was significant at the $p < .05$ level. This performance difference is likely related to the fact that target speech signals presented from these extreme spatial locations are more intense than any of the signals from the other, competing talkers at the ear closest to the target talker (thus affording a very favorable signal-to-noise ratio at the listener's "better ear" [Ericson et al., 2004; Zurek, 1993], as well as a larger interaural time delay), whereas those presented at more medial locations are masked by those from the more intense, interfering talkers at both of the listener's ears.

Overall performance was substantially better for talker locations in the listener's right hemifield than for locations in the listener's left hemifield. On average, performance was almost 25 percentage points better for target talker locations on the listener's right side than it was for target talker locations on the listener's left side. In part, this difference is probably related to the general tendency for the speech-processing centers of the brain to be specialized in the left hemisphere. Previous research has demonstrated left-hemispheric specialization for the auditory-processing centers that govern speech production (Broca, 1861) and speech perception (Kimura, 1961), and a number of previous cocktail party listening experiments have shown that listeners are substantially better at attending to speech presented in the right hemifield than to that presented in the left hemifield (Bolia, Nelson, & Morley, 2001; Brungart & Simpson, 2003; Ericson et al., 2004). The result may also reflect a general bias on the part of the listeners to selectively focus their attention on talkers located to their right.

Transition Type

Figure 5 shows how overall performance varied with the target transition type. The bars in the figure represent the results only from those blocks of trials on which transitions actually did occur (i.e., where the transition probability was greater than 0). In order to test the significance of these results, a two-factor ANOVA with the factors of number of talkers and transition type was conducted on the individual subject scores from the 7 subjects who participated in all the conditions of the experiment. The results of this ANOVA revealed a significant effect of transition type [$F(2,12) = 32.149$, $p = .00002$], and a post hoc analysis of the results (Fisher LSD) indicated that all three transition types were significantly different from one another at the $p < .05$ level.

When the results for the different transition types are compared, it is apparent that the listeners performed best in the Type II transition blocks, in which the same target talker was used throughout the block of trials. This result is not surprising, because the listeners in Type II blocks should have been able to learn the characteristics of the

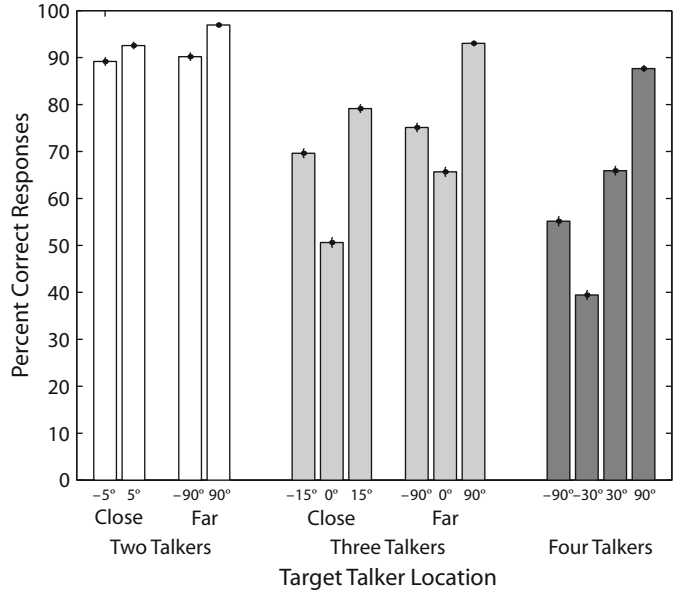

Figure 4. Correct color and number identifications in the coordinate response measure task as a function of the location of the target talker in each spatial configuration tested in the experiment. The error bars show the 95% confidence interval for each data point.

