3-1-2016

Effects of a consistent target or masker voice on target speech intelligibility in two- and three-talker mixtures.

Fabienne Samson

Department of Psychology, The Brain and Mind Institute, Natural Sciences Center, Room 227, The University of Western Ontario, London, Ontario, N6A 5B7, Canada

Ingrid S Johnsrude

Department of Psychology, The Brain and Mind Institute, Natural Sciences Center, Room 227, The University of Western Ontario, London, Ontario, N6A 5B7, Canada

Effects of a consistent target or masker voice on target speech intelligibility in two- and three-talker mixtures

Fabienne Samson, and Ingrid S. Johnsrude

Citation: The Journal of the Acoustical Society of America **139**, 1037 (2016); doi: 10.1121/1.4942589
View online: https://doi.org/10.1121/1.4942589
View Table of Contents: https://asa.scitation.org/toc/jas/139/3
Published by the Acoustical Society of America

Effects of a consistent target or masker voice on target speech intelligibility in two- and three-talker mixtures

Fabienne Samson[a] and Ingrid S. Johnsrude
Department of Psychology, The Brain and Mind Institute, Natural Sciences Center, Room 227, The University of Western Ontario, London, Ontario, N6A 5B7, Canada

(Received 26 November 2015; revised 10 February 2016; accepted 11 February 2016; published online 2 March 2016)

When the spatial location or identity of a sound is held constant, it is not masked as effectively by competing sounds. This suggests that experience with a particular voice over time might facilitate perceptual organization in multitalker environments. The current study examines whether listeners benefit from experience with a voice only when it is the target, or also when it is a masker, using diotic presentation and a closed-set task (coordinate response measure). A reliable interaction was observed such that, in two-talker mixtures, consistency of masker or target voice over 3–7 trials significantly benefited target recognition performance, whereas in three-talker mixtures, target, but not masker, consistency was beneficial. Overall, this work suggests that voice consistency improves intelligibility, although somewhat differently when two talkers, compared to three talkers, are present, suggesting that consistent-voice information facilitates intelligibility in at least two different ways. Listeners can use a template-matching strategy to extract a known voice from a mixture when it is the target. However, consistent-voice information facilitates segregation only when two, but not three, talkers are present. © *2016 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4942589]

[DB] Pages: 1037–1046

I. INTRODUCTION

Segregating a target voice from concurrent sounds in a "cocktail-party" environment is one of the most complex operations performed by the auditory system. This ability involves both sensory processes used to derive individual sound features, and cognitive mechanisms such as attention and working memory that help a listener attend to a target stimulus, ignore masking sounds, and extract sound meaning (Darwin, 1997). The task of understanding one voice that is competing with other speech signals is challenging because the waveforms of the target and masking sounds overlap in time and frequency and thus stimulate overlapping regions of the cochlea and auditory nerve; this phenomenon is referred to as "energetic masking" (Durlach, 2006). Masking can also occur because of the perceptual similarity between the target and masking signals; this is a form of "informational masking." When masking is informational, both the masker and the target are audible, but the listener is either unable to segregate the components of the target signal from those of the masker, or is unable to assign the uttered words to the target talker correctly (Brungart *et al.*, 2001; Durlach *et al.*, 2003; Kidd *et al.*, 2005).

When a masking voice is present, listeners rely on acoustic cues, such as differences in frequency, timbre, onset time, and cues to sound location in order to segregate sounds. For example, different-sex talkers are easier to segregate than same-sex talkers, since the acoustic characteristics of male versus female voices perceptually differentiate them (Brungart *et al.*, 2001), and the spatial separation of target and masking signals also provides substantial release from masking (Hawley *et al.*, 2004). The ability to report what a target talker is saying when a masking talker is present also improves when listeners can rely on non-acoustic cues such as previous knowledge or experience (Nygaard and Pisoni, 1998; Yonan and Sommers, 2000; Freyman *et al.*, 2004; Davis and Johnsrude, 2007; Johnsrude *et al.*, 2013). For instance, Freyman *et al.* (2004) observed significant release from speech-on-speech masking when listeners were exposed to the beginning of each target sentence prior to trial presentation, and then asked to identify the last (unprimed) word.

When trying to extract a target from competing speech signals, listeners can benefit from the consistent presence of a specific target talker. The coordinate response measure (CRM) procedure (Bolia *et al.*, 2000; Brungart *et al.*, 2001) is a common tool for intelligibility multitalker mixtures. The participant listens for a target call sign in a mixture, and reports the color-number coordinate to which that call sign was told to go. Target identification performance on the CRM task was better when listeners were provided with *a priori* information about the vocal characteristics of the target talker (i.e., when this voice was used as the target throughout an experimental block of 180 trials) compared to when the target voice changed from trial to trial (Brungart *et al.*, 2001). This was true for three- and four-talker mixtures and, to a lesser extent, for two-talker mixtures, with greater improvement in different-sex and mixed-sex than in same-sex configurations (Brungart *et al.*, 2001). Similarly, when 1 target and 12 masking speech signals were presented in a sequence of 13 partially overlapping timeslots randomly

[a]Also at: School of Communication Sciences and Disorders, University of Western Ontario, London, Ontario, Canada. Electronic mail: samsonfabienne1@gmail.com

assigned to 13 loudspeakers, listeners' speech-reception thresholds improved significantly when the voice identity or spatial location of the target talker was constrained across trials (Kitterick *et al.*, 2010). In another experiment, 5 sequences of 4 spoken digits from the TIDGIT database (Dehaene and Cohen, 2007; digits 1–9 recorded by 15 different male talkers) were presented simultaneously, with the digits in each temporal position of the sequence coming from 1 of 5 possible spatial locations. Listeners had to report the digits designated as target [cued by a lighted light-emitting diode (LED) on one of the loudspeakers; Best *et al.*, 2008]. In this experiment, listeners were better at extracting a target sequence when all four digits in the sequence came from the same spatial location compared to when the spatial location changed from one digit to the next, and this benefit was enhanced when the identity of the target talker was also held constant, compared to when it changed, between digits. In another study, performance on the CRM task improved when the location of the target talker remained fixed from trial to trial in two-, three-, and four-talker situations, compared to when it varied (Brungart and Simpson, 2007). Last, it was recently observed that when listeners were asked to report a five-digit sequence embedded in competing reversed digits spoken by different talkers, they perform better when the identity of the target talker remained the same across the sequence, compared to when it switched between successive digits, regardless of whether they were informed prior to the task that the target talker would be held constant (Bressler *et al.*, 2014).

