Release from speech-on-speech masking by adding a delayed masker at a different location

Brad Rakerd[a)]
Department of Audiology and Speech Sciences, Michigan State University, East Lansing, Michigan 48824

Neil L. Aaronson and William M. Hartmann
Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan 48824

(Received 27 November 2004; revised 4 December 2005; accepted 5 December 2005)

The amount of masking exerted by one speech sound on another can be reduced by presenting the masker twice, from two different locations in the horizontal plane, with one of the presentations delayed to simulate an acoustical reflection. Three experiments were conducted on various aspects of this phenomenon. Experiment 1 varied the number of masking talkers from one to three and the signal-to-noise (S/N) ratio from −12 to +4 dB. Evidence of masking release was found for every combination of these variables tested. For the most difficult conditions (multiple maskers and negative S/N) the amount of release was approximately 10 dB. Experiment 2 varied the timing of leading and lagging masker presentations over a broad range, to include shorter delay times where room reflections of speech are rarely noticed by listeners and longer delays where reflections can become disruptive. Substantial masking release was found for all of the shorter delay times tested, and negligible release was found at the longer delays. Finally, Experiment 3 used speech-spectrum noise as a masker and searched for possible energetic masking release as a function of the lead-lag time delay. Release of up to 4 dB was found whenever delays were 2 ms or less. No energetic masking release was found at longer delays.

© *2006 Acoustical Society of America.* [DOI: 10.1121/1.2161438]

PACS number(s): 43.66.Pn, 43.66.Qp, 43.55.Hy [AK] Pages: 1597–1605

I. INTRODUCTION

When several people talk at once, listening to just one of them can be difficult. The extent of the difficulty depends upon a number of factors, including the spatial arrangement of the talkers, the similarity of their voices, and the content of the messages (Cherry, 1953; Moray, 1959; Yost, 1997; Bronkhorst, 2000). It might be expected that the difficulty could only increase if the distracting talkers' speech power was to be doubled by an acoustical reflection, but Freyman and colleagues have recently found evidence to the contrary (Freyman *et al.* 1999, 2001, 2004).

In a baseline condition, Freyman *et al.* mixed together independent speech samples from multiple talkers and presented the combined signals from a single loudspeaker directly in front of a listener. The listener's task was to attend to the message of a single target talker and to ignore the competing messages of any other talkers, which were distractors. The surprising finding was that this task became easier when an added copy of the distractors' speech was presented from a second loudspeaker off to the listener's right side, with the copy shifted in time by a few milliseconds. This manipulation produced perceptible changes in the spatial characteristics of the distracting speech, and a substantial reduction in its interfering effects. An important further finding was that the reduction appeared to be based largely on release from informational masking (Watson *et al.*, 1976; Leek *et al.*, 1991; Kidd *et al.*, 1994), as evidenced by the fact that it was obtained with speech distractors but not with other physically comparable sounds.

Most experiments conducted to date have shifted the distractors 4 ms forward in time, so that their presentation leads at the right side relative to the front where the target is located (Freyman *et al.* refer to this as condition F-RF). A shift of this kind can be expected to elicit a robust precedence effect for speech, with its leading and lagging copies fused into a single perceptual image that is localized at or near the location of the leading sound source (Wallach *et al.*, 1949; Litovsky *et al.*, 1999). Specifically, the image of the distractors should be localized at or near the right-hand loudspeaker (+60°) and well away from the target speech location (0°). Freyman *et al.* report just such an effect.

Significant masking release has also been found for a condition where the distractor speech is shifted to lag behind by 4 ms at the side. Freyman *et al.* report that in this condition (their F-FR condition) the distractors should be localized slightly to the right of the front speaker, consistent with a model of the precedence effect (Shinn-Cunningham *et al.*, 1993), and that a small perceptual separation of this kind may provide a basis for differentiating the target and distractor messages. Also, binaural disparities associated with the two-source presentation of the distractors create a relatively diffuse auditory image in this condition. By contrast, the target speech which is presented from a single loudspeaker straight ahead has a more compact image, and this difference in the images may aid a listener in attending to the target.

[a)] Author to whom correspondence should be addressed; electronic mail: rakerd@msu.edu

The primary motivation for the present study was to extend the investigation of this masking release phenomenon to a much broader range of time shifts than have been examined to date. The precedence effect generally operates in rooms, where listeners experience great variation in the timing associated with acoustical reflections. There are differences among rooms owing to their differing physical dimensions, and differences within a room owing to variations in the positioning of talkers and listeners. Our goal in the present study was to represent this reflection delay time variation, and to examine its consequences for speech masking release. Accordingly, perceptual tests were done here at delays comparable to those that have been examined previously (±4 ms), at delays much briefer than this, and at delays much longer.

The speech stimuli used in this study were drawn from the Coordinate Response Measure (CRM) speech stimulus set (Bolia *et al.*, 2000). This corpus is well suited to the study of speech masking and masking release (Brungart, 2001; Brungart and Simpson, 2002; Brungart *et al.*, 2001, 2005). With the CRM, it is possible to create a competition among two or more messages, all spoken by different talkers, and to precisely manipulate the relative intensities of the various talkers' speech. These were the principle variables examined in Experiment 1. Based on the outcome of that experiment, a detailed examination of delay times and speech masking release was then conducted in Experiment 2. Finally, in Experiment 3, speech-spectrum noise was used to look for energetic masking release effects at multiple delay times.

II. EXPERIMENT 1: THE NUMBER OF DISTRACTING TALKERS AND S/N

A computer program and graphical user interface were written to assess speech understanding with the CRM as a function of any of several parameters that have been shown to be relevant to the study of speech masking release. These included the number of distracting talkers present, the delay associated with a multisource presentation of the distractors' speech, and the relative levels of the speech produced by the distractor talkers and by a target talker, which dictates the signal-to-noise ratio (S/N). Experiment 1 examined release from speech masking as a function of the number of distracting talkers and as a function of S/N. All other parameters for the experiment were fixed at values selected to match those employed in previous studies.

