Informational and energetic masking effects in the perception of two simultaneous talkers

Douglas S. Brungart[a)]
Air Force Research Laboratory, Human Effectiveness Directorate, Wright-Patterson AFB, Ohio 45433

(Received 5 June 2000; revised 25 October 2000; accepted 12 December 2000)

Although most recent multitalker research has emphasized the importance of binaural cues, monaural cues can play an equally important role in the perception of multiple simultaneous speech signals. In this experiment, the intelligibility of a target phrase masked by a single competing masker phrase was measured as a function of signal-to-noise ratio (SNR) with same-talker, same-sex, and different-sex target and masker voices. The results indicate that informational masking, rather than energetic masking, dominated performance in this experiment. The amount of masking was highly dependent on the similarity of the target and masker voices: performance was best when different-sex talkers were used and worst when the same talker was used for target and masker. Performance did not, however, improve monotonically with increasing SNR. Intelligibility generally plateaued at SNRs below 0 dB and, in some cases, intensity differences between the target and masking voices produced substantial improvements in performance with decreasing SNR. The results indicate that informational and energetic masking play substantially different roles in the perception of competing speech messages. [DOI: 10.1121/1.1345696]

PACS numbers: 43.66.Pn, 43.66.Rq, 43.71.Gv [DWG]

I. INTRODUCTION

When a speech signal is obscured by a second simultaneous competing speech signal, overall performance is determined by the cumulative effects of two different types of masking (Freyman *et al.*, 1999; Kidd *et al.*, 1998). Traditional ‘‘energetic’’ masking occurs when both utterances contain energy in the same critical bands at the same time and portions of one or both of the speech signals are rendered inaudible at the periphery. Higher-level ‘‘informational masking’’ occurs when the signal and masker are both audible but the listener is unable to disentangle the elements of the target signal from a similar-sounding distracter (Doll and Hanna, 1997; Kidd *et al.*, 1994; Kidd, Mason, and Rohtla, 1995; Neff, 1995; Watson, Kelly, and Wroton, 1976). It is difficult to isolate the informational and energetic elements of speech-on-speech masking, so no previous studies have directly examined their relative contributions to multitalker speech perception. However, the results of previous experiments do provide some insights about each type of masking.

The effects of purely energetic noise masking on speech intelligibility are well documented. The Articulation Index (AI) (French and Steinberg, 1947; Fletcher and Galt, 1950), which is based on a comprehensive series of experiments with non-speech-masking signals conducted in the early days of the telephone industry, is capable of predicting speech intelligibility directly from the long-term rms spectra of the speech and masker signals. Articulation theory has shown that energetic speech masking depends almost exclusively on the spectral overlap between the speech signal and the masker, and that performance decreases monotonically with decreasing signal-to-noise ratio (SNR).

The effects of informational masking on speech intelligibility can only be determined from multitalker experiments. Only a handful of early experiments has examined the effects of SNR in monaural speech-on-speech masking (Miller, 1947; Egan, Carterette, and Thwing, 1954; Dirks and Bower, 1969).[1] The two later studies found that performance generally increased with increasing SNR but was independent of SNR in the region from −10 to 0 dB. Although one would expect the similarity of the target and masker voices to be an important component in informational masking, very few studies have systematically examined the roles of voice characteristics such as talker and masker sex in speech-on-speech masking. The studies that have examined same-sex and different-sex talkers have shown that performance is substantially better in the different-sex condition (Festen and Plomp, 1990). This is consistent with informational masking, which is based on the listener’s ability to segregate perceptually similar sounds, and should increase when the target and masking voices resemble one another.

The majority of recent multitalker research has focused primarily on the binaural effects of spatially separating the target and masking signals (Drullman and Bronkhorst, 2000; Duquesnoy, 1983; Freyman *et al.*, 1999; Hawley, Litovsky, and Colburn, 1999; Festen and Plomp, 1986; Peissig and Kollmeier, 1997; Plomp, 1976; Yost, Dye, and Sheft, 1996). These ‘‘cocktail party’’ studies provide valuable information about the spatial unmasking of speech, but they do not provide many insights into the types of cues listeners use to segregate monaurally or diotically presented competing speech signals. The results of recent experiments do, however, suggest that the informational component of speech-on-speech masking may play a critical role in the ‘‘binaural

[a)] Electronic mail: douglas.brungart@he.wpafb.af.mil

advantage'' of spatially separating the talkers. Kidd and colleagues (1998) have shown that an exceptionally large release from masking can occur when a purely informational masker is moved to a different spatial location than the signal. More recent results have also shown that spatial separation or perceived spatial separation can produce a much larger release from speech-on-speech masking than from speech-on-noise masking (Hawley, Litovsky, and Culling, 2000; Freyman *et al.*, 1999). These results suggest that the informational component of speech-on-speech masking is particularly sensitive to differences in the perceived locations of the target and masker, and may account for a large portion of the spatial unmasking found in cocktail party experiments.

In this experiment, the intelligibility of a target phrase masked by a single competing talker was examined as a function of SNR, target sex, and masker sex under diotic listening conditions. The responses were used to determine the relative contributions of energetic and informational masking to overall performance. The results are discussed in terms of their potential implications in multitalker listening environments.

II. METHODS

A. Stimuli

The stimuli were derived from a publicly available speech corpus for multitalker communications research (Bolia *et al.*, 2000). This corpus, which is based on the coordinate response measure (CRM) first developed by Moore (1981), consists of phrases of the form ''Ready (call sign) go to (color) (number) now'' spoken with all possible combinations of eight call signs (''arrow,'' ''baron,'' ''charlie,'' ''eagle,'' ''hopper,'' ''laker,'' ''ringo,'' ''tiger''), four colors (''blue,'' ''green,'' ''red,'' ''white''), and eight numbers (1–8). Thus, a typical utterance in the corpus would be ''Ready baron go to blue five now.'' Eight talkers (four male, four female) were used to record each of the 256 possible phrases, so a total of 2048 phrases is available in the corpus.