Although research demonstrates that listeners can use the consistent presence of a *target* voice to better segregate and understand a target signal in multitalker environments, the effects of consistency of a *masker* voice are less well established. Studies of non-speech stimuli suggest that listeners can use prior knowledge about maskers to better extract information in complex auditory scenes. For instance, detection of a target narrowband tone-burst sequence embedded in multi-tone maskers was better when, on each trial, listeners were cued with the multi-tone maskers compared to when they heard a notched noise band (notch centered on the center frequency of the target) prior to the pattern-detection task (Kidd *et al.*, 2011). In multitalker situations, presenting a masking speech signal at an expected versus an unexpected spatial location can improve target speech intelligibility (Allen *et al.*, 2011), although this effect is not consistently observed (Jones and Litovsky, 2008). In a recent study examining the effects of voice familiarity on speech segregation using the CRM procedure (but not the CRM corpus of voices), listeners were significantly better at reporting coordinates from voices of strangers, age- and sex-matched to that of their spouse, when their spouse's voice was used as the masker in a two-talker mixture (Johnsrude *et al.*, 2013), compared to when voices from other age- and sex-matched strangers were used as maskers. This indicates that knowledge of the characteristics of a particular voice can be a useful cue to aid intelligibility, not only when that voice is the target, but also when it is the masker and outside the focus of attention.

Another study using the CRM procedure (and voice corpus; Brungart and Simpson, 2004) examined whether monaural target-detection performance was influenced by holding either the voice identity and/or the content of 1 or both masking voices (presented in the same or contralateral ear as the target) constant across blocks of 120 trials. They observed that performance improved when the content of the masking phrase in the target-ear was held constant, but no significant improvement was observed when the voice identity of the masking talker or talkers (one or two maskers, in two- or three-voice conditions) was constant across trials. Better performance was found when the voice identity of the maskers was constant only when the content of the phrases was also held constant, and for conditions where both the identity and content were constant for the two masking signals. These results suggest that listeners might not benefit from the consistent presence of a masking voice, at least in a dichotic task when they are specifically asked to attend to the target presented in one ear. A different pattern of results may be obtained in diotic tasks when participants have to attend to the entire mixture and isolate the target. The question therefore remains as to whether or not listeners benefit from the consistent presence of a masker voice in a diotic multitalker mixture consisting of either two or three voices.

Better intelligibility of speech due to experience with a target voice may be due to better segregation, or to more accurate, or more efficient, matching of the utterance to a learned template (Bregman, 1990). In contrast, better intelligibility of a random target when the masker voice is consistent from trial to trial would suggest an effect on sound segregation itself, since template matching is thought only to occur when the signal being matched is the focus of attention (i.e., the target; Bregman, 1990).

Here, we present two separate experiments that examine how the consistent presence of a particular target or masker voice influences comprehension of a target message in same-sex two-talker (experiment 1) and three-talker (experiment 2) diotic speech mixtures. We expect better performance (word report) when the target voice is held constant compared to when no voice is consistent over trials as previously observed (Brungart *et al.*, 2001; Bressler *et al.*, 2014). If masker consistency improves segregation, we would also expect better report of a non-constant target when the masker voice is constant, compared to when the masker varies over trials. Additionally, if voice consistency does facilitate comprehension, it would be helpful to know whether the consistent voice had to have the same role for a benefit to be realized. Accordingly, another experimental condition was defined in which one voice was held constant, but its role switched from target to masker across successive trials.

II. EXPERIMENT 1: EFFECTS OF THE CONSTANT PRESENCE OF A TARGET OR MASKER VOICE IN A TWO-TALKER MIXTURE

A. Method

1. Listeners

Twenty-five native English speakers (five males; three left-handed; age range 18–21 yr; mean age 19 yr), naive with respect to the test materials and task, participated. All

passed audiometric screening, with pure-tone thresholds over a range of frequencies (250–4000 Hz) in the normal range [group mean 4.1 dB hearing level (HL), range −1.9–12.5 dB HL]. This study was cleared by the Queen's University General Research Ethics Board, and informed consent was obtained from all participants.

2. Stimuli and procedure

On each trial, participants were asked to follow a target voice presented concurrently with one masking voice, using an adaptation of the CRM procedure (Bolia *et al.*, 2000). The sentences were of the form "Ready '*Call Sign*' go to '*Color*' '*Number*' now," and listeners had to indicate on a computer screen the color and number spoken by a target voice (i.e., talker uttering the call sign "Baron"). The masker voice always uttered a different call sign (either "Arrow," "Charlie," or "Eagle"). The color-number coordinates for the target and masker sentences were also always different and were randomly chosen from an array of four colors (white, blue, red, and green) and eight numbers (1–8). The response array at the end of each trial consisted of four colored rows of the numerals 1–8, and the participant indicated the correct color-number coordinate with a mouse click. In-house recordings (44100 Hz sampling rate, 16-bit resolution) from 12 male and 12 female talkers (age range 22–44 yr) were used as stimuli. 128 sentences (4 call signs, 4 colors, and 8 digits) from the CRM database were recorded from each talker and lasted about 3 s (2989 ms on average, 159 ms standard deviation) to maximize temporal alignment of words so that listeners would have to segregate the concurrent phrases to understand them.

The experiment took place in an Eckel (Morrisburg, Ontario, Canada) single-walled soundproof booth. Stimuli were presented through a RME (Haimhausen, Germany) Fireface 400 soundcard at a comfortable listening level [72–82 dB sound pressure level (SPL)], and were delivered diotically over Sennheiser (Wedemark, Germany) HD 265 headphones. MATLAB (The MathWorks, Natick, MA) (www.mathworks.com) was used to present the task and collect responses.

Four within-subject conditions were defined. In the target condition, the target voice was constant from one trial to the next, but the masker voice changed from trial to trial. In the masker condition, the masking voice was constant across successive trials, but the target voice changed from trial to trial. In the switch condition, one voice was constant across successive trials, but its role switched from target to masker and back again. Since different mechanisms may be involved in perception when the constant voice is the focus of attention (i.e., target) and when it is not (i.e., masker), the trials in the switch condition were assigned to switch_T (target constant) and switch_M (masker constant) conditions for the analysis. The other voice in the switch condition varied randomly from trial to trial. In the baseline condition, both voices were different on every trial. Over the course of the experiment, each participant heard all 24 recorded voices. Each recorded voice had a similar probability of occurrence in the different conditions, and the identity of the target and masker voices was counterbalanced across participants to eliminate the possible confound of perceptual differences between voices. Target detection was measured at three target-to-masker ratios (TMRs; −3, 0, +3 dB), with 50 trials of each condition at every TMR. (TMR was varied by changing the amplitude of the target voice relative to a constant-amplitude masker.)

