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], and Vincent L. Gracco[1]

[1]Centre for Research on Brain, Language and Music, McGill University. Rabinovitch House, 3640 rue de la Montagne, Montreal, Quebec, H3G 2A8, Canada.
Electronic mail: mickael.deroche@mcgill.ca

[2]School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT, United Kingdom.

[3]Univ Lyon, ENTPE, Laboratoire Génie Civil et Bâtiment, rue M. Audin, 69518 Vaulx-en-Velin Cedex, France.

(Dated: April 11, 2016)

Running title: Reverberation and informational masking

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. Situations were examined in which reverberation should have no effect on the mechanisms underlying the release 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 non-speech maskers (matched in excitation pattern and temporal envelope to speech maskers) which, as intended, were robust to reverberation. Larger masking releases were obtained for speech maskers, but critically, they were affected by reverberation. The results suggest that reverberation either limits the amount of informational masking there is to begin with, or affects its release by ΔF0s or ΔSLs.

I. INTRODUCTION

In multi-talker situations, listeners can use a difference in fundamental frequency (ΔF0) and a difference in spatial location (ΔSL) 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 where a target sound is made inaudible by a more intense sound of similar spectro-temporal characteristics. Informational masking (Brungart et al., 2001; Durlach et al., 2003; Kidd et al., 2005) refers to the case where a target sound is made unattended, but not necessarily inaudible, by the presence of a competing sound while the latter does not share the same frequency band or occurs at different time windows than the target. A lot of attention has been paid to explore the mechanisms underlying the energetic masking releases offered by a ΔF0 and a ΔSL, and they are generally susceptible to reverberation. In contrast, potential effects of reverberation on the informational masking releases associated with a ΔF0and a ΔSL have been relatively unexplored, and this was partly due to difficulties in disentangling energetic from informational accounts. The present study aims to highlight a detrimental effect of reverberation on the use of ΔF0 and ΔSL while restricting its possible cause to an informational aspect.

A. 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 where sources are monotonized, i.e. have a fixed F0 throughout the entire signal duration, reverberation is harmless. Culling et al. (1994) measured the benefit of a 1-semitone ΔF0 in the case of vowels' 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 required to

achieve 50% intelligibility, for a target voice separated by a 2-semitone ΔF0 from stationary speech-shaped harmonic complexes hereafter referred to as buzzes. They 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 introducing reverberation 1) the masker partials do not move, thereby leaving the exact same spectral dips in between resolved partials as there are in anechoic conditions; and 2) the masker periodicity is not disrupted in resolved channels. Both of these aspects of masker harmonicity seem crucial to the amount of ΔF0 benefit (Deroche et al., 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 et al., 2014c). Thus, while reverberation disrupts the release of energetic masking due to ΔF0s between competing sources in every realistic situation, it is still possible to create a laboratory situation where this is not the case.

B. Reverberation can impair the ΔSL benefit

Reverberation is generally detrimental to the use of a ΔSL between concurrent speech sources (Plomp, 1976; Bronkhorst and Plomp, 1990; Culling et al., 1994, 2003; Beutelmann and Brand, 2006). This impairment has two main causes. First, the sound reflections reduce the acoustic shadowing of the head, i.e. make the target-to-masker ratio relatively more homogenous 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 where the listener has no virtual head (Lavandier and Culling, 2010). Second, reverberation disrupts binaural unmasking, mainly by reducing the interaural coherence of the

masking sounds (Licklider, 1948; Robinson and Jeffress, 1963; Lavandier and Culling, 2007, 2008). Following the equalization-cancellation theory (Durlach, 1972), placed under reverberant conditions, a masker becomes less correlated at the two ears, harder to equalize, and therefore more effective at masking. However, in the particular case where 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 on both ears. In support for this idea, Lavandier and Culling (2010) measured 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 SRT was identical for an asymmetrical listener/noise configuration in an anechoic room and a symmetrical listener/noise configuration in a strongly reverberant room. Thus, while reverberation disrupts the release of energetic masking due to ΔSLs between competing sources in every realistic situation, it is still possible to create a laboratory situation where this is not the case.

C. Reverberation and informational masking

Most studies that investigate informational masking use 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 et al., 2000). The task is to choose which of the simultaneous words is part of the target utterance with a given call sign rather than part of the competing utterances. A specific cue, which is generally the object of investigation, may help listeners to fulfill this task, provided that this cue is sufficiently strong to maintain attention on the appropriate utterance. Since there is a limited set of call signs, colors, and numbers, 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 able to do this, speech mixtures could, in principle, be completely audible, yet incomprehensible. Using the CRM design, Darwin and Hukin (2000) have 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, but did not provide any explanation as to why this would be. For the binaural investigations, the configuration of listener/maskers was not symmetrical in the room, and therefore their results could potentially be explained by an increase in energetic masking (see section B). 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.

D. Goal of the present study

The present study created specific laboratory situations where the energetic masking release from ΔF0 or ΔSL should be robust to reverberation. This was done by using monotonized sources, and by using a symmetrical configuration of listener/maskers, respectively. In both cases, a non-linguistic (i.e., energetic) masker was created with similar spectro-temporal properties (long-term excitation pattern and broadband temporal envelope) as the speech maskers. The ΔF0 benefit and the ΔSL benefit were measured against the two masker types, in anechoic and reverberant conditions. It was expected that the amount of informational masking would be minimal with the non-linguistic maskers, and therefore that reverberation would have very little effect on the benefits of a ΔF0 and a ΔSL. The study investigated whether these respective benefits were nonetheless reduced for speech maskers, thus evaluating the influence of reverberation on informational masking releases. Two methods were used: an adaptive SRT

task presenting unpredictable sentences, and the CRM presenting predictable sentences at fixed target-to-masker ratios, as a way to probe whether the hypothesized effect of reverberation on informational masking would be revealed more or less easily with one method over the other.

