Atten Percept Psychophys
DOI 10.3758/s13414-016-1207-3

Reverberation limits the release from informational masking obtained in the harmonic and binaural domains

Mickael L. D. Deroche[1] · John F. Culling[2] · Mathieu Lavandier[3] · Vincent L. Gracco[1]

Abstract A difference in fundamental frequency (ΔF0) and a difference in spatial location (ΔSL) are two cues known to provide masking releases when multiple speakers talk at once in a room. We examined situations in which reverberation should have no effect on the mechanisms underlying the releases from energetic masking produced by these two cues. Speech reception thresholds using both unpredictable target sentences and the coordinate response measure followed a similar pattern. Both ΔF0s and ΔSLs provided masking releases in the presence of nonspeech maskers (matched in excitation pattern and temporal envelope to the speech maskers) that, as intended, were robust to reverberation. Larger masking releases were obtained for speech maskers, but critically, they were affected by reverberation. These results suggest that reverberation either limits the amount of informational masking that is present to begin with or affects its release by ΔF0s or ΔSLs.

Keywords Speech perception · Psychoacoustics · Perceptual categorization · Perceptual identification

In cocktail party situations (Cherry, 1953), listeners can use a difference in fundamental frequency (ΔF0) and a difference in spatial location (ΔSL) between competing talkers to obtain release from masking. It is generally thought that there are two forms of masking: energetic and informational. *Energetic masking* (Durlach, 2006) refers to the case in which a target sound is made inaudible by a more intense sound of similar spectro-temporal characteristics. *Informational masking* (Brungart, Simpson, Ericson, & Scott, 2001; Durlach et al., 2003; Kidd, Mason, & Gallun, 2005) refers to the case in which a competing sound interferes with the listener's identification of an audible target sound where the competitor does not share the same frequency band or occurs in a different time window than the target. A lot of attention has been paid to the mechanisms underlying the energetic-masking releases offered by ΔF0 and ΔSL, and they are generally susceptible to reverberation, as will be discussed in the following sections. In contrast, the potential effects of reverberation on the informational-masking releases associated with a ΔF0 and a ΔSL remain relatively unexplored. In the present study, we aimed to examine whether reverberation affects the use of ΔF0 and ΔSL while we restricted its possible cause to an informational aspect.

Reverberation can impair the ΔF0 benefit

Reverberation is generally detrimental to the use of ΔF0s between concurrent speech sources. However, in the rather artificial case in which sources are monotonized—that is, have a fixed F0 throughout the entire signal duration—reverberation is harmless. Culling, Summerfield, and Marshall (1994) measured the benefit of a one-semitone ΔF0 in the case of vowel recognition and found that this benefit was reduced by reverberation only when combined with some modulation of F0, but not when F0s were fixed. Deroche and Culling (2011) extended this finding to connected speech, by measuring the speech reception threshold (SRT), defined as the target-to-masker ratio (TMR) required to achieve 50 % intelligibility,

✉ Mickael L. D. Deroche
mickael.deroche@mcgill.ca

[1] Centre for Research on Brain, Language and Music, McGill University, Rabinovitch House, 3640 rue de la Montagne, Montreal, Quebec H3G 2A8, Canada

[2] School of Psychology, Cardiff University, Cardiff, UK

[3] ENTPE Laboratoire Genie Civile et batiment, Université Lyon, F-69518 Vaulx-en-Velin Cedex, rue M. Audin, Lyon, France

Published online: 19 September 2016

for a target voice separated by a two-semitone ΔF0 from stationary speech-shaped harmonic complexes (hereafter referred to as *buzzes*). Deroche and Culling did not measure the ΔF0 benefit directly, but showed that a large elevation of SRT occurred when adding reverberation to a buzz with a modulated F0, whereas no elevation was observed for a buzz with a fixed F0. The rationale is that as long as the masker's F0 is fixed, reverberation may not matter, because when one introduces reverberation, (1) the masker partials do not move, thereby leaving the exact same spectral dips between resolved partials as in anechoic conditions, and (2) the masker periodicity is not disrupted in the resolved channels. Both of these aspects of masker harmonicity seem crucial to the amount of the ΔF0 benefit (Deroche, Culling, Chatterjee, & Limb, 2014a, 2014b). Reverberation also affects the depth of within-channel envelope modulations, particularly in auditory filters centered at high frequencies, but there seems to be little role for such a mechanism unless masker F0s are very low (Deroche, Culling, & Chatterjee, 2014). Thus, although reverberation disrupts the release of energetic masking that is due to ΔF0s between competing sources in most realistic situations, it is still possible to create a laboratory situation in which this is not the case.

Reverberation can impair the ΔSL benefit

Reverberation is generally detrimental to the use of a ΔSL between concurrent speech sources (Beutelmann & Brand, 2006; Bronkhorst & Plomp, 1990; Culling, Hodder, & Toh, 2003; Culling, Summerfield, & Marshall, 1994; Plomp, 1976). This impairment has two main causes. First, the sound reflections reduce the acoustic shadowing of the head—that is, they make the TMR relatively more homogeneous at the two ears—resulting in a smaller advantage of better-ear listening (Plomp, 1976). Although this is an important part of spatial unmasking, it is easy to alleviate this effect by simulating impulse responses without head between the ears (Lavandier & Culling, 2010). Second, reverberation disrupts binaural unmasking, mainly by reducing the interaural coherence of the masking sounds (Lavandier & Culling, 2007, 2008; Licklider, 1948; Robinson & Jeffress, 1963). Following the equalization cancellation theory (Durlach, 1972), when it is placed under reverberant conditions, a masker becomes less correlated at the two ears and harder to equalize, and therefore more effective at masking. However, in the particular case in which the listener and maskers are placed on a symmetrical axis in the room, reverberation should not affect the interaural coherence of the maskers, since all reflections would be exactly identical at both ears. In support for this idea, Lavandier and Culling (2010) measured the SRT for an anechoic target voice against speech-shaped noises in diverse room configurations. They did not measure the ΔSL benefit directly, but showed that a large elevation of SRT occurred when adding reverberation to an asymmetrical listener/noise configuration, whereas no elevation was observed for a symmetrical listener/noise configuration. Thus, although reverberation disrupts the release of energetic masking due to ΔSLs between competing sources in most realistic situations, it is still possible to create a laboratory situation in which this is not the case.

Reverberation and informational masking

Most studies that have investigated informational masking have used very similar competing utterances, so that listeners can confuse the sentence they should attend to. A typical paradigm is known as the *coordinate response measure* (CRM), wherein sentences are of the form "Ready <call sign>, go to <color> <number> now" (Bolia, Nelson, Ericson, & Simpson, 2000). The task is to choose which of the simultaneous words belong to the target utterance with a given call sign, rather than the competing utterance(s). A specific cue, which is generally the object of investigation, may help listeners fulfill this task, provided that this cue is sufficiently strong to maintain attention on the appropriate utterance. Since the sets of call signs, colors, and numbers are limited, the two utterances remain very similar throughout the experiment, and the intelligibility requirement of such a task (identifying the words) is minimal. Such experiments address the question of how listeners decide which words belong to a particular sentence. Unless they are able to do this, speech mixtures could, in principle, be completely audible, yet incomprehensible. Using a design akin to the CRM, but with different stimuli, Darwin and Hukin (2000) found that reverberation reduced both the listeners' ability to use interaural time differences and their ability to use a steady ΔF0 to group the attended words sequentially. For the binaural investigations, the configuration of listener/maskers was not symmetrical in the room, and therefore their results could potentially be explained by energetic masking (see the section above). For the ΔF0 investigations, sources were monotonized, and consequently, the detrimental effect of reverberation on the use of ΔF0 can hardly be interpreted in terms of energetic masking.

Kidd, Mason, Brughera, and Hartmann (2005) used the CRM design to examine the amount of spatial release from masking caused by a 90° separation between a target voice and a masker. Following a method set by Arbogast, Mason, and Kidd (2002) to separate energetic from informational masking, they filtered the target voice into eight out of 15 spectral bands, and used three different masker types: (1) a sum of narrow-band noises whose bands were the same as the target (i.e., primarily energetic), (2) a sum of narrow-band noises whose bands were different from those of the target (i.e., minimizing both energetic and informational masking), and (3) a speech masker whose bands were different from the

target (i.e., primarily informational). In addition, Kidd, Mason, Brughera, & Hartmann, (2005) introduced real reverberation (i.e., not simulated over headphones), presenting stimuli over loudspeakers. For the same-band noise masker, reverberation reduced the spatial benefit due to the loss of head shadow and disruption in the interaural coherence of the masker (energetic effects discussed in the section above). Surprisingly, however, reverberation did not reduce the spatial benefit obtained with the different-band speech masker. Therefore, their results suggested that spatial release from informational masking is robust to reverberation.

Goal of the present study

For the present study, we created specific laboratory situations in which the energetic-masking release from ΔF0 or ΔSL should be robust to reverberation. This was done by using, respectively, monotonized sources and a symmetrical configuration of listener/maskers. In both cases, there were two masker types: a speech masker and a nonlinguistic masker. The nonlinguistic masker (i.e., primarily energetic) was created with similar spectro-temporal properties (long-term excitation pattern and broadband temporal envelope) to those of the speech masker. The ΔF0 benefit and the ΔSL benefit were measured against the two masker types, in anechoic and reverberant conditions. We expected that the amount of informational masking would be minimal with the nonlinguistic maskers, and therefore that reverberation would have very little effect on the benefits produced by ΔF0 and ΔSL. For the speech maskers, thresholds should be overall elevated, from the presence of informational masking in addition to energetic masking. Consequently, we expected that the masking release provided by each cue, harmonic or binaural, would be larger than that obtained with the nonlinguistic maskers, due to this additional informational component. However, we did not have strong predictions as to whether or not reverberation would affect the informational component, given that the literature has presented conflicting evidence (at least in the binaural domain).