A. Subjects

The subjects of this study were two women (ages 20 and 63) and two men (ages 24 and 51). All of the subjects had hearing thresholds within 20 dB of audiometric zero at speech frequencies. All of them were practiced listeners. Two of the subjects were among the authors of this study. The other two had no knowledge of the purpose of the experiment.

B. Stimuli

Stimuli were a subset of the sentences comprising the CRM corpus. This corpus features eight different talkers—four men and four women—each saying 256 different versions of the following sentence:

"Ready ⟨call sign⟩, go to ⟨color⟩⟨number⟩ now . "

The sentence versions differ according to the values assigned to the variables ⟨call sign⟩, ⟨color⟩, and ⟨number⟩. Stimulus files are approximately equal in overall length and time aligned at the onset of the word "Ready."

For the present study we used 64 sentence versions, as recorded by each of the four female talkers (CRM talkers 04-07). The 64 sentences included all possible combinations of four call signs (Charlie, Ringo, Laker, Hopper), four colors (blue, red, white, green), and four numbers (1, 2, 3, 4).

C. Anechoic room and loudspeaker arrangement

All tests were conducted in a 3.0-m-wide × 4.3-m-long × 2.4-m-high anechoic room (IAC 107840). A subject chair was placed at approximately the middle of this room. Two loudspeakers were also placed in the room. One of the loudspeakers was directly in front of the subject chair and 1.5 m away. The other speaker, also 1.5 m away, was 60° to the right. The loudspeakers were set at the approximate ear height of a seated subject. Subjects always sat facing straight ahead.

Stimulus presentation. On test runs referred to here as *Front-Only*, the target speech and the distractors were mixed and presented exclusively from the front loudspeaker location. This was the location that subjects were asked to monitor throughout.

On runs referred to as *Front+Right* the target and distractors were presented from the front speaker, as for Front-Only runs, and in addition the distractors were time shifted and presented from the right-hand loudspeaker.[1] The time-shift associated with the right speaker presentation in this experiment was fixed at +4 ms (where + indicates right speaker leading), to agree with a number of previous studies and to invite precedence effect capture of the distractors and their localization off to the subject's right-hand side.

D. The task

Subjects were instructed to monitor the front loudspeaker throughout testing, always listening for sentences that featured the call sign *Laker*. A sentence with this call sign was presented at the front location on every trial. It could be spoken by any of the four female talkers, and it could feature any of the four colors and any of the four numbers. All of these varied randomly from trial to trial. The subject's task on each trial was to monitor for the target call sign and then to report the color/number combination in the sentence spoken by the talker who addressed *Laker*. Judgments were reported via a custom-designed response box.

Testing was done in runs of 35 trials each. The first 5 trials were treated as practice (without feedback). A percent correct score was calculated based on performance over the remaining 30 trials of the run, with a correct response requiring both the correct color and the correct number. The chance rate for performance of this task is one in sixteen (6.25%).

E. Number of distractors and S/N

In all, there were 24 test conditions in the experiment, based on parameter settings for the number of distractors and for S/N as detailed in the following, and on whether the stimulus presentation was Front-Only or Front+Right. Settings on these parameters were fixed within a test run, and varied randomly across the runs. Each subject completed a total of three runs for each combination of the settings. The ordering of these runs was random and different for every subject.

Prior to the start of formal testing, subjects were given several sample runs to familiarize them with the testing procedures. They received no feedback about response accuracy on these sample runs.

1. Number of distractors

The number of distractor talkers competing with the target talker on a test run was set at one, two, or three. A new set of sentences was generated at the start of each trial of the run, based on this parameter. First, the four female talkers available in CRM were randomly assigned to the target sentence and to the distractor sentences, with the constraint that each talker could be used only once. Next, colors and numbers were randomly assigned to the different sentences, also with the constraint that they could be used only once. Finally, the appropriate sentences were accessed from the CRM database and level adjusted, time shifted, and digitally mixed, as called for. The resulting signals were delivered via a two-channel D/A converter (sample rate=40 ksps), amplifier, and single-driver loudspeakers (Minimus 3.5).

2. S/N ratio

The rms presentation level of individual distractor sentences was fixed in the experiment at 65 dB SPL. Whenever one, two, or three distractor sentences were presented on a test, each sentence was presented at this level. Whenever leading and lagging copies of a distractor sentence were presented from the front and right-hand loudspeakers, both copies were presented at this level. S/N ratio was manipulated by varying the level of the target sentence presentations relative to the fixed level of the individual distractors. Across conditions, targets were presented at levels of 53, 57, 61, 65, and 69 dB, yielding S/N ratios of −12, −8, −4, 0, and +4 dB.[2] For the one-distractor test conditions, runs were conducted at the four lowest values of S/N (−12, −8, −4, and 0 dB). For the two- and three-distractor tests, runs were done at the four highest values (−8, −4, 0, and +4 dB).

F. Results and discussion

Figure 1 shows the results of Experiment 1. Percent correct scores (based on a total of 90 trials per condition per subject completed over three runs) are reported as a function of S/N ratio. There are separate plots for the Front-Only (open symbols, dashed line) and Front+Right (closed symbols, solid line) conditions. Data points are for the average subject ($n=4$). Error bars correspond to the standard deviation over subjects. Results are broken out in different panels according to the number of distractors.