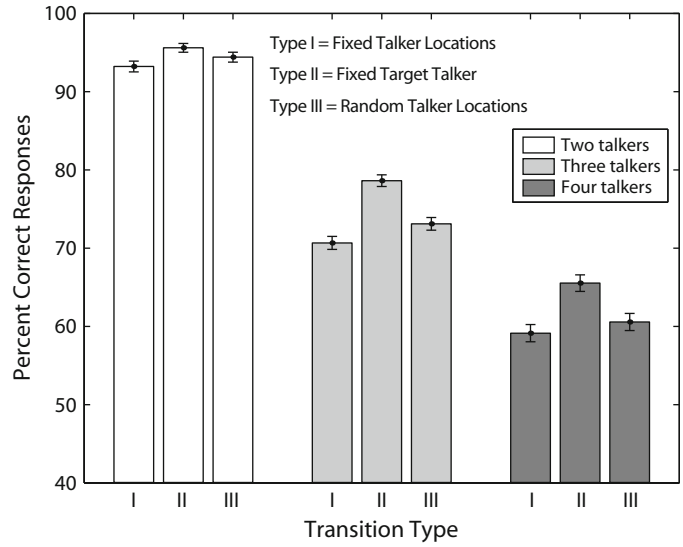

Figure 5. Correct color and number identifications for each type of transition shown in Figure 2. The data are shown only for those conditions in which transitions occurred (i.e., where the transition probability was greater than 0). The error bars show the 95% confidence interval for each data point.

target talker and to use this information as a help in following the target phrase across transitions. Similar performance advantages have been reported in other experiments in which conditions in which listeners had a priori information about the identity of the target talker have been compared with conditions in which no information about the target talker's voice was provided to the listener (Ericson et al., 2004).

A somewhat more unexpected result is that the listeners consistently performed better in the Type III transition blocks (in which the talker locations changed randomly on every trial) than in the Type I transition blocks (in which the talkers remained at fixed locations throughout the block). This may suggest that the listeners were learning to associate the target phrase with the talker assigned to the current target location and that this incorrect association made them less able to shift their attention when the target phrase unexpectedly moved to a new target talker at a different location. Since the target talker changed on every trial in the Type III blocks, the listeners never learned to make an association between the target phrase and any particular talker and, thus, were better able to respond when transitions occurred.

Transition Probability

Figure 6 shows overall performance as a function of the transition probability p in each condition of the experiment. As is suggested by the results of the ANOVA, transition probability had no effect on performance in the two-talker condition of the experiment. The reasons for this are not completely clear. In part, it may be accounted for by a ceiling effect, but it may also be related to the fact that the two-talker CRM task can be performed successfully with a *process of elimination* strategy in which the listener attends to the call sign spoken by one of the talkers and chooses either to maintain attention to the same talker if the call sign is "baron" or to switch attention to the other talker midsentence if the call sign is not "baron." The adoption of this strategy could explain why the listeners in the two-talker configurations were relatively unimpaired in their ability to perform well in the CRM task, even in blocks of trials with high transition probability values.

A process of elimination strategy would not work as well in situations with more than two competing talkers, because the listeners would have no way of knowing which phrase to attend to once they realized that they had been listening to a talker who had spoken the wrong call sign. This may explain why transition probability had such a large impact on performance in the three-talker and four-talker configurations of the experiment. In each case, the percentage of correct responses decreased by roughly 20 percentage points as the transition probability increased from 0 to 1. Roughly half of the total decrease in performance occurred when the transition probability increased from 0 to the smallest positive value tested (from 0 to 1/6 in the three-talker case and from 0 to 1/8 in the four-talker case). This suggests that cocktail party listening is profoundly impaired when the situation changes from one in which there is no uncertainty about the location of the target talker to one in which some uncertainty exists. Once a small amount of uncertainty was added to the situation, additional increases in the transition probability produced more gradual decreases in overall performance.

Adaptation

Up to this point, we have restricted our analyses of the effects of transition probability on multitalker listening to an overall analysis of performance across blocks col-

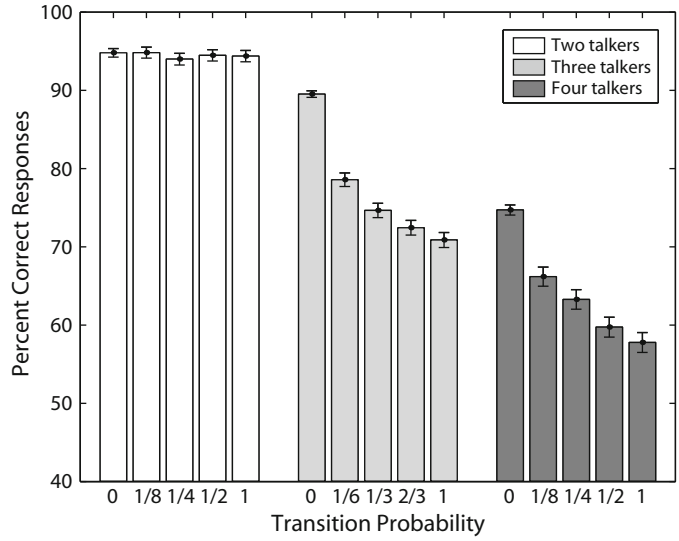

Figure 6. Correct color and number identifications for each type of transition shown in Figure 2. The error bars show the 95% confidence interval for each data point.

