Similar abilities of musicians and non-musicians to segregate voices by fundamental frequency

Mickael L. D. Deroche,[1,a)] Charles J. Limb,[2] Monita Chatterjee,[3] and Vincent L. Gracco[4]
[1]Centre for Research on Brain, Language and Music, McGill University, 3640 rue de la Montagne, Montreal H3G 2A8, Canada
[2]Department of Otolaryngology-Head and Neck Surgery, University of California San Francisco School of Medicine, 2233 Post Street, San Francisco, California 94115, USA
[3]Auditory Prostheses and Perception Laboratory, Boys Town National Research Hospital, 555 North 30th Street, Omaha, Nebraska 68131, USA
[4]Haskins Laboratories, 300 George Street, New Haven, Connecticut 06511, USA

(Received 24 February 2017; revised 8 September 2017; accepted 12 September 2017; published online 3 October 2017)

Musicians can sometimes achieve better speech recognition in noisy backgrounds than non-musicians, a phenomenon referred to as the "musician advantage effect." In addition, musicians are known to possess a finer sense of pitch than non-musicians. The present study examined the hypothesis that the latter fact could explain the former. Four experiments measured speech reception threshold for a target voice against speech or non-speech maskers. Although differences in fundamental frequency ($\Delta F0$s) were shown to be beneficial even when presented to opposite ears (experiment 1), the authors' attempt to maximize their use by directing the listener's attention to the target $F0$ led to unexpected impairments (experiment 2) and the authors' attempt to hinder their use by generating uncertainty about the competing $F0$s led to practically negligible effects (experiments 3 and 4). The benefits drawn from $\Delta F0$s showed surprisingly little malleability for a cue that can be used in the complete absence of energetic masking. In half of the experiments, musicians obtained better thresholds than non-musicians, particularly in speech-on-speech conditions, but they did not reliably obtain larger $\Delta F0$ benefits. Thus, the data do not support the hypothesis that the musician advantage effect is based on greater ability to exploit $\Delta F0$s.

© *2017 Acoustical Society of America*. https://doi.org/10.1121/1.5005496

[VB] Pages: 1739–1755

I. INTRODUCTION

Musicians spend hours and hours practicing their instrument. In many cases, such as wind or string players, the act of playing requires constantly tuning every note. Years of practice progressively tighten the coupling between motor control and auditory feedback. One could therefore reasonably expect musicians to possess a finer sense of pitch than non-musicians, and they do (Sec. I A below). Musicians are also highly trained to listen to individual instruments within an ensemble. Those instruments may have similar pitch ranges and their respective melodies may cross each other. This places a strong emphasis on the ability to segregate one source from a competitor, such as tracking a pitch pattern or a timbre pattern over time without getting distracted by other sources. If skills obtained in the music domain could transfer to the speech domain (e.g., similar neural networks activated in both), one could reasonably expect musicians to make a better use of voice pitch cues in a background of other voices in the so-called "cocktail-party" situation (Cherry, 1953), but it is not clear yet whether this is the case.

A. Better periodicity coding

Pitch sensitivity is perhaps the most intuitive advantage demonstrated by musicians over non-musicians. Two studies (Spiegel and Watson, 1984; Kishon-Rabin *et al.*, 2001) showed that musicians obtained lower fundamental frequency ($F0$) discrimination thresholds than non-musicians, by a factor of about 2. More recently, Micheyl *et al.* (2006) showed that this difference can be larger: musicians exhibited thresholds as low as 2 cents (0.1% of $F0$) whereas non-musicians exhibited thresholds of about 12 cents (0.7% of $F0$), i.e., a factor of 6. This average of 12 cents was derived over the course of 2 h where non-musicians showed a great deal of improvement from about 26 cents at the first block to 8 cents by the last block, while musicians did not (being close to ceiling to begin with). Furthermore, when the testing was repeated across several days, non-musicians actually achieved thresholds of 2 or 3 cents by the second or third day, comparable to musicians. This provocative result suggests that an incredibly fine sense of pitch is actually accessible to anyone provided intense psychoacoustical training. How robust such acute sensitivity would be to the passage of time, or how generalizable it would be to other stimuli, other $F0$ ranges, or more musical tasks, remains an open question. Nonetheless, the differential pitch sensitivity between musicians and non-musicians, which is often taken for granted, may actually not be such a robust finding from a behavioral perspective.