II. EXPERIMENT 1 – SRT with ΔF0s

A. Listeners

Thirty-two listeners (18 females, 14 males, between 18-30 years old) participated in this experiment. They all provided informed consent in accordance with the protocols established by the Institutional Review Board at Cardiff University, and were compensated at an hourly base rate. All listeners reported normal hearing 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 an hour.

B. Stimuli and conditions

Two types of masker were used: speech-modulated buzz and two concurrent masking voices. The speech stimuli came from the Harvard Sentence List (Rothauser et al., 1969), spoken by the same male voice with a mean F0 of 104 Hz. The Praat PSOLA package (Boersma and Weenink, 2013) was used to resynthesize each sentence with a fixed F0 throughout. Eighty target sentences were monotonized at 110 or 174.6 Hz (8 semitones higher). Eight masking sentences (all different from the target sentences) were monotonized at 110 Hz, and then added in pairs to create four 2-voice speech maskers. The buzz maskers were created from a broadband sine-phase harmonic complex with a 110-Hz F0, filtered with a linear-phase FIR filter designed to match the average long-term excitation pattern of the monotonized sentences used as the speech maskers. In addition, the temporal envelopes of the four 2-voice speech maskers were extracted by half-

wave rectification and low-pass filtering (first-order Butterworth with a 3-dB cutoff at 40 Hz) and applied to the complex with a speech-like spectral profile. This manipulation resulted in four speech-modulated buzz maskers. Target and maskers were both heard in anechoic or in reverberant conditions.

[FIG. 1 ABOUT HERE]

Reverberation was added using the ray-tracing method (Allen and Berkley, 1979; Peterson, 1986) as implemented in the |WAVE signal processing package (Culling, 1996). The virtual room 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 and placed along an axis rotated at 25^{o} 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. Reverberation adds irregular perturbations to the stimulus spectrum, known as room coloration. These perturbations were removed using a further FIR filter as part of a package of energetic equalization measures (similar to that used in Deroche and Culling, 2011). 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. Absorption coefficients were all 0.3 for the surfaces of the reverberant room. For the anechoic room, the coefficients were all set to 1. Virtual sources were simulated 2 m straight ahead from the receivers (left panel of Figure 1). Binaural stimuli were produced by generating the impulse responses for the two receivers in virtual space and convolving the sentences or buzzes with these two impulse responses.

The top left panels of Figure 2 show that the two masker types had almost identical excitation patterns in both anechoic and reverberant conditions. They were also very similar across rooms due to the decoloration process. Moreover, the top right panels of Figure 2 show that in the temporal domain, the two masker types had similar waveforms, offering only a few temporal dips, which 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 resulted from 2 masker types × 2 ΔF0s × 2 rooms. All maskers and target stimuli were equalized to the same mean RMS power across the ears. During the adaptive track, changes in target-to-masker ratio (TMR) occurred by adjusting the target level while presenting maskers always at 69 dB SPL.

[FIG. 2 ABOUT HERE]

C. Procedure and equipment

The experimental session began with three practice runs using unprocessed speech, not used in the rest of the experiment, masked by speech-modulated buzz (one run) or 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. While 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 thirty-two listeners resulted in four complete rotations of the conditions.

SRT was measured using a 1-up/1-down adaptive threshold method, 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 maskers. Listeners were instructed to disregard the masking sentences corresponding to the displayed transcripts and to listen to the other (target) sentence. 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 transcript when they could first hear half the target words. 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 3, 4, or 5 correct keywords, and increased by 2 dB if the listener had found 2, 1, or 0 correct keywords. Measurement of each SRT was taken as the mean TMR over the last eight trials. Note that a SRT of 0 dB occurred when the target level was 3 dB higher than each of the two masking sentences.

This experiment was performed in the School of Psychology at Cardiff University. A computer monitor was visible outside the booth window for trial-by-trial feedback and a keyboard was inside for transcript responses. 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.

D. Results

A repeated-measures analysis of variance (ANOVA) was conducted in order to determine the influence of each of the three factors (room × masker type × ΔF0) on the SRTs shown in the left panel of Figure 3. The results are reported in Table I. The three main effects were significant: mean SRTs were lower when the sources were heard in anechoic rather than reverberant

conditions, lower with speech-modulated buzzes than with 2-same-male voices, and lower when sources had different F0s than when they had the same F0. As illustrated in the right panel, the interaction between ΔF0 and masker type was significant, i.e. the masking release provided by the ΔF0 was larger with 2-same-male voices than with buzz maskers, but this was particularly the case in the anechoic rather than reverberant room (3-way interaction).