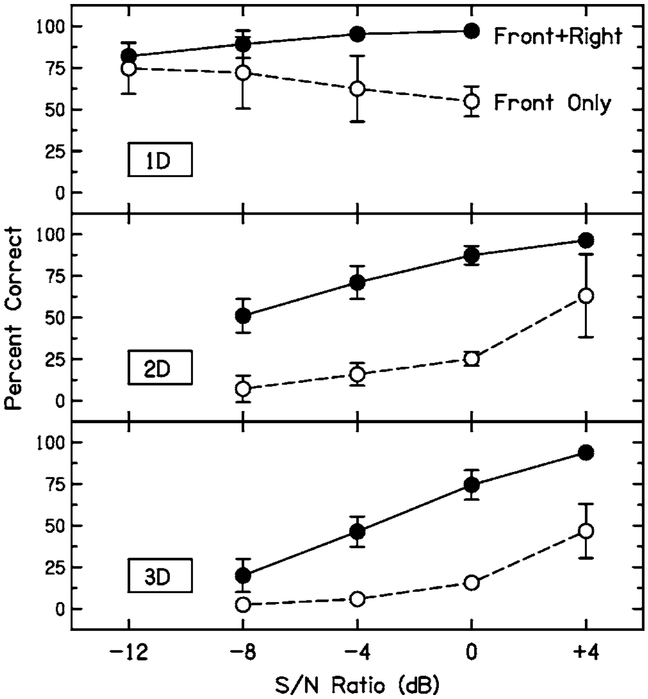

FIG. 1. Results of Experiment 1. Percentage of correct responses to a target presented from a front speaker for two conditions: *Front Only*, where distractors were presented from the same front speaker as the target; and *Front+Right* where distractors were presented from the front speaker and again from a right-hand speaker, with the right leading by 4 ms. The mean score and standard deviation over subjects are plotted as a function of S/N ratio. Top, middle, and bottom panels show results for interference from one distracting talker (1D), two distractors (2D), and three distractors (3D), respectively. Error bars are 2 s.d. in overall length.

1. One distractor

The top panel shows the results for the case of one distractor. There are two findings of note. The first is that Front+Right presentation was found to afford an advantage over Front-Only presentation at every S/N ratio. Hence, with one distractor there was consistent evidence of release from masking for speech.

Also notable is the finding that with one distractor the Front-Only function had a negative slope; that is, performance was inversely related to S/N. For example, at a S/N of 0 dB subjects correctly identified the target color/number combination on 55% of trials, but when the target speech level was reduced by 12 dB (S/N=−12), performance improved significantly to 75% correct [$t(3)=5.15; p<0.02$]. Brungart (2001) also found instances of reversal of this kind, as well as instances of plateauing where performance held constant as S/N was reduced over a considerable interval. Brungart (2001) reports that both such effects can occur on tests where: (a) a listener must attend to one talker's speech and ignore that of a competing talker with similar vocal characteristics; and (b) S/N is reduced over the interval from 0 dB to approximately −10 dB (Egan *et al.*, 1954; Dirks and Bower, 1969; Freyman *et al.*, 1999). The level difference between the two talkers' speech is a cue that aids a listener in differentiating the messages within this regime. The size of that level difference increases as S/N decreases below 0 dB. So long as the S/N degradations do not become too severe, this strengthening level-difference cue can support good listener performance.

2. Two distractors

Results for two distractors are shown in the middle panel of Fig. 1. They show a very clear advantage for two-source presentation at all S/N ratios tested (−8 to +4 dB), with performance for Front+Right conditions exceeding comparable Front-Only conditions by as much as 62 percentage points (two distractors, S/N=0 dB).

The results for two distractors do not show the negative S/N effect for Front-Only presentation that was found with one distractor. Brungart *et al.* (2001) reported a similar result. Reversal/plateauing effects clearly present in their data for one distracting talker were not found with two or three distractors. Apparently when the number of distractors exceeds one a listener can no longer reliably distinguish the target talker's message based on overall level difference. Instead, the limiting factor becomes S/N.

The S/N values corresponding to 50%-correct performance of this task were approximately +3.0 dB for Front-Only presentation and −8.0 dB for Front+Right presentation. Hence, repeated presentation of the distractors afforded 11 dB of masking release in the two-distractor condition. This compares with an estimated release of 10 dB reported by Freyman *et al.* (2001) for tests done with two distractors.

3. Three distractors

Results for three distractors are shown in the bottom panel. They are similar to those for two distractors, with performance reduced somewhat overall. For the Front-Only condition, performance fell to near the chance-level floor (6.25%) for S/N's less than 0 dB. A comparison of the 50% points of the functions (extrapolated for F-Only) shows a masking release of approximately 8 dB with three distractors (+4.5 for F-Only vs −3.5 dB for Front+Right), which is 3 dB less than the release found here with two distractors. This difference between the two and three distractor results is consistent with a recent report by Freyman *et al.* (2004), who found maximal masking release with two distractors and progressively decreasing release thereafter as the number of distractors was increased from three to ten.

4. Overall

Overall, the results of Experiment 1 showed robust evidence of release from speech masking whenever masker signals were presented simultaneously from two locations in the horizontal plane—one front and one right—with the right-hand presentation shifted ahead in time by 4 ms to trigger a strong precedence effect.

III. EXPERIMENT 2: DIFFERENT DELAYS

Previous investigations of release from speech masking, including Experiment 1 of this paper, have employed a distractor delay time constant of 4 ms, which can be expected to elicit a substantial precedence effect for speech. Experiment 2 examined a broad range of delays, to include this 4 ms value as well as a number of values substantially briefer than this and a number substantially longer. Following Freyman *et al.* (1999), the experiment tested both positive delays, where the distractor speech was time shifted to lead in the right-hand speaker relative to the front, and negative delays, where the distractors were shifted to lag in the right-hand speaker.