From a neurophysiological perspective, on the other hand, the evidence is overwhelming. At a cortical level, using mismatch negativity, musicians' brains were shown to

[a)]Electronic mail: mickael.deroche@mcgill.ca automatically detect pitch violations in musical sequences which were undetected by non-musicians' brains (Koelsch *et al.*, 1999; Brattico *et al.*, 2002, 2009). At a subcortical level, using frequency following responses, musicians were shown to have larger and earlier brainstem responses to a syllable or a musical note (Musacchia *et al.*, 2007), and a more precise phase-locking to the periodicity of complex tones (Carcagno and Plack, 2011) or the periodicity/aperiodicity of tuned/detuned musical chords (Bidelman *et al.*, 2011). In other words, musicians have a stronger coding of $F0$ for speech or music stimuli than non-musicians, even at a pre-attentive level. Among many mechanisms that play a role in the recognition of a voice in a noisy background, periodicity of speech is a critical factor (Binns and Culling, 2007; Miller *et al.*, 2010; Deroche *et al.*, 2014a,b). This opened the intriguing possibility that musicians might outperform non-musicians in speech intelligibility tasks.

B. Speech intelligibility: The musician advantage effect

Parbery-Clark *et al.* (2009b) used two clinical tests of speech in noise (QuickSIN and HINT) to show that musicians outperformed non-musicians. However, the effect size was very small (<1 dB), and only observable when sources were collocated, not when they were spatially separated. In a subsequent study, Parbery-Clark *et al.* (2009a) recorded frequency following responses and found that musicians had a faster and enhanced neural representation of the harmonic structure of speech, which correlated with performance on the speech-in-noise tasks. The musician advantage effect, however, has been disputed. Ruggles *et al.* (2014) used the same two clinical tests (QuickSIN and HINT) but while musicians exhibited lower $F0$ discrimination thresholds than non-musicians (about 8 cents versus 30 cents), the authors did not find a difference between musicians and non-musicians in the speech recognition tasks (for both stationary and fluctuating noise). Zendel and Alain (2012) found differences in speech-in-noise performance between musicians and non-musicians but they only emerged beyond 40 yrs of age.

The perception of speech in noise is often limited by energetic masking primarily, i.e., the fact that the voice and the noise energy are present at similar times and frequencies. Informational masking (e.g., Durlach *et al.*, 2003; Kidd *et al.*, 2005), on the other hand, refers to the case where the listener's identification of an audible target, such as a voice, is impaired by the presence of a competing sound which does not share the same frequency band or occurs at different time windows than the target. Informational masking is thought to reflect central limitations and is known to vary widely across listeners with normal hearing (Neff and Dethlefs, 1995; Oh and Lufti, 1998; Lufti *et al.*, 2003). Given that musical training correlates most often with cognitive abilities (whether as a cause or as a consequence, see Corrigall *et al.*, 2013), one may expect musicians to differ from non-musicians in their susceptibility to informational masking, and they do. For example, Oxenham *et al.* (2003) used the random-frequency multitone-bursts method introduced by Kidd *et al.* (1994) and showed that musicians were much less susceptible to informational masking than non-musicians. They also measured frequency selectivity using the notch-noise method, and found no systematic difference between the two populations. It seems therefore that the "musician advantage effect" in speech intelligibility could be more easily observable in situations that involve a large amount of informational masking.

Boebinger *et al.* (2015) examined speech intelligibility in four different types of masker: speech-shaped noise, speech-modulated speech-shaped noise, rotated speech, and clear speech. They found no musician advantage effect in any of them. While this result may not have come as a surprise for noise maskers (given the inconsistencies reported above), it is somewhat surprising for the speech maskers known to involve informational masking. However, their target and masking voice had a different gender, so uncontrolled differences in fundamental frequency ($\Delta F0$s) and in vocal tract length (ΔVTL) could have eliminated most of the informational masking. To determine whether musicians could make a stronger use of $\Delta F0$s and ΔVTLs than non-musicians, Baskent and Gaudrain (2016) measured speech intelligibility against a single interfering talker. The identity of the voice was initially the same, but they manipulated $F0$ and VTL to create differences in vocal characteristics which provided masking releases. Musicians and non-musicians benefited equally from $\Delta F0$s and ΔVTLs, but musicians outperformed non-musicians in all conditions. At first sight, their result would suggest that the musician advantage effect must have other accounts than those related to $F0$ and VTL. In their study, however, the $F0$ contour of the voices still fluctuated. Even when the mean $\Delta F0$ was 0 semitones, there were instantaneous $\Delta F0$s between the two intonated voices, which musicians might have utilized more efficiently than non-musicians, thereby accounting for the musician advantage in all conditions. In fact, a recent study by Leclère *et al.* (2017) showed that the mean $\Delta F0$ between intonated sources is a poor predictor of the amount of masking release. Even more problematic is the fact that instantaneous $\Delta F0$s are themselves a poor predictor of masking release since the same instantaneous $\Delta F0$s resulting from a monotonized target against an intonated masker (at the same mean $F0$) led to no benefit whereas a monotonized masker against an intonated target (at the same mean $F0$) led to almost as much benefit as a fixed 3-semitones $\Delta F0$. Therefore, the question remains somewhat open: is it possible that musicians exploit instantaneous $\Delta F0$s more efficiently than non-musicians, which could (at least partly) account for the musician advantage effect?

