Lexical and indexical cues in masking by competing speech

Karen S. Helfer and Richard L. Freyman
Department of Communication Disorders, University of Massachusetts Amherst, 358 North Pleasant Street, Amherst, Massachusetts 01003

(Received 21 March 2008; revised 30 October 2008; accepted 2 November 2008)

Three experiments were conducted using the TVM sentences, a new set of stimuli for competing speech research. These open-set sentences incorporate a cue name that allows the experimenter to direct the listener's attention to a target sentence. The first experiment compared the relative efficacy of directing the listener's attention to the cue name versus instructing the subject to listen for a particular talker's voice. Results demonstrated that listeners could use either cue about equally well to find the target sentence. Experiment 2 was designed to determine whether differences in intelligibility among talkers' voices that were noted when three utterances were presented together persisted when each talker's sentences were presented in steady-state noise. Results of experiment 2 showed only minor intelligibility differences between talkers' utterances presented in noise. The final experiment considered how providing accurate and inaccurate information about the target talker's voice influenced speech recognition performance. This voice cue was found to have minimal effect on listeners' ability to understand the target utterance or ignore a masking voice.
© *2009 Acoustical Society of America.* [DOI: 10.1121/1.3035837]

PACS number(s): 43.71.Es, 43.71.Bp, 43.66.Dc [MSS] Pages: 447–456

I. INTRODUCTION

Over the past several years, there has been increased interest in studying how listeners are able to understand one talker in the presence of competing conversations. Many of the studies examining this ability have, to some measure, attempted to quantify the relative contributions of two types of masking involved in these listening situations: energetic masking and informational masking. Energetic masking is interference that is produced when a competing signal uses peripheral resources that are necessary to process the target. Informational masking is generally thought to be caused by confusion between the target and masking signals and/or uncertainty regarding the target. Speech maskers have the potential to produce both of these types of masking. This paper will describe a new set of sentence stimuli (the TVM sentences) designed for research on energetic and informational masking in competing speech paradigms and will discuss results of three studies using these sentences.

A. Rationale for the development of the TVM sentences

Many recent investigations of speech perception in a competing speech environment have used the coordinated response measure (CRM) corpus (Bolia *et al.*, 2000) (e.g., Arbogast *et al.*, 2005; Kidd *et al.*, 2005a; Kidd *et al.*, 2005b; Wightman and Kistler, 2005; Brungart *et al.*, 2006; Rakerd *et al.*, 2006; Brungart and Simpson, 2007). These sentences have a number of features that make them particularly useful for studying informational masking. First, each CRM sentence begins with, "Ready *cue name*...," where the cue name is one of eight possible names such as Ringo or Baron. This provides a means of orienting the listener to the target sentence. Second, analyses of error patterns in studies using the CRM sentences suggest that, at least when the target and masker are presented at approximately the same level, informational masking (rather than energetic masking) limits performance (e.g., Brungart *et al.*, 2001). Moreover, because of their closed-set nature and because of the independence of the colors and numbers, the CRM sentences can be used multiple times within and across test sessions with no risk of the listener learning specific stimuli.

While these sentences have proven to be extremely useful for studying a variety of characteristics of speech-on-speech masking, communication interactions outside of the laboratory often involve understanding messages that are not restricted to a small set of alternatives. Given the same set of target stimuli, closed-set tasks are considerably easier than open-set tasks (e.g., Sumby and Pollack, 1954) and, in many instances, do not accurately simulate the demands of finding the target word in lexical memory (e.g., Sommers *et al.*, 1997; Clopper *et al.*, 2006). Although open-set speech materials are available [e.g., the Harvard IEEE corpus (Rothauser *et al.*, 1969), the hearing in noise test (HINT) sentences (Nilsson *et al.*, 1994), and the BKB (Bamford–Kowal–Bench) sentences (Bench *et al.*, 1979)], none of these stimulus sets has the feature of a cue word, which is desirable when conducting studies in competing speech situations. This suggests the need for a large open-set corpus of stimuli that retains the cue word feature.

This paper will introduce the TVM sentences, a new open-set stimulus corpus designed specifically for research questions involving competing speech, and will describe three studies using these stimuli. The focus of these studies was on what cues listeners are able to use to identify and attend to a target message in the presence of competing messages. The primary experimental conditions presented a target TVM sentence from one talker in the presence of two masking TVM sentences with different cue names spoken by other talkers. As with the CRM corpus, if subjects are told the target cue (which, for TVM sentences, is the name Theo, Victor, or Michael) before each trial, they can use that information to find and follow the target message. However, in order to do this, listeners must (1) hear the cue name and (2) somehow connect the key words in later parts of the sentence back to the cue name.

How do listeners make that connection? The most obvious possibility would seem to involve matching the voice reciting the cue name to the voice reciting the rest of the sentence, possibly assisted by additional matching of cue and key word loudness levels in cases where there are large target-to-masker level differences. Brungart *et al.* (2001) found that performance on the CRM stimuli was quite poor when target and masker were from the same talker, particularly at 0 dB target-to-masker ratio and below. Successful selective listening in competing speech may depend to a great degree on following a voice over time. Target voice information may therefore provide as much information as a semantic cue at the start of the target utterance.

B. Voice characteristics and speech masking

In order to successfully negotiate the challenge of listening to one utterance in the presence of other streams of speech, one must first determine the source of the target message and then attend to that signal while ignoring or deselecting the other messages. Faced with this task, an individual may use either semantic information (e.g., a certain word or phrase he/she is trying to find within a mixture of voices) or indexical information (the voice of the person to whom one wants to attend) to find the target utterance. Most of the research paradigms examining speech-on-speech masking use task instructions that direct the listener to a semantic cue, often a key word within the target signal.

Very little attention has been focused on the extent to which listeners can use indexical information in the talker's voice to identify and selectively attend to a target utterance. Experiment 3 in Brungart *et al.* (2001) examined whether knowledge of the target talker's voice (conveyed by blocking trials by the target talker) aided listeners' understanding of CRM sentences. On each trial subjects were instructed to listen for the sentence beginning with "Baron," which, in a given block, was always spoken by the same talker. They found that this indexical information provided no additional benefit besides cueing the listener to the sex of the target talker. That is, when the lexical cue name was available, subjects did not benefit from knowing the target talker's voice if the target and maskers were from same-sex talkers. This is consistent with the idea that the relative importance of any one cue used to distinguish a target utterance from a masking utterance likely depends on which other cues are available (e.g., Kidd *et al.*, 2005a). It is possible that listeners might have benefited from the voice or indexical information if the lexical cue name was not also presented.