In the speech-masker condition of this experiment, the stimuli consisted of two simultaneous phrases from the CRM corpus: a target phrase with the call sign baron and a masker phrase with a randomly selected call sign other than baron. In each trial, the target and masker phrases were selected randomly from the speech corpus with the restriction that different colors and different numbers were used in the two phrases. First, the overall level (rms power) of the masker phrase was set to a comfortable listening level (approximately 60–70 dB SPL). Then, the overall level (rms power) of the target phrase was adjusted relative to the level of the masker phrase to produce one of ten different SNRs ranging from −12 to 15 dB in 3-dB steps. The target and masker signals were then added together, and the combined signal was randomly roved over a 6-dB range (in 1-dB steps) before being presented to the listener over headphones. Within each block of 240 trials, each talker in the corpus was used as the speaker of the target phrase in exactly 30 trials. All other variables, including the masking talker, the masking call sign, the numbers and colors of the target and masker phrases, and the SNR, were chosen randomly with replacement on each trial.

Performance was also measured for two types of noise maskers. The first noise masker consisted of Gaussian noise that was spectrally shaped with a finite impulse response (FIR) filter matching the average long-term rms spectrum of the 2048 sentences in the CRM corpus and rectangularly gated to the same length as the target phrase. SNRs were varied from −18 to +15 dB in 3-dB steps in this condition, which is described in more detail elsewhere (Brungart, 2001). The second noise masker consisted of speech-shaped noise that was modulated with the envelope of a randomly selected competing speech phrase selected from the CRM corpus in the same way as in the speech-masker condition. The envelope was calculated by convolving the absolute value of the competing speech waveform with a 7.2-ms rectangular window. SNRs were varied from −21 to 0 dB in 3-dB steps in this modulated noise condition. As in the speech-masker conditions, the SNRs in the noise-masker conditions were calculated from the overall rms powers of the noise and speech waveforms. The noise-masker conditions were otherwise similar to the speech-masking conditions.

B. Listeners

Nine paid listeners, five male and four female, participated in the experiment. All had normal hearing (15 dB HL from 500 Hz to 6 kHz) and their ages ranged from 21–55. Each had participated in previous auditory experiments, and all but two had previous experience in experiments using the CRM speech materials.

C. Procedure

The listening task was performed while seated at a control computer. In each trial, the speech stimulus was generated by a sound card in the control computer (Soundblaster AWE-64) and presented to the listener over headphones (Sennheiser HD-520). Then an eight-column, four-row array of colored digits corresponding to the response set of the CRM was displayed on the CRT, and the listener used the mouse to select the colored digit corresponding to the color and number used in the target phrase containing the call sign baron. The trials were divided into blocks of 240 trials, with one or two blocks collected on each day of the experiment. A total of 2000 trials for each of the nine listeners was collected in the speech-masker condition of this experiment.[2] The speech-masker condition was followed by 960 trials per subject in the continuous-noise-masker condition, and then by 240 trials per subject in the speech-envelope modulated noise-masker condition.

III. RESULTS AND DISCUSSION

A. The dominance of informational masking

The results in Fig. 1 show substantial differences between the speech-shaped noise masker and the three speech maskers used in the experiment. These differences reflect the different mechanisms involved in the two masking condi-

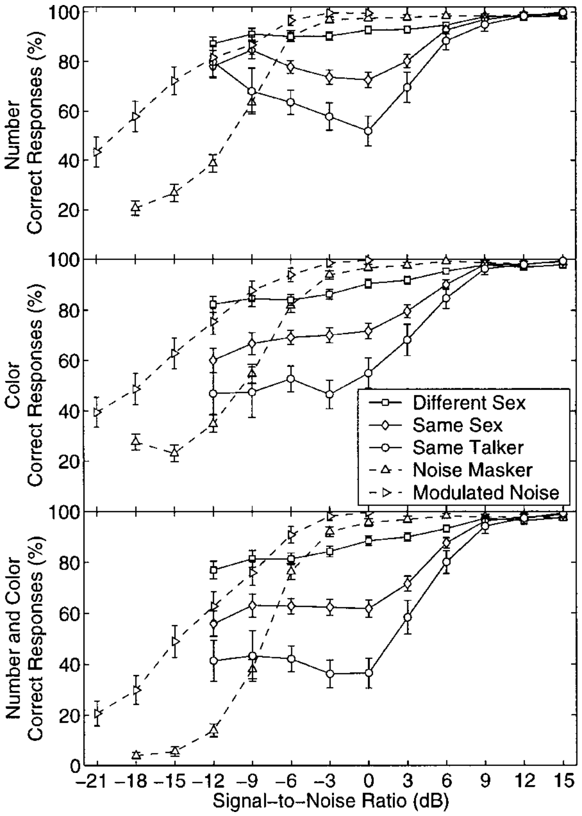

FIG. 1. Percentage of correct identifications as a function of SNR. The top panel shows the percentage of trials where the listener correctly identified one of eight different numbers. The middle panel shows the percentage of correct identifications of one of four colors. The bottom panel shows the percentage of trials in which the listener correctly identified both the number and color. The data are shown separately for trials where the competing talkers were of different sexes, where the talkers were different but were both male or female, and for trials where the same talker was used for both the target and masker sentences. Results are also shown for a speech-shaped noise masker and an envelope-modulated speech-shaped noise masker. In each case, the data at each SNR value were averaged across the nine listeners used in the experiment. Note that each data point represents approximately 900 trials for the different-sex condition, 675 trials for the same-sex condition, 225 trials for the same-talker condition, 720 data points for the noise-masker condition, and 270 trials for the modulated noise-masker condition. The error bars represent the 95%-confidence intervals for each point. Note that the SNR ratios were calculated from the overall rms power of each signal.