C. Goal of the present study

In light of the aforementioned literature, the musician advantage effect is clearly not a robust phenomenon. The primary aim of this study was thus concerned with observing the musician advantage effect in the first place, i.e., whether musicians could obtain lower speech reception thresholds (SRTs) than non-musicians at least in some situations. We hypothesized that this would be primarily the case for situations that involved a large amount of informational masking, e.g., when using maskers made of a mix of two sentences spoken by the same male talker as the target (2-same-male maskers), but not necessarily for situations that involved little informational masking, e.g., when using speech-modulated buzzes (which provided similar amounts of energetic masking since they were equated in both a long-term excitation pattern and broadband temporal envelope to the speech maskers). The question that followed directly was to test whether musicians could obtain larger masking releases from $\Delta F0$s than non-musicians. If this were the case, it would be most likely observed in those same situations that involved a lot of informational masking where musicians could make a greater use of $\Delta F0$ as a streaming cue and exhibit greater selective attention than non-musicians. We also reasoned that this might depend on the size of $\Delta F0$s considered. Given that musicians have a finer sense of pitch than non-musicians, we hypothesized that differences between the two populations might be maximized at small $\Delta F0$s (such as 2 semitones), while non-musicians would have caught up to some extent at large $\Delta F0$s (such as 8 semitones), explaining why a group difference may not have been seen in past situations where $F0$s were uncontrolled.

In addition to the primary hypothesis (observing the musician advantage effect and determining whether this is partly due to a stronger benefit drawn from $\Delta F0$s) which was examined throughout the entire study, a number of secondary hypotheses were tested specifically in each experiment, all somewhat related to the issue of selective attention. Experiment 1 contrasted classic situations of diotic stimuli with more unorthodox situations where the target and maskers were presented to opposite ears, with and without $\Delta F0$. In the complete absence of energetic masking, SRTs should be extremely low but potentially limited by contralateral masking. We tested whether a small but measurable benefit could arise from such "binaural $\Delta F0$s." Experiment 2 examined whether it was possible to maximize the benefits of $\Delta F0$s by providing listeners with a beeping tone, prior to the speech stimuli, indicating the pitch of the target voice. Experiments 3 and 4 examined whether it was, on the contrary, possible to hinder the benefits of $\Delta F0$s by generating uncertainty about the competing $F0$s (roving or swapping their relative positions across trials). If enhancements (experiment 2) or impairments (experiments 3 and 4) could be observed, this would suggest that listeners were able to make some predictions about the pitch of the competing voices to be heard, opening the possibility that this might be how musicians would outperform non-musicians in their ability to utilize $F0$ cues.

II. GENERAL METHODS

A. Listeners

Twenty-four listeners participated in experiment 1, 12 musicians and 12 non-musicians. Twenty-four listeners participated in experiment 2, 12 musicians (eight of whom had taken part in experiment 1 and four new subjects) and 12 non-musicians (five of whom had taken part in experiment 1 and seven new subjects). Thirty-two listeners participated in experiments 3 and 4, 16 musicians (14 of whom had taken part in either experiment 1 or 2 and two new subjects) and 16 non-musicians (14 of whom had taken part in either experiment 1 or 2 and two new subjects). The number of listeners in each population was chosen to cover one complete rotation of the number of experimental conditions, necessary to counterbalance the effect of speech material. Among musicians, there were 4 males and 14 females aged between 18 and 28 yrs old (mean of 21.9 yrs and standard deviation of 2.6 yrs). All subjects included in the musician group had (1) begun musical training at or before 8 yrs old, (2) played an instrument (including singing) for at least 8 yrs, and (3) were still practicing their instrument on a regular basis at the time of testing. Among non-musicians, there were 7 males and 14 females, aged between 18 and 35 yrs old (mean of 25.1 yrs and standard deviation of 5.9 yrs). All these subjects identified themselves as non-musicians; they had not received musical training for more than 2 years (several of them had played the flute casually in high school) and were not practicing at the time of testing. All subjects 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 were native speakers of North-American English and had audiometric thresholds less than 20 dB hearing level at octave frequencies between 250 Hz and 8 kHz. None of them were familiar with the sentences used during the test. Including practice, experiments 1 and 2 lasted about 60 min each, and experiments 3 and 4 lasted about 90 min each.