150 trials of each condition were presented in 6 blocks of 100 trials, with 25 trials of each condition in every block. To maximize efficiency, we used a dynamic stochastic design in which the probability of occurrence of every condition varied in a sinusoidal fashion over time (Friston *et al.*, 1999). Each condition was tested in sets of between three and seven consecutive trials. For each condition, clusters of five trials were most common (probability 0.4), followed by clusters of four and six trials (probability 0.2), and clusters of three and seven trials were least common (probability 0.1). TMR was held constant within a cluster, but changed across clusters. Clusters of trials of every condition were presented in pseudo-randomized order, with the limitation that no two successive clusters were the same condition. In order to effectively use all trials, the last trial of every cluster served as a "voice prime" for the next condition when required, so that listeners gained experience with the constant voice (in the target, masker, and switch conditions) prior to the first trial of the next cluster. For instance, the target voice in the trial preceding a target cluster defined the target voice for all the trials of the target cluster. An additional dummy trial was added at the beginning of every block to give prior exposure to the constant voice for the first cluster of trials. In cases where the block started with a baseline cluster (i.e., no constant voice); both voices in the dummy trial were different from those of the first trial of the first cluster. Within each block, target and masker voices were always of the same sex so that the role of the constant voice could be switched from target to masker in the switch condition. Male voices were used for three blocks and female voices for the other three blocks. The order of the six blocks was counterbalanced across participants and optional breaks were offered between blocks. Prior to the start of the experiment, participants were familiarized with the experimental paradigm although they were not told about the different conditions, and were therefore not aware that they could potentially use voice consistency as a cue. They all completed a short training session of five trials, with feedback, to ensure that they understood the task and knew how to indicate their response.

3. Data analysis

Responses were considered correct if participants identified both the color and the number uttered by the target voice. Data for the different conditions were collapsed across male and female experimental blocks as we found no significant interaction with sex. For the switch condition, trials where the constant voice was the target (switch_T) were analyzed separately from trials where the constant voice was the masker (switch_M). Also, in order to examine performance as a function of the number of successive trials of the same condition, accuracy scores were computed separately for the first, second, third, and fourth trials in each cluster from a

given condition, with a fifth "bin" that collapsed across the fifth, sixth, and seventh trials in a cluster (since the probability of occurrence of this many trials of the same condition in a row was relatively low). There were 10 trials in the first, second, and third positions, 9 in the fourth position, and 11 in the fifth position for each condition at each TMR. Data were entered into a repeated-measures analysis of variance (ANOVA), and the Huynh-Feldt correction for sphericity violation was used when necessary. Hypothesis-driven comparisons between the target and the baseline conditions and the masker and the baseline conditions are reported with no correction (least significant difference); otherwise, for comparisons with no *a priori* hypothesis, Sidak correction was used to control type I error.

Errors were classified into one of three types according to which voice uttered the selected color and number coordinates. Errors were labeled "wrong-voice" when listeners selected both coordinates from the masker voice. In "mixed-voice" errors, participants selected the number spoken by one of the two voices in the mixture, and the color spoken by the other voice. "Other-voice" errors occurred when at least one of the reported dimensions of the coordinate (color and/or number) was not present in the trial stimulus phrase. Since participants made other-voice errors on fewer than 2% of the trials (such errors comprised <7% of the errors), this type of error was not included in the analysis. Data (proportion of trials in which a particular type of error was made, out of all the trials in a particular condition, collapsed across TMR and trial position) were entered into a repeated-measures ANOVA, and the Huynh-Feldt correction for sphericity violation was used when necessary. For *post hoc* comparisons, Sidak correction was used to control type I error.

B. Results

Since they were not told about the objective of the experiment or the different conditions prior to the task, participants were asked in debriefing whether or not they noticed the presence of a consistent talker during the experiment. Only 4 out of 25 participants did, and 2 out of the 4 specifically found this consistency to be helpful. (Note that the pattern of results remained the same even when these participants were excluded from the analysis.)

The repeated-measures ANOVA on accuracy with three within-subject factors [TMR, with three levels: −3, 0, +3; condition, with five levels: baseline, target, masker, switch_T, and switch_M; and trial position, with five levels: first, second, third, fourth, and higher (fifth, sixth, seventh)] revealed significant main effects of TMR [$F(1.39,33.36) = 23.09$, $p < 0.001$], condition [$F(3.26,78.24) = 10.67$, $p < 0.001$], and trial position [$F(4.00,96.00) = 6.15$, $p < 0.001$]. The repeated-measures ANOVA revealed no significant two-way or three-way interactions among the TMR, condition, and trial position factors (all $p \geq 0.108$).

Post hoc pairwise comparisons examining the significant main effects revealed that performance increased with increasing TMRs across conditions, as expected. As shown in Fig. 1(A), performance was significantly better at +3 dB compared to −3 dB ($p < 0.001$) and 0 dB ($p < 0.001$), with no significant difference in performance between the −3 dB and the 0 dB conditions ($p = 0.931$). Pairwise comparisons examining the expected effects of target, masker, and baseline conditions revealed significantly better performance in the target condition ($p < 0.001$) and in the masker condition ($p = 0.006$) compared to the baseline condition. On average, target-detection performance improved by 5% in the target condition and 2.6% in the masker condition compared to the baseline condition. Sidak-corrected comparisons between the remaining pairs of conditions revealed significantly better performance in the target condition compared to the switch_T ($p < 0.001$) and the switch_M ($p < 0.001$) conditions. Performance in the masker condition was significantly better than in the switch_M condition ($p = 0.038$).

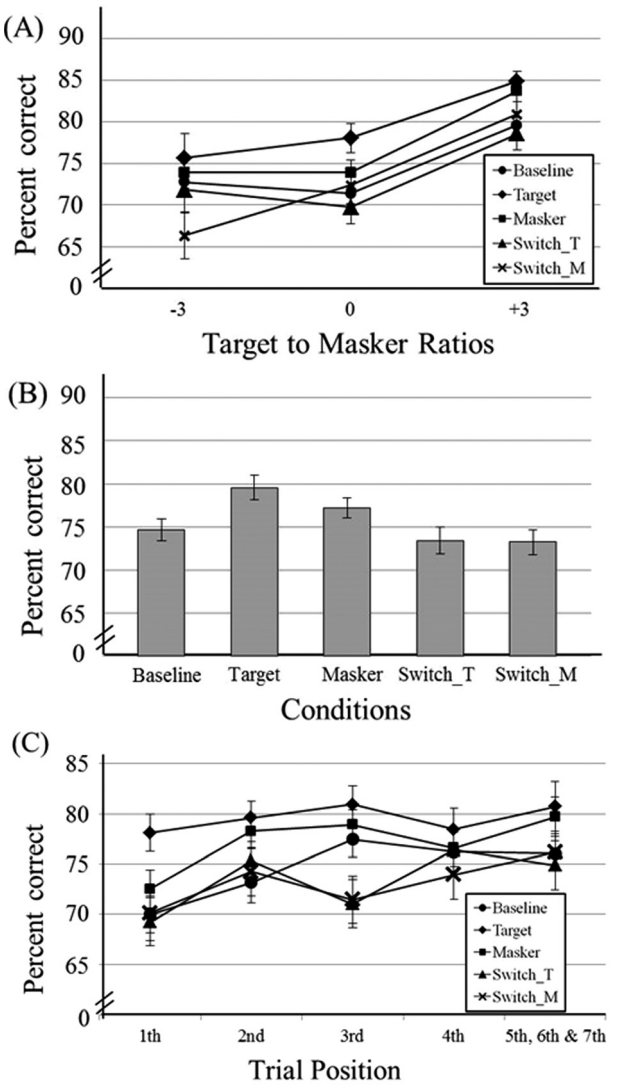

FIG. 1. (A) Percentage of trials in which participants correctly selected the color and number uttered by the target talker as a function of TMR for the five conditions in two-talker configurations. (B) Performance in the baseline, target, masker, switch_T, and switch_M conditions in two-talker situations. (C) Percentage of trials in which participants correctly selected the color and number uttered by the target talker as a function of trial position for the five conditions in the two-talker configuration. The error bars represent standard error of the mean.