lected with different underlying *p* values. However, additional insights can be obtained by examining performance with a finer grained analysis of the effects that transitions had on performance in the CRM task over the course of a block of trials. For example, it is possible to examine how listeners adapted to sudden changes in the position of the target phrase by collapsing the data across all the different transition probabilities tested and plotting overall performance as a function of the number of consecutive preceding trials with the same target talker location as that in the current trial (as illustrated by A in Figure 7). Figure 8 shows the results of this analysis. The individual data points in the figure were generated by sorting all of the individual trials in each talker configuration by the number of preceding trials with the same target location and dividing those trials into bins containing a minimum of 2,000 trials each. The mean performance levels for each bin were then plotted as a function of the mean number of preceding fixed-target trials across all the trials included in that bin. Thus, the data points with an abscissa of 0 represent the mean performance level that was obtained in those trials immediately following a transition (the first trials presented at a new target location after a transition). The data points with an abscissa of 1 represent the second trials after transitions, those with an abscissa of 2 represent the third trials, and so on. A number of observations can be made from these results.

Overall performance was worst in trials that immediately followed a transition. In all three talker configurations tested, trials that occurred immediately after a transition produced significantly fewer correct responses than did those that occurred two or more trials after a transition.

Overall performance improved rapidly for the first three to four trials after a transition. In all the configurations tested, mean performance increased quickly for the first few trials following a change in target location. This increase was greatest for the three-talker configuration (an increase of approximately 15 percentage points), slightly smaller in the four-talker configuration (10 percentage point increase), and smallest in the two-talker configuration (roughly 3 percentage points). This relatively rapid learning curve likely represents the number of trials that the listeners required to realize that the target talker had moved from the previous location and to determine the new location of the target talker.

In the three- and four-talker configurations, performance continued to improve systematically until roughly 30 consecutive trials had occurred with the same target talker location. In the conditions with more than two talkers, a slow learning process seemed to continue for dozens of trials after the listeners had already adapted to the change in the location of the talker.[2] This slow adaptation process suggests the possibility that the listeners were gradually shifting their listening strategy from one that was based on the assumption that the target talker might move to a different location at any time (and thus required them to divide their attention across multiple source locations) to one that was based on the assumption that the talker would never move from the current location (thus allowing them to selectively focus their attention only on this location). More evidence for this kind of gradual shift in strategy will be provided in the next two sections.

Negative Priming

The data shown in Figure 8 demonstrate the effects that prior trials with the *same* target talker location had on performance in the dynamic CRM listening task. However, it is also helpful to look at the influence that prior trials with a *different* target talker location had on a listener's ability to react to a change in the location of the target talker. Figure 9 shows the effect that exposure to consecutive

preceding trials with one target talker location had on performance in the first trial immediately following a transition to a different target location. These data points were generated by sorting all of the individual trials in each talker configuration by the number of preceding trials with the same target location and dividing those trials into bins containing a minimum of 500 trials each. The mean performance levels for each bin were then plotted as a function of the mean number of preceding fixed-target trials across all the trials included in that bin. In other words, the data points in Figure 9 show performance on trials immediately following a transition as a function of the number of consecutive trials with the same target location as the one that occurred prior to that transition (as shown by B in Figure 7). In all the configurations tested, there was a strong negative correlation between the number of consecutive trials with the same target talker location prior to the transition and the performance level achieved in the trial immediately following the transition. This is a classic *negative priming effect*: The consecutive trials with the same target location gradually built up an expectation that the target phrase would remain in the same location in subsequent trials, and this expectation made the listeners less able to react quickly when there was an unexpected change in the location of the target talker (Tipper, Brehaut, & Driver, 1990).

In the two-talker and three-talker configurations, the effects of negative priming increased monotonically with the number of static trials prior to the transition. However, in the four-talker configuration, the effects of negative priming plateaued after approximately four consecutive trials. No further decrease in performance was observed in transition trials preceded by sequences of more than four consecutive stimulus presentations with the same target location.