B. Stimuli

All target stimuli were sentences taken from the Harvard Sentence List (Rothauser *et al.*, 1969), and spoken by one single male talker. There were 120, 120, 160, and 160 target sentences in experiments 1, 2, 3, and 4, respectively. Each experiment had a different set of sentences such that the same subject could participate in several of them without having been exposed to the stimuli before. The Praat PSOLA package (Boersma and Weenink, 2013) was used to resynthesize each sentence with a fixed $F0$ (which changed depending on the experimental case) throughout the entire duration. These stimuli were finally filtered following a procedure of energetic masking equalization (described in detail in the appendix of Deroche *et al.*, 2014b). Briefly, this procedure adjusts the spectral envelope of monotonized speech, such that at an equal root-mean-square (RMS) level, a given sentence has the same excitation level in unresolved regions regardless of its $F0$ (see right panel of Fig. 1), a characteristic that is generally not true at the output of Praat when $F0$ is changed substantially.

Two types of masker were generated: speech and non-speech maskers. Speech maskers came from 20 sentences, spoken by the same talker but different from, and slightly longer than, the target sentences. They were $F0$-processed through Praat and filtered in the same way as targets, and then added in pairs to create 2-voice maskers. Experiments 1 and 2 had six of those 2-voice maskers, respectively. Experiments 3 and 4 had the remaining four of those 2-voice maskers, respectively. Non-speech maskers were speech-modulated

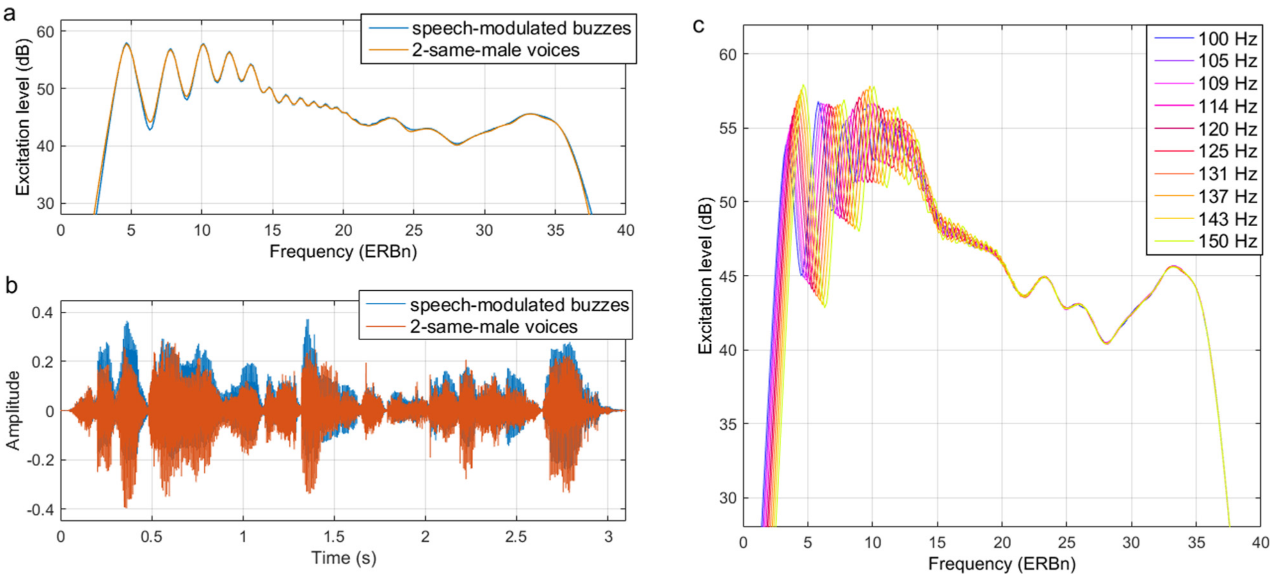

FIG. 1. (Color online) Averaged excitation patterns (a) and examples of broadband waveform (b) for the two masker types at an $F0$ of 150 Hz, as used in experiments 1 and 2. Averaged excitation patterns of the 2-voice maskers used in experiments 3 and 4, whose $F0$ varied in ten logarithmic steps between 100 and 150 Hz (c).