It is well established that the amount of masking (both energetic and informational) produced by speech is related to the similarity of the speech target to the speech masker. In most situations, same-sex maskers produce greater masking (especially informational masking, which is related to confusion between the target and the masker) than do maskers produced by individuals differing in sex from the target talker (e.g., Festen and Plomp, 1990; Brungart, 2001; Brungart *et al.*, 2001; Darwin *et al.*, 2003). Moreover, intelligibility of a target utterance in the presence of a speech masker can be enhanced by introducing fundamental frequency differences between the masker and the target (e.g., Brokx and Nooteboom, 1982; Darwin *et al.*, 2003). Few studies have sought to quantify and account for the differences in masking effectiveness when the target and masker are spoken by people of the same sex. Some data suggest that certain voices appear to be more resistant to same-sex speech masking than others (e.g., Brungart, 2001). A recent study from our laboratory found substantial differences in the amount of informational masking produced by various combinations of two-talker female maskers (Freyman *et al.*, 2007). It is reasonable to assume that at least some of the variability in the amount of informational masking produced by specific voices is related to the listener's ability to differentiate the target voice from the masking voice(s).

Research has clearly demonstrated that voice information and phonetic information from an utterance are intertwined and are processed together. For example, a number of studies have documented a reduction in speech understanding ability when there is uncertainty regarding the talker from trial to trial as compared to performance measured when the same talker is used throughout the experiment (Mullennix *et al.*, 1989; Nygaard *et al.*, 1995). Performance on speeded classification tasks (where one must attend to a talker's voice while ignoring the lexical content of the message) shows that listeners cannot ignore irrelevant changes in one of these dimensions while attending to the other (e.g., Mullennix and Pisoni, 1990). Another piece of evidence of the connectivity of voice and phonetic information comes from investigations of the influence of talker familiarity on speech recognition. These studies show that listeners find it easier to perceive words in noise that are spoken by familiar (versus unfamiliar) talkers (e.g., Nygaard *et al.*, 1994). Hence, voice information appears to be an important component of speech recognition and, as such, may play a significant role in how well listeners can understand speech in adverse listening conditions.

One purpose of the studies described in this paper is to examine how listeners use lexical and indexical information in a competing speech situation. It has been decades since Broadbent (1952) reported that listeners can use information about the talker's voice to help resolve a single-talker-interference task. Since then, surprisingly little attention has been devoted to examining the degree to which listeners can use voice information in a competing speech task as a cue to identifying the target talker. The fact that trial-by-trial target talker uncertainty has a detrimental effect on speech perception in the presence of speech maskers (e.g., Brungart *et al.*, 2001) suggests that information about the target talker's voice is somehow used by listeners in competing speech situations.

This paper will first describe the development of the TVM sentence corpus. We then will summarize the results of three studies using the TVM sentence stimuli. The first study examined how the type of cue to which the subject was

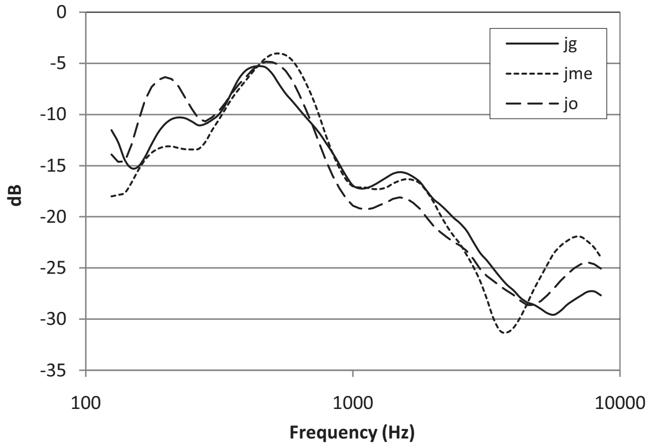

FIG. 1. Long-term third-octave smoothed spectra derived from a sample of 20 sentences from each talker.

asked to attend (key word or target talker voice) influenced speech recognition. Since differences in intelligibility between talkers' utterances were noted in the first study, a second experiment was conducted to determine the source of this talker variability. In this study, each sentence was presented in steady-state noise, and data analysis focused on identifying differences in intelligibility among the three talkers' recordings. The third study was designed to further examine the role of exposure to a voice on the ability to attend to or ignore a target or masking talker.

II. DEVELOPMENT OF THE TVM SENTENCE CORPUS

Each of the stimuli in the TVM corpus has the format, "*Name* discussed the ____ and the ____ today," where *Name* is the cue name *Theo*, *Victor*, or *Michael* and blanks correspond to one- or two-syllable nouns used for scoring. The cue names were chosen to be distinctive in both the auditory and the visual (i.e., lipreading) domains. The scoring words were common nouns, most of which were taken from the Thorndike–Lorge lists (Thorndike and Lorge, 1952). A total of 1080 unique sentences were created (360 beginning with each of the three cue names), with no scoring word repeated across the corpus. Examples of TVM stimuli are, "Theo discussed the swamp and the whisper today" and "Victor discussed the plant and the book today."

The sentences were recorded onto digital videotape from three male talkers who had no discernible regional dialect. Each talker was recorded saying each of the 1080 sentences. Recordings were generated in a sound-treated audiometric chamber. A remote microphone (Shure MX 183) was clipped to the talker's shirt approximately 6 in. below the mouth. The output of the microphone was routed to a preamplifier (PreSonus TubePre) and then sent to a digital video camera (Panasonic PV-DV953). Talkers were instructed to speak in a conversational manner but to attempt to put equal emphasis on the cue name and the two scoring words within each sentence. The studies described in the present paper used only the audio portions of these stimuli. Long-term third-octave smoothed spectra derived from the same 20 sentences from each talker are displayed in Fig. 1.

Stimuli were transferred to a PC for editing, storage, and presentation. Each sentence was saved in a separate file and scaled to produce utterances equal in rms amplitude. Three college-aged adults independently verified that the cue name and scoring words in each sentence were intelligible when heard in quiet. Sentences deemed unacceptable were discarded and then re-recorded.

III. EXPERIMENT 1: TASK INSTRUCTIONS AND SPEECH-ON-SPEECH MASKING

This study examined performance on the TVM corpus in a speech-on-speech masking task where listener instructions were manipulated. Specifically, on some trials subjects were instructed to repeat the sentence beginning with one of the cue names; on other trials subjects were given a preview of the target voice and told to repeat the sentence spoken by that talker. The primary purpose of this experiment was to compare listeners' use of these two types of cues (semantic and indexical) in both spatially coincident and spatially separated listening conditions.