tions: with the speech-shaped noise masker, the masking is primarily energetic and results from acoustic degradation of the speech signal; with the competing speech maskers, energetic masking occurs when the two fluctuating speech signals overlap in time and frequency, and higher-level informational masking occurs when the listener tries to segregate the two similar-sounding talkers into distinct utterances. While there is no question that some energetic masking occurs in the simultaneous speech conditions, there is substantial evidence that informational masking effects dominated performance in the two-talker conditions.

An upper bound on the influence of energetic masking in the two-talker conditions is provided by the percentages of correct responses at the lowest SNR tested, −12 dB. At this level, the listeners were still able to correctly identify 80% of the target numbers and 50% of the target colors, even when the same talker was used for the target and masker phrases. The effects of energetic masking decrease monotonically with increasing SNR, so these performance levels reflect the greatest possible influence of energetic masking in this experiment. Any reduction in performance below these levels, such as the significant reductions in correct number identifications that occur when the SNR is near 0 dB in the same-sex and same-talker conditions, can only be attributable to informational masking effects.

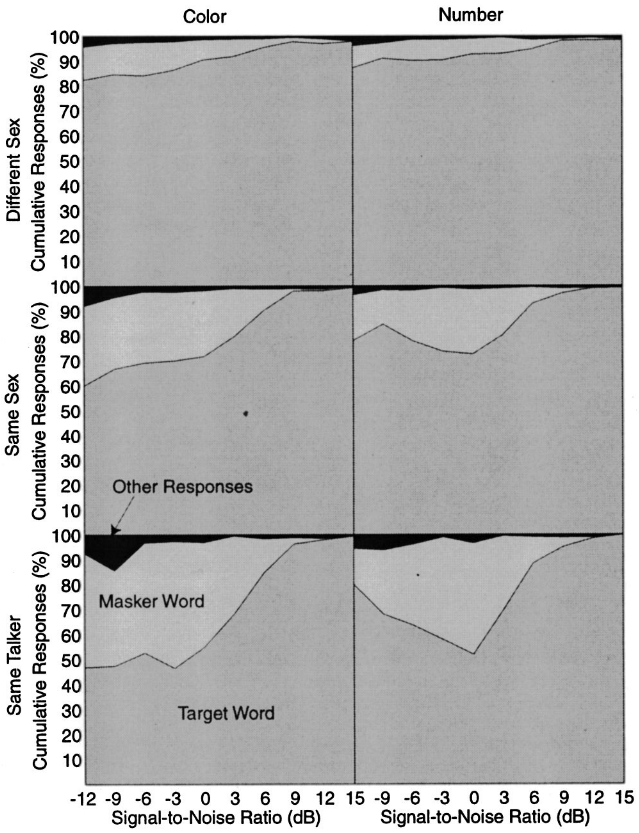

FIG. 2. These area graphs show the distribution of the listeners' responses among the correct word spoken by the target talker, the incorrect word spoken by the masking talker, and all other possible incorrect responses. In all conditions, the masking word occurred in the overwhelming majority of the incorrect responses.

The distributions of incorrect responses in the experiment provide additional evidence that informational masking dominates the perception of two-talker speech. When energetic masking interferes with the perception of a speech stimulus, the utterance is rendered inaudible by the noise and the listener must randomly choose the response from among all words in the vocabulary. Thus, energetic masking should produce roughly random distributions of incorrect responses. There is, however, no evidence of this type of error distribution in the two-talker conditions of this experiment. An analysis of the errors in the experiment (Fig. 2) shows that the listeners who failed to correctly identify the words in the target phrase were much more likely to respond to the words heard in the masker phrase than to the other possible words in the response set. At negative SNRs, where the target phrase was difficult to comprehend, it is not surprising that the listeners tended to respond to the coordinates in the easily understood masker phrase. At large positive SNRs, however, it is much harder to explain the masker-word responses in the context of energetic masking. In the same-talker condition at 9 dB SNR, where substantial energetic masking of the masker phrase by the target phrase should have occurred, approximately 66% of the incorrect color responses and 75% of the incorrect number responses were assigned to the masker word (compared to the 33% and 14% expected for randomly distributed errors). This suggests that the listeners were usually able to detect the color–number coordinates, even when they were energetically masked by a 9-dB more intense voice. At 0 dB SNR, where energetic masking should have been roughly equivalent for the two talkers, the target or masker coordinates appear in virtually all the responses. It is interesting to note that the percentages of nonmasker, nontarget responses were not much higher at −6 dB than they were at 15 dB, where the target phrase should have been clearly audible and the incorrect responses were probably the result of extraneous factors such as subject inattention. These extraneous errors should be evenly distributed across all the SNR levels in the experiment, so there is little evidence that energetic masking is causing any of the incorrect responses at SNRs above −6 dB.