buzzes. Buzzes were created from a broadband random-phase harmonic complex with a specified $F0$ (that varied depending on the experimental case), which was also steady throughout the entire duration. This complex tone was then filtered with a linear-phase finite impulse response filter designed to match the averaged 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. Thus, experiments 1 and 2 had six of those speech-modulated buzzes. Experiments 3 and 4 had four different speech-modulated buzzes. Note that the same masker types were used in a recent study (Deroche *et al.*, 2017) which showed a strong contrast in terms of error types (i.e., random errors for speech-modulated buzzes versus confusions with masking words for 2-same-male voices) in addition to a large elevation of thresholds in the case of 2-same-male voices. Therefore, the speech maskers were certainly expected to involve much more informational masking than speech-modulated buzzes.

Panel (a) of Fig. 1 shows the average excitation pattern for the two masker types, at an $F0$ of 150 Hz as used in experiments 1 and 2. Panel (b) shows the waveforms of one example of speech-modulated buzz with one example of the respective 2-voice masker used in experiment 1. In both spectral and temporal domains, the two masker types were similar and should have produced roughly similar amounts of energetic masking. Panel (c) shows the average excitation pattern of the 2-voice maskers used in experiments 3 and 4, where $F0$ ranged among ten discrete values between 100 and 150 Hz. Note that the average excitation patterns of the corresponding speech-modulated buzzes were almost identical. This figure simply illustrates the fact that the position of peaks and dips necessarily changed with $F0$ in resolved regions of the masker (up to 800–1315 Hz for $F0$s between 100 and 150 Hz, i.e., roughly up to 14–18 equivalent-rectangular-bandwidth number, ERBn), but the excitation level was unaffected in unresolved regions (above 1444–2281 Hz for $F0$s between 100 and 150 Hz, i.e., roughly above 18–22 ERBn), thanks to the procedure used after the processing via Praat.

All stimuli were equalized to the same mean RMS level of 65 dB sound pressure level (SPL). In this study, the masker level is always defined as the combined level of two maskers together, and similarly, the target-to-masker ratio (TMR) is defined as the ratio between the level of a single target talker relative to the combined level of the maskers. A TMR of 0 dB thus corresponded to a situation where the level of the target talker was 3 dB above that of each masker. During the adaptive track, changes in TMR occurred by adjusting the target level while presenting maskers always at 65 dB.

C. Procedure

Each new listener began the study with three practice runs using unprocessed speech, not used in the rest of the experiment, masked by the speech-modulated buzz (one run) or the speech maskers (two runs). The following runs measured one SRT for each experimental condition. While each of the 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. SRT was measured using a 1-up/1-down adaptive threshold method (Plomp and Mimpen, 1979; Culling and Colburn, 2000), 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 buzzes. 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. 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 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 (while being instructed to disregard errors in verb tense, singular/plural, and words with the same root, see the Appendix for a complete analysis of self-scoring reports). 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 identified 3, 4, or 5 keywords correctly, and increased by 2 dB if the listener had identified 2, 1, or 0 keywords correctly. Measurement of each SRT was taken as the mean TMR over the last eight trials, and targeted a performance level of 50% intelligibility.

D. Equipment

Experiments were performed at two sites between 2014 and 2016. About 16% of the data were collected at the Music Perception Laboratory at Johns Hopkins, and about 84% were collected at the School of Communication Sciences and Disorders at McGill University. In both setups, signals were sampled at 44.1 kHz and 16 bits, digitally mixed, D/A converted by a sound card (either 24-bit Edirol UA-25, manufactured by Roland Corporation, U.S., for the Johns Hopkins site; or Scarlett 2i4, manufactured by Focusrite in Canada, for the site at McGill). They were presented binaurally over Sennheiser HD 280 headphones. The user-interface was displayed on a monitor, inside an audiometric booth. Listeners used a keyboard to type their transcript.

III. EXPERIMENT 1—Δ*F*0 ON OPPOSITE EARS

A. Rationale and design

The first experiment investigated whether a Δ*F*0 presented to opposite ears could lead to a measurable masking release on SRTs. If part of the Δ*F*0 effect was purely informational, one might be able to observe it in the most extreme case, where energetic masking was completely absent. Furthermore, it was of interest to see (1) how this phenomenon would vary with the size of Δ*F*0 (small or large), (2) whether it could also occur with non-speech maskers, or only with speech maskers, and (3) whether musicians would differ from non-musicians in these respects. There were 12 experimental conditions, resulting from 2 presentation modes (diotic or dichotic) × 2 masker types (speech-modulated buzzes or 2-voice maskers) × 3 Δ*F*0s (0, −2, and −8 semitones).