A. Methods

1. Subjects

There were six subjects in this experiment. Four of the subjects (S1–S4) were the same individuals who participated in Experiment 1. They completed all testing for that experiment before beginning this one. Two new subjects (S5 and S6), one male (age 22), one female (age 24), were tested here as well. Both of the new subjects were experienced listeners. Both had hearing thresholds within normal limits. Neither had any knowledge of the purpose of this study.

2. Stimuli

As in Experiment 1, the target (call sign Laker) was always presented from the front speaker together with distractors, and an additional copy of the distractors was optionally presented from the right-hand speaker with an intervening time delay. The number of distractors and S/N were both fixed for this experiment, based on the results of Experiment 1. The number of distractors was fixed at two and the S/N was fixed at −4 dB. In Experiment 1, this combination was found to yield a substantial separation in performance between the Front-Only (16% correct) and Front+Right (71% correct) conditions, with no evidence of either floor or ceiling effects.

3. Delays and test runs

Front+Right test runs were done in Experiment 2 at each of 16 different delays. The delays tested were −64, −32, −16, −4, −0.5, −0.25, 0.0, +0.25, +0.5, +1, +2, +4, +8, +16, +32, and +64 ms, where negative delays correspond to time shifts in which the right-hand loudspeaker presentation lagged relative to the front speaker, and positive delays to shifts where the right-hand loudspeaker led. In all, subjects completed a total of three runs at each delay. They also completed a set of three runs in a Front-Only condition to establish a baseline comparison for this experiment. The order of these runs was random, and different for every subject.

B. Results and discussion

1. Front-Only tests

Figure 2 shows the results of Experiment 2. In Front-Only testing with two distractors and a −4-dB S/N, the six subjects averaged 15.9% correct. The hatched area in Fig. 2 brackets a 95% confidence interval around this mean score for comparison with the various Front+Right conditions.

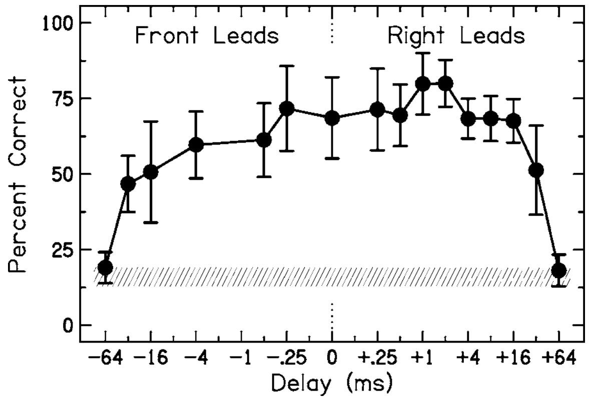

FIG. 2. Results of Experiment 2. Percent-correct mean and standard deviation over subjects, plotted as a function of the delay between the front and right loudspeakers. For negative delays, the distractors' speech led at the front speaker, i.e., at the target speaker. For positive delays it led at the right speaker. The hatched area shows a 95% confidence interval around the mean score (15.9%) when stimuli were presented from the front speaker only.

2. Front + Right tests

Results for the Front+Right tests are plotted as a function of delay. Data points correspond to the subjects' mean percent correct score. Error bars show ±1 s.d. over subjects. Performance with Front+Right presentations fell into two regimes, depending on the duration of the delay.

For the two longest delays tested, one forward in time (+64 ms) the other backward (−64 ms), there was no evidence of release from masking, with scores in both instances falling well within the 95% confidence interval for Front-Only presentations.

For all of the remaining tests, which were done at shorter delays ranging from −32 to +32 ms, there was consistent evidence of release from masking for speech. Scores were substantially greater than for the Front-Only condition in every instance, always falling well outside the Front-Only 95% confidence interval.

An analysis of variance on the results of Experiment 2 showed a significant difference among the conditions overall $[F(16,80)=58.86; p<0.001]$. Paired comparisons (Bonferroni-protected, $p<0.05$ experimentwise) between the Front-Only condition and each of the Front+Right conditions with time delays between −32 and +32 ms found a significant difference in every case.

C. Conclusions

The present findings support several new conclusions regarding speech masking release elicited by adding a delayed masker at a different location in the horizontal plane.

1. A wide-ranging effect

The first conclusion is that the effect is quite general. Significant release from masking was seen here for delays brief enough to fall within the regime of summing localization, where the leading copy of the distractors and the lagging copy both contribute substantially to perceived location. Summing localization occurs for delays less than 1 ms (Blauert, 1971). The present study found masking release for five such delays (−0.5, −0.25, 0, +0.25, and +0.5 ms). Significant masking release was also found here for a family of longer delays for which the precedence effect is fully engaged and sound localization depends mostly on cues associated with the leading presentation only (Wallach *et al.*, 1949; Litovsky *et al.*, 1999). Altogether, the set of delays found to exhibit significant release from speech masking in this study ranged from −32 to +32 ms. This spans the full range of acoustical reflection delay times that a listener is likely to encounter on an everyday basis in rooms.

2. Release at negative delays

A second conclusion to be drawn from the results of Experiment 2 is that masking release for speech is substantial for negative delays where the repeated distractors lag the distractors presented with the target. The effect for negative delays was found to reach out to at least −32 ms. This confirms and extends the report of Freyman *et al.* (1999), who found significant masking release at −4 ms. They pointed out that at −4 ms listeners may have been aided in distinguishing targets from distractors by a relative difference in diffuseness of the images, with images of the distractors being the more diffuse. In the present study, diffuse distractor images were also noted by listeners at negative delay times.

A temporal asymmetry. Figure 2 shows some evidence of an asymmetry between percent-correct scores obtained with negative delays and those obtained with positive delays, with the former generally somewhat lower. For example, the subjects' mean percent-correct score was 8.6% lower at −4 ms than at +4 ms (59.7% vs 68.3%), and it was 16.9% lower at −16 ms than at +16 ms (50.7% vs 67.6%).[3] We examined the individual subject records to learn more about the details of this asymmetry.