A. Procedures

On a given trial, three sentences were presented simultaneously, each beginning with a different cue name and spoken by a different talker. One of these sentences was designated as the target sentence, while the other two were maskers. All three sentences began simultaneously and ended at approximately the same time (depending on the specific sentence length).

Testing took place in an IAC sound-treated room that has been used in previous competing speech experiments (Helfer and Freyman, 2005; Freyman *et al.*, 2007). The reverberation time in this chamber ranges from 0.12 s in the high frequencies to 0.24 s in the low frequencies (Nerbonne *et al.*, 1983). Our previous studies in this room have demonstrated similar amounts of spatial release from masking to that obtained in an anechoic chamber (e.g., Freyman *et al.*, 1999).

On half of the trials, the target and masking sentences were presented from a front loudspeaker located at 0° azimuth and at a distance of 1.3 m from the subject's head and at a height of 1.2 m from the floor (ear height for the average adult when seated); this spatial condition is hereafter referred to as F-F (for front-front). On the other trials (F-RF for front–right front), the target sentence was presented from the front loudspeaker, while the masking sentences were presented from the front and from a loudspeaker located 60° to the right of the subject at the same distance and height as the front, with a 4-ms time lead favoring the right. Due to the precedence effect, this spatial configuration produces the perception of the masker being located toward the right, well separated from the target. Comparing data obtained in the F-F and F-RF configurations allows us to identify the relative contributions of energetic and informational masking, as will be described in Sec. III B.

Target and masking sentences were mixed by summing the respective digital waveforms and then were presented from the computer's sound board. The stimuli were attenu-

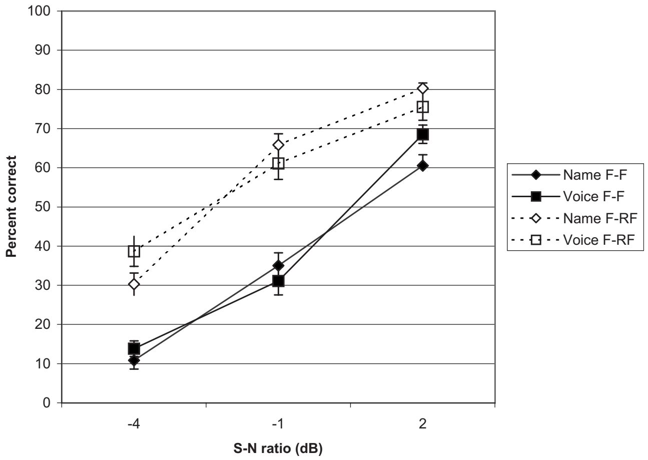

FIG. 2. Percent-correct recognition of key words in the TVM sentences. Filled symbols represent performance in the F-F (spatially coincident) condition, and open symbols display performance in the F-RF (spatially separated) condition. Performance for the name instruction mode is denoted by the diamond-shaped symbols; identification accuracy for the voice instruction mode is shown with the square symbols. Error bars represent one standard error.

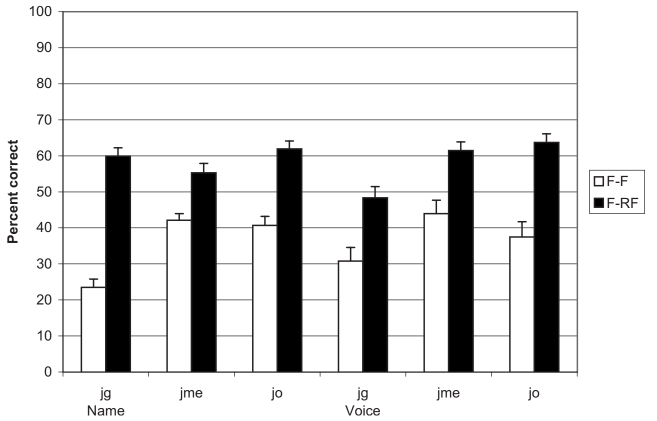

FIG. 3. Percent-correct recognition of key words in the TVM sentences with data aggregated by target talker and instruction mode. Open bars represent performance in the F-F (spatially coincident) condition, and filled bars display performance in the F-RF (spatially separated) condition. Data are averaged across S-N ratios. Error bars represent one standard error.

ated (TDT PA4), amplified (TDT HBUF5), and then power amplified (TOA P75D) before being sent to the loudspeaker(s) (Realistic Minimus 7). The target sentence was presented at a level of 56 dBA (re speech peaks).

Twelve conditions were run in this experiment: all possible combinations of two spatial configurations (F-F and F-RF), three signal-to-noise (S-N) ratios (−4, −1, and +2 dB), and two instruction modes. In the present study, the nominal S-N ratio was based on the level of the combination of the two maskers (i.e., a 0 dB S-N ratio was produced by presenting the target and the mixture of the two maskers at an equal level). No adjustment in the expressed S-N ratio was made for the F-RF condition even though the masker was presented from an additional loudspeaker (and, as a result, was 3 dB higher than in the F-F condition). In the *name cue* instruction mode, subjects were told to repeat the sentence beginning with a specified cue name (e.g., Theo, Victor, or Michael). This cue was presented to the subject (via text) on a screen of a laptop computer immediately before sentence presentation for a period of approximately 3 s. For the *voice cue* instruction mode, subjects heard a preview of the target talker's voice (saying, "this is the target sentence") prior to being presented with the three simultaneous sentences. They were instructed to repeat the sentence spoken by the target voice.

All variables were randomized on a trial-by-trial basis. Nine young normal-hearing subjects (age: 21–33 years; mean: 23.50 years) each participated in one session lasting approximately 2 h. Subjects heard 30 trials per condition. Data described below are based on 540 responses per condition (30 trials × 2 scoring words per trial × 9 listeners).

B. Results

Accuracy of identification of scoring words in the two instruction modes is displayed in Fig. 2. Several patterns can be observed in the data. First, there was little difference in performance between the two instruction conditions. It appears that listeners were, in general, equally adept at using semantic and lexical cues to resolve this listening task. Second, when three sentences were presented together, the TVM stimuli produced a substantial amount of informational masking. We assume that the F-RF condition, in which the masker is clearly heard in a location different from that of the target, produces little or no informational masking. Hence, our premise is that better performance in the F-RF condition versus the F-F condition indicates the presence of informational masking as differences in the amount of energetic masking (which is slightly greater in the F-RF condition since the masker comes from two loudspeakers rather than one) cannot explain this result (e.g., Freyman *et al.*, 1999). Repeated-measure analysis of variance (ANOVA) on these data [transformed into rationalized arcsine units (rau) (Studebaker, 1985)] confirmed these trends. Significant main effects were found for both S-N ratio [$F(2,7)=251.50$, $p<0.001$] and spatial condition [$F(1,8)=115.90$, $p<0.001$], but the main effect of instruction mode was not significant. The interaction of S-N ratio *x* spatial condition was significant [$F(2,7)=12.00$, $p=0.005$], as were the instruction mode *x* S-N ratio interaction [$F(2,7)=6.82$, $p=0.023$] and the three-way interaction [$F(2,7)=4.96$, $p=0.046$]. Hence, there was a small effect of instruction mode that depended on S-N ratio.