In addition, an extra condition was run for diotic speech in silence, simply to determine the "floor SRT" in the absence of masking. It was not part of the main statistical analysis, but provided us with a baseline to compare the dichotic SRTs where performance could be impaired by contralateral distractors in the absence of energetic masking. Note that the TMR is obviously not defined when there is no masker, but it was still possible to locate this measurement on the same scale simply by considering the target level relative to 65 dB SPL.

B. Results

A repeated-measures analysis of variance (ANOVA) was conducted with one between-subjects factor (population) and three within-subjects factors (masker type, presentation, and Δ*F*0) in order to determine the influence of each factor on the SRTs shown in Fig. 2. The results are reported in Table I.

There was no main effect of population, nor did it interact in 2-, 3-, or 4-ways. The three other factors led to significant main effects and interactions. The main effect of masker type reflected that SRTs were higher (by 4 dB on average) for 2-voice maskers than for buzzes (circles versus squares in Fig. 2). The main effect of presentation reflected that SRTs were much higher (by 28 dB on average) for diotic than dichotic conditions (filled versus empty symbols in Fig. 2). The main effect of Δ*F*0 reflected that SRTs were lowered when the target *F*0 differed from the masker *F*0 (variation along the abscissa in Fig. 2). The 3-way interaction was further examined by testing the simple effect of Δ*F*0 at each factorial combination of the other two factors. As expected, there was a significant Δ*F*0 benefit for the two diotic conditions [buzzes: $F(2,21) = 79.7$, $p < 0.001$; 2-voice: $F(2,21) = 74.3$, $p < 0.001$]. More surprisingly, the Δ*F*0 benefit was also significant for the contralateral buzzes [$F(2,21) = 5.5$, $p = 0.012$], as well as the contralateral 2-voice

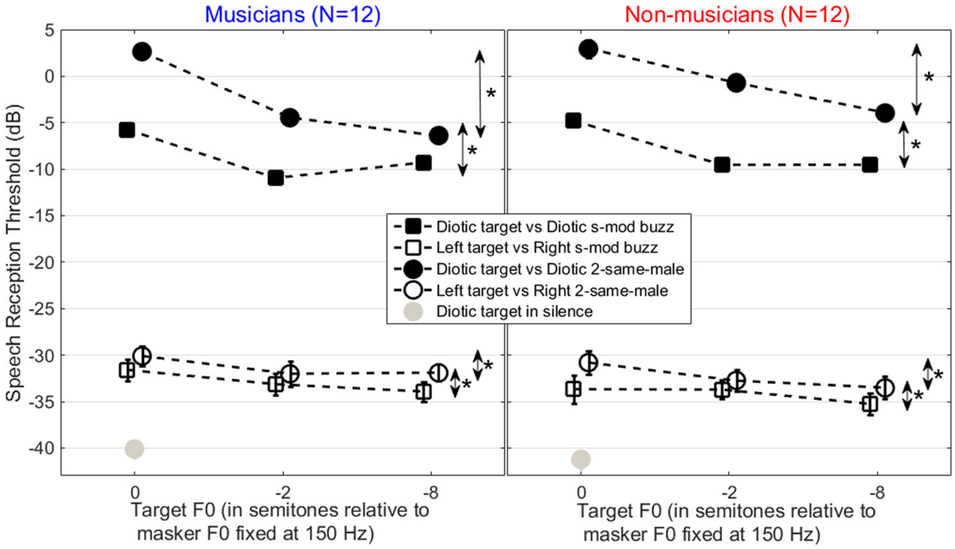

FIG. 2. (Color online) Mean SRTs measured in experiment 1, obtained by musicians (left panel) and non-musicians (right panel), for a monotonized target voice against speech-modulated buzzes (squares) or against two interfering voices (circles), presented either diotically (i.e., same input on both ears, filled symbols) or dichotically (i.e., different input on opposite ears, empty symbols). Lower thresholds indicate greater intelligibility. Here and later, error bars are +/− one standard error of the mean over listeners, but may not be visible given the scale of SRTs. Here, asterisks indicate a significant effect of Δ*F*0.