If reverberation interacts with informational masking, this phenomenon might have nothing to do with harmonicity per se, or with binaural hearing per se. Therefore, we intended to examine reverberation within the same framework but with the cue that induced masking release being ΔF0 or ΔSL. We also wanted to see whether the effect of reverberation would generalize across listening tasks. Thus, two methods were used: an adaptive SRT task presenting unpredictable sentences (Exps. 1 and 2), and the CRM presenting predictable sentences at fixed TMRs (Exps. 3 and 4). Using the CRM design, Brungart et al. (2001) made an extensive investigation of the roles of the sex and identity of competing voices in two-, three-, and four-talker mixtures, as a function of TMR. They showed that the psychometric functions could in some cases display unexpected shapes. For instance, with a two-talker mixture—that is, a single masking voice—performance could plateau (well above chance, and at different levels of performance depending on the characteristics of the masking voice) as the TMR decreased below 0 dB. This represents a major problem for an adaptive task, such as the SRT procedure, that is designed to present stimuli around a given point—for example, 50 %. A plateau in the vicinity of that point could make the measured threshold very unreliable. With a three- or four-talker mixture, this plateau disappeared and the psychometric functions displayed a more typical S shape. In the present study, the speech mixture consisted of three talkers, so we did not expect to see any plateau. However, the speech-modulated buzzes/noises were likely to form a single source perceptually, so it was unclear whether the psychometric function would have a standard shape for this masker type. This is another reason why we intended to use both an adaptive task and the task of fixed TMRs to reconstruct the whole psychometric functions and examine their shapes across conditions.

General method

Listeners

For the two experiments using the SRT procedure, the effect of reverberation on informational masking release was assumed to be weak (effect size = 0.2). An a priori power of 0.8 required a sample size of 32 subjects, with alpha = 0.05 and a correlation among repeated measures of 0.7 (using G-Power, ver. 3.1.9.2). Thus, 32 listeners (18 females, 14 males, 18–30 years old) participated in Experiment 1, and 32 other listeners (23 females, nine males, 18–43 years old) participated in Experiment 2. They all provided informed consent in accordance with the protocols established by the Institutional Review Board at the respective institutions, and were compensated at an hourly base rate. All listeners reported normal hearing (audiometric thresholds less than 20 dB HL at octave frequencies between 250 and 8 kHz) and English as their native language. None of them were familiar with the sentences used during the test. Each listener attended a single experimental session that lasted about 60 min.

For the two experiments using the CRM procedure, it was assumed that variance would be limited, because (1) thresholds were extracted from a fit on the whole psychometric function rather than obtained adaptively, and (2) no variance was induced by the different speech materials present in the SRT procedure. Furthermore, the CRM design is particularly suited to informational masking effects; therefore the effect of reverberation on informational masking release was expected to be stronger. Assuming

an effect size of 0.4, achieving an a priori power of 0.8 required a sample size of ten subjects, with alpha = .05 and a correlation among repeated measures of 0.7 (using G-Power). Ten listeners (eight females, two males; 19–26 years old) participated in Experiment 3, and ten other listeners (eight females, two males; 18–34 years old) participated in Experiment 4. They were recruited and screened in a similar manner as for the SRT experiments. Each listener attended three experimental sessions that lasted about 50, 50, and 65 min.

Stimuli and conditions

The speech stimuli used in Experiments 1 and 2 came from the Harvard Sentence List (Rothauser et al., 1969), all spoken by the same male voice, which had a mean F0 of 104 Hz. Eighty sentences were used as the targets, and eight different sentences as maskers. The speech stimuli used in Experiments 3 and 4 came from Bolia et al. (2000). For a given voice, there were 256 combinations of eight call signs ("Charlie," "Ringo," "Laker," "Hopper," "Arrow," "Tiger," "Eagle," and "Baron"), four colors ("blue," "red," "white," and "green"), and eight numbers (1 to 8). The stimuli were presented in four different male voices, resulting in a total of 1,024 sentences in the original materials. In Experiments 1 and 3, the Praat PSOLA package (Boersma & Weenink, 2013) was used to resynthesize each sentence with a fixed F0 throughout, at either 110 or 174.6 Hz (eight semitones higher). In Experiments 2 and 4, all sentences were left unprocessed—that is, naturally intonated.

Two types of masker were generated: two concurrent sentences and nonspeech maskers. Masking sentences were monotonized at 110 Hz (Exps. 1 and 3) or left unprocessed (Exps. 2 and 4), and then added in pairs to create two-voice speech maskers. Nonspeech makers were either speech-modulated buzz (Exps. 1 and 3) or speech-modulated noise (Exps. 2 and 4). Buzz maskers were created from a broadband sine-phase harmonic complex with a 110-Hz F0; noise maskers were created from Gaussian white noise. Both were filtered with a linear-phase FIR filter designed to match the average long-term excitation pattern of the sentences used as speech maskers in each experiment, respectively. In addition, the temporal envelopes of the speech maskers were extracted by half-wave rectification and low-pass filtering (first-order Butterworth with a 3-dB cutoff at 40 Hz) and multiplied with the buzz/noise. Target and maskers were both heard in anechoic and in reverberant conditions.

The virtual room used in all four experiments was 5 m long × 3.2 m wide × 2.5 m high. The listener was simulated as two receivers (omnidirectional microphones) at 1.65 m from the ground, separated by 18 cm. In Experiments 1 and 3, as is depicted in the left panel of Fig. 1, the listener was placed along an axis rotated 25° from the plane parallel to the 5-m wall, on either side of a center point located 1.2 m from the 5-m wall and 2 m from the 3.2-m wall. Sources were all simulated 2 m straight ahead of the listener. In Experiments 2 and 4, as depicted on the right panel of Fig. 1, the listener was located 2.5 m from the 3.2-m wall and 1.0 m from the 5-m wall. The two ears were placed on either side of the axis parallel to the 3.2-m wall, halving the room symmetrically. The maskers were always located 2 m away from the listener on that same axis, and the target was either collocated with the maskers or placed at an equal distance (2 m), but on an axis rotated at 60° from the listener–masker axis.

Reverberation was added using the ray-tracing method (Allen & Berkley, 1979; Peterson, 1986), as implemented in the |WAVE signal processing package (Culling, 1996). Reverberation adds irregular perturbations to the stimulus spectrum, known as room coloration. These perturbations were removed using a FIR filter as part of a package of energetic equalization. In addition, the receivers were suspended in the air with no head between them. The head-shadow and pinna effects generated by the use of a dummy head would have produced another spectral coloration, but, since such effects were all removed from the final stimuli, there was no point in including them in the room model. The absorption coefficients were all .3 for the surfaces of the reverberant room. For the anechoic room, the coefficients were all set to 1. Binaural stimuli were produced by generating the impulse responses for the two receivers in virtual space and convolving the sentences or speech-modulated buzzes/noises with these two impulse responses.

The left panels of Fig. 2 show that the two masker types had almost identical excitation patterns. In the top panels, peaks and dips are observable due to the fixed harmonic structure of the maskers, whereas the excitation is smooth in the bottom panels due to the natural F0 fluctuations (or noise). These excitation patterns were also very similar across rooms, due to the decoloration process. The right panels of Fig. 2 show that in the temporal domain, the two masker types had similar waveforms, in which a few temporal dips were "filled in" to some extent by reverberation. Thus, the two masker types should have produced very similar amounts of energetic masking.

The eight experimental conditions of Experiments 1 and 3 resulted from 2 masker types × 2 ΔF0s × 2 rooms. The eight experimental conditions of Experiments 2 and 4 resulted from 2 masker types × 2 ΔSLs × 2 rooms. All maskers and target stimuli were equalized to the same mean RMS power. Note that in this study, the masker level is always defined as the combined level for the two maskers together (e.g., two competing talkers), and similarly, the TMR is defined as the ratio between the level of a single target talker relative to the combined level of the maskers. This is quite critical as this definition differs in the literature. Therefore, a TMR of 0 dB corresponded here to a situation in which the level of the target talker was 3 dB above that of each masking talker. During the

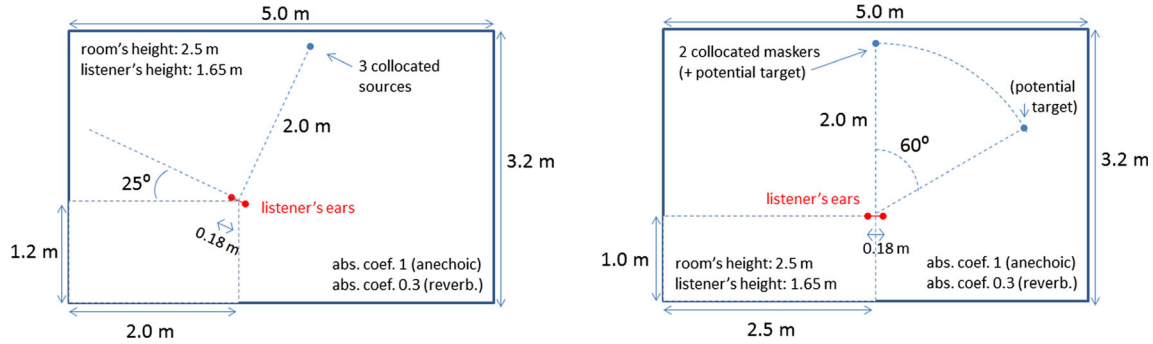

Fig. 1 Spatial configurations and virtual room considered in Experiments 1 and 3 (left panel) and Experiments 2 and 4 (right panel)

adaptive track, changes in TMR occurred by adjusting the target level while presenting maskers always at 69 dB SPL.

SRT procedure

The experimental session began with three practice runs using unprocessed speech, not used in the rest of the experiment, masked by the nonspeech masker (one run) or the speech maskers (two runs), also not used in the rest of the experiment. The following eight runs measured one SRT for each of the eight experimental conditions. Although each of the 80 target sentences was presented to every listener in the same order, the order of the conditions was rotated for successive listeners, to counterbalance effects of order and material. The 32 listeners resulted in four complete rotations of the conditions.