[FIG. 3 ABOUT HERE]

E. Discussion

1. Reverberation only affected the masking release obtained with speech maskers

On one hand, for speech-modulated buzz maskers, the ΔF0 benefit was about 5 dB in both anechoic and reverberant conditions. As it was intended by keeping all sources monotonized, the release from masking (presumably largely energetic for this masker type) was robust to reverberation. For speech maskers on the other hand, SRTs were substantially elevated, despite presenting similar amount of energetic masking (Fig. 2). Since three utterances were presented simultaneously, and especially since they were spoken by the same male talker, there was uncertainty as to which sentence listeners should have attended to. Without ΔF0, SRT was 5 dB higher with the 2 voices than with buzzes in anechoic conditions. A major part of this elevation is presumably due to additional informational masking. Listeners could use an 8-semitones ΔF0 to release from energetic as well as informational masking, and this is why the ΔF0 benefit was greater with 2 voices than with buzzes. The focus of the present study, however, was to examine a potential effect of reverberation on this latter benefit. The right panel of Figure 3 illustrates that the ΔF0 benefit obtained with 2 masking voices was reduced in reverberant compared to anechoic conditions. This 3-way interaction would therefore suggest that

reverberation affects the informational component of the ΔF0 benefit, consistent with the results obtained by Darwin and Hukin (2000).

2. 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 (Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985). Being independent of any other masking effects involved, this loss of modulation transmission should have occurred similarly whether there was a ΔF0 or not, and whatever the masker type. Note that the magnitude of this effect was quantified at 2 dB, by Deroche and Culling (2011, Figure 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 and Plomp, 1983; Festen and Plomp, 1990; Hawley et al., 2004). Although the present maskers consisted of 2 simultaneous utterances, there remained a few dips that listeners could potentially have exploited. 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 and Lavandier, 2013). Reverberation, however, "fills-in" to some extent the temporal dips in the masker waveforms, which prevents to some extent their exploitation (Bronkhorst and Plomp, 1990; George et al., 2008; Beutelmann et al., 2010; Collin and Lavandier, 2013). 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 anti-phasic in adjacent frequency channels (Howard-Jones and 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 2-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 2-voice maskers. Regardless of its magnitude, the "filling-in" effect of reverberation should have occurred similarly whether the ΔF0 was 0 or 8 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.

III. EXPERIMENT 2 – SRT with ΔSLs

If the effect of reverberation observed in experiment 1 was indeed associated with informational masking, then it might have nothing to do with the harmonicity of voices or F0 processing at all. If so, one might still observe it when binaural cues provide masking release.

A. Listeners

Thirty-two listeners (23 females, 9 males, between 18-43 years old) participated in this second experiment. They provided written informed consent in accordance with the protocols established by the McGill University Institutional Review Board, and were compensated at an hourly base rate. They all had audiometric thresholds less than 20 dB HL at octave frequencies between 250 Hz and 8 kHz, and reported 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 an hour.

B. Stimuli and conditions

Two types of masker were used: speech-modulated noise and two concurrent voices. The speech stimuli came again from the Harvard Sentence List spoken by the same male voice. Sentences were, this time, naturally intonated (not processed through Praat). Eight masking sentences were added in pairs to create four 2-voice speech maskers. Four noise maskers were created from a Gaussian white noise, filtered with a linear-phase FIR filter designed to match the long-term excitation pattern of the 2-voice maskers, and modulated by their broadband temporal envelopes. Target and maskers were both heard in anechoic or in reverberant conditions.

Reverberation was added using the processing package described earlier. The virtual room and source/listener configuration is shown in the right panel of Figure 1. The rooms (anechoic/reverberant) and height of sources and receivers (ears) were identical but 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 (still separated by 18 cm with no head between them) were placed on either side of the axis parallel to the 3.2-m wall halving the room symmetrically. The maskers were located 2-m away from the listener on that same axis. The target, however, was either collocated with the maskers, or placed at equal distance (2 m) but on an axis rotated at 60^{o} from the listener-masker axis. As in experiment 1, the room coloration was removed. Figure 2 shows that the two masker types had very similar excitation patterns (bottom left panel) and very similar waveforms (bottom right panel), such that they should have produced very similar amounts of energetic masking. The eight experimental conditions resulted from 2 masker types × 2 ΔSLs × 2 rooms. All maskers and target stimuli were equalized to the same mean RMS power across the ears. During the adaptive track, maskers were always presented at 69 dB SPL and the relative target level was adjusted.

C. Procedure and equipment

The procedure was the SRT task described earlier. Thirty-two listeners resulted in four complete rotations of the conditions. The experiment was performed at the School of Communication Sciences and Disorders at McGill University. A user-interface was displayed on a monitor, inside a sound-attenuating audiometric booth. Signals were sampled at 44.1 kHz and 16-bit resolution, digitally mixed, D/A converted by a Focusrite Scarlett 2i4 sound card and presented binaurally over Sennheiser HD 280 headphones.

D. 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 Figure 4. The results are reported in Table I. The three main effects were significant: mean SRTs were lower when the sources were heard in anechoic rather than reverberant conditions, lower with speech-modulated noise than with 2 interfering voices, and lower when sources had different SLs than when they were collocated. As illustrated in the right panel, the interaction between ΔSL and masker type was significant, i.e. the spatial release from masking was larger with 2 interfering voices than with noise maskers, but this was particularly the case in the anechoic rather than reverberant room (3-way interaction).