Figure 3 shows each subject's percent-correct scores for the Front+Right tests, plotted as a function of delay. Also shown, for comparison, is the score obtained in the baseline Front-Only condition (the hatched area). The error bars (and the 95% confidence interval for Front-Only) are based on a subject's standard deviation over test runs (3 runs per condition). Quintic functions, indicated in the figure by the solid-line plots, were fitted to each subject's delay series. These functions accounted for over 90% of the variance in the individual subject data. The area under each function was integrated over all negative delay times and over all positive delays to compare the strengths of the perceptual effects in these two regimes.

For one subject (S1), the negative and positive areas differed by less than 1%. But for the other subjects (S2 through S6) the areas differed by 5% or more, and in every case the negative area was smaller than the positive area. A statistical comparison based on the results for all six subjects showed the negative and positive areas under the functions to be significantly different $[t(5)=3.77;\ p<0.01]$, with the negative area 8.8% smaller on average. It appears that negative delay times, although effective at eliciting significant masking release relative to baseline (Front-Only), are nevertheless somewhat less effective than the corresponding positive delays.

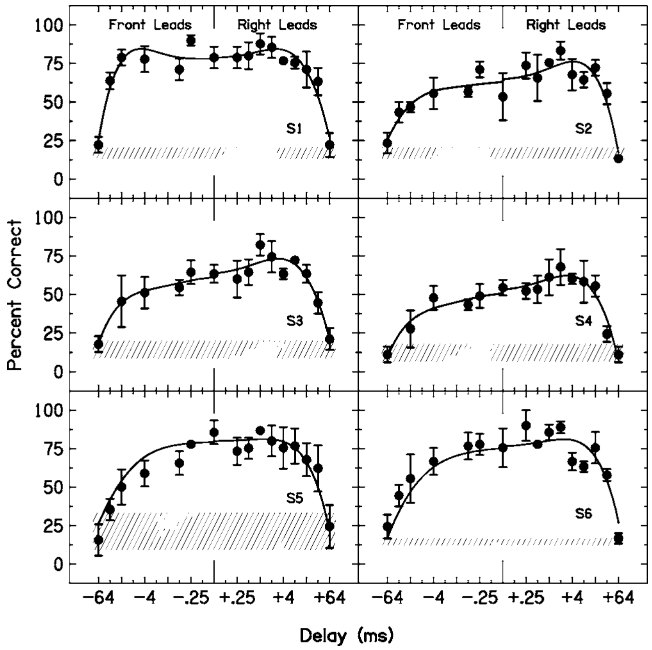

FIG. 3. Individual subject results for Experiment 2. Percent-correct mean and standard deviation over runs ($n=3$), plotted as a function of the delay between the front and right loudspeakers, and fitted with a best-fit fifth-order function (solid line). The hatched area shows a 95% confidence interval around the mean score when stimuli were presented from the front speaker only.

3. No release at very long delays

Sixteen different delay times were tested in this experiment and evidence of significant masking release was found for all but two of them. For both of the failing conditions (−64 and +64 ms) there was a 64-ms time delay between leading and lagging presentations of the distractors' speech. If acoustical reflections of speech are delayed by 64 ms in a room there is a good probability that listeners will become aware of them as echoes, so much so that the reflections may become disruptive (Haas, 1951; Kutruff, 1979, pp. 174–177). The present results suggest that once speech reflections become noticeable in this way they can no longer participate in speech masking release.

IV. EXPERIMENT 3: SPEECH-SPECTRUM NOISE MASKERS

In Experiments 1 and 2, masking release was found for a number of tests in which speech maskers were presented from two locations in the horizontal plane (front and right) with an intervening time delay. This masking release could have taken either or both of two forms. First, it is possible that there was release from energetic masking (EM), which takes place at levels as low as the auditory periphery. Second, alternatively or additionally, there could have been release from informational masking (IM), which takes place more centrally.

Freyman *et al.* (1999) conducted an extensive search for EM release. They fixed masker-signal delay at each of two values, +4 and −4 ms, and then replaced their speech maskers with continuous noise. This noise had a power spectrum that resembled the speech, and was therefore subject to a similar level of EM release if any was present. The experiment found no evidence of EM release, however. To the contrary, subjects performed somewhat worse with the noise in the Front+Right condition than in the Front-Only condition. Virtually all of the release observed with speech maskers in −4 and +4-ms tests could therefore be attributed to informational masking release only.

Experiment 2 of the present study found significant release with speech maskers for the delay times of −4 and +4 ms and for a number of additional delays that had not been previously tested. Experiment 3 was therefore conducted to see whether EM release could explain any part of the results obtained at these new delay times.

A. Methods

1. Matched noise maskers

The maskers for this experiment were 192 noise files, created to individually match each of the 192 CRM speech files that served as maskers in Experiments 1 and 2. For each CRM file, we created a noise file equivalent by: (a) Fourier transforming the entire original signal as a single function; (b) randomizing the phases of all components; and (c) inverse Fourier transforming the derived spectrum to generate the noise waveform. The resulting noise signals matched the CRM originals in duration, power, and long-term amplitude spectrum. Lacking were characteristic speech temporal envelopes and any speech-like information. Everyone agreed that the maskers sounded like a swarm of bees.

2. Subjects

There were four subjects in the experiment. All four (S1, S2, S4, S6) had previously participated in Experiment 2, and all but one (S6) had participated in Experiment 1 as well.