SRT was measured using a one-up/one-down adaptive threshold method (Plomp & Mimpen, 1979), in which an individual measurement is made by presenting successively ten target sentences against the same masker. For the speech maskers, the two transcripts of masking sentences were displayed on a computer screen and nothing was displayed for the buzz/noise maskers. Listeners were specifically instructed not to type the words displayed visually as they belonged to the interfering sentences but to listen to the third sentence. It is useful to remember that all speech stimuli came from the same talker (i.e., there was no difference in the voice characteristics); they all had the same onset; and in the most extreme cases (e.g., in Exp. 1), they could all come from the same position in the same room, being manipulated to have the same steady F0 throughout their whole duration. Therefore, the only cues that remain to define a target from maskers come from the semantic content of the different utterances. Since, in the SRT procedure, there is also no call sign to inform subjects to listen to the words that follow, we had to provide subjects with at least one piece of information to define what the target was. This is why interfering sentences were displayed on the screen in front of them. The TMR was initially at –32 dB and listeners had the opportunity to listen to the first sentence a number of times, each time with a 4-dB increase in TMR. Listeners were instructed to type a

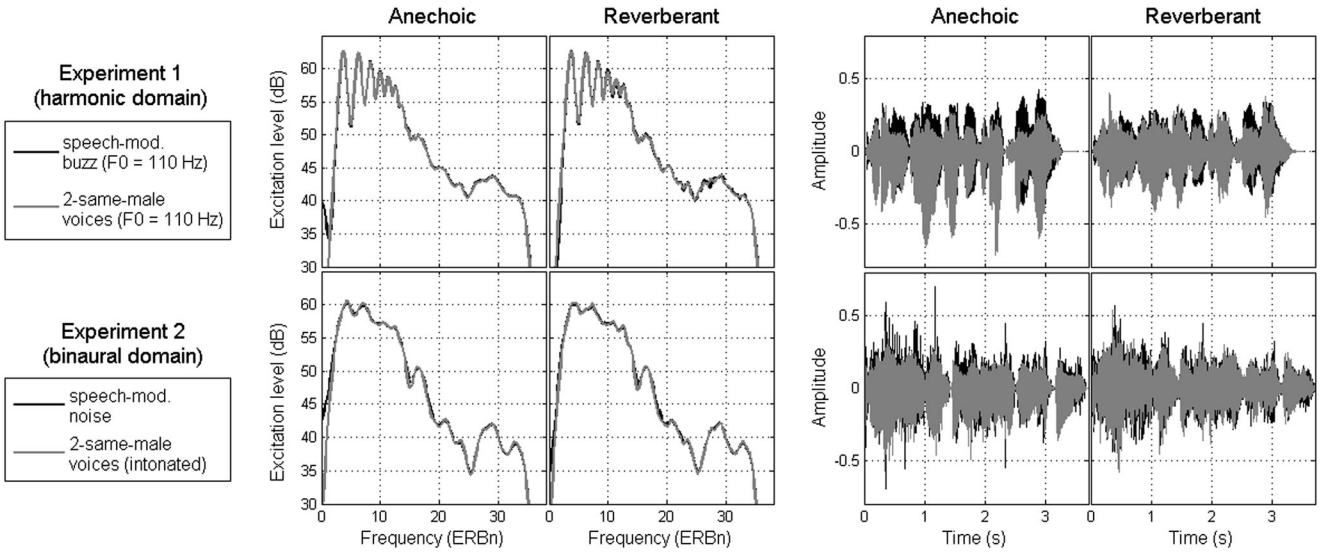

Fig. 2 Averaged excitation patterns (left panels) and example broadband waveforms (right panels) for the two maskers used in Experiment 1 (top panels) and the two maskers used in Experiment 2 (bottom panels), in anechoic and reverberant conditions. For simplicity, only the signals at the right ear are shown. Note that the excitation patterns of the targets shifted by 60° in Experiment 2 were essentially the same as in the bottom panels, due to the room decoloration

transcript when they could first hear about half of the target sentence. The correct transcript was then displayed and the listener self-marked how many key words he/she got correct. Subsequent target sentences were presented only once and self-marked in a similar manner. The level of the target voice decreased by 2 dB if the listener had found three, four, or five correct keywords, and increased by 2 dB if the listener had found two, one, or no correct keywords. Measurement of each SRT was taken as the mean TMR over the last eight trials, and targeted a performance level of 50 % intelligibility.

CRM procedure

Listeners were asked to follow the target voice, which always spoke the call sign "Baron," and to report its coordinates (color and number), chosen randomly, with a mouse click on a monitor that displayed all 32 possible answers. The target voice was always presented concurrently with two maskers, either two speech-modulated buzzes/noises, or two sentences. The call signs, colors and numbers of the two maskers were randomly chosen but were different from each other and different from those of the target. Each of the eight experimental conditions was measured at six different TMRs (chosen from pilot data to cover the full psychometric functions). Performance was measured over 50 trials in each of these 48 conditions. Thus, each subject had to complete a total of 2,400 trials, which were divided into ten experimental blocks (of approximately 240 trials each, taking about 15 min each). Subjects came on three different days, to complete three, three, and four blocks, respectively.

A dynamic stochastic design was used in which the same condition (at a fixed TMR) was presented in clusters of consecutive trials: clusters of three and seven trials occurred once; clusters of four and six trials occurred twice; clusters of five trials occurred four times (for a total of 50 trials). This design enabled us to examine performance as a function of the trial position within a cluster. The rationale was that listeners might take a few trials, every time a new condition was presented, to realize what characteristics of the target voice would be most efficient to track. One might therefore expect to find performance improving with trial position in those particular conditions in which streaming played a great role. Within an experimental block, both the order of the conditions and the cluster sizes were randomized. The last condition of a given block also had to differ from the first condition of the next block. Each subject received a different randomization of condition order and cluster size. Furthermore, the identity of the male talker was kept constant for all sources in one block, but it changed randomly from one block to the next, as well as across subjects, among the four male voices available in the original materials (Bolia et al., 2000), simply ensuring that the results were not tightly dependent upon the specific characteristics of a given voice.

Prior to the start of the experiment, subjects were familiarized with the stimuli and experimental paradigm, by completing 20–40 trials on any of the experimental conditions at random, but making sure that some trials presented the two speech-modulated buzzes/noises and some trials presented the two interfering voices. Within each session, breaks were offered in between blocks.

Equipment

Experiment 1 was performed in the School of Psychology at Cardiff University. Signals were sampled at 20 kHz and 16 bits, digitally mixed, D/A converted by an Edirol UA-20 sound card and amplified by a MTR HPA-2 Headphone Amplifier. They were presented binaurally to listeners over Sennheiser HD650 headphones in a single-walled IAC sound-attenuating booth within a sound-treated room. A computer monitor was visible outside the booth window and a keyboard was inside for transcript responses.

Experiments 2, 3, and 4 were performed at the School of Communication Sciences and Disorders at McGill University. Signals were sampled at 44.1 kHz and 16 bits, digitally mixed, D/A converted by a Focusrite Scarlett 2i4 sound card. They were presented binaurally over Sennheiser HD 280 headphones. The user interface was displayed on a monitor, inside an audiometric booth.

Experiment 1: SRT with ΔF0s

Results

A repeated measures analysis of variance (ANOVA) was conducted to determine the influence of each of the three factors (Room × Masker Type × ΔF0) on the SRTs, shown in the left panel of Fig. 3. The results are reported in Table 1. The three main effects were significant: The mean SRTs were lower when the sources were heard in anechoic rather than reverberant conditions, lower with speech-modulated buzzes than with two-same-male voices, and lower when sources had different F0s than when they had the same F0. As is illustrated in the right panel, the interaction between ΔF0 and masker type was significant; that is, the masking release provided by the ΔF0 was larger with two-same-male voices than with buzz maskers, but this was particularly the case in the anechoic relative to the reverberant room (three-way interaction).

Discussion

Reverberation only affected the masking release obtained with speech maskers For the speech-modulated buzzes, the ΔF0 benefit was about 5 dB in both the anechoic and reverberant conditions. As was intended with keeping all sources

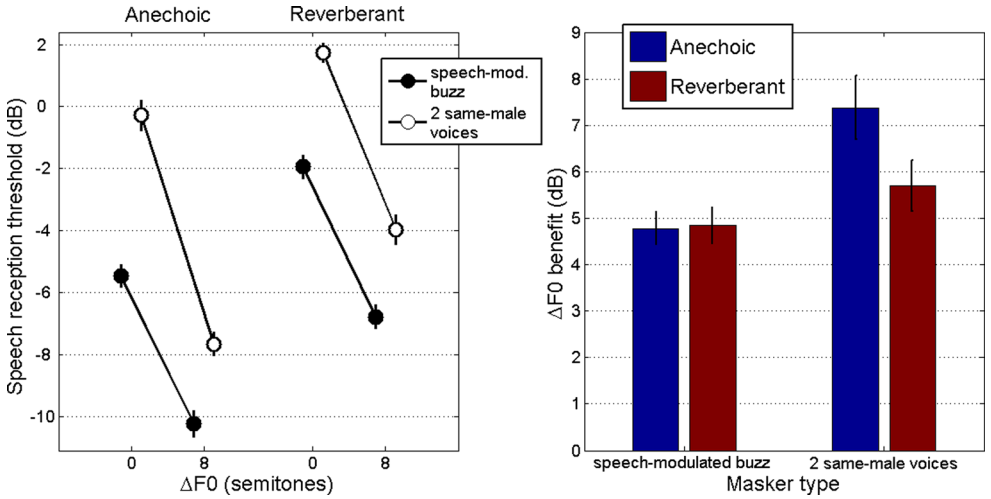

Fig. 3 (Left) Mean speech reception thresholds measured in Experiment 1, in the anechoic and reverberant conditions, for the two types of masker (speech-modulated buzz and two monotonized voices), with and without a ΔF0 with the target. Lower thresholds indicate greater intelligibility. (Right) Mean ΔF0 benefits for each masker type and each room. Error bars indicate ±1 standard error of the mean across subjects

monotonized, the release from masking (presumably largely energetic for this masker type) was robust to reverberation. For speech maskers, as expected, the SRTs were substantially elevated, despite presenting similar amounts of energetic masking (Fig. 2). Without ΔF0, the SRT was 5 dB higher with the two voices than with buzzes in anechoic conditions. A major part of this elevation was presumably due to informational masking. Exactly what form of informational masking occurred, though, is arguably difficult to determine. One may think of it as attention capture (e.g., Colflesh & Conway, 2007). Here, for example, since three utterances were spoken simultaneously by the same male talker, there was great uncertainty as to which sentence listeners should attend to. Another way to think of this is semantic confusion (e.g., Sörqvist & Rönnberg, 2012). Here, for example, the two interfering utterances written on the screen could have caused cross-modal distraction, whereas this would not have happened with the buzz maskers since nothing was displayed on the screen for this masker type. Note, however, that the latter form of distraction would presumably remain constant, whether or not a ΔF0 was present. Therefore the release from masking obtained here from the eight-semitones ΔF0 between the competing sentences is unlikely to come from a reduction in semantic confusion, and must be some combination of reduced energetic masking by harmonicity-based processes and an enhanced selective attention. The focus of the

Table 1 Statistics for the thresholds and slopes extracted at 50 % intelligibility in each experiment, following an ANOVA with three within-subjects factors: Room (anechoic vs. reverberant), Masker Type (speech-modulated buzz/noise vs. two-same-male voices), and Cue (ΔF0 in Exps. 1 and 3, or ΔSL in Exps. 2 and 4)