[FIG. 4 ABOUT HERE]

E. Discussion

1. Reverberation only affected the masking release obtained with speech maskers

On one hand, for speech-modulated noise, 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 masker and listener both positioned with the room symmetrical about them, so that the masker interaural coherence was intact in reverberation, and 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 detrimental effect of reverberation on binaural unmasking as being mediated by the masker coherence (Lavandier and Culling, 2007, 2008). For speech maskers on the other hand, there was some uncertainty as to which sentence one should attend to: without ΔSL, 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 Figure 4 illustrates that the ΔSL benefit obtained with speech maskers was reduced in reverberant compared to anechoic conditions. Therefore, just as it did in the harmonic domain in experiment 1, this 3-way interaction would suggest that reverberation affects the informational component of the ΔSL benefit.

2. 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 could perhaps 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 and 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.

INTERIM CONCLUSION

The first two experiments showed that, for both ΔF0s and ΔSLs, the benefits associated with these cues in the presence of masking voices were smaller in reverberant than in anechoic conditions. Although this is generally the case in every realistic situation, it is in the present study very puzzling because great care had been taken to avoid any effect of reverberation on the energetic components of these benefits. This was confirmed by observing that the respective benefits obtained with non-linguistic analogs of the speech maskers were unaffected by reverberation. At first sight, therefore, it could be concluded that reverberation is detrimental to the informational masking releases provided by ΔF0s and ΔSLs.

Using the CRM design, Brungart et al. (2001) made an extensive investigation of the roles of sex and identity of competing voices in a 2, 3, and 4-talker mixtures, as a function of target-to-masker ratio. They showed that psychometric functions could in some cases display unexpected shapes. For instance, with a 2-talker mixture, i.e. 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 designed to present stimuli around a given point, e.g. 50%. A plateau in the vicinity of that point could make the measured threshold very unreliable. With a 3- or 4-talker mixture, this plateau disappeared and the psychometric functions displayed a more typical S shape. This is reassuring for the present study which used a three-talker mixture. However, the speech-modulated buzzes (despite being modulated by two utterances) may perceptually form a single harmonic masker and it is unclear whether the psychometric function would have a standard shape for this masker type. To our knowledge, no study has ever used the CRM with buzz maskers.

Brungart et al. (2001) also highlighted a clear distinction between the cases where the target talker was more intense than any masking voice and cases where 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 there was more than one masking voice. Crudely, the rationale is that performance has more to do with selective attention at positive TMRs and more to do with peripheral mechanisms at negative TMRs. This may represent a concern for our first two experiments because the conditions with reverberant interfering voices without ΔF0 or ΔSL, which are at the heart of the observed 3-way interaction, were the only conditions that displayed a positive SRT, between +1 and +2 dB. This raises the possibility that a ceiling effect could have prevented the SRT from increasing higher. The target voice might have been sufficiently loud to become too easily recognizable. This is a plausible

scenario considering that at a SRT of -3 dB all three voices had equal intensity increasing their confusability and consequently the amount of informational masking. As TMR increased beyond -3 dB and reached positive values, the target voice progressively stood out from the two other sentences, potentially grasping the listener's attention. 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 by the salience of the target voice.

These concerns were investigated by measuring performance in the conditions tested in experiments 1 and 2, but at fixed TMRs and using the CRM design which is particularly suited to phenomena related to informational masking. A range of six TMRs was chosen from pilot data to cover the full psychometric function, different in each condition. By having access to psychometric functions, we could 1) verify whether a plateau existed in any condition, and 2) test whether a ceiling effect had somehow distorted the adaptive SRT procedure at positive TMRs, for the specific conditions of the reverberant interfering voices without ΔF0 or ΔSL.

IV. EXPERIMENT 3 – CRM with ΔF0s

A. Listeners

Ten listeners (8 females, 2 males, between 19-26 years old) participated. They were recruited and screened in a similar manner to those of Exp. 2. Each listener attended three experimental sessions that lasted about 50, 50, and 65 minutes.

B. Stimuli

The stimuli came from Bolia et al. (2000). The sentences were of the form of "*Ready <call sign> go to <color> <number> now"*. For a given voice, there were 256 combinations of

eight call signs (‘Charlie’, ‘Ringo’, ‘Laker’, ‘Hopper’, ‘Arrow’, ‘Tiger’, ‘Eagle’, ‘Baron’), four colors (‘blue’, ‘red’, ‘white’, ‘green’), and eight numbers (1 to 8). There were four different male voices, resulting in a total of 1024 sentences in the original material. The preparation of the materials was identical to experiment 1. All sentences were monotonized at 110 or 174.6 Hz. For each sentence, equivalent speech-shaped, speech-modulated buzzes were created. All stimuli (sentences and buzzes) were then convolved with the anechoic and reverberant impulse responses and filtered for room decoloration.

C. Procedure and equipment

Listeners were asked to follow the target voice, always following the 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, or two sentences. The call signs, colors and numbers of the two maskers were randomly chosen but were different from each other and from those of the target. The compound of the two maskers was then equalized to 69 dB SPL. As in experiment 1, a TMR of 0 dB occurred when the target level was 3 dB higher than the level of each of the two maskers.