Performance did not significantly differ among the baseline, switch_T, and switch_M conditions [see Fig. 1(B)]. *Post hoc* pairwise comparisons on the trial position factor revealed that, across conditions, performance was poorer on the first trial (the first trial in which the voice was repeated) compared to most of the other trials of the cluster [second ($p = 0.002$), third (tendency; $p = 0.097$), fourth ($p = 0.010$), and higher (5th, 6th, and 7th; $p < 0.001$); see Fig. 1(C)]. Performance did not differ among any of the other trial positions.

The percentage of trials (out of a total of 150 trials in the baseline, target, and masker conditions and 75 trials in the switch_T and switch_M conditions) in which wrong- and mixed-voice errors were committed were entered in a repeated-measures ANOVA with error type (two levels: wrong-voice, mixed-voice) and condition (five levels: baseline, target, masker, switch_T, and switch_M) as within-subject factors. This analysis revealed a significant interaction between error type and condition [$F(4,96.00) = 3.89$, $p = 0.006$]. As shown in Fig. 2, there were no differences across conditions for the mixed-voice errors (all comparisons $p \geq 0.384$) whereas listeners made fewer wrong-voice errors in the target compared to the baseline, switch_M, and switch_T conditions (all comparisons $p \leq 0.001$) and in the masker compared to the baseline condition ($p = 0.015$). This result for the masker condition suggests that a constant voice does not necessarily lead to an attentional bias toward that voice (such that listeners are tempted to report it, instead of the correct target).

C. Discussion

In this experiment, performance improved when the target voice was constant compared to when no voice was held constant across trials (baseline), consistent with previous reports (Brungart *et al.*, 2001; Best *et al.*, 2008; Kitterick *et al.*, 2010; Bressler *et al.*, 2014). Target recognition was higher and fewer errors were committed in the target condition overall, with a particularly low incidence of wrong-voice errors. Improved performance with a constant-voice target may be due to better segregation, or it could be that the consistent voice becomes a learned template (Bregman, 1990), and the listener is able to use a template-matching strategy. However, the template would have to have been established very quickly (as soon as the target voice was heard once) since we see advantages in this condition as early as the second trial position, and we did not see evidence of a template building up over successive trials.

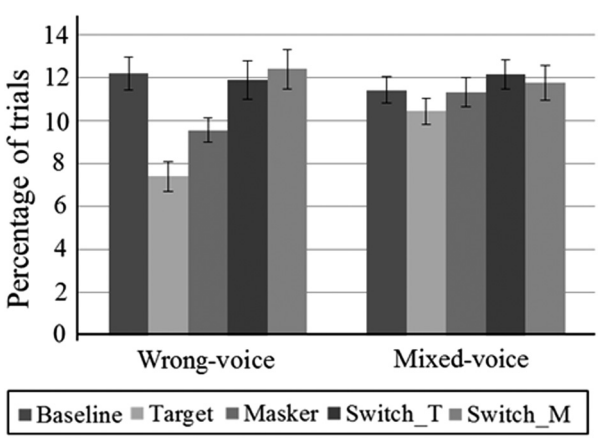

FIG. 2. Percentage of trials with mixed-voice and wrong-voice errors for the five conditions in the two-talker configuration (experiment 1). The error bars represent the standard error of the mean.

Our findings also demonstrate that speech perception is improved when the voice of the interfering talker is constant across successive trials, and this improved performance can be explained, as in the constant-target condition, by a drop in wrong-voice, compared to mixed-voice, errors. Thus, the constant masking voice did not appear to capture listeners' attention. Instead, these results are in line with reports of better perceptual segregation after priming with a multi-tone masker (Kidd *et al.*, 2011) or with spatial location (Allen *et al.*, 2011) or when the masker is a highly familiar voice (Johnsrude *et al.*, 2013). Since the masker is outside of the focus of attention, such improved performance due to masker consistency cannot be due to template-matching (Bregman, 1990), but may be due to improved segregability of the two voices.

As for the effect of TMR, we observed no difference in performance between the −3 dB and the 0 dB conditions, but better performance in the +3 dB conditions. These results are consistent with previous studies showing that target detection performance on the CRM task is independent of TMR for values of −3 dB or 0 dB in a two-talker context (Brungart *et al.*, 2001; Johnsrude *et al.*, 2013). This indicates that listeners are benefiting from the level difference at −3 dB and are able to attend to the less intense target when only one masking talker is present. We found no interaction between loudness and voice-consistency cues; this indicates that reliable voice-consistency effects can improve speech segregation even when listeners can also rely on level differences.

In this experiment, target identification improved significantly when the target or the masker voice was held constant across trials; however, this benefit seemed to be specific to situations where the constant voice consistently played the same role. We observed no benefit when the consistent voice switched roles across trials; there were no significant differences in performance among the baseline, switch_T, and switch_M conditions. It appears that the role alternation for the consistent voice in the switch conditions (from target to masker and back again) prevented listeners from using their knowledge about this voice to better extract information. Why the consistent voice in the switch condition did not provide a benefit is not clear. Listeners might have been tempted to follow that voice, no matter what role it played, but, if that were the case, we would have observed a greater proportion of wrong-voice errors in the switch_M condition and better performance in the switch_T condition, which we did not see. In any case, the lack of benefit in the switch condition indicates that the benefit arising from a consistent voice is not simply due to the familiarity of the voice. It seems that not only does the voice have to be consistent across successive trials, but its role, either as a target or a masker, also needs to be consistent for performance to improve.

In sum, experiment 1 demonstrated improved segregability of the voices in two-talker mixtures when listeners are presented with a consistent target or masking voice although the simple presence of a consistent voice, when the role alternated from trial to trial between target and masker, was not helpful. In experiment 2, we examine whether the benefits associated with a constant target or masker voice could be observed for three-talker mixtures, with two masking voices. With two talkers, the masker voice is also segregated once the target voice is segregated, and such "automatic" segregation may be a necessary prerequisite for benefit from a consistent masking voice to be realized. With three talkers, segregation of the target leaves a potentially unsegregated mixture of two maskers. If a consistent masker benefit is observed in such situations, it would mean that the consistent masker must have been successfully segregated from the novel one.