	Thresholds				Slopes	
	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 3	Exp. 4
Room	$F(1, 31) = 139.3$ $p < .001$	$F(1, 31) = 89.8$ $p < .001$	$F(1, 9) = 96.1$ $p < .001$	$F(1, 9) = 85.0$ $p < .001$	$F(1, 9) = 0.3$ $p = .587$	$F(1, 9) = 19.6$ $p = .002$
Masker Type	$F(1, 31) = 227.3$ $p < .001$	$F(1, 31) = 45.5$ $p < .001$	$F(1, 9) = 2{,}262.1$ $p < .001$	$F(1, 9) = 498.9$ $p < .001$	$F(1, 9) = 87.4$ $p < .001$	$F(1, 9) = 17.8$ $p = .002$
Cue	$F(1, 31) = 538.0$ $p < .001$	$F(1, 31) = 401.2$ $p < .001$	$F(1, 9) = 204.5$ $p < .001$	$F(1, 9) = 192.4$ $p < .001$	$F(1, 9) = 41.1$ $p < .001$	$F(1, 9) = 38.5$ $p < .001$
Room × Masker Type	$F(1, 31) = 1.7$ $p = .204$	$F(1, 31) = 0.5$ $p = .495$	$F(1, 9) = 65.7$ $p < .001$	$F(1, 9) = 21.1$ $p = .001$	$F(1, 9) = 3.6$ $p = .088$	$F(1, 9) = 0.2$ $p = .697$
Room × Cue	$F(1, 31) = 2.9$ $p = .098$	$F(1, 31) = 3.6$ $p = .066$	$F(1, 9) = 10.7$ $p = .010$	$F(1, 9) = 8.7$ $p = .016$	$F(1, 9) = 3.1$ $p = .111$	$F(1, 9) = 1.6$ $p = .243$
Masker Type × Cue	$F(1, 31) = 7.4$ $p = .011$	$F(1, 31) = 76.8$ $p < .001$	$F(1, 9) = 66.1$ $p < .001$	$F(1, 9) = 1.9$ $p = .198$	$F(1, 9) = 25.9$ $p = .001$	$F(1, 9) = 11.0$ $p = .009$
Three-way interaction	$F(1, 31) = 4.4$ $p = .045$	$F(1, 31) = 5.0$ $p = .033$	$F(1, 9) = 50.0$ $p < .001$	$F(1, 9) = 13.0$ $p = .006$	$F(1, 9) = 1.1$ $p = .330$	$F(1, 9) = 0.8$ $p = .388$

present study was to examine a potential effect of reverberation on this latter benefit—that is, the informational component. The right panel of Fig. 3 illustrates that the ΔF0 benefit obtained with two masking voices was reduced in reverberant as compared to anechoic conditions. This three-way interaction therefore suggests that reverberation affects the informational component of the ΔF0 benefit, consistent with the results obtained by Darwin and Hukin (2000).

Known effects of reverberation It is known that the intelligibility of a voice is degraded in reverberation. The delayed reflections from the walls reduce the modulations of the within-channel temporal envelopes. To put it more simply, the voice is temporally blurred and loses articulation in reverberation (Houtgast & Steeneken, 1985; Steeneken & Houtgast, 1980). Being independent of any other masking effects involved, this loss of modulation transmission should have occurred similarly, whether or not there was a ΔF0 and whatever the masker type. Note that the magnitude of this effect was quantified at 2 dB by Deroche and Culling (2011, Fig. 4), who used an identical room configuration.

It is also well established that listeners can "listen in the dips" of a temporally fluctuating masker (de Laat & Plomp, 1983; Festen & Plomp, 1990; Hawley, Litovsky, & Culling, 2004). Although the present maskers consisted of two simultaneous utterances, listeners could have exploited remaining dips. Furthermore, this exploitation is known to be facilitated when the same maskers are used throughout a block of sentences, as here, because listeners have an expectation of when dips will happen (Collin & Lavandier, 2013). Reverberation, however, "fills-in" to some extent the temporal dips in the masker waveforms, which prevents their exploitation (Beutelmann, Brand, & Kollmeier, 2010; Bronkhorst & Plomp, 1990; Collin & Lavandier, 2013; George, Festen, & Houtgast, 2008). This represents a second, detrimental, effect of reverberation, but its magnitude is not trivial to estimate. Particularly, it is not clear whether or not this "filling-in" effect should have occurred similarly for the two masker types. Some evidence suggests that at least for modulated noises, the synchronization of dips across frequency will provide more benefit than dips set to be antiphasic in adjacent frequency channels (Howard-Jones & Rosen, 1993). So it seems plausible that dip-listening is a little more advantageous for the modulated buzzes in which dips are co-timed across frequency than for the two-voice maskers in which dips are more randomly distributed across frequency. It follows that the "filling-in" effect of reverberation could, in turn, be slightly more detrimental for the modulated buzzes than for the two-voice maskers. Regardless of its magnitude, the "filling-in" effect of reverberation should have occurred similarly whether the ΔF0 was zero or eight semitones, and therefore this phenomenon does not stand either as a potential candidate to explain the reduction in the ΔF0 benefit observed with interfering voices when introducing reverberation.

Experiment 2: SRT with ΔSLs

Results

A repeated measures ANOVA was conducted to determine the influence of each of the three factors (Room × ΔSL × Masker Type) on the SRTs, shown in the left panel of Fig. 4. The results are reported in Table 1. The three main effects were significant: The mean SRTs were lower when the sources were heard in anechoic rather than reverberant conditions, lower with speech-modulated noise than with two interfering voices, and lower when the sources had different SLs than when they

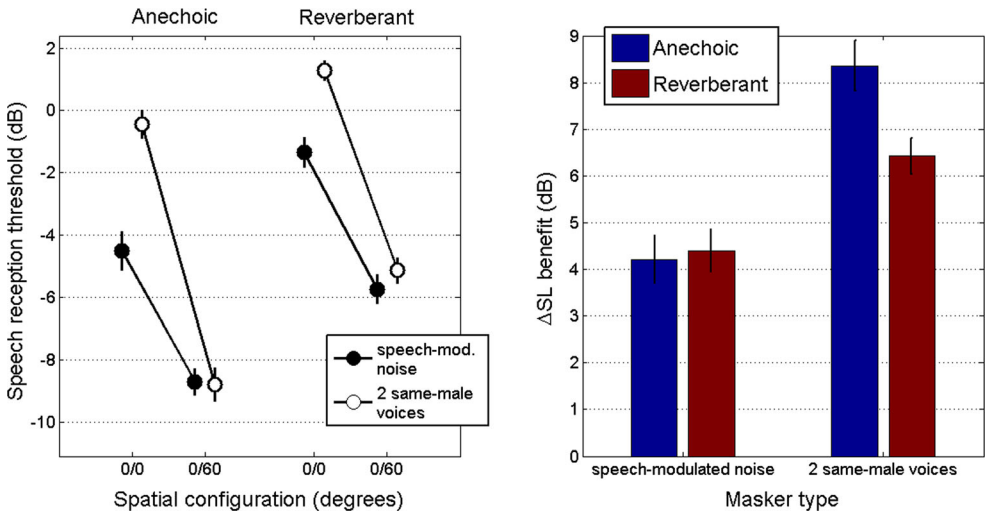

Fig. 4 (Left) Mean speech reception thresholds measured in Experiment 2, in the anechoic and reverberant conditions, for the two types of masker (speech-modulated noise and two naturally intonated voices), with and without a ΔSL with the target. (Right) Mean ΔSL benefits for each masker type and each room. Error bars indicate ±1 standard error of the mean across subjects

were collocated. As is illustrated in the right panel, the interaction between ΔSL and masker type was significant; that is, the spatial release from masking was larger with two interfering voices than with noise maskers, but this was particularly the case in the anechoic relative to the reverberant room (three-way interaction).

Discussion

Reverberation only affected the masking release obtained with speech maskers For speech-modulated noises, the ΔSL benefit was about 4 dB in both anechoic and reverberant conditions. The spatial release from masking (presumably largely energetic for this masker type) was therefore robust to reverberation. Note that this result was not trivial to obtain; it required a very specific listening configuration with the masker and listener both positioned with the room symmetrical about them, so that the masker interaural coherence was intact in reverberation, and it also required removing any effect of interaural level differences (i.e., having no virtual head and cancelling the room coloration). This result provides strong support for the idea that the detrimental effect of reverberation on binaural unmasking is at least partly mediated by disruption in the masker coherence (Lavandier & Culling, 2007, 2008). For speech maskers, on the other hand, there was some uncertainty as to which sentence one should attend to: Without ΔSL, the SRT was 4 dB higher with speech maskers than with noise maskers in anechoic conditions, but listeners could use the 60° separation to release from energetic as well as informational masking, resulting in a greater ΔSL benefit with speech maskers than with noise. The focus of the present study was to examine the potential effect of reverberation on this latter benefit. The right panel of Fig. 4 illustrates that the ΔSL benefit obtained with speech maskers was reduced in reverberant as compared to anechoic conditions. Therefore, just as it did in the harmonic domain in Experiment 1, this three-way interaction would suggest that reverberation affects the informational component of the ΔSL benefit.

Known effects of reverberation As before, the main effect of reverberation reflected (1) the degradation in articulation of the target voice and (2) a possible "filling-in-the-dips" effect, which perhaps could be more detrimental for noise maskers than for speech maskers, since the co-timing of dips in modulated noise could have more of an influence (Howard-Jones & Rosen, 1993). But in any case, these expected effects of reverberation would have occurred similarly whether the sources were collocated or spatially separated, and therefore they do not stand as a potential candidate to explain the reduction in the ΔSL benefit observed with speech maskers when introducing reverberation.

Experiment 3: CRM with ΔF0s

Results

The symbols displayed in Fig. 5 represent performance averaged over the ten subjects for each of the eight experimental conditions spanning six different TMRs. In each condition, performance was as low as 30 % or less at the lowest TMR, and as high as 90 % or more at the highest TMR, confirming that the range of TMRs chosen for each condition was sufficiently broad to cover most of the psychometric function and to get reliable estimates of the thresholds and slopes at 50 %. A maximum-likelihood technique with Gaussian priors was used to fit a logistic function to the data collected for each subject individually. The lines and areas in Fig. 5 are the means and standard errors of the fits in each condition. From the individual fits, a TMR corresponding to 50 % performance was extracted and served as the basis for the statistical analysis. The corresponding mean thresholds are shown in the left panel of Fig. 6.