Examination of the data suggests that the voice cue was slightly more effective than the name cue at −4 dB S-N ratio, more so for the F-RF condition than for the F-F condition. The most parsimonious explanation for this finding is that at −4 dB S-N ratio, listeners might fail to correctly perceive the cue name (which was only briefly presented at the beginning of the target sentence), while they had a longer opportunity to capture the target talker's voice, which was available throughout the trial. For the F-RF spatial condition, the name cue was slightly more effective than the voice cue for the more advantageous S-N ratios. The reverse was true at +2 dB S-N ratio for the F-F condition.

Data were also analyzed to examine differences in performance among the talkers. Results of this analysis can be seen in Fig. 3. ANOVA on these data (averaged across S-N

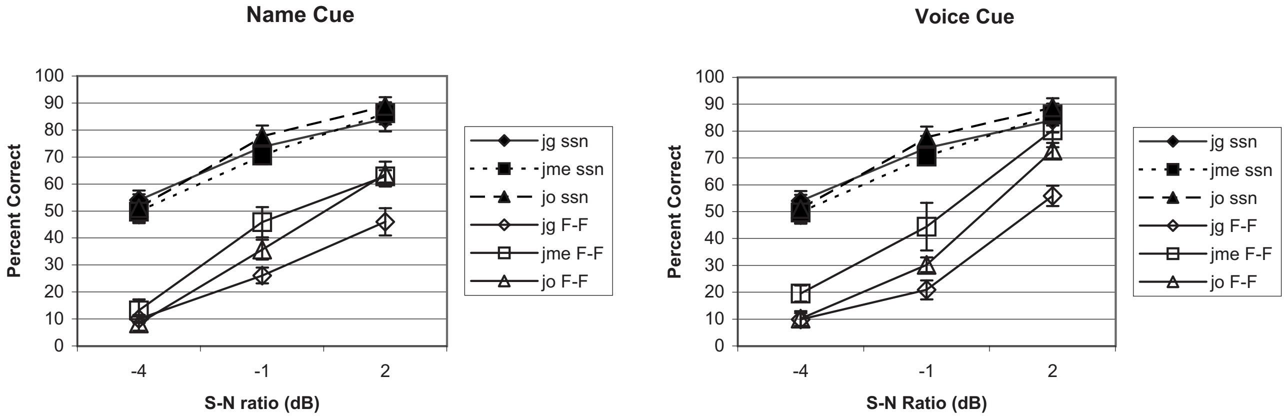

FIG. 4. Comparison of percent-correct identification of key words in the TVM sentences presented in SSN and in the presence of speech maskers. Filled symbols are data from experiment 2 (SSN), and open symbols are from experiment 1 (competing speech maskers) for the name instruction mode (left panel) and the voice instruction mode (right panel). Data are aggregated by target talker (jg, jme, and jo). Error bars represent one standard error.

ratios after being converted to rau) revealed significant main effects for talker $[F(2,7)=18.33,\ p=0.002]$ and for spatial condition $[F(1,8)=138.23,\ p<0.001]$ as well as a significant interaction between these two factors $[F(2,7)=6.74,\ p=0.023]$. There were no significant main or interaction effects involving instruction mode except for the three-way interaction, which just reached significance $[F(2,7)=4.81,\ p=0.049]$. When the utterances of these three talkers were presented simultaneously, talker JG was more difficult to understand than the other two talkers, and talker differences were greater in the spatially colocated F-F condition than when presented with spatial cues. This result suggests that voices that are equally intelligible when presented with what is assumed to be a purely energetic masker (i.e., the F-RF condition) may be differentially affected by informational masking (which is presumed to contribute in the F-F condition). *Post hoc* t-tests indicated that intelligibility of all three talkers was significantly different from one another during F-F presentation. For the F-RF configuration, talker JO was significantly easier to understand than the other two talkers; intelligibility differences between JME and JG were not significant in this spatial configuration.

IV. EXPERIMENT 2: TVM SENTENCE INTELLIGIBILITY IN SPEECH-SHAPED NOISE

The purpose of this experiment was to determine whether the talker differences noted in experiment 1 (with speech maskers) were also obtained when noise was used as competition. Because talker differences were greater in the first experiment in the spatially coincident F-F condition than when stimuli were spatially separated, we had reason to believe that informational masking played a key role in this pattern of results; in other words, stimuli from the different talkers varied in intelligibility in experiment 1 because their voices were more or less confusable. We were interested in determining the extent to which talker differences were seen when the masker was purely energetic.

A. Procedure

Each sentence was played in the presence of speech-shaped noise (SSN) that was derived from the TVM sentences. To produce the SSN, 15 of the scaled sentences from each talker were concatenated. A custom software program was used to extract the spectral envelope from this waveform and to shape a white noise with this derived envelope. Ten young college-aged listeners (age: 20–31 years; mean: 23.2 years) with normal hearing (verified via pure-tone screening) participated in this experiment. None of these subjects participated in experiment 1.

Testing was completed in an IAC sound-treated chamber. Across all subjects, each sentence from each of the three talkers was presented at three S-N ratios (−4, −1, and +2 dB). The experimental setup was identical to that used in experiment 1, with the exception that the speech and SSN stimuli were always delivered from the front loudspeaker. The speech stimuli were played at a level of 56 dBA (re speech peaks), and the noise level was adjusted to produce the desired S-N ratios.

Each subject listened to 324 sentences (36 sentences for each of the nine talker/S-N ratio combinations). The target talker and S-N ratio were varied on a trial-by-trial basis. Subjects were instructed to verbally repeat the target sentence, and an experimenter, seated in a control room, scored the sentences online. Testing took place in one session lasting approximately 1.25 h.