Strategic Adaptation

Figure 10 shows the mean level of performance in trials immediately following a transition as a function of the transition probability used in each block of trials. In the two-talker and three-talker configurations, there was a clear positive correlation between the level of performance achieved in trials immediately following a transition and the transition probability used in each block of trials. A two-factor ANOVA conducted on the scores of the individual subjects for the factors of number of talkers and transition probability revealed that this main effect was significant at the $p < .001$ level [$F(13,18) = 11.007$]. This result stands in stark contrast to the systematic decrease in *overall* performance with increasing transition probability shown in Figure 6. This result is important because it implies that the listeners in the experiment were able to adapt their listening strategies in order to maximize overall performance in each of the dynamic listening environments tested in the experiment. In listening environments in which transitions were relatively infrequent (e.g., $p = 1/8$), the listeners adopted a listening strategy that performed well when the target talker stayed

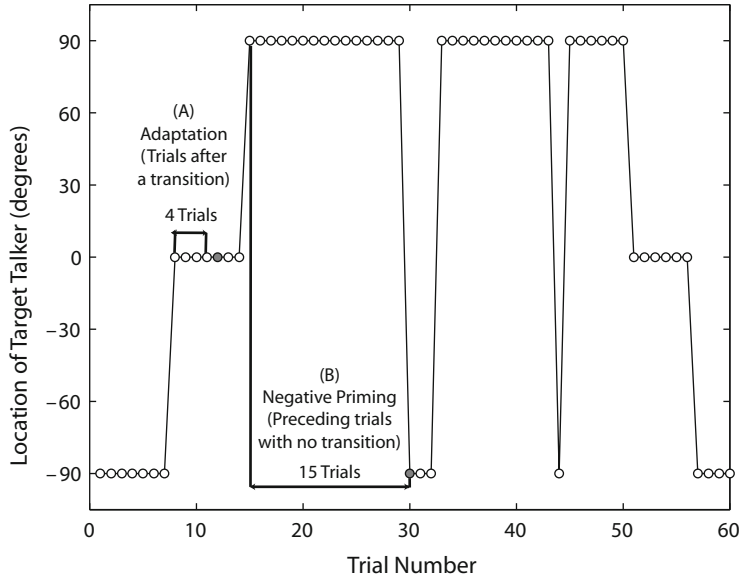

Figure 7. Talker location as a function of trial number for a block of trials in the three-talker-far configuration with a transition probability of 1/6. Point A illustrates how the data were categorized for the analysis of adaptation shown in Figure 8, which plots performance as a function of the number of preceding trials with the same target talker location. Point B illustrates how the data were categorized for the analysis of negative priming shown in Figure 9, which plots performance for trials immediately following a transition as a function of the number of preceding trials with the same target location. In each case, the shaded circle represents the analyzed trial, and the arrow indicates how the number of preceding trials was calculated for type of analysis.

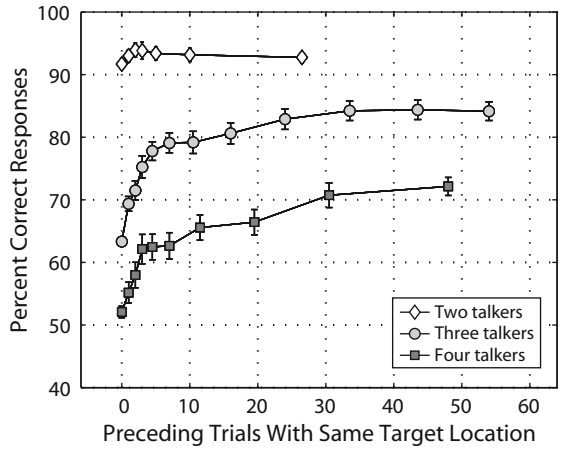

Figure 8. Correct color and number identifications as a function of the number of consecutive preceding trials with the target phrase in the same location as the current trial. The error bars show the 95% confidence intervals for each data point.

in the same location for more than one trial but performed relatively poorly on trials in which the target talker happened to unexpectedly move to a new location. In listening environments with very frequent target talker transitions (e.g., $p = 1$), the listeners adopted a strategy that was optimized to accommodate changes in the location of the target talker. Although the details of these listening strategies are difficult to determine from these results, the results from the three-talker configurations, in particular, suggest that listeners are quite good at analyzing the dynamic properties of the current listening environment and using this information to choose a strategy that maximizes overall performance.

One intriguing aspect of the results shown in Figure 10 is that the listeners did not seem to be able to adapt their listening strategies to account for different dynamic listening environments with four simultaneous talkers. Performance on trials that immediately followed a transition in the four-talker configuration was virtually identical at all the transition probabilities tested. This may suggest that there is some fundamental limitation that prevents listeners from adopting strategies that efficiently cope with frequent changes in target talker locations in more complex environments with more than three simultaneous talkers.