The results of the modulated noise condition provide a final indication that energetic masking plays a relatively small role in the speech-masker conditions. With the modulated noise masker, the energy in the masking signal is located in roughly the same spectral region as in the speech maskers, and the level of the noise signal fluctuates in the same way as the level of the speech maskers. Thus, the temporal distribution of energy in the modulated noise should be very similar to the temporal distribution of energy in the speech.[3] In the same-sex condition, performance does seem to be bounded by the modulated noise condition. At −9 dB SNR, there is a local maximum in the percentage of correct number identifications where the same-sex performance curve intersects with the modulated noise masker. Elsewhere, however, performance was substantially better in the modulated noise condition than in any of the two-talker conditions. Again, this suggests that energetic masking had relatively little impact on the intelligibility of the color and number coordinates in the two-talker conditions, and that the color–number coordinates of both masker phrases were audible across most of the range of SNRs tested.

A final comment should be made about the relationship between energetic masking and the type of speech intelligibility test used. From articulation theory, it is known that the relationship between speech intelligibility and noise level is highly dependent on the type of test used to measure intelligibility. Previous work relating the CRM task to the AI (Brungart, 2001) has shown that the CRM test is relatively insensitive to energetic masking by noise: 50% performance with the CRM requires an AI of only 0.08, compared to 0.17 with the well-known Modified Rhyme Test (MRT). Indeed, the insensitivity of the CRM to noise is readily apparent from the performance curves in Fig. 1: substantial noise masking does not occur until the noise is at least 6 dB more intense than the speech. It is likely that the amount of energetic masking that occurs in multiple-talker speech is directly related to the noise sensitivity of the speech utterances tested. Thus, one would expect to see larger energetic-masking effects in a multitalker experiment based on a more sensitive speech test such as the MRT than were found in this CRM-based experiment. The relative absence of energetic effects with the CRM speech corpus makes it a relatively good choice for experiments focusing on the effects of informational masking in speech.

B. Talker and masker voice characteristics

The informational masking in this experiment is related to the listener's inability to segregate similar-sounding speech signals, so one would expect the effects of informational masking to increase when the target and masking voices have similar characteristics. This is clearly seen in the results of this experiment (Fig. 1). At negative SNRs, correct identifications were 15%–20% lower in the same-sex condition than in the different-sex condition, and 15%–20% lower in the same-talker condition than in the same-sex condition. These large differences in performance with the talker and masker voice characteristics are consistent with the concept of informational masking. A listener's ability to segregate two simultaneous voice signals is dependent on the distinctive characteristics of the two voices. Male and female voices are so different that the listener has little difficulty discriminating between the target and masker phrases. The voices of two different talkers of the same sex are much more difficult to segregate, but still provide the listener with substantial acoustic cues about the identity of the two talkers. Two voices from the same talker provide only minor acoustic discrimination cues and make the multitalker listening task extremely difficult. Clearly, in experiments focusing on informational masking, same-talker voices are preferred and different-sex voices should be avoided.

Of course, energetic masking should also increase with the similarity of the target and masker voices because the energy is concentrated in the same frequency ranges when same-sex and same-talker voices are used. However, the informational masking component appears to dominate this increase in energetic masking. Festen and Plomp (1990) used noise maskers that matched the long-term rms spectra of male and female voices and found much larger differences in performance between speech-on-speech masking with same-sex and different-sex talkers than they found between noise-on-speech masking with same-sex and different-sex speech-shaped noise maskers.

The relationship between the sexes of the target and masking talkers had a much larger impact on overall performance than did the particular sex of the target talker (Fig. 3). In each masking condition, performance was similar for male and female target talkers. However, there were some differences in performance across the eight individual talkers used in the CRM speech corpus. One male talker consistently produced more accurate responses than the others did (Fig. 4). When Talker 3 was used for the target phrase, the listeners correctly identified the number and color in 11% more trials than the average of the other seven talkers. When Talker 3 was used for the masker, the listeners correctly identified the color and number in 9% more trials than the average of the other talkers. This result implies that distinctive individual voice characteristics can improve the intelligibility of a par-

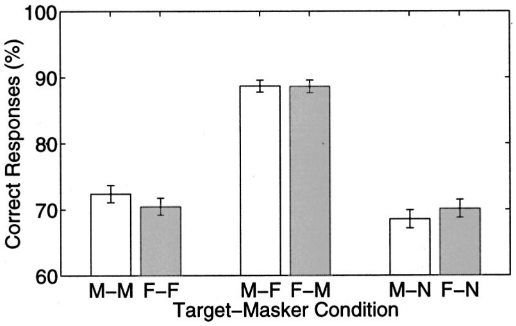

FIG. 3. Overall percentages of correct identifications of both color and number for each talker–masker combination. The conditions are shown as A–B where A is the target and B is the masker. M=male speech; F=female speech; N=speech-shaped noise. The error bars represent the 95%-confidence intervals of each point. The noise-masker data have been averaged over a wider range of SNR values than the speech-masker data (−18 to +15 dB for the noise versus −12 to +15 dB for the speech), so these results should not be used to compare overall performance in the noise-masking and speech-masking conditions.

ticular voice in multichannel communications, both when the voice serves as the target and when the voice serves as the masker.

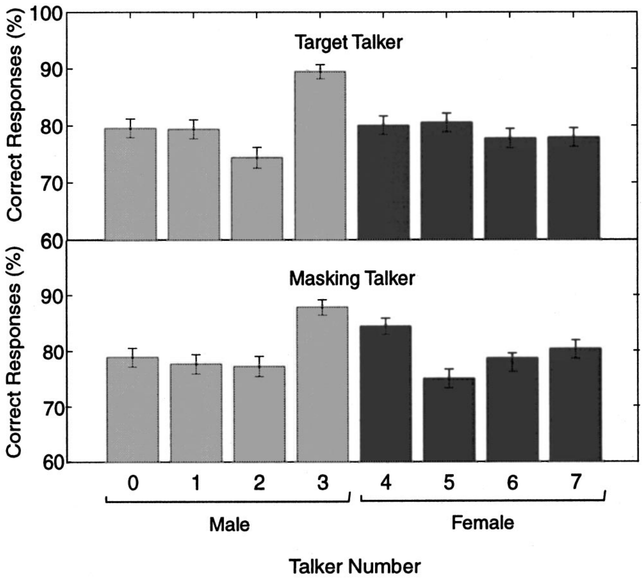

FIG. 4. Percentage of correct overall responses as a function of the talker used in the target phrase (top panel) and the masker phrase (bottom panel). Error bars show the 95%-confidence intervals for each data point.