TABLE I. Statistics for experiments 1 and 2, using a mixed between/within ANOVA. Population consisted of musicians or non-musicians. Masker type was either speech-modulated buzzes or 2-voice maskers (same male voice as the target). The $\Delta F0$ was 0, −2, or −8 semitones. Presentation referred to either diotic/dichotic presentation in experiment 1, or the presentation with/without priming in experiment 2. Mauchly's tests of sphericity were never significant [$\chi 2(2) < 2.2$, $p > 0.345$ in experiment 1; $\chi 2(2) < 4.4$, $p > 0.111$ in experiment 2]. *indicates p values less than 0.05.

Factors	Experiment 1	Experiment 2
population	$F(1,22) < 0.1$ $p = 0.851$	$F(1,22) = 3.7$ $p = 0.066$
masker type	$F(1,22) = 220.8$ $p < 0.001$*	$F(1,22) = 1574.0$ $p < 0.001$*
presentation	$F(1,22) = 1297.5$ $p < 0.001$*	$F(1,22) = 21.0$ $p < 0.001$*
$\Delta F0$	$F(2,44) = 82.6$ $p < 0.001$*	$F(2,44) = 170.5$ $p < 0.001$*
population × masker type	$F(1,22) = 2.6$ $p = 0.123$	$F(1,22) < 0.1$ $p = 0.799$
population × presentation	$F(1,22) = 2.9$ $p = 0.100$	$F(1,22) = 0.9$ $p = 0.351$
population × $\Delta F0$	$F(2,44) = 2.3$ $p = 0.115$	$F(2,44) = 0.9$ $p = 0.411$
masker type × presentation	$F(1,22) = 108.8$ $p < 0.001$	$F(1,22) = 17.5$ $p < 0.001$*
masker type × $\Delta F0$	$F(2,44) = 6.4$ $p = 0.004$*	$F(2,44) = 9.1$ $p = 0.001$*
presentation × $\Delta F0$	$F(2,44) = 25.8$ $p < 0.001$*	$F(2,44) = 2.2$ $p = 0.123$
population × masker type × presentation	$F(1,22) = 1.6$ $p = 0.221$	$F(1,22) = 0.9$ $p = 0.352$
population × masker type × $\Delta F0$	$F(2,44) = 0.3$ $p = 0.743$	$F(2,44) = 0.8$ $p = 0.445$
population × presentation × $\Delta F0$	$F(2,44) = 0.5$ $p = 0.636$	$F(2,44) = 0.6$ $p = 0.553$
masker type × presentation × $\Delta F0$	$F(2,44) = 7.1$ $p = 0.002$*	$F(2,44) = 4.4$ $p = 0.018$*
4-way	$F(2,44) = 2.6$ $p = 0.087$	$F(2,44) < 0.1$ $p = 0.916$

maskers [$F(2,21) = 6.4$, $p = 0.007$], although the effect size was modest, about 2 dB.

Finally, for the diotic conditions, the pattern of the masking release differed depending on the masker type: for diotic buzzes, SRT decreased from 0 to −2 semitones ($p < 0.001$) but did not decrease further ($p = 0.266$); for diotic 2-voice maskers, SRT decreased from 0 to −2 semitones ($p < 0.001$) and decreased further from −2 to −8 semitones ($p < 0.001$). In other words, the $\Delta F0$ benefit was steep with buzzes (4–5 dB as soon as −2 semitones), whereas it was more gradual for 2-voice maskers (5.3 and 7.9 dB, respectively, at −2 and −8 semitones).

C. Discussion

1. The musician advantage effect

For the dichotic conditions (empty symbols), both populations performed similarly well, with extremely low SRTs between −30 and −35 dB. Thus, in the absence of energetic masking, there was no musician advantage effect. For that matter, we can also analyze the SRTs for the extra condition of speech in quiet (gray dots). An independent-samples t-test revealed no effect between the two populations [$t(22) = 0.8$, $p = 0.420$]. Both populations reached an SRT of −40 dB or less, meaning that all listeners, whether they were musically trained or not, could comprehend speech in quiet, at 50% intelligibility, at a presentation level of only 25 dB SPL. What is interesting about this measure is that it is at least 5 dB lower than the lowest SRT obtained among dichotic conditions. Admittedly, in the dichotic conditions, the target voice was only presented on one side, while it was presented on both for speech in quiet. This binaural summation for speech intelligibility in quiet should amount to 3 dB at most (Shaw *et al.*, 1947). Anything beyond that can be taken as evidence that the contralateral presence of maskers did produce some distraction. Listeners could not completely ignore their right ear in which no useful information was to be heard, and in this respect musicians were as bad as non-musicians. This was confirmed by restricting the data to a subset of SRTs corresponding to the dichotic conditions only. Population did not lead to a main effect, nor did it interact with masker type or $\Delta F0$.