Learning

In the previous section, we noted that the listeners in the two-talker and three-talker configurations seemed to be able to adopt listening strategies that were optimized for different dynamic listening environments that were encountered in each block of trials. Because the listeners were not provided with any a priori information about these dynamic listening environments, this result implies that the subjects were able to learn the properties of the environment over the course of each block. Figure 11 shows mean performance as a function of trial position within a block of trials (in effect, the learning curve) for each configuration of the experiment. The data are shown separately for the 1st, 2nd, 3rd, and 4th trials of each block, and the data for trials that occurred 5 or more trials into each block have been pooled together into bins with a minimum of 5,000 trials in each bin.

In the two-talker configurations, there was no indication that any significant amount of learning occurred over the course of each block of trials, which may have been the result of a ceiling effect in the data. In the three- and four-talker configurations, the results indicate that the listeners improved quickly for the first 8–10 trials in each block but that they received no further benefit from obtaining more than 10 trials of listening experience with the same dynamic listening environment. This suggests that listeners require only a brief exposure to successfully tailor their listening strategies to the dynamic properties of an unfamiliar listening environment.

Spatial Filtering

In general, increasing the spatial separation between the competing talkers tends to improve performance in cocktail party listening tasks. However, there is some evidence to suggest that increasing the spatial separation between the talkers might, in some cases, actually *decrease* performance on trials that immediately follow a change in the location of the target talker. Rhodes (1987), for example, has shown that the time required to shift auditory attention to a new spatial location is a linearly increasing function of the angular displacement between the new and the old source locations. On the basis of Rhodes's result, one might expect target talker transitions in the present experiment to cause a greater drop in performance when they involve large changes in the angular location of the target talker than when they involve smaller changes. However, there is little evidence in the data to support this conclusion. Figure 12 shows mean performance on trials that immediately followed a transition as a function of the size of

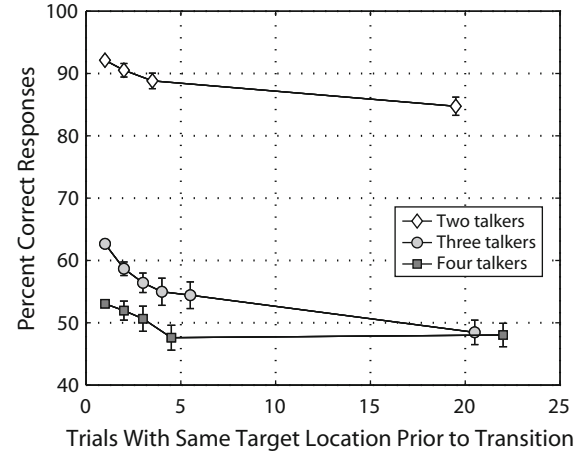

Figure 9. Correct color and number identifications in trials immediately following a transition as a function of the number of consecutive preceding trials with the target phrase in the same location as that in the trial immediately preceding the transition. The error bars show the 95% confidence interval for each data point.

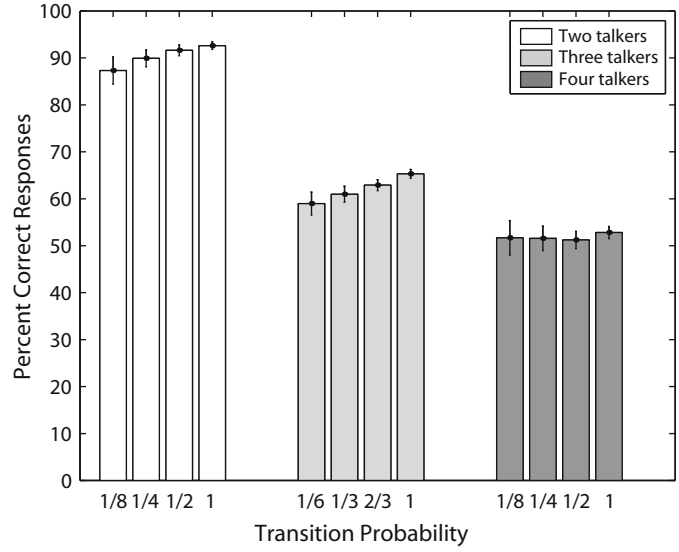

Figure 10. Percent correct color and number identifications in trials immediately following a transition as a function of transition probability. The error bars show the 95% confidence interval for each data point.

the angular displacement from the previous target location to the new target location. In order to balance for the effect of talker location (e.g., Figure 4), the data were averaged only across those trials on which the target was located at the leftmost or rightmost talker locations in the current spatial configuration. As an example, the numbered arrows on the diagram on the right side of the figure show how the angular displacement would be determined in the four-talker configuration in transitions in which the target talker moved to the leftmost source location (A) from one of the three other possible source locations (B, C, or D).