Each of the eight conditions of experiment 1 was measured at six different TMRs, resulting in forty-eight conditions. For each of these forty-eight conditions, performance was measured over 50 trials. Thus, each subject had to complete a total of 2400 trials, which were divided into ten experimental blocks (of approximately 240 trials each, taking about 15 minutes 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 to examine performance as a function of the trial position within a cluster. The rationale is that listeners may take a few trials, every time a new condition is presented, to realize what characteristics of the target voice could be most efficient to track. One might therefore expect to find performance improving with trial position in those particular conditions when streaming plays 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 conditions' order and clusters' size. Furthermore, the identity of the male talker was kept constant for all sources in one block, but changed randomly from one block to the next, as well as across subjects, among the four male voices available in the original material (Bolia et al., 2000), simply ensuring that the results were not tightly dependent upon specific characteristics of a given voice. Prior to the start of the experiment, subjects were familiarized with the stimuli and experimental paradigm, by completing between 20-40 trials on any of the experimental conditions at random, but making sure that some trials presented the two speech-modulated buzzes and some trials presented the two interfering voices. Within each session, breaks were offered in between blocks. This experiment was performed using the same equipment as in experiment 2.

D. Results

For each subject, correct responses were calculated over 50 trials for each of the 48 conditions. The symbols displayed on Figure 5 are the performance averaged over the 10 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 get reliable estimates of 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 surfaces on Figure 5 are the mean fits in each condition. From individual fits, a TMR corresponding to 50% performance was extracted and served as basis for the statistical analysis. The corresponding mean thresholds are shown on the left panel of Figure 6.

The results of the ANOVA are reported in Table I. Main effects were all significant, reflecting that 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 and most importantly the 3-way interaction. As illustrated on 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 compared to reverberant conditions.

[FIG. 5, 6, and 7 ABOUT HERE]

For each subject, the slope of the logistic fits at 50% performance was also extracted, and submitted to a similar ANOVA (Table I). There was a main effect of masker type, a main effect of ΔF0, and both strongly interacted. As shown in the left panel of Figure 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 type of errors made for each experimental condition, 2) a potential effect of trial position within clusters. These results were

somewhat beside the present focus (i.e., the 3-way interaction), and therefore postponed to the Appendix.

E. Discussion

The results of experiment 3 were qualitatively similar to those of experiment 1. Perhaps, the most obvious difference is the scale of thresholds obtained with buzzes, ranging between -9 and -16 dB in Fig. 6 (compared to -2 and -10 dB in Fig. 3), while the scale of thresholds obtained with masking voices was relatively constant. This is 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 demands in terms of intelligibility because the same utterances are presented over and over again. In the absence of any confusion between sources, i.e., buzz maskers, glimpsing very little information such as a phoneme <e> followed by a phoneme <ʊ> could be sufficient in reconstructing "red two" and potentially getting a correct response. This is why thresholds for buzzes could be much lower in the CRM than in the SRT task. Of course, the more masking there is to start with, the more masking release there can be, and 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 thresholds further, while it did in experiment 1. This, again, is 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 2 competing voices was reduced in reverberation, while from an energetic-masking perspective, there is no reason this should have happened.

By having access to the full psychometric function of each experimental condition, we could verify that they all displayed monotonic S shapes. There was no plateau which could have prevented the adaptive procedure from working properly in experiment 1, therefore this potential confound can be discarded. We also take a closer look at the idea that a ceiling effect could have been responsible for the 3-way interaction observed in experiment 1. The two cases of interfering voices monotonized at the same F0 as the target voice (with and without reverberation) displayed steeper slopes than any other condition (most-right curves in Fig. 5 and left panel of Fig. 7). Why is this so? These two conditions were extremely challenging to the listeners, and this is not surprising given that all three utterances were spoken by the same person, at the same position and same F0, and were all very similar sentences. There were indeed very few cues left for listeners to do the task, so they probably had to rely on level differences almost exclusively. As the TMR went beyond -3 dB, the target voice became progressively louder than any of the two masking voices. Not only could listeners start performing the task but they rapidly achieved high performance. So, the psychometric function is indeed very steep when listeners listen to the loudest voice in a crowd of clones. The hypothesis of a ceiling effect, however, makes a particular statement, that there would have been a specific TMR (or at least a localized range of TMRs) for which any detrimental effect of reverberation would have been counteracted by the clear salience of a loud target voice. In other words, one could have imagined a psychometric function for the reverberant case that would have started with a right-ward horizontal shift relative to the anechoic case but converged with the anechoic function above a given TMR. That is not what the data showed. The slopes did not differ between the anechoic and reverberant case (for interfering voices without ΔF0). Perhaps a more convincing argument is to look at performance at fixed TMR. For instance, at a TMR of -1 dB (Fig. 5), the target voice was thus a

little louder than each masking voice; this loudness cue was identical whether a ΔF0 was present or not and whether the room was anechoic or reverberant, and yet the effect of interest was present: the 8-semitones ΔF0 provided a 60% increase in performance in anechoic conditions but only 45% increase in performance in reverberant conditions. Thus, the idea that the 3-way interaction is caused by a ceiling effect must also be discarded.

As aforementioned, 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). 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 acts similarly to increasing the number of interfering utterances. Moreover, this effect would be constant across TMRs, because the reverberation characteristics were fixed and that the combined masker level was also fixed at 69 dB SPL. It would thus apply to the whole psychometric function, not just on a localized range of TMRs, which is closer to the effect observed in the present data. This leads to a different interpretation of this 3-way interaction: the reason why the ΔF0 benefit obtained with competing voices was smaller in reverberation may not necessarily be because reverberation breaks apart F0-streaming, it could also be because

there is less informational masking to begin with in the presence of reverberant interfering voices than in the presence of anechoic ones. Note that these two interpretations are not mutually exclusive.