C. Signal-to-noise ratio

From direct comparison of the speech masker results to those with the speech-shaped noise masker, it is immediately apparent that SNR does not have the same effect in informational masking as it does in energetic masking. In this experiment, the continuous speech-shaped noise serves as a classical energetic masker (Fig. 1). Intelligibility is essentially unaffected by the noise until the noise becomes loud enough to mask the energy in the speech signal (at about −3 dB). Once this masking starts to occur, performance degrades very rapidly. In just a 9-dB drop in SNR (from −3 to −12 dB), the speech signal is swamped by the masker and the number of correct responses falls by 90%. As the SNR continues to decrease, the target phrase becomes inaudible and the listener responses approach random guessing (this occurs for the color response at SNRs less than −15 dB). The speech-shaped noise masker is spectrally similar to the speech, so energetic masking causes a very rapid drop-off in performance with decreasing SNR in this experiment. When the noise has a different frequency distribution than the speech, or its level varies with time, different spectral and temporal regions of the speech are masked at different SNR levels and the onset of energetic masking is more gradual. This is exactly the case with the speech-modulated noise masker shown in Fig. 1. The listener is able to hear the speech during the quiet portions of the modulated masker, and performs substantially better than with the continuous noise despite the higher peak levels that must occur in the modulated noise to maintain the same overall SNR. As SNR decreases, the less intense regions of the modulated masker gradually become loud enough to mask the target phrase and overall performance slowly decreases. However, energetic masking always causes monotonic decreases in performance with decreasing SNR: increasing the noise level can only decrease the amount of information in the stimulus and, as long as there is some masker energy in all the frequency bands and temporal regions of the signal, energetic masking will eventually reduce the signal to inaudibility.

The role of SNR in informational masking is quite different. At high SNR values (above +9 dB), the target phrase is so loud relative to the masker signal that near-perfect performance occurred regardless of the masker signal used. At lower SNR values, however, there are substantial differences among the three speech-masking conditions. Performance in the different-sex condition was largely independent of the SNR—performance dropped by approximately 1%/dB at SNRs less than 9 dB, and correct overall identifications occurred in approximately 80% of the trials even at the lowest SNR value tested (−12 dB). In the same-sex and same-talker conditions, overall correct identifications decreased sharply as SNR fell from 9 to 0 dB, then plateaued well above chance level as SNRs decreased from 0 to −12 dB. Correct number identifications actually *increased* substantially (as much as 25% in the same-talker condition) as SNR decreased from 0 to −12 dB. Clearly, SNR has a much different impact on speech intelligibility in the information-masking dominated same-sex and same-talker conditions than it does in the energetic-masking dominated speech-shaped noise condition.

These results are consistent with previous studies that have examined the effect of SNR on the intelligibility of speech masked by a single competing talker. Egan, Carterette, and Thwing (1954) and Dirks and Bower (1969) examined a wide range of SNRs and found performance curves very similar to the ones found in this experiment: performance fell rapidly as SNR dropped from 10 to 0 dB, plateaued or increased slightly as SNR dropped from 0 to −10 dB, and fell rapidly again at SNRs lower than −10 dB. Stubbs and Summerfield (1990) examined only positive SNRs and found a rapid increase in performance from 0 to 9 dB and near-perfect performance above 9 dB. Freyman *et al.*

(1999) examined only negative SNRs and found that performance decreased much less rapidly when the speech was masked by colocated speech from the same talker (decreasing approximately 2.5%/dB) than when the speech was masked by a colocated speech-shaped noise masker (decreasing 6.7%/dB).

Taken together, the data from this experiment and the results of these previous studies tell a consistent story about speech intelligibility in the presence of a single competing talker: (1) Speech intelligibility is unimpeded by a competing talker when the target phrase is at least 10 dB more intense than the masker; (2) Speech intelligibility drops substantially as the SNR drops from 10 to 0 dB; (3) Speech intelligibility is roughly independent of SNR at SNRs from 0 to −10 dB.

The somewhat paradoxical third result can be explained in the context of informational masking. In order to understand multiple simultaneous talkers, the listener needs some way to differentiate the target voice from the masking voice. In the case where both talkers have the same voice and speak at the same level, the cues available for this discrimination are minimal (Egan *et al.*, 1954). This is exactly the case in the same-talker condition at 0 dB, where both the target and masker phrases were spoken by the same voice at the same level (Fig. 1). Correct number and color identifications occurred in only about half of the trials in that condition, and nearly all of the incorrect responses included the colors and numbers used in the masker phrase (Fig. 2). From an informational masking standpoint, this is effectively chance performance: the listeners were able to hear both the target and masker phrases, but were unable to determine which color–number pair was addressed to the call sign baron.