The results of this analysis show no reliable correlation between the size of the angular displacement that occurred during a target talker transition and the impact that the transition had on identification performance in the next trial following that transition. Although there is no way to know for certain why angular displacement did not have an effect on performance, it is likely that the relatively slow time course of the CRM task, which consists of phrases that are roughly 2 sec in duration and total trial times on the order of 5.25 sec, may allow listeners more than ample time to successfully shift their attention to a new spatial location over the course of a single stimulus presentation.

SUMMARY AND CONCLUSIONS

Although dynamic effects have been largely ignored in previous studies of cocktail party listening, the results of the experiment presented here demonstrate that the dynamic properties of the listening environment can significantly impact a listener's ability to extract information from a target speech signal masked by one or more simultaneous interfering voices. Not surprisingly, the results indicate that listeners generally perform better in listening environments in which the target talker tends to remain fixed in one location than in environments where the target talker changes locations frequently. This improvement in performance for infrequently moving target talkers can be attributed primarily to the listener's ability to learn the location of the target talker and selectively focus attention on that location. However, the number of trials required to complete this process was surprisingly long: In the three- and four-talker configurations of this experiment, performance improved systematically until the listener had heard as many as 30 consecutive trials with the same target talker location. Also, although the selective focusing of attention on a single talker location clearly improved performance overall, this improvement came at a cost: Focusing on one target location generally degraded the listener's ability to correctly respond

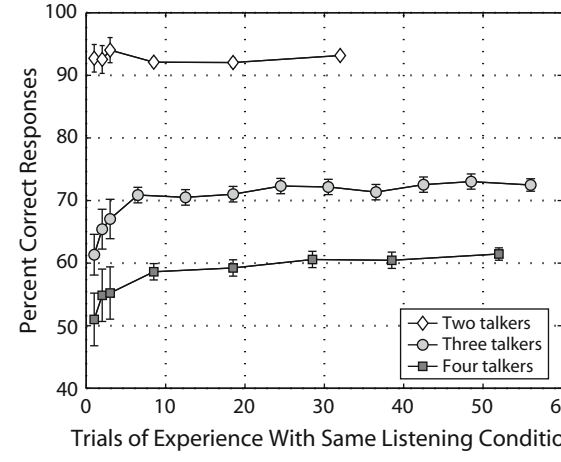

Figure 11. Overall performance as a function of the trial position within a block of 40 or 60 consecutive trials. The error bars show the 95% confidence interval around each data point.