B. Results

Percent-correct performance for experiment 2 is displayed in Fig. 4; for comparison purposes, scores from experiment 1 are also plotted in this figure. The left panel of Fig. 4 shows performance in the presence of SSN compared to data from experiment 1 in the name cue conditions, while the right panel contrasts performance in SSN to that using the voice cue. Data collected in experiment 2 (converted to rau) were analyzed via repeated-measure ANOVA with talker and S-N ratio as the independent variables. The main effect of S-N ratio $[F(2,8)=33.36,\ p<0.001]$ was significant, but neither the main effect of talker nor the S-N ratio by talker interaction reached statistical significance. Hence, differences in intelligibility among the three talkers' recordings were very small (and nonsignificant) when the sentences were presented in steady-state noise. Sentence perception in the steady-state SSN (filled symbols) was better than performance in the presence of a two-talker masker (open symbols). This result is consistent with the speech masker producing informational masking.

Results of these two experiments demonstrate that talker differences that are not apparent in the presence of steady-state noise can be seen when three sentences are presented simultaneously, especially when target and masker are spatially coincident. Moreover, at least for the present recordings, listeners can use either indexical or semantic cues in speech-on-speech masking, although the extent to which these cues are effective appears to depend on the S-N ratio as well as on the particular talker. Our next experiment was designed to further examine subjects' use of indexical information in speech-on-speech masking.

V. EXPERIMENT 3: THE EFFECT OF A VOICE CUE ON TVM SENTENCE PERCEPTION

This experiment was an exploration of the effect of listeners' exposure to voice information. Experiment 1 showed that knowledge of the talker's voice helped the listeners locate and attend to the target talker. However, it is unclear whether the processing of voice cues in competing speech is automatic or mandatory. Data from previous research suggest that listeners have difficulty ignoring indexical information during speech recognition. For example, Mullennix and Howe (1999) presented a same- or different-voice prime either before or after a target word (presented in quiet at a low intensity level). They found that a different-voice prime degraded performance, but only when the listener was required to attend to the prime. Our experience with presenting a semantic prime prior to stimulus presentation (Freyman *et al.*, 2004) suggests that it causes the target utterance to "pop out" from within a mixture of other voices, but that study did not include a prime that provided only indexical or voice information. The present experiment was designed to determine the extent to which presentation of a voice cue prior to a trial automatically draws the listener's attention to that voice. Specifically, we were interested in determining whether giving a voice cue or prime from a *masking* talker draws the subject's attention to that voice and makes it difficult to ignore.

A. Procedures

The same equipment, spatial conditions, and S-N ratios used in experiment 1 were also used in this experiment. On each trial the listener was given the name cue (Theo, Victor, or Michael) corresponding to that in the target sentence. As in experiment 1, this cue was displayed on the screen of a computer immediately prior to stimulus presentation. On 1/3 of the trials (hereafter called name trials), subjects were given only this name cue. On 1/3 of the trials, listeners were also given the voice cue corresponding to the target talker's voice; that is, they heard a preview of the target talker saying "this is the target sentence" immediately prior to stimulus presentation (this will be referred to as congruent trials). On the remaining trials, subjects were given the name cue but also heard a voice cue from one of the two masking talkers (these will be called incongruent trials). Subjects were told that on any given trial, the voice preview may or may not be the target voice. Each of ten young college-aged subjects (age: 19–25 years; mean: 20.88 years) listened to an average of 30 trials in each of the 12 conditions (all combinations of three S-N ratios, two spatial conditions, and three trial types). None of these individuals had participated in experiment 1 or experiment 2. The precise number of trials per condition varied slightly from subject to subject, and conditions were randomized on a trial-by-trial basis.

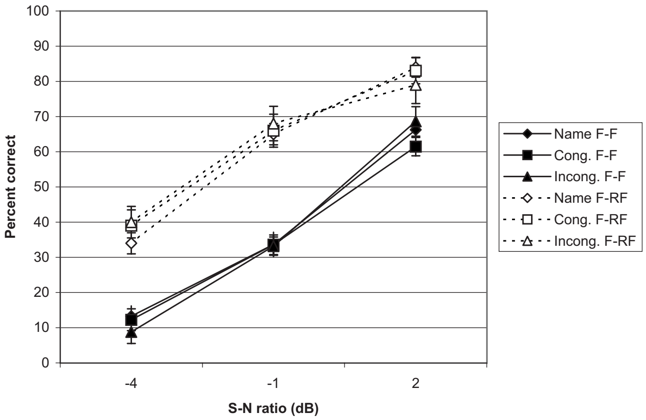

FIG. 5. Percent-correct identification of key words in the TVM sentences in the name, congruent (cong.), and incongruent (incong.) conditions. Filled symbols represent performance in the F-F (spatially coincident) condition, and open symbols display performance in the F-RF (spatially separated) condition. Error bars represent one standard error.

B. Results

One subject was not able to complete this experiment, and data from another showed clear outliers in the F-F condition. Hence, the analyses discussed below are based on data from eight subjects. Performance in percent correct for experiment 3 is shown in Fig. 5. It is clear that the presentation of an incongruent voice cue does not adversely affect performance. Averaged across S-N ratios and spatial conditions, performance for the three types of conditions was almost identical, between 49% and 50% correct. Repeated-measure ANOVA (on the data transformed into rau) confirmed this trend: significant main effects of spatial condition $[F(1,7)=102.70,\ p<0.001]$ and S-N ratio $[F(2,6)=505.42,\ p<0.001]$ and their interaction $[F(2,6)=8.36,\ p=0.018]$ were found, with insignificant main effects and interactions involving condition type. It appears that the presence of a congruent voice cue in addition to a name cue does not provide benefit, and the presentation of an incongruent voice cue does not hinder performance. It should be kept in mind that subjects could choose to ignore the voice prime since they were told that it might or might not contain useful information. Therefore, results of this study could be interpreted as suggesting that voice information is not automatically encoded while listening to speech. Although subjects were not asked about their subjective impressions, it is likely that providing a voice cue from a masking talker did not cause that voice to pop out, at least not to the extent that it interfered with the perception of the target sentence.

VI. TRENDS ACROSS EXPERIMENTS

A. Error patterns

We examined error patterns in the data from experiments 1 and 3 in terms of the types of responses listeners made when their perception was incorrect. Errors were classified into two categories. Nonconfusion errors were those in which subjects' responses were either errors of omission (e.g., the subject said "I don't know" or something similar) or (usually) close acoustic approximations to the target (e.g., "face" for "pace"). Confusion errors were those in which the incorrect response was a word from one of the maskers, which presumably reflects some confusion between the target and the maskers.