III. EXPERIMENT 2: EFFECTS OF THE CONSTANT PRESENCE OF A TARGET OR MASKER VOICE IN A THREE-TALKER MIXTURE

A. Method

1. Listeners

Thirty-eight native English speakers (three males; one left-handed; age range 18–23 yr; mean age 19 yr; normal hearing: group mean 3.1 dB HL, range −5.0–11.3 dB HL), naive to the CRM stimuli and task, participated. Exclusion criteria, ethics clearance, and consent procedures were the same as in experiment 1.

2. Stimuli and procedure

The stimuli and procedure were the same as in experiment 1, with the exception that the target voice was presented concurrently with *two* masking voices. As in experiment 1, the target call sign was always Baron, and the masker call signs were different from Baron and different from each other (either Arrow, Charlie, or Eagle). The color-number coordinates for the target and the two masker sentences were also always different. In the masker condition, one of the masker voices was constant from trial to trial, whereas the other masker voice changed randomly from trial to trial. Similarly, in the switch condition, one constant voice alternated between the roles of target and masker while the identity of the second masking voice always changed from trial to trial (as did the other—target or masker—voice). For each trial, the three phrases were first normalized to the same root-mean-square (RMS) power. Then the amplitude of the target was scaled by the TMR value for each specific trial (−3 dB, 0 dB, or 3 dB) and the three sounds were added together and the mixture presented to participants.

3. Data analysis

As in experiment 1, responses were considered correct if participants identified both the color and the number uttered by the target voice. Again, data were collapsed across male and female blocks as there was no significant interaction with sex. Errors were classified as wrong-voice if both coordinates were selected from one of the two maskers; as mixed-voice if the two coordinates were spoken by different talkers (either the target and one of the masking talkers, or one coordinate from each of the masking talkers); or other-voice, if at least one of the reported coordinates was not produced by any talker in the mixture. Participants made other-voice errors on <3% of the trials (<5% of the errors); therefore, this type of error was again not included in the analysis. Analysis procedures were essentially the same as for experiment 1.

B. Results

1. Experiment 2

After the completion of the task, participants were debriefed and asked whether they noticed the presence of a constant voice. Out of the 38 participants, 4 noticed the presence of a consistent talker during the experiment; 3 of these 4 participants found the repetition helpful and 1 specifically mentioned trying to focus on the constant voice when it was present for more than 2 trials in a row. (Note that, as in experiment 1, the results were unchanged when these participants were excluded from the analysis.)

A repeated-measures ANOVA on accuracy with three within-subject factors [TMR, with three levels: −3, 0, +3; condition, with five levels: baseline, target, masker, switch_T, and switch_M; and trial position, with five levels: first, second, third, fourth, and higher (fifth, sixth, and seventh)] revealed significant main effects of all three factors: TMR [$F(1.96,72.63) = 633.83$, $p < 0.001$], condition [$F(3.52, 130.30) = 5.52$, $p = 0.001$], and trial position [$F(3.99,147.57) = 3.99$, $p = 0.008$].

Post hoc pairwise comparisons examining the significant main effects revealed that, across conditions, performance significantly increased as TMR increased (all comparisons significant $p < 0.001$) [see Fig. 3(A)]. Pairwise comparisons examining the predicted effects among the target, masker, and baseline conditions revealed significantly better performance in the target ($p = 0.009$), but not in the masker condition ($p = 0.185$) compared to the baseline condition. There was, on average, a 2.5% improvement in performance for the target condition compared to the baseline. Sidak-corrected comparisons between the remaining pairs of conditions revealed significantly better performance in the target condition compared to the masker ($p = 0.009$) and the switch_M ($p < 0.001$) conditions. Performance in the switch_M ($p = 0.107$) and switch_T ($p = 1.000$) conditions did not differ from that in the baseline condition [see Fig. 3(B)]. Finally, *post hoc* pairwise comparisons on the trial position factor revealed that, across conditions, performance was better in the latter trials of a cluster (fifth, sixth, and seventh) compared to the first trial of the cluster ($p = 0.006$); no other trial position effects were significant.

The three-way (TMR by condition by trial position) interaction was not significant [$F(26.78,990.77) = 1.23$, $p = 0.199$]. The TMR factor did not significantly interact with condition ($p = 0.227$) or trial position ($p = 0.282$), but the condition by trial position interaction was significant

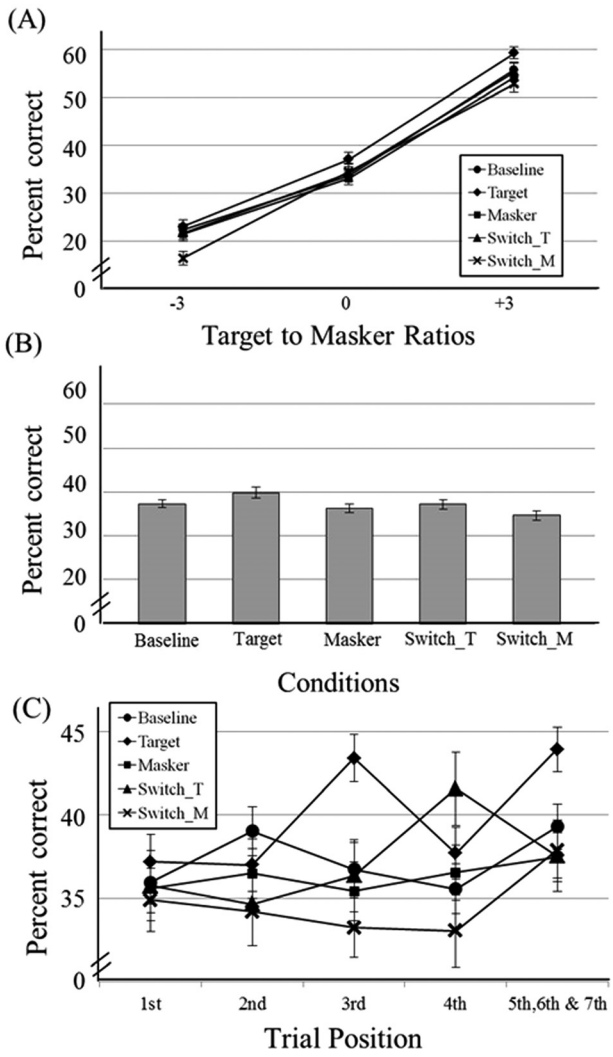

FIG. 3. (A) Percentage of trials in which participants correctly selected the color and number uttered by the target talker as a function of TMR for the five conditions in three-talker configurations. (B) Performance in the baseline, target, masker, switch_T, and switch_M conditions in three-talker situations. (C) Percentage of trials in which participants correctly selected the color and number uttered by the target talker as a function of trial position for the five conditions in three-talker configuration. The error bars represent the standard error of the mean.