3. Procedure

The noise masker test was modeled on the two-distractor speech masker test described earlier. The methods were, in fact, identical except that the two randomly selected CRM speech masker files were replaced by their corresponding noise masker files on each trial. S/N for this experiment was set at −10 dB, based on pilot testing which showed that at this level subjects performed the baseline Front-Only test better than chance and also at a low enough level to allow headroom for a search for possible energetic masking release.

4. Front-Only runs

A set of baseline runs was conducted in the Front-Only condition with the number of distractors set at two, and the S/N set at −10 dB. Also, for a second point of reference, Front-Only runs were done with the S/N increased by 4 dB compared to baseline (S/N=−6 dB).

5. Front + Right runs

For the Front+Right condition, test runs were done at nine different values of time delay, covering the range from −32 to +32 ms where significant speech-masker release was

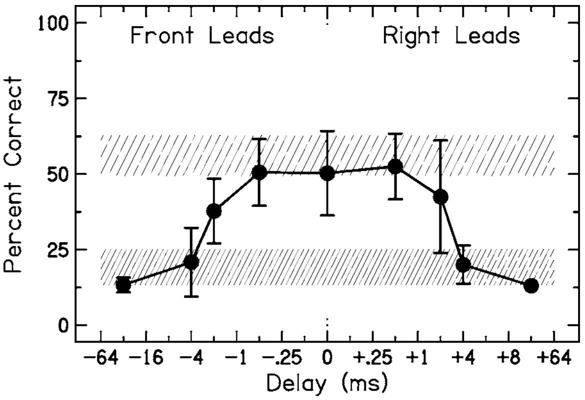

FIG. 4. Results of Experiment 3. The masker signals were noises with long-term spectra matched to the speech of Experiments 1 and 2. Closed circles connected by a solid line give percent-correct mean and standard deviation over subjects, as a function of the delay between the front and right loudspeakers. The hatched areas show 95% confidence intervals around the mean scores obtained when stimuli were presented from the front speaker only at a baseline level (S/N=−10 dB; hatched area at the bottom) and with the target signal level increased by four dB (S/N=−6 dB; hatched area at the top).

found in Experiment 2. Specifically, tests were done at delays of −32, −4, −2, −0.5, 0.0, +0.5, +2, +4, and +32 ms. Subjects completed three runs (90 total trials) for each test condition, with run order randomized differently for each subject.

B. Results and discussion

The results of Experiment 3 are given in Fig. 4. The lower hatched area shows a 95% confidence interval based on the four subjects' performance in the baseline Front-Only condition (−10-dB S/N). The upper hatched area gives a confidence interval around the mean score for the second Front-Only test, in which S/N was increased by 4 dB (S/N =−6 dB). Results for the Front+Right condition tests are plotted as a function of delay, with closed circles connected by a solid line.

Freyman *et al.* found no evidence of EM release for the delay times of −4 and +4 ms. The present results confirm that finding. We also found no EM release at the two longer delay times tested here, −32 and +32 ms. In all of these instances, the subjects' mean percent correct score fell well within or below the confidence interval for baseline Front-Only presentations (−10-dB S/N). It follows that when delay times are several milliseconds long or longer, in either a positive or a negative direction, subjects experience informational masking release only.

A very different pattern of results emerged here for short delays. For all five of the short delays tested (−2, −0.5, 0.0, +0.5, and +2 ms), the mean percent-correct score was above the Front-Only baseline, and in all but one case (+2 ms) the error bars were above it as well. An analysis of variance on these results was significant overall [$F(5,60)=12.91$; $p<0.001$], reflecting significant pairwise differences between the Front-Only baseline score and the scores obtained at −2, −0.5, 0.0, and +0.5 ms (Bonferoni-protected paired comparisons, $p<0.05$). A separate statistical comparison with the Front-Only condition in which S/N had been increased by 4 dB (the −6-dB S/N test), found the Front+Right mean score for the delay time of −2 ms to be significantly lower and the remaining scores to be not significantly different ($p>0.05$).

Altogether, these results indicate: (a) that there is consistent EM release for speech maskers when time delays are brief (2 ms or less); and (b) that the resulting masking release can be equivalent to an increase in signal level of 4 dB.

1. Individual differences

The EM release effects described here were, for the most part, very consistent across listeners. The only notable individual differences were at the delay times of −2 and +2 ms. Three of the four listeners showed diminished but still measurable release at one or both of these delays relative to the release seen at briefer delays (−0.5, 0.0, and +0.5 ms). For the fourth subject (S6) there was no diminution. Release at ±2 ms was the same as at the other brief delays.

2. Cause of release from energetic masking

Conjectures as to the origin of the EM release seen here begin by recognizing that when the masker is presented simultaneously from two different sources with one source delayed, and when the sources are differently located and the signals are diffracted around the head, there is an effective filtering of the masker. Therefore, we performed a calculation of the overall masker transfer functions at the two ears. Calculations were done using a spherical head model as described by Kuhn (1977) and implemented by Constan and Hartmann (2003). The model describes diffraction of plane waves around the head and automatically includes frequency-dependent interaural timing and level differences. For each ear, computations of complex transfer functions were made independently for front and right-hand sources. An additional delay ($\tau=+0.5$, 0.0, or −0.5 ms) was added to the right source, and the complex functions for the two sources were summed to obtain the transfer function at the ear.

Figure 5 shows the amplitude response functions (magnitude of the complex functions) for the left and right ears, depending on the value of τ (top panel: $\tau=+0.5$ ms; middle: $\tau=0.0$ ms; bottom: $\tau=-0.5$ ms). The low-frequency limit is 2.0 as expected, corresponding to the addition of two identical signals at a point in free field located at the center of the head. The low-frequency functional form is consistent with delay-and-add filtering where the delay is the sum of τ and a head-related delay. These amplitude response functions lead to the first conjecture about the origin of EM release. The functions show deep spectral valleys in the vicinity of speech formants caused by the interference effects from the speaker on the right. It is possible that the appearance of these valleys in the masker enabled listeners to hear out important features of the target talker. Further, the valleys at the left ear were normally different from those at the right, giving listeners twice the opportunity to hear out target features.