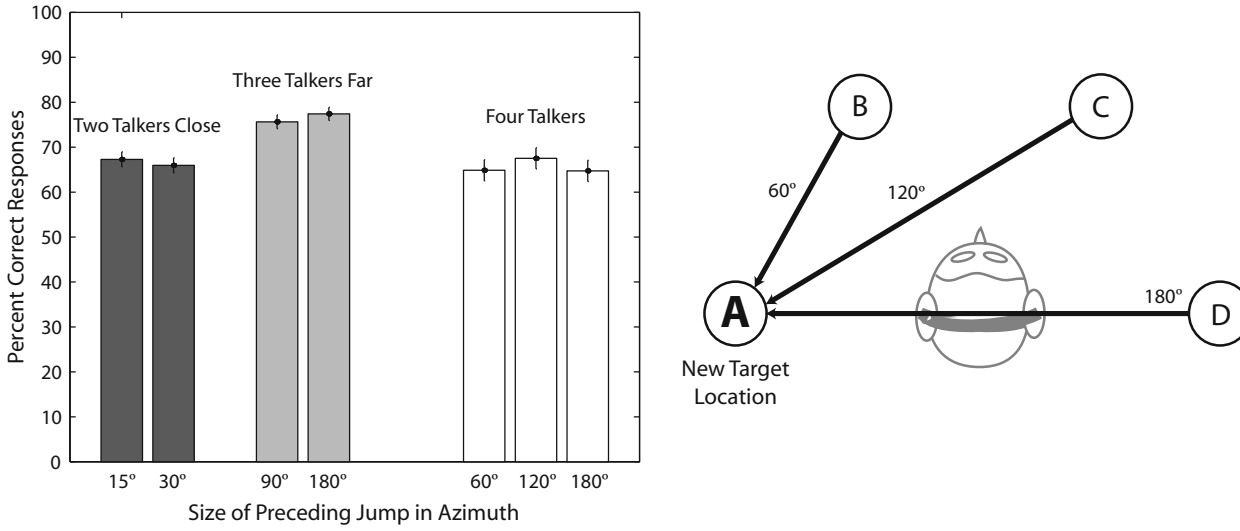

Figure 12. Mean performance in trials immediately following a transition as a function of size of the angular displacement from the previous target talker location to the new target talker location. See the text for details.

when the target speech signal unexpectedly moved to a different target location. The listeners did, however, appear to have some control over this adaptation process: In blocks of trials in which the target talker moved frequently, the listeners seemed to be able to divide their attention across multiple talker locations in order to respond appropriately to changes in target location. They also seemed to be able to learn the optimal strategy for the current listening environment relatively quickly; their performance generally plateaued after only 10 trials of exposure to a given dynamic environment. Overall, these results suggest that the strategies listeners use in cocktail party situations are relatively fluid. Listeners constantly assess the dynamic properties of their surrounding environments and use this information to shape their listening strategies in order to optimize their overall performance. Future models that attempt to characterize the role of attention in the cocktail party problem will need to account for these dynamics.

AUTHOR NOTE

Portions of this research were supported by AFOSR Grant 01-HE-01-COR. Correspondence concerning this article should be addressed to D. S. Brungart, AFRL/HECB, 2610 Seventh St., Wright-Patterson Air Force Base, OH 45433-7901 (e-mail: douglas.brungart@wpafb.af.mil).

REFERENCES

BEST, V. (2004). *Spatial hearing with simultaneous sound sources: A psychophysical investigation*. Unpublished doctoral thesis, University of Sydney.

BOLIA, R. S., NELSON, W. T., ERICSON, M. A., & SIMPSON, B. D. (2000). A speech corpus for multitalker communications research. *Journal of the Acoustical Society of America*, **107**, 1065-1066.

BOLIA, R. S., NELSON, W. T., & MORLEY, R. M. (2001). Asymmetric performance in the cocktail party effect: Implications for the design of spatial audio displays. *Human Factors*, **43**, 208-216.

BROADBENT, D. E. (1958). *Perception and communication*. New York: Pergamon.

BROCA, P. (1861). Nouvelle observation d'aphémie produite par une lésion de la moitié postérieure des deuxième et troisième circonvolutions frontales gauches [New observations of aphemia produced by a lesion of the posterior half of the second and third left frontal circumvolutions]. *Bulletins de la Société Anatomique de Paris*, **36**, 398-407.

BRUNGART, D. S. (2001a). Evaluation of speech intelligibility with the coordinate response measure. *Journal of the Acoustical Society of America*, **109**, 2276-2279.

BRUNGART, D. S. (2001b). Informational and energetic masking effects in the perception of two simultaneous talkers. *Journal of the Acoustical Society of America*, **109**, 1101-1109.

BRUNGART, D. S., ERICSON, M. A., & SIMPSON, B. D. (2002). Design considerations for improving the effectiveness of multitalker speech displays. In *Proceedings of the International Conference on Auditory Display (ICAD 2002)* (pp. 169-174). Kyoto.

BRUNGART, D. S., & SIMPSON, B. D. (2002). The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. *Journal of the Acoustical Society of America*, **112**, 664-676.

BRUNGART, D. S., & SIMPSON, B. D. (2003). Optimizing the spatial configuration of a seven-talker speech display. In *Proceedings of the International Conference on Auditory Display* (pp. 188-191). Boston.

CHERRY, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, **25**, 975-979.

CRISPIEN, K., & EHRENBERG, T. (1995). Evaluation of the "cocktail party effect" for multiple speech stimuli within a spatial audio display. *Journal of the Audio Engineering Society*, **43**, 932-940.

DRULLMAN, R., & BRONKHORST, A. W. (2000). Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *Journal of the Acoustical Society of America*, **107**, 2224-2235.

ERICSON, M. A., BRUNGART, D. S., & SIMPSON, B. D. (2004). Factors that influence intelligibility in multitalker speech displays. *Journal of Aviation Psychology*, **14**, 313-334.