For each subject, the proportions of each error type (out of the total number of errors) were calculated for each condition, aggregated across S-N ratios and talkers. By far, the most frequent type of error in the data from both experiments 1 and 3 were nonconfusion errors. This type of error accounted for 78%–95% of all errors across listening conditions. It should be noted that the overwhelming prevalence of nonconfusion errors found here is quite different from that found with the CRM corpus, where most of the errors involve responses from a masker (Brungart, 2001; Brungart and Simpson, 2002; Brungart and Simpson, 2007; Kidd *et al.*, 2005a; Wightman and Kistler, 2005). The difference in error patterns most likely reflects the fact that the CRM is a closed-set measure, while listeners are not given alternatives when using the TVM sentences.

The proportion of nonconfusion errors was substantially higher in the spatially separated F-RF condition than in the F-F condition. In other words, subjects' confusions between the target and masker were more common in the spatially coincident F-F condition than when the target and masker were spatially separated, as shown in Fig. 6, consistent with results of studies using the CRM stimuli (e.g., Arbogast *et al.*, 2002; Arbogast *et al.*, 2005; Brungart and Simpson, 2002; Brungart and Simpson, 2007).

It should also be noted that error patterns varied substantially among listeners, which could reflect individual differences in response style and/or differences in perception. Particularly notable was the observation that some listeners made a relatively large number of confusion errors, while other subjects rarely responded in this way. For example, within the data from one target talker in one listening condition (F-F for name trials with talker JO in experiment 3), the percentage of confusion errors across listeners ranged from 6% to 38%. This could indicate that some listeners had a more difficult time than others in distinguishing the target from the masker. Conversely, it could also reflect differences in response bias, wherein some individuals were more likely than others to respond with a word that they heard within the mixture even if they were unsure of whether or not it came from the target, while other subjects chose not to respond when unsure.

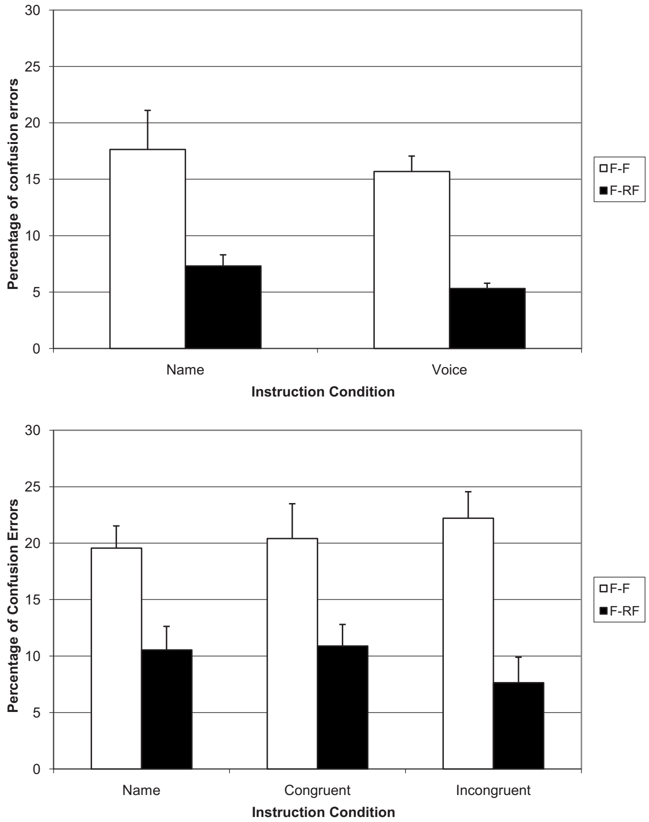

FIG. 6. Proportions of errors that involved responses from maskers in experiment 1 (top panel) and experiment 3 (bottom panel), aggregated by instruction condition in each study. Open bars represent performance in the F-F (spatially coincident) condition, and filled bars display performance in the F-RF (spatially separated) condition. Error bars represent the standard error.

B. Learning effects

Perceptual differences between target and masking voices may be used by listeners to resolve a competing speech situation. Listeners can learn to differentiate and identify voices with fairly high accuracy (e.g., Nygaard *et al.*, 1994), and the learning of voices can occur incidentally (that is, without conscious intention). We were interested in examining the extent to which incidental learning of the voices of the three talkers used in these experiments affected performance. We expected to see a small learning effect over the course of the experiment as subjects gained experience with the task and determined the effectiveness of certain strategies. Our premise was that listeners would become familiar with the three talkers' voices over the course of the experiment and that this exposure would contribute to a larger learning effect in certain conditions than in others. Specifically, we theorized that repeated exposure to voice information would provide greater benefit in conditions in which confusion between the target talker and masking talkers limited performance (that is, in the F-F condition) versus that in conditions in which informational masking is minimized (in the F-RF condition for speech maskers or when the masker was noise).

Performance on the first and last 25 trials (that is, the first and last 50 scored items, with two scored words per trial) presented to each listener for each condition (F-F and

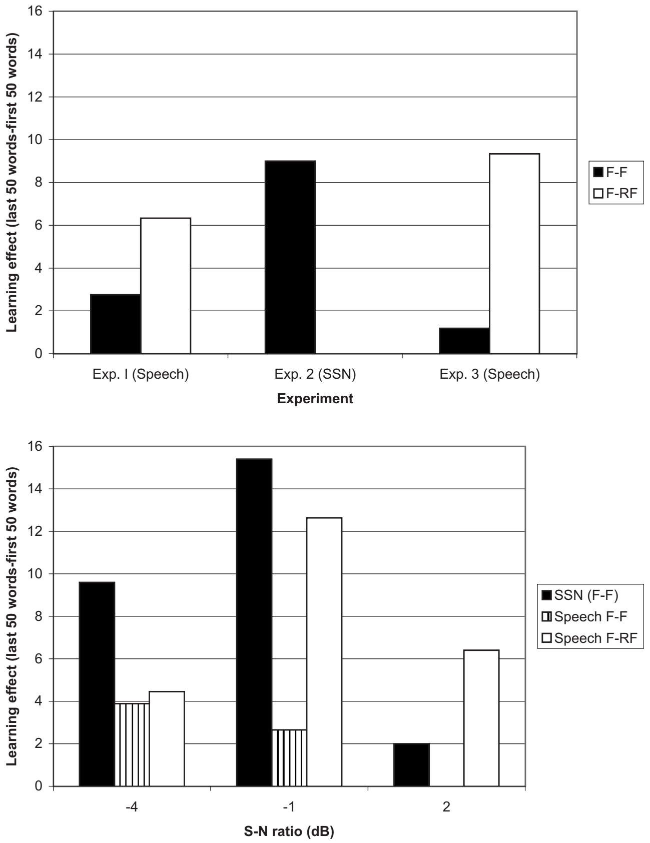

FIG. 7. The difference in recognition ability between the first 25 sentences (i.e., first 50 key words) and last 25 sentences (last 50 key words). The top panel displays data aggregated by experiment, averaged across S-N ratios. The bottom panel shows learning effects aggregated by condition (F-F/F-RF and S-N ratio) with data in the presence of speech maskers averaged across experiments 1 and 3.