Figure 6 shows the interaural phase response, which leads to a second conjecture. It is possible that release from EM resulted from a binaural masking level difference of the

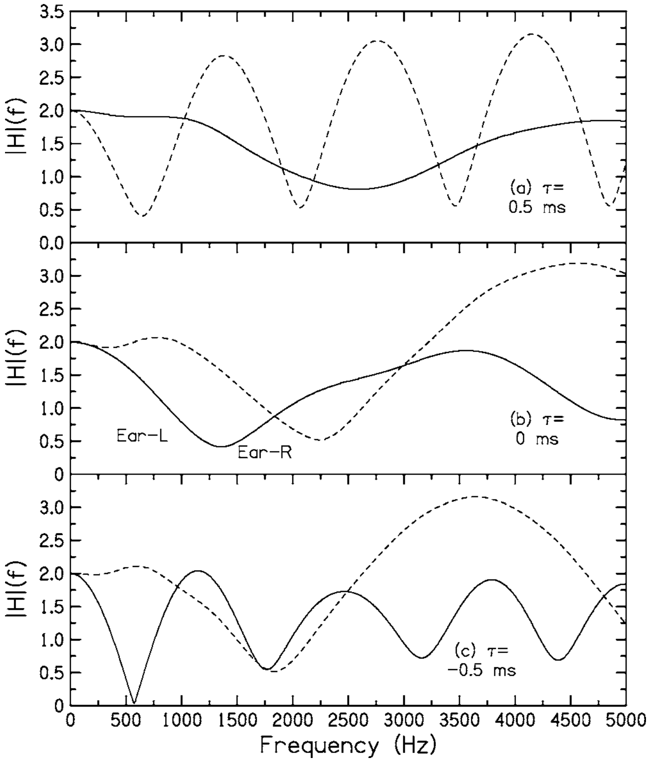

FIG. 5. Filter amplitude transfer functions for the masker delivered from two separated sources with time delay of +0.5, 0.0, or −0.5 ms, corresponding to test conditions where release from energetic masking was observed.

form $N_{\phi}S_0$. The figure shows that the interaural phase (Φ) of the maskers is near ±180° in a number of frequency regions, including the important region near 500 Hz, where binaural sensitivity is greatest. An interaural phase of ±180° corresponds to the condition $N_{\pi}S_0$.

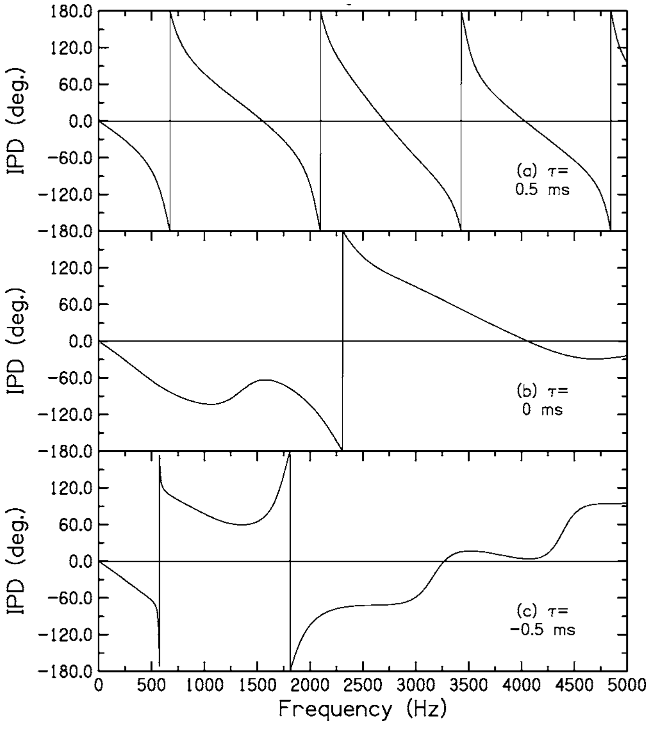

FIG. 6. Filter interaural phase response for the masker delivered from two separated sources with time delay of +0.5, 0.0, or −0.5 ms, corresponding to test conditions where release from energetic masking was observed.

Similar calculations for values of delay greater than ±0.5 ms show similar structure in both the amplitude responses and the interaural phase responses but the structure varies more rapidly with frequency. The above-noted conjectures would continue by suggesting that as the delays increase to 2.0 ms and beyond, the spectral or interaural features of the masker are confined to narrow, disconnected frequency regions that offer minimal advantage. Further investigation of these conjectures is beyond the scope of this report.

3. EM release and IM release

EM release of up to 4 dB was found at several different short delay times in this experiment. These effects, though consistent, were relatively small compared to the release seen in Experiment 1 with comparable speech maskers. There, we found 11 dB of masking release on a test done with two competing talkers, and 8 dB of release on a test with three talkers. It seems likely, therefore, that in speech-on-speech masking situations listeners will experience two kinds of masking release at very short delay times. There will be an EM release component of up to 4 dB, and an IM component that may be as large or larger.

V. SUMMARY

The CRM speech corpus was used here to test for release from speech masking when the masker was presented from two locations in the horizontal plane (front and right), with an intervening time delay. Key variables were (a) the number of distractor talkers who competed with a target talker, (b) the S/N for target and distractors, and (c) the delay associated with a multi-source presentation of the distractors' speech. Two experiments employed CRM speech maskers. One fixed the delay time at +4 ms and varied the number of distractors (1, 2, or 3) and S/N (−12, −8, −4, 0, or +4 dB). The other fixed the number of distractors (2) and S/N (−4 dB) and varied the delay over a wide range (16 different values between −64 and +64 ms). A third experiment was conducted with noise maskers matched to the spectra of CRM speech stimuli. It searched for energetic masking release at a range of delay times. The experiments support the following conclusions.