When the listener is faced with two similar or identical voices, any differences in the characteristics of the two talkers can help improve performance. When the intensity of the target is lowered relative to the masker, the intensity difference provides a means to discriminate the target and masking talkers, and the advantages of this level difference outweigh the increase in energetic masking caused by decreasing the SNR. In other words, the listeners were able to recognize that the baron call sign was spoken by the less-intense talker and could selectively tune their attention to the quieter voice. This explains why number identifications in the same-sex conditions of this experiment were least accurate at an SNR of 0 dB (Fig. 1) and increased substantially (by as much as 25%) when the intensity of the target voice was reduced relative to the masker.[4]

Note that the gains in performance afforded by level differences between the target and masking voices were not symmetric: positive SNRs always produced a larger performance benefit than negative SNRs of the same magnitude. Not surprisingly, it was always easier to attend to the louder of two simultaneous talkers than to the quieter one.

D. Statistical dependence of color and number errors

Within each trial in the CRM, the listener is required to respond independently to the color and number used in the target phrase. It is therefore informative to know whether the errors in the color and number responses are evenly distributed across all trials or grouped into the same trials. This can be determined by checking the responses for statistical independence (Fig. 5). If the errors are statistically independent, the probability of correctly identifying the color across all trials at a given SNR $[P(C)]$ should be the same as the probability of correctly identifying the color in trials where the number was also correctly identified $[P(C|N)]$. As seen in the figure, this is true for the energetic noise masker. Thus, the ability to correctly identify the number in a given noise-masker trial was independent of the ability to correctly identify the color. With the informational speech maskers, however, the listeners were substantially more likely to correctly identify the color when they were also able to correctly identify the number. Thus, with the speech maskers, the color and number errors tended to be grouped into the same trials. As shown in the right panel of Fig. 5, this effect was strongest in the same-talker condition and weakest (but still substantial) in the different-sex condition. The tendency to group color and number trials together in the speech-masker conditions relates directly to the listeners' attempts to segregate the target and masker phrases. Because the colors and numbers occur sequentially in the CRM phrases, there are strong prosodic cues that link together the color–number pairs spoken by the same talkers. Also, when there are differences in speaking rates across the talkers, the color–number pairs will occur at different times in the target and masker phrases. These temporal and prosodic cues influence the listeners to respond to the color–number pair spoken by one of the two talkers (thus getting both coordinates correct or missing both) rather than responding to the color spoken by one talker and the number spoken by the other (and getting only one of the two coordinates correct).

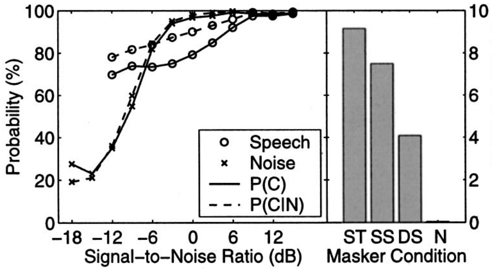

FIG. 5. Statistical dependence of correct responses for the color and number in the target phrase. The left panel shows the overall probability of a correct color response $[P(C)]$ and the probability of a correct color response when the listener correctly identified the number in the target phrase $[P(C|N)]$ averaged across all the target and masking talkers used in the experiment. The right panel shows the average value of $P(C|N)-P(C)$ across all SNRs for the same-talker (ST), same-sex (SS), different-sex (DS), and noise-masker (N) conditions.

E. Color and number effects

The specific colors and numbers used in the target and masker phrases had relatively little impact on performance in the experiment (Fig. 6). The percentages of correct responses varied by less than 8% across the eight numbers and four colors used in the experiment. Note, however, that there was a strong negative correlation between performance when the

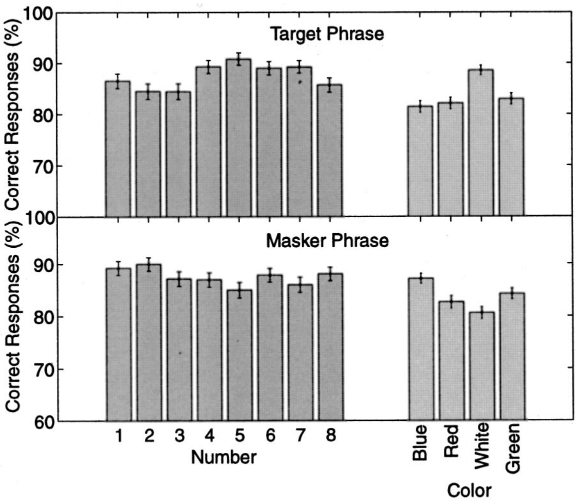

FIG. 6. Correct identifications for the different color and number words used in the CRM corpus. The data on the left show the percentages of correct number identifications as a function of the number coordinate used in the target phrase (top panel) and in the masker phrase (bottom panel). The data on the right show the percentages of correct color identifications as a function of the color coordinate used in the target phrase (top panel) and in the masker phrase (bottom panel). The error bars represent 95%-confidence intervals.

individual words appeared in the target and masker phrases. The color ''white,'' for example, produced the best overall performance when it appeared in the target phrase and the worst overall performance when it appeared in the masker phrase. The correlation coefficient between the percentage of correct responses when each word appeared in the target phrase and the percentage of correct responses when each word appeared in the masker phrase was −0.74 for the number coordinates and −0.80 for the color coordinates. This negative correlation indicates that the words that were easy to identify in the target phrase were especially effective maskers when they appeared in the masker phrase. Note that this is in direct contrast to the effects of the individual talker voices on performance. Talker 3 produced better performance than any other talker, both as a target talker and as a masking talker (Fig. 4), and the correlation coefficient between overall performance when each talker was used for the target phrase or the masking phrase was +0.74. These results suggest that the intelligibility of the individual coordinate words is based primarily on their ability to overpower the coordinates in the competing phrase, while the intelligibility of the individual talkers is based primarily on the acoustic similarities between the target and masking voices.