F-RF for each of the S-N ratios) was compared for the three experiments. These data are shown in Fig. 7. Learning effects were greatest for conditions that were unlikely to cause informational masking (in the presence of the SSN masker and in the F-RF presentation mode with speech maskers) and smallest in the F-F condition with speech maskers.

These trends were confirmed with repeated-measure ANOVA (with the data transformed to rau). For experiment 2, the difference in recognition of the first versus last 50 target words was significant $[F(1,9)=35.06,\ p<0.001]$, as was the interaction between order and S-N ratio $[F(2,8)=9.72,\ p=0.007]$. Analysis of the data from experiment 1 also showed significant effects of order $[F(1,8)=12.46,\ p=0.008]$ and significant interactions between spatial condition and order $[F(1,8)=6.14,\ p=0.038]$ and between the three variables of spatial condition, order, and S-N ratio $[F(2,7)=12.48,\ p=0.005]$. ANOVA on data from experiment 3 revealed no significant main or interaction effects involving order.

Results of this analysis did not support the idea that learning of the talkers' voices would provide greater benefit in conditions where confusion between talkers is likely to play a role in performance. It is possible that learning effects could have been related simply to task difficulty rather than to spatial condition or masker. However, there was no systematic influence of S-N ratio on the order effect (see the bottom panel of Fig. 7), suggesting that a simple explanation of task difficulty cannot explain the data. This finding is somewhat contrary to a recent report of Van Engen and Bradlow (2007), who found learning effects in their task of speech-on-speech masking using native and non-native maskers. They did not analyze their data in terms of learning effects in maskers that did and did not produce informational masking, so a direct comparison between our results and theirs cannot be made.

VII. DISCUSSION

The studies described in this paper investigated the types of cues listeners use in a competing speech situation. Experiment 1 demonstrated that listeners can use both semantic information and indexical (or voice) information to identify and attend to a target message in a multitalker environment. The differences that were noted in listeners' ability to use these types of cues were small and depended on both the specific talker and the S-N ratio. There was some indication that the voice cue was slightly more effective than the name cue at the lowest S-N ratio and that the reverse was true at higher S-N ratios (although this trend was not entirely consistent).

The relative equivalence in performance when using the two types of cues (name versus voice) was perhaps an unexpected finding, as the ability to remember an unfamiliar voice and use that information to identify the target utterance is inherently a more difficult task than simply finding the sentence beginning with a specific name. Hence, it might be expected that performance using the voice cue would be poorer than that obtained using the name cue. It is possible that this result was not found because, at least at the most adverse S-N ratio, the cue name itself was masked to such an extent—during the brief period it was available—that listeners could not always determine the target utterance. The observation that the difference in instruction cues was larger in the F-RF than in the F-F condition supports this premise as energetic masking was greater in the F-RF trials (because the masker energy was 3 dB higher). In essence, a voice cue prior to sentence presentation may give the listener the same information as specifying a semantic cue at the beginning of a sentence, assuming that listeners are using the voice they hear reciting the cue name to find the stream corresponding to the target key words.

The recordings produced by the three talkers varied in intelligibility when presented in a multitalker situation but not when played in steady-state noise. Moreover, talker differences were larger when the voices were presented in the spatially coincident F-F condition than when spatially separated. Taken together, these two findings suggest that informational masking may contribute more than energetic masking to the differences in intelligibility found among the talkers' utterances.

It is possible that talker differences occurred primarily because the talkers' productions varied in intelligibility when presented together (that is, one of the voices "stood out" from the others or, conversely, was more confusable with one of the other voices). This result would be consistent with that reported by Brungart (2001), whose data suggest that some voices are more resistant to (or produce more) informational masking than others. It is also feasible that talker differences were caused by the relative masking effectiveness of the specific two-talker masker combinations. This explanation is consistent with data collected previously in our laboratory, where we found that different pairs of two-talker maskers produced greater differences in informational masking than in energetic masking (Freyman *et al.*, 2007). Because the same two-talker masker was always used for a given target voice in the present study, we are unable to tease apart these factors.

Although it was clear that listeners could use voice information to help identify and attend to the target, such information did not appear to be automatically encoded. Results from experiment 3 demonstrated that subjects could readily ignore a voice prime presented just prior to stimulus presentation. The presentation of a prime that was consistent with the target voice did not help the listener when the name cue was also available. This result agrees with data from Brungart *et al.* (2001), who also found that knowing the target talker's voice provided little benefit in same-sex masking conditions when the cue name was presented. Moreover, the presence of an incongruent prime did not hinder the subjects' ability to find and understand the target message. It is likely that because subjects were told that the voice cue might or might not be useful on any given trial, they chose to ignore it. If voice information is encoded automatically, it could be argued that subjects would not be able to ignore a voice cue, even if they knew it was irrelevant. Future research needs to be done to determine whether encoding of indexical information is mandatory in a competing speech situation.

One unexpected finding in the present studies had to do with learning effects. Listeners' performance did improve slightly over the course of the experiment, but more so in conditions that produced only or predominantly energetic masking (in the presence of steady-state noise in experiment 2 and in the F-RF modes in experiments 1 and 3). Learning effects were minimal in the F-F condition with speech maskers, which produced both informational and energetic masking. This result is in opposition to what we had anticipated: that learning or experience would play a greater role in reducing informational masking than energetic masking.

Although studies of informational masking using tonal stimuli (Neff and Callaghan, 1988; Neff and Dethlefs, 1995; Oxenham *et al.*, 2003) or speech detection (Balakrishnan and Freyman, 2008; Freyman *et al.*, 2008) have noted little or no evidence of learning effects, to our knowledge only one study (Van Engen and Bradlow, 2007) has examined the extent to which this finding persists in a competing speech recognition task. Van Engen and Bradlow (2007) did indeed note that experience with the task and/or the target talker's voice improved subjects' performance over the course of their study. Intuitively, repeated exposure to the talkers' voices should make it easier to differentiate one voice from another. Assuming that informational masking is caused (in part) by uncertainty regarding the target utterance versus the masking signals, enhancing the ability to differentiate between voices within the mixture should lead to reduced informational masking. We currently can offer no explanation for the pattern of learning effects found in the present study.