(1) Masking release occurs in a wide variety of listening conditions. For the most challenging conditions, such as multiple distracting talkers with negative S/N, the amount of release is of the order of 10 dB.
(2) Masking release is robust over variations in the masker delay time. Significant release was found here for delays spanning a wide range (−32 to +32 ms), consistent with the wide variation in acoustical reflection delays that is present in everyday rooms.
(3) Significant masking release takes places at negative delays, where the repeated distractors lag the distractors presented with the target. The diffuseness of distractor images may provide a cue that aids listeners in these instances.
(4) The upper limit on delays that elicit masking release falls somewhere between 32 and 64 ms. In 1927, Petzold de-

fined the "threshold of masking" as a delay for which a noticeable deterioration of the acoustical impression of speech occurs. The value was 50±10 ms. The conclusions of the celebrated paper by Haas (1951) confirmed the perceptual significance of 50 ms. As noted by Blauert (1983, p. 226), "At delay times less than 50 ms, echoes are no longer perceived as annoying even if the reflection is considerably stronger than the primary sound." The results of Experiment 2, wherein masking release was seen for a delay of 32 ms but not seen for a delay of 64 ms, are consistent with the idea that release from speech-on-speech masking fails when the delay of the delayed masker is long enough that the delayed masker no longer fuses with the primary masker and becomes annoying.

(5) For delay times longer than 2 ms, virtually all of the masking release that takes place for speech can be attributed to release from informational masking. For delay times of 2 ms or less, there is also a component of release from energetic masking.

ACKNOWLEDGMENTS

We thank Associate Editor Armin Kohlrausch and three anonymous reviewers of this manuscript for their many useful comments. We also thank the authors of the CRM for sharing this very useful stimulus set. This research project was supported by Grant No. DC00181 from the NIDCD of the NIH.

[1] The Front-Only presentation condition of the present study corresponds to a condition that Freyman *et al.* (1999) refer to as Front-Front (F-F). Our Front+Right condition corresponds to their Front-Right/Front condition (F-RF).

[2] The noise level reference for calculations of S/N (65 dB) conforms to a single distractor sentence presented from a single loudspeaker. Whenever multiple distractors were presented and/or whenever distractors were presented from two loudspeakers (front and right), noise power increased and S/N was effectively worsened.

[3] Both of these group differences were statistically significant [±4 ms: $t(5)=2.87$, $p<0.035$; ±16 ms: $t(5)=3.24$, $p<0.023$].

Blauert, J. (**1971**). "Localization and the law of the first wavefront," J. Acoust. Soc. Am. **50**, 466–470.

Blauert, J. (**1983**). *Spatial Hearing* (MIT, Cambridge, MA).

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

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

Brungart, D. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

Brungart, D., and Simpson, B. (**2002**). "Within-ear and across-ear interference in a cocktail-party listening task," J. Acoust. Soc. Am. **112**, 2985–2995.

Brungart, D., Simpson, B., Darwin, C., Arbogast, T., and Kidd, G. (**2005**). "Across-ear interference from parametrically degraded synthetic speech signals in a dichotic cocktail-party listening task," J. Acoust. Soc. Am. **117**, 292–304.

Brungart, D., Simpson, B., Ericson, M., and Scott, K. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527–2538.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech with one and two ears," J. Acoust. Soc. Am. **25**, 975–979.

Constan, Z. A., and Hartmann, W. M. (**2003**). "On the detection of dispersion in head-related transfer functions," J. Acoust. Soc. Am. **114**, 998–1008.

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

Egan, J. P., Carterette, E. C., and Thwing, E. J. (**1954**). "Some factors affecting multi-channel listening," J. Acoust. Soc. Am. **26**, 774–782.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**, 2112–2122.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2004**). "Effect of number of masking talkers and auditory priming on informational masking in speech recognition," J. Acoust. Soc. Am. **115**, 2246–2256.

Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3588.

Haas, H. (**1951**). "Uber den Einfluss eines Einfachechos auf die Horsamkeit von Sprache [On the influence of a single echo on the intelligibility of speech]," Acustica **1**, 49–58.

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

Kuhn, G. F. (**1977**). "Model for interaural time differences in the azimuthal plane," J. Acoust. Soc. Am. **62**, 157–167.

Kuttruff, H. (**1973**). *Room Acoustics* (Applied Science, London).

Leek, M. R., Brown, M. E., and Dorman, M. F. (**1991**). "Informational masking and auditory attention," Percept. Psychophys. **50**, 205–214.

Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. J. (**1999**). "The precedence effect," J. Acoust. Soc. Am. **106**, 1633–1654.

Moray, N. (**1959**). "Attention in dichotic listening: Affective cues and the influence of instructions," Q. J. Exp. Psychol. **11**, 56–60.

Petzold, E. (**1927**). *Elementare Raumakustik* (Bauwelt, Berlin), p. 8.

Shinn-Cunningham, B. G., Zurek, P. M., and Durlach, N. I. (**1993**). "Adjustment and discrimination measurements of the precedence effect," J. Acoust. Soc. Am. **93**, 2923–2932.

Wallach, H., Newman, E. B., and Rosenzweig, M. R. (**1949**). "The precedence effect in sound localization," Am. J. Psychol. **57**, 315–336.

Watson, C. S., Kelly, W. J., and Wroton, H. W. (**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. A. (**1997**). "The cocktail party problem: Forty years later," in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. H. Gilkey and T. R. Anderson (Erlbaum, Hillsdale, NJ), pp. 329–347.