[$F(12.93,478.41) = 1.78$, $p = 0.043$]. *Post hoc* pairwise comparisons investigating this interaction revealed a significant effect of trial position only in the target condition. Performance was best in this condition for the third (all comparisons $p < 0.018$) and higher (fifth, sixth, and seventh positions; all comparisons $p < 0.010$) compared to the first, second, and fourth positions. These three trial positions did not differ significantly from one another (all comparisons $p = 1.000$). This pattern is broadly consistent with benefit from a constant target voice building up over time [see Fig. 3(C)], except for the odd reversal at time points 3 and 4, with performance at point 3 higher than at point 4 in the target condition.

The percentage of trials with wrong- and mixed-voice errors out of the total number of trials in each condition (150 for the target, masker, and baseline conditions; 75 for the switch_M and switch_T conditions) were entered in a repeated-measures ANOVA with error type (two levels: wrong-voice, mixed-voice) and conditions (five levels: baseline, target, masker, switch_T, and switch_M) as within-subject factors. This analysis revealed a significant interaction between error type and condition [$F(4,148.00) = 4.63$, $p = 0.002$]. As shown in Fig. 4(A), fewer wrong-voice errors were committed in the target compared to all other conditions (all comparisons $p < 0.025$), while listeners made fewer mixed-voice errors in the switch_T compared to the masker condition ($p = 0.047$).

Whereas both masker voices changed from trial to trial in the baseline, target, and switch_T conditions, one of the two masking voices was held constant in the masker and the switch_M conditions. Therefore, there were really two kinds of wrong-voice errors for these two conditions: the wrong-voice errors made when listeners selected the random masker versus when they selected the constant masker. If listeners were to randomly pick one of the masking voices in the mixture, we should observe no difference between the two types of errors. Paired *t*-tests on the percentage of trials on which these two types of wrong-voice errors were committed revealed significantly fewer errors involving the

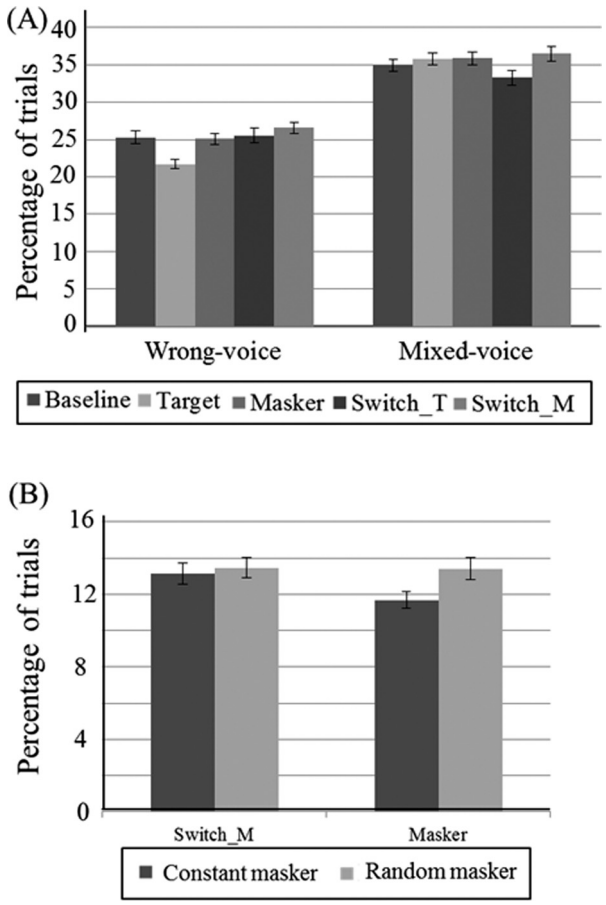

FIG. 4. (A) Percentage of trials with mixed-voice and wrong-voice errors for the five conditions in the three-talker configuration (experiment 2). The error bars represent the standard error of the mean. (B) Percentage of trials with wrong-voice errors for which participants selected the constant or the random masker for the masker and the switch_M (constant voice alternating to the masker position) conditions. The error bars represent the standard error of the mean.

constant voice compared to the random-voice masker in the masker condition [$p = 0.031$; see Fig. 4(B)], whereas the proportion was not significantly different in the switch_M condition ($p = 0.710$). This suggests that, although we did not observe a benefit of the masker condition in terms of accuracy, listeners seem to try to avoid mistaking a constant-voice masker for the target.

2. Comparison of the results of experiments 1 and 2

In experiment 1, we found that performance on the target identification task improved, on average, 5% in the target condition and 2.6% in the masker condition compared to the baseline condition. In experiment 2, there was, on average, a 2.5% improvement in performance for the target condition and no significant improvement in the masker condition compared to the baseline condition.

In order to test whether performance in the target, masker, and baseline conditions was different in two- and three-talker situations, we conducted a repeated-measures ANOVA on accuracy scores with number of talkers (two levels: two-talker experiment 1, three-talker experiment 2) as a between-subjects factor and condition (three levels: baseline, target, masker) as a within-subjects factor. This analysis revealed a significant number-of-talkers by condition interaction [$F(1.97,120.05) = 3.82$, $p = 0.025$] due to better performance in the target ($p < 0.001$) and masker ($p = 0.015$) conditions compared to the baseline in experiment 1 (two talkers), and better performance in the target ($p = 0.016$), but not the masker ($p = 0.432$) condition, compared to the baseline, in experiment 2 (three talkers). Performance was significantly better in the target compared to the masker ($p = 0.001$) condition in the three-talker, but not the two-talker experiment.

We wanted to verify that the benefit associated with the presence of a constant target talker was similar across a number of talkers in a mixture, but that the benefit associated with the presence of a constant masker was not. Target benefit and masker benefit were calculated by subtracting baseline accuracy scores from target and masker accuracy scores, within subjects. As shown in Fig. 5, two-group *t*-tests revealed that there was no significant difference in benefit for the target condition between experiments ($p = 0.083$), whereas the benefit (in the masker condition) was significantly larger in experiment 1, with two talkers ($p = 0.003$).

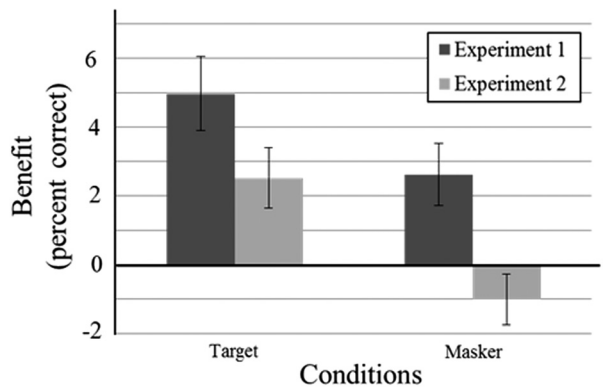

FIG. 5. Comparison of the benefit (percent correct) in the target and masker condition between the two experiments. The error bars represent the standard error of the mean.