V. EXPERIMENT 4 – CRM with ΔSLs

A. Listeners

Ten listeners (8 females, 2 males, all different from Exp. 3, between 18-34 years old) took part in this last experiment. They were recruited and screened in a similar manner to those of Exp. 2. Each listener attended three experimental sessions that lasted about 50, 50, and 65 minutes.

B. Stimuli, procedure and equipment

The 1024 sentences (256 × 4 male voices) were the same as in experiment 3, but were not monotonized. The 1024 equivalent speech-shaped, speech-modulated noise maskers were created from the speech materials in a similar manner to Exp. 2 and the CRM procedure was identical to experiment 3.

C. Results

The data were analyzed in the same way as those of experiment 3. Figure 8 shows mean performance (symbols) and fits (lines) averaged over the 10 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 I. As shown in

the left panel of Figure 9, 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 3-way interaction was significant: as illustrated on the right panel, the ΔSL benefit was larger with masking voices than with buzzes, in anechoic conditions but not in reverberant conditions.

[FIG. 8 and 9 ABOUT HERE]

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

D. Discussion

The results of experiment 4 (Fig. 9) were qualitatively similar to those of experiment 2. The main difference was the lower scale of thresholds obtained for noise maskers. The ΔSL benefit obtained for noise maskers tended to increase when introducing reverberation (also this trend did not reach significance here, $F(1,9)=4.3$, $p=0.067$). 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 (most-right in Figure 8 and right panel in Figure 7) were steeper than the other six functions, suggesting like in experiment 2 that the task was heavily dependent upon level differences or the relative loudness of the three competing voices in the mixture. But critically, the slope of these two functions did not differ.

These loudness cues would have played the same role in anechoic and reverberant conditions. Rather than a ceiling effect which would have differentially changed their slope within a narrow range of TMRs, the observed data rather point towards a shift of the whole psychometric function spanning a 12-dB range or so. It seems plausible that, by blurring the masking voices, reverberation makes them more noise-like and reduces the potency of each one to attract the listener's attention. In other words, there may be less informational masking to begin with in reverberant compared to anechoic conditions, and so, regardless of the cue being utilized (ΔF0 or ΔSL), there is less informational masking to be released from in reverberation.

VI. SUMMARY

This study presented four experiments intentionally designed to have a very similar format, using two different methods (SRT or CRM) and two different perceptual segregation cues (harmonicity and spatial separation). In each experiment, similar effects were found. Thresholds were considerably elevated in the presence of interfering voices compared to non-linguistic analogs, presumably because speech maskers involved informational masking whereas non-linguistic maskers did not (or very little). This distinction was supported by the analysis of error types in experiments 3 and 4 (Appendix). The cue under investigation, a ΔF0 or a ΔSL, provided a masking release for non-linguistic 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 in this latter case. 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 where reverberation did not impair the benefits obtained with non-linguistic maskers. Yet, the benefits obtained with speech maskers were reduced by reverberation. Somehow, reverberation does not allow as much informational masking release to take place as in anechoic conditions.

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-subject factors used in each individual experiment, and two between-subject factors (task and domain). The 3-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 material, and whether masking releases were provided by ΔF0s or ΔSLs.

Two interpretations seem plausible, and not mutually exclusive, to account for the fact that reverberation reduced the masking releases obtained in a three-talker mixture. On one hand, the reflections in a reverberant room may, to some extent, duplicate the interfering sentences and blur them. By artificially increasing the number of masking voices and making them less intelligible, the combined masker could be getting closer to the percept of a multi-talker babble where each masking source would be less likely to interfere with the listener's ability to track the target voice. Reverberation would therefore limit the amount of informational masking to begin with, and consequently reduce any release from this masking.

On the other hand, it may be that the reverberation adversely impacts the informational masking releases provided by a ΔF0 or a ΔSL. Although it is somewhat speculative, one could imagine that reverberation produces heavy cognitive demands, leaving fewer resources for streaming mechanisms. 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 et al. (2008) presented orally 50 one-syllable words to listeners either in quiet or in a background noise. Words were separated by 3 or 4 seconds, 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,

while again words were correctly identified. So there may be such 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 be harder to attend to certain characteristics of a reverberant voice in the context of competitors.

ACKNOWLEDGMENTS

This research was partly supported by a UK EPSRC grant awarded to J. F. Culling and partly supported by a NSERC grant awarded to V. L. Gracco. We are grateful to the eighty-four participants for their time and effort.

APPENDIX

A.1 Error types

In order to better appreciate why performance in the CRM decreased with TMR in the different conditions, errors were categorized into three types. Errors were labelled 'wrong-voice' when listeners selected both coordinates from the maskers. Errors were labelled 'mixed-voice' when listeners selected one of the coordinates (color or number) from the target, and one from one of the maskers. Errors were labelled 'other' when at least one of the coordinates was not present in the trial. Figure A1 shows percentage of these 3 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, i.e. for 2-same-male voices and particularly in the absence of ΔF0 or ΔSL (left panel). One must bear in mind that the probability of making a wrong-voice error, simply by chance, is the probability of picking a masker color (2/4) by the probability of picking a masker number (2/8), i.e. 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 (1200 trials) these maskers were never perceived as a phonetic content to any subject. It was simply chance if listeners responded both number and color corresponding to the sentence from which the buzz/noise was constructed. Errors at low TMRs were primarily random for buzzes and noises (right panel). This striking contrast in the type of errors strengthens the idea that performance was limited by audibility, or energetic masking, in the case of speech-modulated buzzes and noises, but limited by informational masking or difficulties in focusing attention on the target source in the case of 2-same-male voices. For the 'mixed-voice' error category, there was no obvious contrast between the two masker types. This can be understood considering that three out of 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.