F. Intersubject differences

In general, the differences across the listeners were small relative to the large effects of SNR and the sexes of the target and masker. The pattern of responses (with respect to the SNR, as shown in Fig. 1) was roughly similar for each listener, and the percentage of correct overall identifications ranged from 72% to 86% across the nine listeners used in the experiment.

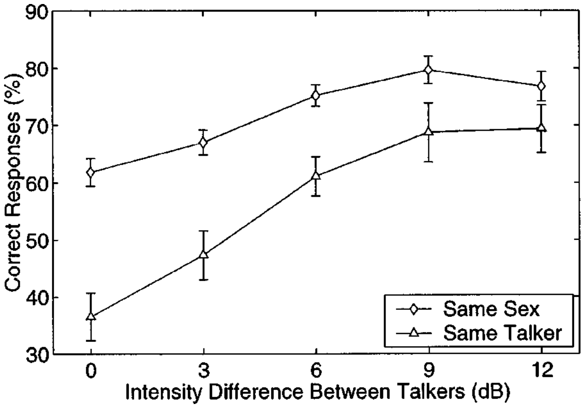

FIG. 7. Percentage-correct identifications of both color and number for two competing phrases spoken by different same-sex talkers and by the same talker when the target sentence is equally likely to come from either of the two voices. The results are shown as a function of the difference in intensity between the two talkers. The error bars in the top panel represent 95%-confidence intervals.

IV. THE IMPACT OF SNR ON THE OVERALL INTELLIGIBILITY OF TWO-TALKER SPEECH

The plateauing in performance that occurs in the same-sex conditions at SNRs less than 0 dB has important implications in monaural or diotic multitalker listening situations. Consider, for example, a listener who is monaurally monitoring two radio channels for information that is equally likely to originate from either channel. If the listener is able to control the levels of the two channels independently, how should the radio be configured for optimal performance in this task? Because both channels are equally important, intuition would suggest that both should be set at the same level. However, the results of this experiment suggest not only that setting the channels to the same level is less than optimal, but that it is among the worst possible strategies. The explanation for this apparent paradox lies in the performance curves shown in Fig. 1. Assume that two different talkers are used on the two channels and that both are males or both are females. If both channels are set to the same level, overall correct identifications should occur in approximately 60% of all trials. Now, if one channel is set to a 3-dB-higher SNR than the other channel, a substantial (10%) improvement in intelligibility occurs when the target phrase occurs in the more intense channel, while performance is effectively unchanged when the target phrase occurs in the less intense channel. The number of correct overall identifications in this scenario will be the average of performance in the two channels, or 65%. Thus, by setting the overall levels of the two channels 3 dB apart, a 5% gain in overall intelligibility has been achieved. As the level difference between the two channels is increased (Fig. 7), overall performance continues to improve until reaching its peak value when the channel separation is 9 dB. At separations greater than 9 dB, intelligibility approaches 100% in the more intense channel and begins degrading due to energetic masking in the less-intense channel. Consequently, no further performance gains can be achieved by separating the channels by more than 9 dB. Note that this improvement is driven more by the rapid intelligibility increase at positive SNRs than by the increase in performance that occurs in number identifications at negative SNRs: performance is maximized for both the color and number coordinates when the channel separation is 9 dB.

The improvements in performance obtained by varying the levels of the talkers are substantial. When the intensities of two different same-sex talkers differ by 9 dB, the percentage of overall correct identifications is more than 20% higher than when the talkers are at the same level. The improvement is even larger when both competing messages are spoken by the same talker. These improvements are roughly equivalent to increasing the SNR by 5 dB in both channels.

V. CONCLUSIONS

This experiment has examined the factors that influence a listener's ability to selectively attend to one of two competing speech messages in the coordinate response measure paradigm. There is substantial evidence that informational masking, rather than energetic masking, dominated performance in this experiment. Thus, it is possible to draw some general conclusions about the differences between informational and energetic masking in a two-competing-talker task:

(i) The voice characteristics of the target and masking talkers have a profound effect on the intelligibility of competing speech messages. Performance was substantially better when the talkers were of different sexes than when they were the same sex, and substantially better with different talkers of the same sex than when the same talker was used for both the target and masker phrase.

(ii) The SNR generally has a much smaller influence on speech intelligibility with a speech masker than with a noise masker. When an energetic noise masker is used, performance decreases monotonically with the SNR, and falls off rapidly to chance level when the masker begins to overpower the target speech signal. When a speech masker is used, performance decreases when the SNR is reduced from 9 to 0 dB, but plateaus at well above chance level and, in some cases, increases when the SNR is reduced from 0 to −9 dB.

(iii) In some cases, the intelligibility of two competing talkers can be substantially improved by introducing a level difference in the two voices. With the CRM phrases, overall intelligibility could be improved from about 60% to about 80% by introducing a 9 dB level difference between two same-sex talkers.

The results also suggest that the phrases in the CRM speech corpus are ideally suited for speech intelligibility experiments that focus on the informational component of speech on speech masking. When the CRM phrases are used at SNRs near 0 dB, there is strong evidence that both the target and masker voices are clearly audible and that the vast majority of incorrect responses were due to the listeners' inability to correctly determine which color and number coordinates were directly addressed to their assigned call sign. Thus, it appears that the CRM phrases would be an excellent choice for researchers who want to isolate the informational masking component of speech in future two-talker cocktail party experiments.