C. Discussion

In experiment 2 (three talkers), best performance was again obtained in the target condition, as predicted (Brungart *et al.*, 2001; Best *et al.*, 2008; Kitterick *et al.*, 2010; Johnsrude *et al.*, 2013; Bressler *et al.*, 2014). The accompanying reduction in errors, compared to the baseline condition, was accounted for by a reduction in wrong-voice errors, but not in mixed-voice errors. We examined performance as a function of the number of successive trials of the same condition and observed generally better performance for later trials in the target condition only, suggesting that when listeners can exploit knowledge of a consistent target voice in three-talker mixtures, the benefit generally increases over time. Improved intelligibility associated with such buildup may reflect enhanced segregation or, alternatively, a template-matching strategy whereby listeners define a template to which they can match signal on subsequent trials (Bregman, 1990). Bregman (1990) has suggested that template-matching can work only for signals that are the focus of attention (i.e., targets) and, indeed, we observe a pattern consistent with template matching only in the target condition.

We found no significant difference in performance between the masker and baseline conditions in this experiment, and performance in the target condition was significantly better than in the masker condition; this is significantly different to what we observed in experiment 1 with two talkers. Our results suggest that the consistent presence of one masker in three-talker mixtures does not enhance segregation of the target from the masking signal. The contrasting pattern of results suggests that the lack of benefit may be related to the addition of a second interfering talker, since all other factors were held constant between the two experiments. The lack of improvement in three-talker mixtures with one constant masking voice may be due to an attentional bias, which would manifest as people tending to report the constant-voice coordinates even when the constant voice was a masker, i.e., an elevated incidence of wrong-voice errors in the masker condition. However, specifically for the masker condition, we found that participants made significantly fewer wrong-voice errors involving the constant masker compared to the random masker, indicating that the two masking voices must have been segregated at least partially. This suggests that participants may have been trying *not* to select the coordinates uttered by the constant masking voice and that they might be able to use the constant presence of a non-target voice (which is presumably not the focus of attention) to better ignore it, even in three-talker mixtures. Although significant, this effect is not very strong possibly because the masker voice was only held constant for 3–7 trials in the current experiment. It has been previously shown that, especially for situations with more than one interfering talker, the benefit associated with the constant presence of a (target) voice improves systematically up to about 30 consecutive trials (Brungart and Simpson, 2007).

Providing information about the masking signal might indeed improve segregability between the maskers but not enough to significantly benefit target detection performance, at least when the voices are held constant for a relatively small number of trials. Different mechanisms may be involved when target voice is held constant (i.e., template-matching of the target voice) compared to when one masker voice is constant across trials (i.e., enhanced segregation of the two maskers) in three-talker mixtures.

In experiment 2, as in experiment 1, we observed no benefit for the switch condition, in which a constant voice alternated roles between target and masker. In fact, in experiment 2, compared to the baseline condition without a constant voice, we observed *poorer* performance for trials on which the constant voice was the masker (switch_M) and no difference when it was the target (switch_T). This confirms that listeners can only benefit from the consistent presence of a voice when the role of that voice is also held constant across successive trials. Again, since benefit was not observed in the switch_T condition, and listeners did not commit more wrong-voice errors when the constant voice was the masker (switch_M), and they did not tend to select the coordinates from the constant masker more than those spoken by the random masker in the switch_M condition, it does not appear that listeners' attention is biased toward the constant voice.

IV. SUMMARY AND GENERAL DISCUSSION

We conducted two experiments to test whether listeners could better segregate and follow a target message when the identity of one of the talkers in a two- or three-talker mixture was held constant across 3–7 trials while all other voices changed from trial to trial. Our findings add to the growing body of literature (Brungart *et al.*, 2001; Best *et al.*, 2008; Kitterick *et al.*, 2010; Bressler *et al.*, 2014) documenting the benefit that listeners receive from the consistent presence of a specific target voice in multitalker mixtures. Target speech intelligibility also improves when a masker voice is consistent in a two-talker mixture, but this benefit largely disappears in three-talker mixtures. However, some minor benefit was noted in the fact that the listeners were significantly less likely to report the coordinates spoken by the consistent-voice masker (compared to those spoken by a concurrent random-voice masker) in the masker condition, suggesting that listeners may derive some benefit from having a familiar masking voice in a three-voice mixture.

In both experiments, listeners almost always perceived the words uttered by talkers in the mixture, but were not always able to correctly link the color and number coordinates to the appropriate talker, suggesting a high degree of informational, not energetic, masking. The closed-set nature of the CRM sentences forces listeners to recognize all, or at least part, of the keywords to correctly perform the task. In natural speech, by contrast, meaningful context is a helpful cue to understand a target sentence even if some words are not identifiable from the signal. Our results indicate that the consistent presence of a particular voice in multitalker situations can reduce informational masking, similar to other non-acoustic cues such as context and previous knowledge (Freyman *et al.*, 2004; Johnsrude *et al.*, 2013).

These experiments showed that the benefit of voice consistency develops rapidly. However, performance may have continued to improve if the talkers' identity had been held constant for more than seven trials in a row. Despite this relatively short exposure time, we observed significant benefit related to the consistent presence of a voice and the benefits of consistency appeared to grow over time, particularly, for a consistent target voice in a three-talker mixture where we observed a significant effect of trial position (i.e., increasing benefit as the number of constant trials increases). This suggests that, in three-talker situations at least, the attended constant target voice may become a learned template to which participants match on subsequent trials (Bregman, 1990).

In the two experiments, we observed significant effects of TMR. In two-talker mixtures, listeners benefited from level differences both when the target was more intense and when it was less intense than the masking voice. In three-talker mixtures, listeners did not seem to be able to use level differences to help segregate a less intense target voice from a more intense masking signal; instead, performance seemed to depend more on the level of the target. In both experiments, we found no interaction between TMR and voice-consistency conditions. Our results show that voice-consistency effects can improve speech segregation even when listeners can also exploit level differences. Although the effects associated with the constant presence of a voice are not large, they are significant and would probably have been larger if voices had been held constant for more than our maximum seven consecutive trials (Brungart and Simpson, 2007).

Although a consistent masker was decidedly helpful in two-talker mixtures, holding one of two masking voices constant in three-talker mixtures did not benefit intelligibility. This statistically significant interaction suggests that the way the brain extracts information from voice mixtures differs depending on whether one or two masking voices are present. One consistent masker in two-talker situations enhances the segregability of the target and masker voices and, in turn, improves target detection performance. However, in three-talker mixtures, one consistent masker may enhance the perceptual separation of this voice from the others (since participants seemed to specifically avoid reporting the consistent masker), but does not seem to enhance the perceptual segregation of the target from the maskers, since no improvement in target intelligibility was observed, compared to when both maskers were novel. It is possible that, if we had held voices constant over a greater number of trials (i.e., more than seven), we may have observed a benefit (Brungart and Simpson, 2007).