The results of the ANOVA are reported in Table 1. The main effects were all significant, reflecting that the thresholds were lower in anechoic than in reverberant conditions, lower with buzzes than with masking voices, and lower with than without ΔF0. All interactions were significant, including, most importantly, the three-way interaction. As is illustrated in the right panel of Fig. 6, the ΔF0 benefit was larger with masking voices than with buzzes, but this was particularly the case in anechoic relative to reverberant conditions.

For each subject, the slope of the logistic fits at 50 % performance was also extracted and was submitted to a similar ANOVA (Table 1). We observed a main effect of masker type, a main effect of ΔF0, and both strongly interacted. As is shown in the left panel of Fig. 7, the psychometric functions for the conditions of interfering voices monotonized at the same F0 as the target were almost twice as steep as the functions for the other six conditions.

Further analyses were performed to examine (1) the types of errors made for each experimental condition and (2) the potential effect of trial position within clusters. These results were somewhat irrelevant to the present focus (i.e., the three-way interaction), and therefore are presented in the Appendix.

Discussion

The results of Experiment 3 were qualitatively similar to those of Experiment 1. Perhaps, the most obvious difference was the scale of thresholds obtained with buzzes, ranging between –9 and –16 dB in Fig. 6 (as compared to –2 and –10 dB in Fig. 3), whereas the scale of thresholds obtained with masking voices was relatively constant. This was very likely due to the predictability of the sentences of the CRM corpus and the closed-set characteristics of the task. The CRM poses very few

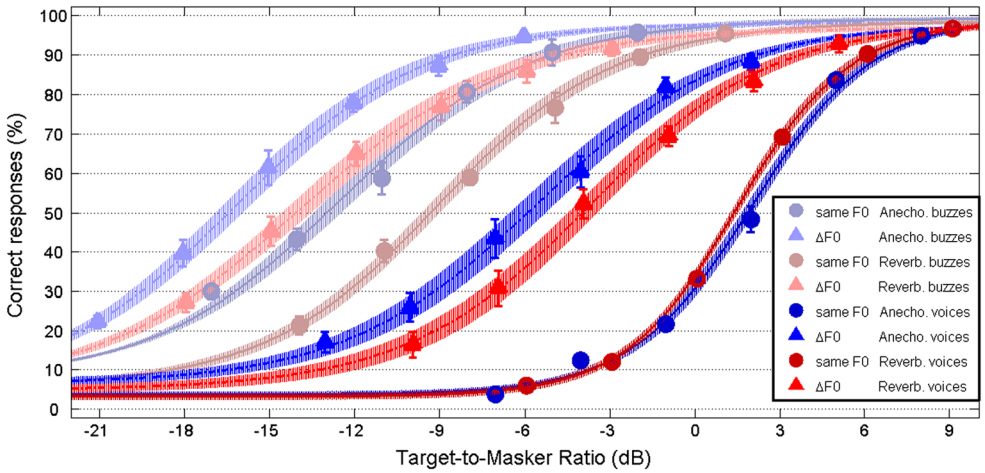

Fig. 5 Mean performance (symbols) collected with the CRM design in Experiment 3, for each of the eight experimental conditions (Anechoic vs. Reverberant Room × Buzz vs. Masking Voices × ΔF0 vs. Same F0 as the Target). Using the maximum-likelihood technique, logistic functions were fitted to the individual-subject data measured at six different target-to-masker ratios, chosen to span most of the function for each condition. Error bars on the symbols indicate ±1 standard error of the mean across subjects, and the widths of the logistic fits indicate ±1 standard error of the mean fit

demands in terms of intelligibility, because the same utterances are presented over and over again. In the absence of any confusion between sources—that is, with buzz maskers—decoding very little information, such as a phoneme <e> followed by a phoneme <ʊ>, could be sufficient to reconstructing "red two" and potentially produce a correct response. This is why the thresholds for buzzes could be much lower in the CRM than in the SRT task. At these very low TMRs (e.g., –16 dB), a floor effect might have limited the release seen in the anechoic condition. This may simply be why the ΔF0 provided a larger masking release in reverberation than in anechoic conditions in this experiment, an effect that did not occur in Experiment 1. Another notable difference concerns the interfering voices in the absence of ΔF0: Introducing reverberation did not elevate the thresholds further, where it had, by 2 dB, in Experiment 1. This, again, was very likely due to the fact that listeners did not attempt to decipher the target utterance; they knew roughly what it was supposed to say. Therefore, one should perhaps not expect any detrimental effect of the temporal smearing of the target speech by reverberation. These differences set aside, the key result was that the ΔF0 benefit obtained in the presence of two competing voices was reduced in reverberation, which from an energetic-masking perspective should not have happened.

By having access to the full psychometric function of each experimental condition, we could verify that all

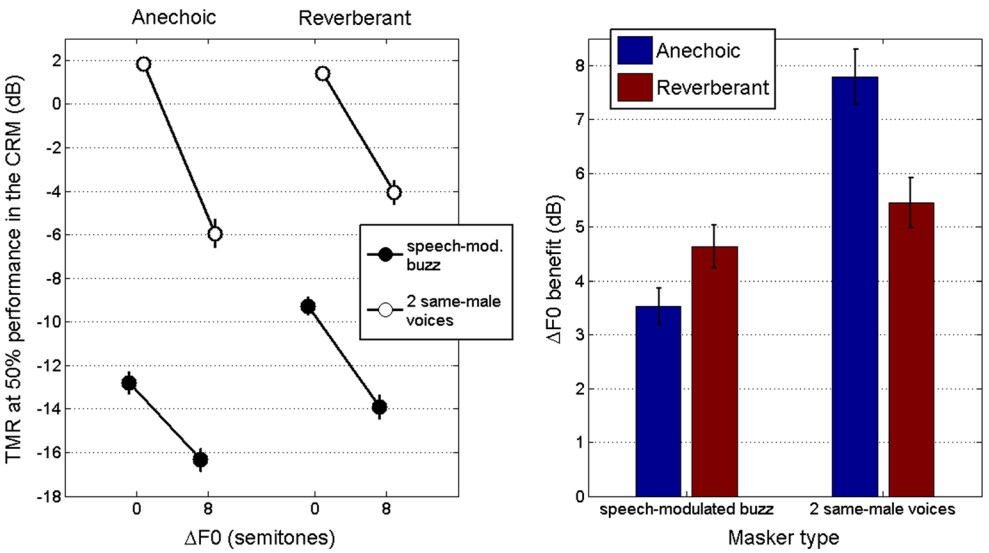

Fig. 6 (Left) Mean CRM thresholds obtained in Experiment 3, extracted from the logistic fits of each subject at 50 % performance. Lower thresholds indicate better performance. (Right) Mean ΔF0 benefits for each masker type and each room. Error bars indicate ±1 standard error of the mean across subjects

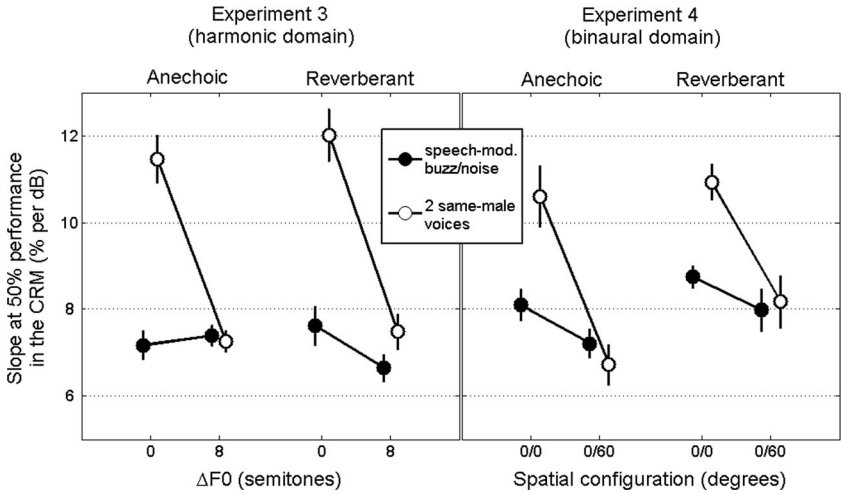

Fig. 7 Mean CRM slopes extracted from the logistic fits of each subject at 50 % performance, in the harmonic domain (left) and the binaural domain (right)

displayed monotonic S shapes. There was no plateau that could have prevented the adaptive procedure from working properly, as in Experiment 1; therefore, this potential confound can be discarded.

We also take a closer look at the range of TMRs covered by each experimental condition. What is striking is that, in the two cases of interfering voices monotonized at the same F0 as the target voice (with and without reverberation), performance was so poor that both functions lay mostly beyond –3 dB. This is the boundary beyond which the target voice started to be louder than the two other sentences. These two functions displayed steeper slopes than any other condition (rightmost curves in Fig. 5 and the left panel of Fig. 7), providing compelling evidence that loudness cues enhanced performance abnormally quickly as TMR increased beyond –3 dB. This result supports a distinction made by Brungart et al. (2001), in their investigation of multitalker mixtures, between cases in which the target talker was more intense than any masking voice and cases in which it was less intense than at least one masking voice. First, performance was much more dependent on the similarities between competing voices at positive TMRs than at negative TMRs. Second, performance unexpectedly increased with the number of talkers at positive TMRs (defined, as here, from the combined masker level), whereas it dropped considerably when more than one masking voice was presented at negative TMRs. Crudely, the rationale is that performance has more to do with selective attention at positive TMRs but more to do with peripheral mechanisms at negative TMRs. This distinction raises the possibility that loudness cues (beyond –3 dB TMR in the present study) could have been used to release from informational masking. But what is critical here is that this sort of ceiling effect occurred similarly for both anechoic and reverberant conditions, and therefore it could hardly have caused the three-way interaction. To clarify, in Experiment 1 it seemed possible that the ceiling effect was stronger in the reverberant than in the anechoic condition, because there was a 2-dB difference in the SRTs between the two (Fig. 3). This was no longer the case in Experiment 3, since, in fact, the reverberant threshold was 0.5 dB *lower* than the anechoic threshold (Fig. 6). Another way to unfold the argument is to look at performance at a fixed TMR. At a TMR of –1 dB, for example, the target voice was a little louder than each masking voice; this loudness cue was identical, whether or not a ΔF0 was present and whether the room was anechoic or reverberant, yet the effect of interest was still present: The eight-semitone ΔF0 provided a 60 % increase in performance in anechoic conditions, but only a 45 % increase in performance in reverberant conditions (Fig. 5). Thus, the idea that the three-way interaction was caused by a ceiling effect is unconvincing, at least in this experiment using the CRM design.