In summary, results of the studies described in this paper suggest several potentially important factors related to competing speech perception. Listeners can use either talker voice (lexical) or cue name (semantic) information to resolve this listening task; voices that do not differ in susceptibility to energetic masking may still vary in intelligibility when presented under conditions with informational masking; and incidental learning of the talkers' voices across an experimental session may play a greater role in reducing energetic masking than informational masking.

Finally, these studies suggest that the TVM sentences may prove to be a viable corpus for use in competing speech research. These stimuli have the advantage of incorporating a cue name but are open-set in nature, allowing the listener to use higher-level linguistic skills to aid in understanding. Perhaps because of their open-set nature, the TVM sentences are more difficult than the CRM sentences at equivalent S-N ratios and so may be appropriate for future research where closed-set stimuli such as the CRM are likely to be too easy, such as during auditory-visual presentation. Future studies will examine the extent to which individual sentences can be repeated across (and between) experimental sessions without the risk of remembering specific stimuli, as well as measure statistical properties of the corpus.

ACKNOWLEDGMENTS

We thank Sugata Bhattacharjee, Beth Ann Jacques, J. Ackland Jones, Amanda Lepine, and April Teehan for their contributions to this project. Portions of these data were presented at the Acoustical Society of America Conference, Providence, RI, June 2006. This work was supported by NIDCD Grant No. 01625.

Arbogast, T. L., Mason, C. R., and Kidd, G. K., Jr. (**2002**). "The effect of spatial separation on informational and energetic masking of speech," J. Acoust. Soc. Am. **112**, 2086–2098.

Arbogast, T. L., Mason, C. R., and Kidd, G. K., Jr. (**2005**). "The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **117**, 2169–2180.

Balakrishnan, U., and Freyman, R. L. (**2008**). "Speech detection in spatial and non-spatial speech maskers," J. Acoust. Soc. Am. **123**, 2680–2691.

Bench, J., Kowal, A., and Bamford, J. (**1979**). "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," Br. J. Audiol. **13**, 108–112.

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

Broadbent, D. E. (**1952**). "Listening to one of two synchronous messages," J. Exp. Psychol. **44**, 51–55.

Brokx, J. P. L., and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics **10**, 23–36.

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

Brungart, D. S., Iyer, N., and Simpson, B. D. (**2006**). "Monaural speech segregation using synthetic speech signals," J. Acoust. Soc. Am. **119**, 2327–2333.

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

Brungart, D. S., and Simpson, B. D. (**2007**). "Effect of target-masker similarity on across-ear interference in a dichotic cocktail-party listening task," J. Acoust. Soc. Am. **122**, 1724–1734.

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

Clopper, C. G., Pisoni, D. B., and Tierney, A. T. (**2006**). "Effects of open-set and closed-set task demands on spoken word recognition," J. Am. Acad. Audiol. **17**, 331–349.

Darwin, C., Brungart, D., and Simpson, B. D. (**2003**). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**, 2913–2922.

Festen, J. M., 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.

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., Balakrishnan, U., and Helfer, K. S. (**2008**). "Spatial release from masking with noise-vocoded speech," J. Acoust. Soc. Am. **124**, 1627–1637.

Freyman, R. L., Helfer, K. S., and Balakrishnan, U. (**2007**). "Variability and uncertainty in masking by competing speech," J. Acoust. Soc. Am. **121**, 1040–1046.

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

Helfer, K. S., and Freyman, R. L. (**2005**). "The role of visual speech cues in reducing energetic and informational masking," J. Acoust. Soc. Am. **117**, 842–849.

Kidd, G., Jr., Arbogast, T. L., Mason, C. R., and Gallun, F. J. (**2005a**). "The advantage of knowing where to listen," J. Acoust. Soc. Am. **118**, 3804–3815.

Kidd, G., Jr., Mason, C. R., Brughera, A., and Hartmann, W. M. (**2005b**). "The role of reverberation in release from masking due to spatial separation of sources for speech identification," Acta. Acust. Acust. **91**, 526–536.

Mullennix, J. W., and Howe, J. N. (**1999**). "Selective attention in perceptual adjustments to voice," Percept. Mot. Skills **89**, 447–457.

Mullennix, J. W., and Pisoni, D. B. (**1990**). "Stimulus variability and processing dependencies in speech perception," Percept. Psychophys. **47**, 379–390.

Mullennix, J. W., Pisoni, D. B., and Martin, C. S. (**1989**). "Some effects of talker variability on spoken word recognition," J. Acoust. Soc. Am. **85**, 365–378.

Neff, D. L., and Callaghan, B. P. (**1988**). "Effective properties of multicomponent simultaneous maskers under conditions of uncertainty," J. Acoust. Soc. Am. **83**, 1833–1838.

Neff, D. L., and Dethlefs, T. M. (**1995**). "Individual differences in simultaneous masking with random-frequency, multicomponent maskers," J. Acoust. Soc. Am. **98**, 125–134.

Nerbonne, G. P., Ivey, E. S., and Tolhurst, G. C. (**1983**). "Hearing protector evaluation in an audiometric testing room," Sound Vib. **17**, 20–22.

Nilsson, M., Soli, S. D., and Sullivan, J. (**1994**). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.

Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (**1994**). "Speech perception as a talker-contingent process," Psychol. Sci. **5**, 42–46.

Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (**1995**). "Effects of stimulus variability on perception and representation of spoken words in memory," Percept. Psychophys. **57**, 989–1001.

Oxenham, A. J., Fligor, B. J., Mason, C. R., and Kidd, G. (**2003**). "Informational masking and musical training," J. Acoust. Soc. Am. **114**, 1543–1549.

Rakerd, B., Aaronson, N. L., and Hartmann, W. M. (**2006**). "Release from speech-on-speech masking by adding a delayed masker at a different location," J. Acoust. Soc. Am. **119**, 1597–1605.

Rothauser, E. H., Chapman, W. D., Guttman, N., Norby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. (**1969**). "I.E.E.E. recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 227–246.

Sommers, M. S., Kirk, K. I., and Pisoni, D. B. (**1997**). "Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners I: The effects of response format," Ear Hear. **18**, 89–99.

Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455–462.

Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**, 212–215.

Thorndike, K. I., and Lorge, I. (**1952**). *The Teacher's Word Book of 30,000 Words* (Columbia University Press, New York).

Van Engen, K. J., and Bradlow, A. R. (**2007**). "Sentence recognition in native- and foreign-language multi-talker background noise," J. Acoust. Soc. Am. **121**, 519–526.

Wightman, F. L., and Kistler, D. J. (**2005**). "Informational masking of speech in children: Effects of ipsilateral and contralateral distracters," J. Acoust. Soc. Am. **118**, 3164–3176.