[FIG. A1 ABOUT HERE]

A.2 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 10 times), the second trial (which occurred 10 times), the third trial (which occurred 10 times), the fourth trial (which occurred 9 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 mean and standard deviation obtained with the performance averaged across trials (shown in Fig. 5 and 8). In other words, it was 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 analysis of variance was then performed including trial position as a fourth within-subject factor. Mauchly's test of sphericity was never significant ($\chi2(9)<13.5$, $p>0.148$ in experiment 3; $\chi2(9)<14.9$, $p>0.100$ in experiment 4), so the assumption of homogeneity of variance was not violated. All results mentioned earlier and reported in Table I (third and fourth columns) remained similar, with smaller p values due to the increase in statistical power caused by 5-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 [$F(4,36)=3.3$, $p=0.020$ in experiment 3; $F(4,36)=5.6$, $p=0.001$ in experiment 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 the second trial). In experiment 3, trial position interacted with masker type [$F(4,36)=3.0$, $p=0.030$]. Indeed, the simple effect of trial position was not significant for buzzes [$F(4,6)<0.1$, $p=0.960$] but significant for masking voices [$F(4,6)=8.7$, $p=0.011$]. Trial position also interacted with room, $\Delta F0$, and masker type [$F(4,36)=3.8$, $p=0.011$].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 being 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 $\Delta F0$ or ΔSL against masking voices, perhaps with different strength in anechoic or 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 compared to the differences observed between experimental conditions.

References

Allen, J. B. and Berkley, D. A. **(1979)**. "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am. **65**, 943-950.

Beutelmann, R., and Brand, T. (**2006**). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **120**, 131–342.

Beutelmann, R., Brand, T., and Kollmeier, B. **(2010)**. "Revision, extension, and evaluation of a binaural speech intelligibility model," J. Acoust. Soc. Am. **127**, 2479-2497.

Boersma, P., and Weenink, D. **(2013)**. Praat: doing phonetics by computer [Computer program]. Version 5.3.57, retrieved 27 October 2013 from http://www.praat.org/

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

Bregman, A. S. **(1990)**. "Auditory Scene Analysis", M.I.T. Press, Cambridge, MA, pp. 773.

Bronkhorst, A., and 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. and Scott, K. **(2001)**. "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527-2538.

Collin, B., and Lavandier, M. **(2013)**. "Binaural speech intelligibility in rooms with variations in spatial locations of sources and modulation depth of noise interferers," J. Acoust. Soc. Am. **134**, 1146-1159.

Culling, J. F. **(1996)**. "Signal processing software for teaching and research for psychoacoustics under UNIX and X windows," Behav. Res. Methods Instrum. Comput. **28**, 376-382.

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

Culling, J. F., Hodder, K., and Toh, C. **(2003)**. "Effects of reverberation on perceptual segregation of competing voices," J. Acoust. Soc. Am. **114**, 2871-2876.

Darwin, C. J. and Hukin, R. W. **(2000)**. "Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention," J. Acoust. Soc. Am. **108**, 335-342.

Deroche, M. L. D., and Culling, J. F. **(2011)**. "Voice segregation by difference in fundamental frequency: Evidence for harmonic cancellation," J. Acoust. Soc. Am., **130**, 2855-2865.

Deroche, M. L. D., Culling, J. F., Chatterjee, M., and Limb, C. J. **(2014a)**. "Speech recognition against harmonic and inharmonic complexes: spectral dips and periodicity", J. Acoust. Soc. Am. **135**, 2873-2884.

Deroche, M. L. D., Culling, J. F., Chatterjee, M., and Limb, C. J. **(2014b)**. "Roles of target and masker fundamental frequency in voice segregation," J. Acoust. Soc. Am. **136**, 1225-1236.

Deroche, M. L. D., Culling, J. F., and Chatterjee, M. **(2014c)**. "Phase effects in masking by harmonic complexes: Detection of bands of speech-shaped noise," J. Acoust. Soc. Am. **136**, 2726-2736.

Durlach, N. I. **(1972)**. "Binaural signal detection: Equalization and cancellation theory," *Foundations of Modern Auditory Theory*, edited by J. Tobias (Academic, New York), Vol. II, pp. 371-462.

Durlach, N. **(2006)**. "Auditory masking: need for improved conceptual structure," J. Acoust. Soc. Am., **120**, 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**, 368-379.

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

George, E. L. J., Festen, J. M., and Houtgast, T. **(2008)**. "The combined effects of reverberation and nonstationary noise on sentence intelligibility," J. Acoust. Soc. Am. **124**, 1269-1277.

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

Houtgast, T. and Steeneken, H. **(1985)**. "A review of the MTF concept in room acoustics and its use for estimatingspeech intelligibility in auditoria," J. Acout. Soc. Am. **77**, 1069-1077.

Howard-Jones, P. A., and Rosen, S. **(1993)**. Uncomodulated glimpsing in 'checkerboard' noise. J. Acoust. Soc. Am. **93**, 2915–2922.