As we mentioned above, reverberation does blur the modulations of speech, but it does so equally for the target and the masking voices. Its impact on the target is generally detrimental because intelligibility of a voice relies upon the transmission of these modulations. Its effect on the masking voices, however, could well be beneficial. By making the interfering voices less intelligible, in a way more "noise-like," reverberation also makes them less efficient as informational maskers. This phenomenon could be equivalent to the effect of number of talkers at positive TMR observed by Brungart et al. (2001), mentioned earlier. As the number of masking voices increases, each voice is made progressively less intelligible and merges into babble. This reduces the chances that listeners would switch their selective attention into any one of them, which could explain why performance at positive TMR actually increases with more masking voices. Reverberation duplicates several slightly different versions of the same masking voices, so it might perhaps act similarly to increasing the number of interfering utterances. However, if this were so, one might

have expected reverberation to reduce the number of wrong-voice errors, but this was not apparent in the data (see Fig. 10).

Experiment 4: CRM with ΔSLs

Results

Figure 8 shows the mean performance (symbols) and fits (lines), averaged over the ten subjects. In each condition, performance was measured as low as 25 % or less at the lowest TMR, and as high as 90 % or more at the highest TMR, confirming that the range of TMRs chosen for each condition was sufficiently broad to cover most of the psychometric function. Thresholds at 50 % were extracted for each subject and submitted to the ANOVA whose results are reported in Table 1. As is shown in the left panel of Fig. 9, the thresholds were lower in anechoic than in reverberant conditions, lower with noises than with masking voices, and lower with than without ΔSL, resulting in three main effects. Most importantly, the three-way interaction was significant: As is illustrated on the right panel, the ΔSL benefit was larger with masking voices than with noises in anechoic but not in reverberant conditions.

Slopes were also extracted at the 50 % point and submitted to a similar ANOVA whose results are also reported in Table 1. The three main effects were significant, and masker type interacted with ΔSL. As is shown in the right panel of Fig. 7, the psychometric functions for the two conditions of collocated interfering voices were steeper than the functions for the other six conditions.

Discussion

The results of Experiment 4 (Fig. 9) were qualitatively similar to those of Experiment 2. The main difference was the lower scale of the thresholds obtained for noise maskers. The ΔSL benefit obtained for noise maskers tended to increase when introducing reverberation (although this trend did not reach significance here, p = .067). Visual inspection of the psychometric functions revealed no indication of any plateau in any of the tested conditions. They all displayed typical S shapes, within which the adaptive task used in Experiment 2 seems perfectly appropriate. The functions for the collocated voices (rightmost curves in Fig. 8 and right panel in Fig. 7) were steeper than the other six functions, suggesting that the loudness of the target voice at these relatively high TMRs might have served to release from informational masking, but, critically, it would have done so similarly in both conditions, and therefore cannot account for the 2-dB difference that emerged when ΔSL was present. The proportions of wrong-voice errors were also similar (Fig. 10) in the reverberant and anechoic conditions, providing no support for the idea that reverberation produced less informational masking by blurring the interfering voices.

As we mentioned in the introduction, Kidd, Mason, Brughera, and Hartmann (2005) used the CRM design to examine the effect of reverberation on spatial release from masking. Their study is therefore most relevant to this Experiment 4. They found that, for the same-band noise masker (i.e., primarily energetic), reverberation reduced the spatial benefit. This was due to the loss of head-shadow and disruption in interaural coherence of the masker. These energetic effects (discussed in the introduction) were avoided in the present study, and this is why our results differed for the noise maskers. Interestingly, however, reverberation did not reduce the spatial benefit obtained with the different-band speech masker (i.e., primarily informational) in their study. This result is in straight contradiction with the present results as it suggests that the spatial release from informational masking is largely insensitive to reverberation. One possible explanation for this apparent discrepancy may come from the distinct forms

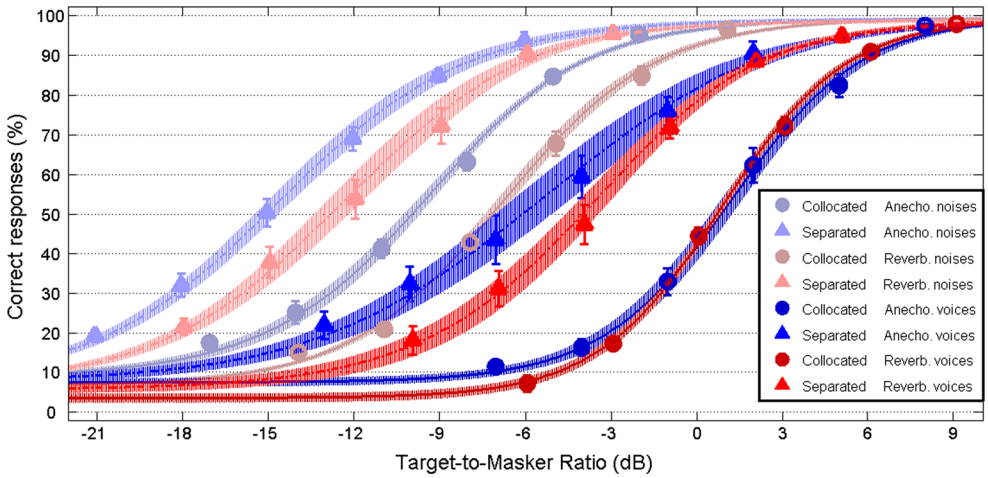

Fig. 8 Same as Fig. 5, but in the binaural domain (Exp. 4)

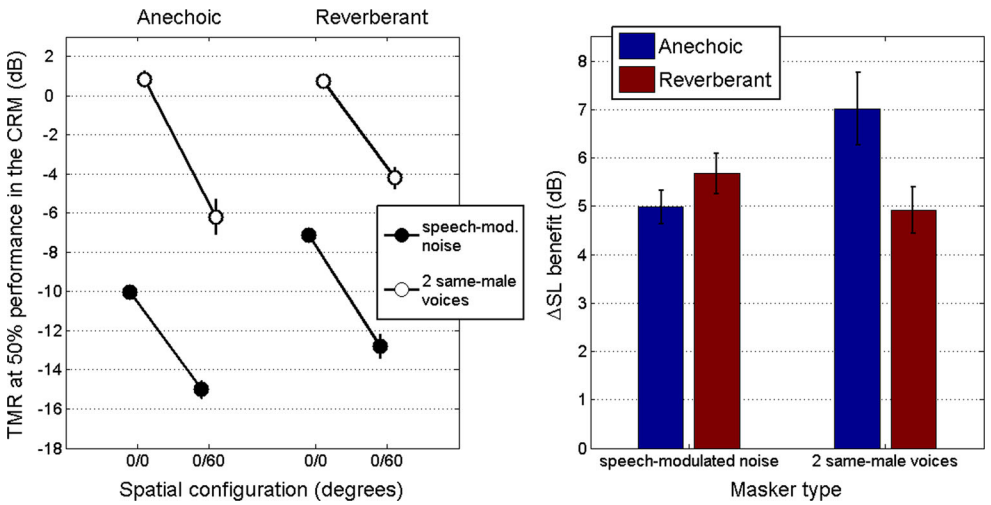

Fig. 9 Same as Fig. 6, but in the binaural domain (Exp. 4)

of reverberation. In realistic settings like the one used by Kidd, Mason, Brughera, and Hartmann, it may be that listeners can learn the acoustic characteristics of different reverberant rooms and make a stronger use of binaural cues. Slight movements of the head and immersion in the room may make listeners more robust to reverberation than when reverberation is imposed directly on the stimuli and presented over headphones (as in here). Some evidence is emerging that different degrees of reverberation between speech sources could itself act as a cue to release from informational masking (Westermann & Buchholz, 2015). So, it may be that knowledge of the listening environment helps in localizing sources and enhances the precedence effect (Freyman, Helfer, McCall, & Clifton, 1999) compared with a simulated environment.

General discussion

This study presented four experiments intentionally designed to have a very similar framework, using two different methods (SRT or CRM) and two different perceptual segregation cues (harmonic or binaural). The strength of this study is that a similar pattern of results was observed in all four experiments. Thresholds were considerably elevated in the presence of interfering voices relative to nonlinguistic analogs, presumably because speech maskers involved informational masking whereas nonlinguistic maskers did not (or very little). This distinction was certainly supported by the analysis of error types in the CRM design (Exps. 3 and 4; see the Appendix), which can be taken as evidence that attention capture occurred in the

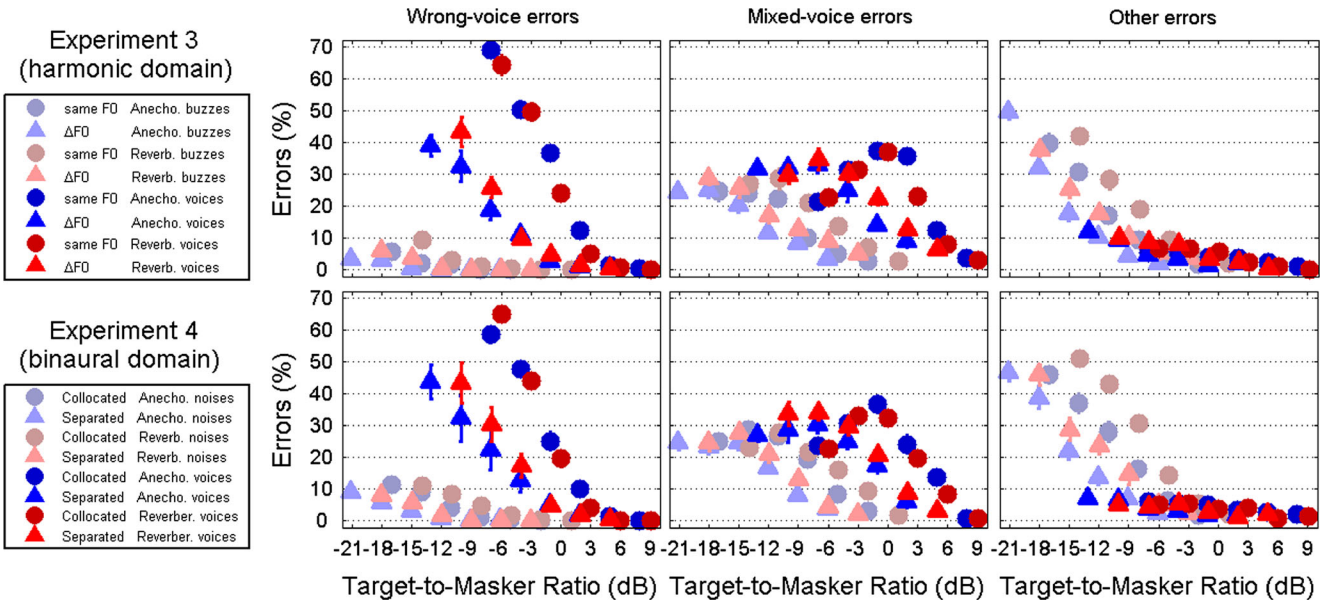

Fig. 10 Analysis of the types of errors made in the CRM tasks of Experiments 3 (top panels) and 4 (bottom panels), as a function of target-to-masker ratio. Errors were categorized into three types: "wrong-voice," "mixed-voice," and "other"