ACKNOWLEDGMENTS

The author would like to thank Brian Simpson, Mark Ericson, Todd Nelson, and Bob Bolia both for the use of their CRM corpus and for their helpful comments during the preparation of this manuscript. This work was supported in part by AFOSR Grants F49620-98-1-0108 and 01-HE-01-COR.

[1] See Bronkhorst (2000) for a recent review of studies examining monaural or diotic multitalker speech perception.

[2] Due to a technical error, some trials were collected with the same number or color in the target and masker phrases. These trials were discarded, and the first 2000 valid trials for each subject were used in the analyses.

[3] Note that the spectral distribution of energy is fixed in the modulated noise masker but time variant in speech.

[4] It is not clear why the advantage of negative SNRs was limited to the number response. It may be related to the placement of the number after the color in the carrier phrase, where variations in the speaking rates of the talkers are more likely to reduce the temporal overlap of the numeric coordinates in the target and masking phrases. It may also be related to the greater sensitivity of the color coordinate to energetic masking (Brungart, 2000). The fact that color identification performance did not decrease at negative SNRs may imply that the decrease in informational masking caused by the difference in the target and masker levels was offset by a corresponding increase in energetic masking in these conditions.

Bolia, R., Nelson, W., Ericson, M., and Simpson, B. (**2000**). ''A speech corpus for multitalker communications research,'' J. Acoust. Soc. Am. **107**, 1065–1066.

Bronkhorst, A. (**2000**). ''The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions,'' Acustica **86**, 117–128.

Brungart, D. (**2001**). ''Evaluation of speech intelligibility with the coordinate response measure,'' J. Acoust. Soc. Am. (to be published).

Dirks, D., and Bower, D. (**1969**). ''Masking effects of speech competing messages,'' J. Speech Hear. Res. **12**, 229–245.

Doll, T., and Hanna, T. (**1997**). ''Directional cueing effects in auditory recognition,'' in *Binaural and Spatial Heating in Real and Virtual Environments*, edited by R. H. Gilkey and T. R. Anderson (Erlbaum, Hillsdale, NJ).

Drullman, R., and Bronkhorst, A. (**2000**). ''Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation,'' J. Acoust. Soc. Am. **107**, 2224–2235.

Duquesnoy, A. (**1983**). ''Effect of a single interfering noise or speech source on the binaural sentence intelligibility of aged persons,'' J. Acoust. Soc. Am. **74**, 739–943.

Egan, J., Carterette, E., and Thwing, E. (**1954**). ''Factors affecting multichannel listening,'' J. Acoust. Soc. Am. **26**, 774–782.

Festen, J., and Plomp, R. (**1986**). ''Speech reception threshold in noise with one and two hearing aids,'' J. Acoust. Soc. Am. **79**, 465–471.

Festen, J., and Plomp, R. (**1990**). ''Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing,'' J. Acoust. Soc. Am. **88**, 1725–1736.

Fletcher, H., and Galt, R. (**1950**). ''The perception of speech and its relation to telephony,'' J. Acoust. Soc. Am. **22**, 89–151.

French, N., and Steinberg, J. (**1947**). ''Factors governing the intelligibility of speech sounds,'' J. Acoust. Soc. Am. **19**, 90–119.

Freyman, R., Helfer, K., McCall, D., and Clifton, R. (**1999**). ''The role of perceived spatial separation in the unmasking of speech.'' J. Acoust. Soc. Am. **106**, 3578–3587.

Hawley, M., Litovsky, R., and Culling, J. (**2000**). ''The 'cocktail party' effect with four kinds of maskers: Speech, time-reversed speech, speech-shaped noise, or modulated speech-shaped noise,'' in Proceedings of the Midwinter Meeting of the Association for Research in Otolaryngology, 31.

Hawley, M., Litovsky, R., and Colburn, H. (**1999**). ''Speech intelligibility and localization in a multisource environment,'' J. Acoust. Soc. Am. **105**, 3436–3448.

Kidd, G. J., Mason, C., Deliwala, P., Woods, W., and Colburn, H. (**1994**). ''Reducing informational masking by sound segregation,'' J. Acoust. Soc. Am. **95**, 3475–3480.

Kidd, G. J., Mason, C., and Rohtla, T. (**1995**). ''Binaural advantage for sound pattern identification,'' J. Acoust. Soc. Am. **98**, 1977–1986.

Kidd, G. J., Mason, C., Rohtla, T., and Deliwala, P. (**1998**). ''Release from informational masking due to the spatial separation of sources in the identification of nonspeech auditory patterns,'' J. Acoust. Soc. Am. **104**, 422–431.

Miller, G. (**1947**). ''The masking of speech,'' Psychol. Bull. **44**, 105–129.

Moore, T. (**1981**). ''Voice communication jamming research,'' in AGARD Conference Proceedings 331: Aural Communication in Aviation, pp. 2:1–2:6, Neuilly-Sur-Seine, France.

Neff, D. (**1995**). ''Signal properties that reduce masking by simultaneous random-frequency maskers,'' J. Acoust. Soc. Am. **96**, 1909–1921.

Peissig, J., and Kollmeier, B. (**1997**). ''Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners,'' J. Acoust. Soc. Am. **35**, 1660–1670.

Plomp, R. (**1976**). ''Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of the azimuth of a single competing sound source (speech or noise),'' Acustica **34**, 325–328.

Stubbs, R., and Sommerfield, Q. (**1990**). ''Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners,'' J. Acoust. Soc. Am. **87**, 359–372.

Watson, C., Kelly, W., and Wroton, H. (**1976**). ''Factors in the discrimination of tonal patterns. II. Selective attention and learning under various levels of stimulus uncertainty,'' J. Acoust. Soc. Am. **60**, 1176–1185.

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