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

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

de Laat, J.A.P.M., and Plomp, R. **(1983)**. "The reception threshold of interrupted speech," in *Hearing: Physiological Bases and Psychophysics*, edited by Kinke, R. and Hartman, R. (Springer, Berlin) pp. 359-363.

Lavandier, M. and Culling, J. F. **(2007)**. "Speech segregation in rooms: effects of reverberation on both target and interferer," J. Acoust. Soc. Am. **122**, 1713-1723.

Lavandier, M. and Culling, J. F. **(2008)**. "Speech segregation in rooms: Monaural, binaural and interacting effects of reverberation on target and interferer," J. Acoust. Soc. Am. **123,** 2237-2248**.**

Lavandier, M., and Culling, J. F. **(2010)**. "Prediction of binaural speech intelligibility against noise in rooms," J. Acoust. Soc. Am. **127**, 387–399.

Licklider, J. (**1948)**. "The influence of interaural phase relations upon masking of speech by white noise," J. Acoust. Soc. Am. **20**, 150–159.

Ljung, R., and 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," J. Acoust. Soc. Am. **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.

Robinson, D., and Jeffress, L. (**1963**). "Effect of varying the interaural noise correlation on the detectability of tonal signals," J. Acoust. Soc. Am. **35**, 1947–1952.

Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. **(1969)**. "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225-246.

Steeneken, H. J. M. and Houtgast, T. **(1980)**. "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, 318-326.

List of Tables

TABLE I. Statistics for the thresholds and slopes extracted at 50% intelligibility, in each experiment, following an ANOVA with three within-subjects factors: room (anechoic versus reverberant), masker type (speech-modulated buzz/noise versus 2-same-male voices), and cue (a ΔF0 in experiment 1 and 3, or a ΔSL in experiment 2 and 4).

	Thresholds				**Slopes**	
	Exp.1	**Exp.2**	**Exp.3**	**Exp.4**	**Exp.3**	**Exp.4**
room	F(1,31)=139.3 p<0.001	F(1,31)=89.8 p<0.001	F(1,9)=96.1 p<0.001	F(1,9)=85.0 p<0.001	F(1,9)=0.3 p=0.587	F(1,9)=19.6 p=0.002
masker type	F(1,31)=227.3 p<0.001	F(1,31)=45.5 p<0.001	F(1,9)=2262.1 p<0.001	F(1,9)=498.9 p<0.001	F(1,9)=87.4 p<0.001	F(1,9)=17.8 p=0.002
cue	F(1,31)=538.0 p<0.001	F(1,31)=401.2 p<0.001	F(1,9)=204.5 p<0.001	F(1,9)=192.4 p<0.001	F(1,9)=41.1 p<0.001	F(1,9)=38.5 p<0.001
room × masker type	F(1,31)=1.7 p=0.204	F(1,31)=0.5 p=0.495	F(1,9)=65.7 p<0.001	F(1,9)=21.1 p=0.001	F(1,9)=3.6 p=0.088	F(1,9)=0.2 p=0.697
room × cue	F(1,31)=2.9 p=0.098	F(1,31)=3.6 p=0.066	F(1,9)=10.7 p=0.010	F(1,9)=8.7 p=0.016	F(1,9)=3.1 p=0.111	F(1,9)=1.6 p=0.243
masker type × cue	F(1,31)=7.4 p=0.011	F(1,31)=76.8 p<0.001	F(1,9)=66.1 p<0.001	F(1,9)=1.9 p=0.198	F(1,9)=25.9 p=0.001	F(1,9)=11.0 p=0.009

3-way	F(1,31)=4.4 p=0.045	F(1,31)=5.0 p=0.033	F(1,9)=50.0 p<0.001	F(1,9)=13.0 p=0.006	F(1,9)=1.1 p=0.330	F(1,9)=0.8 p=0.388

List of Figures

FIG. 1 Spatial configurations and virtual room considered in experiments 1-3 (left panel) and experiments 2-4 (right panel).

FIG. 2 Averaged excitation patterns (left panels) and an example of broadband waveform (right panels) for the two maskers used in experiment 1 (top panels) and 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 experiments 2 were essentially the same as in the bottom panels due to the room decoloration.

FIG. 3 (left panel) Mean speech reception thresholds measured in experiment 1, in anechoic and reverberant conditions, for two types of masker (speech-modulated buzz and 2 monotonized voices), with and without a ΔF0 with the target. Lower thresholds indicate greater intelligibility. (right panel) Mean ΔF0 benefits for each masker type and each room. Error bars are ±1 standard error of the mean across subjects.

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

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 TMRs chosen to span most of the function for each condition. Error bars are ±1 standard error of the mean across subjects.

FIG. 6 (left panel) Mean CRM thresholds obtained in experiment 3 extracted from the logistic fits of each subject, at 50% performance. Lower thresholds indicate greater performance. (right panel) Mean ΔF0 benefits for each masker type and each room. Error bars are ±1 standard error of the mean across subjects.

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

FIG. 8 Same as Fig. 5 but in the binaural domain (experiment 4).

FIG. 9 Same as Fig. 6 but in the binaural domain (experiment 4).

FIG. A1 Analysis of the type of errors made in the CRM task of experiment 3 (top panels) and 4 (bottom panels), as a function of TMR. Errors were categorized into three types: “wrong-voice”, “mixed-voice”, and “other”.