presence of masking voices but not in the presence of nonlinguistic maskers. Whether attention capture occurred in the SRT experiments as well is less certain. Listeners were specifically instructed not to type the masking words written on the screen, and consequently errors in their transcripts rarely contained masking words. The cue under investigation, a ΔF0 or a ΔSL, provided a masking release for nonlinguistic maskers, between 3.5 and 6 dB. This benefit was larger for speech maskers, between 5 and 8.5 dB, because the cue provided a release from both energetic and informational masking. The objective of the study was to examine the effect of reverberation on these benefits, while limiting any energetic-based account for this effect. This was done by presenting a specific room/source configuration and keeping F0s steady. These manipulations were successful in presenting listening situations in which reverberation did not impair the benefits obtained with nonlinguistic maskers. Yet, the benefits obtained with speech maskers were reduced by reverberation.

Since each Experiment followed a similar format, it was possible to analyze the thresholds of the four experiments together to investigate the potential influences of the task and domain of investigation. A repeated measures ANOVA was performed with five factors, the three within-subjects factors used in each individual experiment, and two between-subjects factors (Task and Domain). The three-way interaction between room, masker type, and cue did not interact with the task, did not interact with the domain, and did not interact with Task × Domain. In other words, the key finding occurred similarly regardless of the task/speech materials and whether masking releases were provided by ΔF0s or ΔSLs.

The meaning of these three-way interactions, however, remains unclear, because several interpretations to account for the fact that reverberation reduced the masking releases obtained in a three-talker mixture seem plausible and are not mutually exclusive. First, perhaps the most speculative hypothesis is that reverberation has a genuine impact on selective auditory attention. Ultimately, the voice segregation task requires listeners to store the target message temporarily. Working memory must presumably have a limited processing capacity: the more resources are allocated to word identification, the fewer resources are left for storage. For instance, Kjellberg, Ljung, and Hallman (2008) presented orally 50 one-syllable words to listeners either in quiet or in a background noise. Words were separated by 3 or 4 s, during which listeners were asked to repeat aloud each word to check for their intelligibility. At the end of a set, listeners were asked to write down all the words they could recall. Recall was impaired by the background noise although the words were all identified correctly. Ljung and Kjellberg (2009) used a similar reasoning but tested the influence of reverberation rather than background noise. They found that listeners recalled a smaller number of words spoken in reverberation, although again words were correctly identified. Thus, there may be a trade between the processing of a degraded speech signal and more cognitive mechanisms. Tracking a voice over time on the basis of its F0 or its SL is certainly different from the early consolidation of long-term memory but some form of attention may be necessary in both. The more degraded a voice, the harder it may be to attend to it. Speech being degraded in reverberation, it may thus be harder to attend to certain characteristics of a reverberant voice in the context of competitors.

The second hypothesis was that reflections in a reverberant room may duplicate slightly different copies of the interfering sentences and blur them, such that the combined masker could be getting closer to the percept of a multitalker babble in which each masking source would be less likely to interfere with the listener's ability to track the target voice. Although the present data did not offer any direct evidence for this interpretation (since there were similar proportions of wrong-voice errors in the CRM experiments; see Fig. 10), its rationale is consistent with the pattern of results observed.

A third hypothesis is that uncertainty about which voice to attend to at a cocktail party diminishes as soon as the target voice becomes louder than the masking voices. A salient loudness cue would therefore serve as a way to release from informational masking. The reason why this phenomenon would affect the reverberant conditions more than the anechoic conditions is that speech intelligibility tasks are generally harder under reverberant conditions, and thus require higher TMRs to achieve a similar level of performance as in anechoic conditions. For instance, in Experiments 1 and 2, SRT was 2 dB higher in reverberant than in anechoic conditions, in the three-talker mixture in the same-F0/same-SL conditions. Consequently, it is plausible that the loudness of the target voice at +2-dB TMR was more effective at releasing informational masking than at 0-dB TMR. Thus, any detrimental effect of reverberation at this point (be it in the form of temporal smearing of the target or filling-in the masker dips) could have been counteracted to some extent by the salience of the target voice, causing the three-way interaction. The problem is that, in Experiments 3 and 4, performance in the same-F0/same-SL conditions was similar in both anechoic and reverberant conditions, and yet a difference was still observed in the ΔF0/ΔSL conditions. In other words, none of those interpretations is fully convincing.

Note that both the second and third interpretations share a common idea, that there is less informational masking to begin with in the same-F0/same-SL reverberant conditions, and consequently less room for masking release to occur, than in anechoic conditions. In contrast, the first interpretation is more straightforward in that the amount of informational masking is

similar to start with, regardless of the room, but is not as effectively released in reverberation.

Several studies in the literature offer further support for the third interpretation. Using the CRM design, Arbogast et al. (2005) examined the amount of spatial release from masking caused by a 90° separation between a target voice and a masker. They used three masker types: a same-band noise masker, a different-band noise masker, and a different-band speech masker. They recruited both normal-hearing listeners and listeners with sensorineural hearing loss. For normal-hearing listeners, they found thresholds of –25 and –3 dB, respectively for the different-band noise and speech masker condition collocated with the target. This 22-dB difference stemmed largely from informational masking, and it was partly released by the 90° separation, which, in addition to the energetic benefit (about 6 dB, seen from the same-band noise masker), resulted in a total of 15 dB spatial release from masking. The interesting result came from the hearing-impaired listeners: they displayed thresholds of –12 and 0 dB, respectively for the different-band noise and speech masker condition collocated with the target. Thus, instead of 22 dB, they observed only 12 dB of informational masking to begin with, and consequently obtained a smaller informational benefit from the 90° separation, although their energetic benefit was actually similar to that of normal-hearing listeners. The authors argued that this phenomenon might be due to the ceiling effect occurring in the vicinity of 0-dB TMR. When the level of the target voice exceeded that of the masking voice, the salient loudness of the target might have resolved any confusion preexisting between the two talkers.

Later on, Freyman, Balakrishnan, and Helfer (2008) followed up on the same reasoning, by using noise-vocoded sentences rather than recruiting hearing-impaired listeners, in order to drive performance into a positive range of TMR. Using the precedence effect with a spatial separation that was known to produce a large release from informational masking (Freyman et al., 1999), they found no spatial benefit at all in this very high range of TMR between +3 and +24 dB, necessary to understand vocoded speech in this condition. In a second experiment, listeners were asked to simply detect the presence of target words. This less demanding task was performed at negative TMRs, and large spatial benefits were once again observed. In line with the interpretation of Arbogast et al. (2005), the authors suggested that at positive TMR, the loudness of the target voice might be such that any confusion with the masking voice is already resolved, and therefore a release from informational masking is unlikely to occur.

More recently, Best, Marrone, Mason, and Kidd (2012) used the CRM design with a target voice against two masking voices or two time-reversed masking voices with and without spatial separation. They found that the spatial release from masking was indeed smaller for hearing-impaired than for normal-hearing listeners, and more so for the forward maskers than the reversed maskers. This phenomenon seemed to be due, again, to a ceiling of threshold at positive TMRs. In a second experiment, they asked normal-hearing listeners to do the task with several degrees of noise-vocoding. Reducing the number of spectral channels generally increased thresholds but had progressively less effect in conditions under which threshold was already high. As a consequence, the spatial benefit was smaller with more degraded sentences and also smaller with forward than with reversed maskers.

Taken together, these studies show quite convincingly that at positive TMRs performance quickly asymptotes, creating interactions between spatial release from masking and hearing loss or masker type. Our present results (particularly the steeper slopes of the psychometric functions, lying mostly above –3 dB) are generally consistent with this idea. But, as we mentioned earlier, this interpretation can no longer explain the interaction in cases in which baseline performance is identical between anechoic and reverberant conditions, as in Experiments 3 and 4. Therefore, we believe that several phenomena could be at play to explain why the ΔF0 and ΔSL benefits are smaller in reverberant speech mixtures.

Author note This research was partly supported by a UK EPSRC grant awarded to J.F.C. and partly supported by a NSERC grant awarded to V.L.G. We are grateful to the 84 subjects for their time and effort.

Appendix

Error types

To better appreciate why performance in the CRM decreased with TMR in the different conditions, errors were categorized into three types. Errors were labeled “wrong-voice” when listeners selected both coordinates from the maskers. They were labeled “mixed-voice” when listeners selected one of the coordinates (color or number) from the target, and one from one of the maskers. Finally, errors were labeled “other” when at least one of the coordinates was not present in the trial. Figure 10 shows percentages of these three error types in the two experiments that used the CRM. It is apparent that, as TMR decreased, listeners responded with the coordinates of one of the two maskers, only when these maskers possessed a linguistic content—that is, for two-same-male voices—and particularly in the absence of ΔF0 or ΔSL (left panels). One must bear in mind that the probability of making a wrong-voice error simply by chance was the probability of picking a masker color (2/4) multiplied by the probability of picking a masker number (2/8)—that is, 12.5 %. The percentage of wrong-voice errors never exceeded 12.5 % in the case of speech-modulated buzzes or noises, suggesting that, even after so many repetitions (1,200 trials), these maskers were never perceived as a phonetic content by any subject. It was

simply chance if listeners responded to both the number and color corresponding to the sentence from which the buzz/noise was constructed. The errors at low TMRs were primarily random for buzzes and noises (right panels). This striking contrast in the types of errors strengthens the idea that performance was limited by audibility, or energetic masking, in the case of speech-modulated buzzes/noises, but was limited by informational masking or difficulties in focusing attention on the target source in the case of two-same-male voices. For the "mixed-voice" error category, we found no obvious contrast between the two masker types. This can be understood, considering that three out of the four possible colors were presented on each trial. So, it ought to occur that listeners often picked the color of one source (target or masker) with, by chance, the number of another.