Listeners benefit from the consistent presence of a voice whether they are exposed *a priori* to the voice and explicitly told that this specific talker will be the target for the trials to come (Brungart *et al.*, 2001; Yang *et al.*, 2007; Huang *et al.*, 2010), or whether the consistency simply happens, without explicit instructions. In our experiments, the transitions between the different conditions were not obvious since clusters of trials were presented successively without any breaks, and clusters varied in length. Listeners were not told about the experimental conditions until after they completed

the study; they were simply asked to follow the voice saying the call sign Baron on every trial, and identify the color and number coordinates spoken by that specific talker. In fact, in post-experiment debriefing where participants were simply asked to give their thoughts about the task, they generally reported that they did not notice the presence of the consistent voice, and only a few spontaneously noted that they found it to be helpful, and the pattern of results remained the same when these participants were excluded from the analyses. Interestingly, a recent study looking at the effects associated with consistency of target talker reported a similar pattern of better target intelligibility when the previous target was uttered by the same talker whether listeners were aware that the voice would repeat or whether they were not told prior to the task that the target talker would ever repeat (Bressler *et al.*, 2014).

In sum, our experiments demonstrate that listeners can use voice consistency as a cue to enhance speech intelligibility in multitalker mixtures. We demonstrate a clear benefit associated with the constant presence of a target voice in both two- and three-talker mixtures, but benefit of a consistent masker voice only in two-talker mixtures. This interaction suggests that different strategies are involved in exploiting voice consistency in situations with one versus multiple interfering talkers. Although the constant presence of a specific talker appears to improve segregation of two voices in two-talker situations (given the constant-masker benefit and the lack of trial position effects), listeners seem to use a template-matching strategy to extract the target (without necessarily improving perceptual segregation of all voices) in three-talker mixtures. However, the two maskers in the three-talker mixtures must have been partially segregated from each other, since listeners made fewer wrong-voice errors involving the constant compared to the random talker. This partial segregation was not sufficient to provide benefit for target identification, however.

In future experiments, it would be useful to examine whether, and how, a constant voice can improve speech intelligibility in populations of individuals with difficulty understanding speech in multitalker situations; such groups include older adults (Helfer and Freyman, 2008) and individuals with autism spectrum disorders (Stiegler and Davis, 2010).

ACKNOWLEDGMENTS

The authors would like to thank Bob Carlyon for providing helpful comments on the experimental design and Paul Plante for helping to program the task. This work was made possible by operating grants from the Natural Sciences Research Council of Canada and the Canadian Institutes of Health Research to I.S.J., a postdoctoral grant from the Fonds de la Recherche en Santé du Québec to F.S., and a postdoctoral grant from the Canadian Institutes of Health Research to F.S.

Allen, K., Alais, D., Shinn-Cunningham, B., and Carlile, S. (**2011**). "Masker location uncertainty reveals evidence for suppression of maskers in two-talker contexts," J. Acoust. Soc. Am. **130**(4), 2043–2053.

Best, V., Ozmeral, E. J., Kopco, N., and Shinn-Cunningham, B. G. (**2008**). "Object continuity enhances selective auditory attention," Proc. Natl. Acad. Sci. U.S.A. **105**(35), 13174–13178.

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**(2), 1065–1066.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT Press, Cambridge, MA), pp. 1–773.

Bressler, S., Masud, S., Bharadwaj, H., and Shinn-Cunningham, B. (**2014**). "Bottom-up influences of voice continuity in focusing selective auditory attention," Psychol. Res. **78**(3), 349–360.

Brungart, D. S., and Simpson, B. D. (**2004**). "Within-ear and across-ear interference in a dichotic cocktail party listening task: Effects of masker uncertainty," J. Acoust. Soc. Am. **115**(1), 301–310.

Brungart, D. S., and Simpson, B. D. (**2007**). "Cocktail party listening in a dynamic multitalker environment," Percept. Psychophys. **69**(1), 79–91.

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**(5), 2527–2538.

Darwin, C. J. (**1997**). "Auditory grouping," Trends Cogn. Sci. **1**(9), 327–333.

Davis, M. H., and Johnsrude, I. S. (**2007**). "Hearing speech sounds: Top-down influences on the interface between audition and speech perception," Hear. Res. **229**(1-2), 132–147.

Dehaene, S., and Cohen, L. (**2007**). "Cultural recycling of cortical maps," Neuron **56**(2), 384–398.

Durlach, N. (**2006**). "Auditory masking: Need for improved conceptual structure," J. Acoust. Soc. Am. **120**(4), 1787–1790.

Durlach, N., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., and Kidd, G., Jr. (**2003**). "Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity," J. Acoust. Soc. Am. **114**(1), 368–379.

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**(5), 2246–2256.

Friston, K. J., Zarahn, E., Josephs, O., Henson, R. N., and Dale, A. M. (**1999**). "Stochastic designs in event-related fMRI," Neuroimage **10**(5), 607–619.

Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (**2004**). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," J. Acoust. Soc. Am. **115**(2), 833–843.

Helfer, K. S., and Freyman, R. L. (**2008**). "Aging and speech-on-speech masking," Ear Hear. **29**(1), 87–98.

Huang, Y., Xu, L., Wu, X., and Li, L. (**2010**). "The effect of voice cuing on releasing speech from informational masking disappears in older adults," Ear Hear. **31**(4), 579–583.

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., and Carlyon, R. P. (**2013**). "Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice," Psychol. Sci. **24**(10), 1995–2004.

Jones, G. L., and Litovsky, R. Y. (**2008**). "Role of masker predictability in the cocktail party problem," J. Acoust. Soc. Am. **124**(6), 3818–3830.

Kidd, G., Jr., Mason, C. R., and Gallun, F. J. (**2005**). "Combining energetic and informational masking for speech identification," J. Acoust. Soc. Am. **118**(2), 982–992.

Kidd, G., Jr., Richards, V. M., Streeter, T., Mason, C. R., and Huang, R. (**2011**). "Contextual effects in the identification of nonspeech auditory patterns," J. Acoust. Soc. Am. **130**(6), 3926–3938.

Kitterick, P. T., Bailey, P. J., and Summerfield, A. Q. (**2010**). "Benefits of knowing who, where, and when in multi-talker listening," J. Acoust. Soc. Am. **127**(4), 2498–2508.

Nygaard, L. C., and Pisoni, D. B. (**1998**). "Talker-specific learning in speech perception," Percept. Psychophys. **60**(3), 355–376.

Stiegler, L. N., and Davis, R. (**2010**). "Understanding sound sensitivity in individuals with autism spectrum disorders," Focus Autism Other Dev. Disabil. **25**(2), 67–75.

Yang, Z., Chen, J., Huang, Q., Wu, X., Wu, Y., Schneider, B. A., and Li, L. (**2007**). "The effect of voice cuing on releasing Chinese speech from informational masking," Speech Commun. **49**, 892–904.

Yonan, C. A., and Sommers, M. S. (**2000**). "The effects of talker familiarity on spoken word identification in younger and older listeners," Psychol. Aging **15**(1), 88–99.