HAWLEY, M. L., LITOVSKY, R. Y., & COLBURN, H. S. (1999). Speech intelligibility and localization in a multi-source environment. *Journal of the Acoustical Society of America*, **105**, 3436-3448.

HAWLEY, M. L., LITOVSKY, R. Y., & CULLING, J. F. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *Journal of the Acoustical Society of America*, **115**, 833-843.

HIRSH, I. J. (1950). Relation between localization and intelligibility. *Journal of the Acoustical Society of America*, **22**, 196-200.

HOWARD-JONES, P. A., & ROSEN, S. (1993). Uncomodulated glimpsing

in "checkerboard" noise. *Journal of the Acoustical Society of America*, **93**, 2915-2922.

KIDD, G., JR., ARBOGAST, T. L., MASON, C. R., & GALLUN, F. J. (2005). The advantage of knowing where to listen. *Journal of the Acoustical Society of America*, **118**, 3804-3815.

KIMURA, D. (1961). Some effects of temporal-lobe damage on auditory perception. *Canadian Journal of Psychology*, **15**, 156-165.

KOEHNKE, J., BESING, J. M., ABOUCHACRA, K. S., & TRAN, T. V. (1998). Speech recognition for known and unknown target message locations. In *Proceedings of the 1998 Midwinter Meeting of the Association for Research in Otolaryngology*. St. Petersburg, FL.

MORAY, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, **11**, 56-60.

NELSON, W. T., BOLIA, R. S., ERICSON, M. A., & MCKINLEY, R. L. (1999). Spatial audio displays for speech communication: A comparison of free-field and virtual sources. In *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting* (pp. 1202-1205). Santa Monica, CA: Human Factors & Ergonomics Society.

RHODES, G. (1987). Auditory attention and the representation of spatial information. *Perception & Psychophysics*, **42**, 1-14.

SHINN-CUNNINGHAM, B. G., & IHLEFELD, A. (2004). Selective and divided attention: Extracting information from simultaneous sound sources. In *Proceedings of the International Conference on Auditory Display*. Sydney.

SHINN-CUNNINGHAM, B. G., SCHICKLER, J., KOPCO, N., & LITOVSKY, R. (2001). Spatial unmasking of nearby speech sources in a simulated anechoic environment. *Journal of the Acoustical Society of America*, **110**, 1118-1129.

SPIETH, W., CURTIS, J. F., & WEBSTER, J. C. (1954). Responding to one of two simultaneous messages. *Journal of the Acoustical Society of America*, **26**, 391-396.

TIPPER, S. P., BREHAUT, J. C., & DRIVER, J. (1990). Selection of moving and static objects for the control of spatially directed action. *Journal of Experimental Psychology: Human Perception & Performance*, **16**, 492-504.

TREISMAN, A. M. (1964). Verbal cues, language, and meaning in selective attention. *American Journal of Psychology*, **77**, 206-219.

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

WIGHTMAN, F. L., & KISTLER, D. J. (1989a). Headphone simulation of free-field listening: I. Stimulus synthesis. *Journal of the Acoustical Society of America*, **85**, 858-867.

WIGHTMAN, F. L., & KISTLER, D. J. (1989b). Headphone simulation of free-field listening: II. Psychophysical validation. *Journal of the Acoustical Society of America*, **85**, 868-878.

YOST, W. A., DYE, R. H., JR., & SHEFT, S. (1996). A simulated "cocktail party" with up to three sound sources. *Perception & Psychophysics*, **58**, 1026-1036.

ZUREK, P. M. (1993). Binaural advantages and directional effects in speech intelligibility. In G. A. Studebaker & I. Hochberg (Eds.), *Acoustical factors affecting hearing aid performance* (2nd ed., pp. 255-276). Boston: Allyn & Bacon.

NOTES

1. Although these HRTFs were nonindividualized and, thus, would be expected to produce somewhat less accurate broadband source localization than would be achieved with individualized HRTFs or with free-field sound presentations (Wenzel, Arruda, Kistler, & Wightman, 1993), there is evidence that nonindividualized HRTFs, such as the ones used here, perform nearly as well as individualized ones in multitalker speech perception tasks that rely primarily on low-frequency speech stimuli (Drullman & Bronkhorst, 2000; Nelson et al., 1999).

2. No such effect was seen in the two-talker configurations, but this may have been the result of a ceiling effect in the data.

(Manuscript received August 12, 2005;
revision accepted for publication February 24, 2006.)