Trial position within clusters

Correct performance was also examined as a function of the trial position within a cluster for each condition. Scores were computed separately for the first trial (which occurred ten times), the second trial (which occurred ten times), the third trial (which occurred ten times), the fourth trial (which occurred nine times), and a fifth "bin" collapsing across the fifth, sixth, and seventh trial in a cluster (which occurred 11 times together). Although the resolution of performance specific to position within a cluster was poorer than the resolution of performance averaged across trials (9 %–11 % instead of 2 %), it was still possible to fit a logistic function for position-specific performance by constraining fits to have priors for the inflection point and slope shaped with the means and standard deviations obtained with the performance averaged across trials (shown in Figs. 5 and 8). In other words, we considered that each of the position-specific fits had to result in thresholds in the vicinity of the final thresholds to which they contributed. An ANOVA was then performed that included Trial Position as a fourth within-subjects factor. Mauchly's test of sphericity was never significant [$\chi^2(9) < 13.5, p > .148$, in Exp. 3; $\chi^2(9) < 14.9, p > .100$, in Exp. 4], so the assumption of homogeneity of variances was not violated. All results mentioned earlier and reported in Table 1 (third and fourth columns) remained similar, with smaller p values due to the increase in statistical power caused by five-fold replication of very similar thresholds in each experimental condition. More to the point of this analysis, the main effect of trial position was significant in both experiments [$F(4, 36) = 3.3, p = .020$, in Exp. 3; $F(4, 36) = 5.6, p = .001$, in Exp. 4], reflecting that on average, performance improved over successive presentation of the same condition and, as a result, thresholds decreased by 0.4–0.5 dB (with most of the effect arising between the first and second trials). In Experiment 3, trial position interacted with masker type [$F(4, 36) = 3.0, p = .030$]. Indeed, the simple effect of trial position was not significant for buzzes [$F(4, 6) < 0.1, p = .960$], but it was significant for masking voices [$F(4, 6) = 8.7, p = .011$]. Trial position also interacted with room, ΔF0, and masker type [$F(4, 36) = 3.8, p = .011$]. Unfortunately, these interactions were not observed in Experiment 4, casting doubt on their possible interpretation. In principle, the effect of trial position within clusters could have been a sign that a particular condition was easier to perform after successive presentation of the same acoustic cue, tapping into the "building-up" hypothesis of streaming (Bregman, 1990). For instance, one could have hoped to see the effect of trial position arising specifically in the presence of a ΔF0 or ΔSL against masking voices, perhaps with different strengths in anechoic and reverberant conditions. But this was not the case in Experiment 4, and even in Experiment 3 those differences never amounted to more than 2 dB. Instead, the effect of trial position in this study may be better appreciated in terms of consistency effects and was overall negligible, as compared to the differences observed between the experimental conditions.

References

Allen, J. B., & Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America, 65,* 943–950.

Arbogast, T. L., Mason, C. R., & Kidd, G., Jr. (2002). The effect of spatial separation on informational and energetic masking of speech. *Journal of the Acoustical Society of America, 112,* 2086–2098.

Arbogast, T. L., Mason, C. R., & Kidd, G., Jr. (2005). The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America, 117,* 2169–2180.

Best, V., Marrone, N., Mason, C. R., & Kidd, G., Jr. (2012). The influence of non-spatial factors on measures of spatial release from masking. *Journal of the Acoustical Society of America, 13,* 3103–3110.

Beutelmann, R., & Brand, T. (2006). Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America, 120,* 131–342.

Beutelmann, R., Brand, T., & Kollmeier, B. (2010). Revision, extension, and evaluation of a binaural speech intelligibility model. *Journal of the Acoustical Society of America, 127,* 2479–2497.

Boersma, P., & Weenink, D. (2013). Praat: Doing phonetics by computer (Version 5.3.57) [Computer program]. Retrieved October 27, 2013, from www.praat.org/

Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multitalker communications research. *Journal of the Acoustical Society of America, 107,* 1065–1066.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.

Bronkhorst, A., & Plomp, R. (1990). A clinical test for the assessment of binaural speech perception in noise. *Audiology, 29,* 275–285.

Brungart, D., Simpson, B., Ericson, M., & Scott, K. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *Journal of the Acoustical Society of America, 110,* 2527–2538.

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

Colflesh, G. J. H., & Conway, A. R. A. (2007). Individual differences in working memory capacity and divided attention in dichotic listening. *Psychonomic Bulletin & Review, 14,* 699–703. doi:10.3758/BF03196824

Collin, B., & Lavandier, M. (2013). Binaural speech intelligibility in rooms with variations in spatial locations of sources and modulation depth of noise interferers. *Journal of the Acoustical Society of America, 134,* 1146–1159.

Culling, J. F. (1996). Signal processing software for teaching and research for psychoacoustics under UNIX and X Windows. *Behavior Research Methods, Instruments, & Computers, 28,* 376–382.

Culling, J. F., Hodder, K., & Toh, C. (2003). Effects of reverberation on perceptual segregation of competing voices. *Journal of the Acoustical Society of America, 114,* 2871–2876.

Culling, J. F., Summerfield, Q., & Marshall, D. (1994). Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels. *Speech Communication, 14,* 71–95.

Darwin, C. J., & Hukin, R. W. (2000). Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention. *Journal of the Acoustical Society of America, 108,* 335–342.

de Laat, J. A. P. M., & Plomp, R. (1983). The reception threshold of interrupted speech. In R. Kinke & R. Hartman (Eds.), *Hearing: Physiological bases and psychophysics* (pp. 359–363). Berlin, Germany: Springer.

Deroche, M. L. D., & Culling, J. F. (2011). Voice segregation by difference in fundamental frequency: Evidence for harmonic cancellation. *Journal of the Acoustical Society of America, 130,* 2855–2865.

Deroche, M. L. D., Culling, J. F., Chatterjee, M., & Limb, C. J. (2014a). Speech recognition against harmonic and inharmonic complexes: Spectral dips and periodicity. *Journal of the Acoustical Society of America, 135,* 2873–2884.

Deroche, M. L. D., Culling, J. F., Chatterjee, M., & Limb, C. J. (2014b). Roles of target and masker fundamental frequency in voice segregation. *Journal of the Acoustical Society of America, 136,* 1225–1236.

Deroche, M. L. D., Culling, J. F., & Chatterjee, M. (2014). Phase effects in masking by harmonic complexes: Detection of bands of speech-shaped noise. *Journal of the Acoustical Society of America, 136,* 2726–2736.

Durlach, N. I. (1972). Binaural signal detection: Equalization and cancellation theory. In J. Tobias (Ed.), *Foundations of modern auditory theory* (Vol. II, pp. 371–462). New York, NY: Academic Press.

Durlach, N. (2006). Auditory masking: Need for improved conceptual structure. *Journal of the Acoustical Society of America, 120,* 1787–1790.

Durlach, N., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., & Kidd, G., Jr. (2003). Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. *Journal of the Acoustical Society of America, 114,* 368–379.

Festen, J. M., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *Journal of the Acoustical Society of America, 88,* 1725–1736.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2008). Spatial release from masking with noise-vocoded speech. *Journal of the Acoustical Society of America, 124,* 1627–1637.

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *Journal of the Acoustical Society of America, 106,* 3578–3588.

George, E. L. J., Festen, J. M., & Houtgast, T. (2008). The combined effects of reverberation and nonstationary noise on sentence intelligibility. *Journal of the Acoustical Society of America, 124,* 1269–1277.

Hawley, M., Litovsky, R., & Culling, J. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *Journal of the Acoustical Society of America, 115,* 833–843.

Houtgast, T., & Steeneken, H. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *Journal of the Acoustical Society of America, 77,* 1069–1077.

Howard-Jones, P. A., & Rosen, S. (1993). Uncomodulated glimpsing in "checkerboard" noise. *Journal of the Acoustical Society of America, 93,* 2915–2922.

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

Kidd, G., Jr., Mason, C. R., & Gallun, F. J. (2005). Combining energetic and informational masking for speech identification. *Journal of the Acoustical Society of America, 118,* 982–992.

Kjellberg, A., Ljung, R., & Hallman, D. (2008). Recall of words heard in noise. *Applied Cognitive Psychology, 22,* 1088–1098.

Lavandier, M., & Culling, J. F. (2007). Speech segregation in rooms: Effects of reverberation on both target and interferer. *Journal of the Acoustical Society of America, 122,* 1713–1723.

Lavandier, M., & Culling, J. F. (2008). Speech segregation in rooms: Monaural, binaural and interacting effects of reverberation on target and interferer. *Journal of the Acoustical Society of America, 123,* 2237–2248.

Lavandier, M., & Culling, J. F. (2010). Prediction of binaural speech intelligibility against noise in rooms. *Journal of the Acoustical Society of America, 127,* 387–399.

Licklider, J. (1948). The influence of interaural phase relations upon masking of speech by white noise. *Journal of the Acoustical Society of America, 20,* 150–159.

Ljung, R., & Kjellberg, A. (2009). Long reverberation time decreases recall of spoken information. *Building Acoustics, 16,* 301–312.

Peterson, P. M. (1986). Simulating the response of multiple to a single source in a reverberant room. *Journal of the Acoustical Society of America, 80,* 1527–1529.

Plomp, R. (1976). Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise). *Acustica, 34,* 200–211.

Plomp, R., & Mimpen, A. M. (1979). Speech-reception threshold for sentences as a function of age and noise level. *Journal of the Acoustical Society of America, 66,* 1333–1342.

Robinson, D., & Jeffress, L. (1963). Effect of varying the interaural noise correlation on the detectability of tonal signals. *Journal of the Acoustical Society of America, 35,* 1947–1952.

Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., & Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio Electroacoustics, 17,* 225–246.

Sörqvist, P., & Rönnberg, J. (2012). Episodic long-term memory of spoken discourse masked by speech: What is the role for working memory capacity? *Journal of Speech, Language, and Hearing Research, 55,* 210–218.

Steeneken, H. J. M., & Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *Journal of the Acoustical Society of America, 67,* 318–326.

Westermann, A., & Buchholz, J. M. (2015). The effect of spatial separation in distance on the intelligibility of speech in rooms. *Journal of the Acoustical Society of America, 137,* 757–767.