In the diotic conditions, musicians tended to obtain lower thresholds than non-musicians, but none of the interactions involving musicianship reached significance. Therefore, the musician advantage effect was overall not observed in experiment 1.

2. $\Delta F0$ alleviates contralateral distraction

Regardless of musicianship, this experiment succeeded in showing a $\Delta F0$ effect for contralateral maskers. The intelligibility of a voice presented to the left ear improved when its $F0$ was set further apart from the $F0$ of maskers presented to the right ear. As mentioned earlier, these dichotic SRTs were higher than 3 dB above the floor SRT (gray dot). Therefore, contralateral maskers must have produced some form of distraction. This distraction may be referred to as informational masking and has been observed in a similar paradigm before. For example, Wightman *et al.* (2010) used the coordinate response measure with a target voice presented monaurally and masked by speech-shaped noise (contrary to our experiment where the voice presented monaurally on the left side was in silence). On the contralateral ear, they presented either nothing (which formed a baseline condition), or a speech-modulated noise, or a single male talker, or a single female talker. The latter three conditions were subtracted from the baseline to estimate the amount of informational masking that resulted from each masker type. The speech-modulated noise produced no informational masking across all listeners, whereas speech maskers produced about 4 dB (in adults, and as much as 20 dB in children), but the gender of the interfering voice did not matter. Several aspects can be outlined to compare the two studies. First, here, SRTs for the contralateral speech-modulated buzzes (empty squares) were on average 2 dB lower than those for the contralateral voices (empty circles). This is consistent with the idea that speech maskers are indeed more effective at producing informational masking than speech-modulated buzzes, and one could perhaps speculate further that speech-modulated buzzes would presumably be more effective than speech-modulated noise at producing informational masking. Second, $F0$s were not controlled in the study by Wightman *et al.*, so there must have been $\Delta F0$s (at least instantaneous $\Delta F0$s if not mean $\Delta F0$s) available to listeners. This might have been a constant factor across both of their speech masker conditions, and given their lack of gender effect, it is most likely that binaural $\Delta F0$ benefits as observed here were absent in their data. Nonetheless, what is particularly surprising in the present data is (1) that this "binaural $\Delta F0$ benefit" occurred equally with both masker types, and (2) that it occurred with just 2 semitones $\Delta F0$s. Since speech maskers produced more informational masking than buzzes, it would have seemed likely that a binaural $\Delta F0$ benefit would have occurred for speech maskers more than buzzes. Similarly, a small $\Delta F0$ such as 2 semitones is not an effective cue to release from informational masking (Darwin *et al.*, 2003), while 8 semitones is. It would have seemed more likely that this binaural $\Delta F0$ benefit would have occurred at 8 but not 2 semitones.

Perhaps a better way to interpret the surprising traits of this phenomenon is not so much that listeners used $\Delta F0$s between the two ears to alleviate contralateral distraction, but rather that the case of a contralateral masker presented with exactly the same $F0$ generated a binaural fusion (Cramer and Huggins, 1958; Bilsen, 1976; Pantev *et al.*, 1996) with information from the other ear which would not have occurred otherwise. Indeed, it is difficult to find a realistic situation where binaural fusion of identical $F0$s is undesirable, at least for normally-hearing listeners (this can certainly happen for users of cochlear implants, for example, where the two ears can have different tonotopic maps). Therefore, it makes perfect sense for the brain to fuse identical harmonic sources coming from opposite ears, as they are most likely coming from the same speaker. From the perspective of binaural $F0$ fusion, (1) there is no need for maskers to be linguistic or not, and (2) this phenomenon would be avoided every time a $\Delta F0$ exists regardless of its size. In other words, the effect seemed more of a same-$F0$ impairment (caused by binaural fusion) than a $\Delta F0$ benefit. Of course, it is worth noting that this binaural fusion could in principle occur in the diotic conditions as well, but it would presumably be swamped by large energetic masking effects.

IV. EXPERIMENT 2—PRIMING TO THE TARGET *F0*

A. Rationale and design

From the results of experiment 1, it seemed that if a musician advantage effect could be seen, it would be first observed in diotic conditions. Thus, the following experiments were all presented in diotic conditions. However, it remains unclear how much of the listener's attention is needed to make the most use of $\Delta F0$s. Experiment 2 investigated whether the use of $\Delta F0$s could be enhanced by drawing the listener's attention toward the target $F0$ in the form of a prime. If the prime could help tracking the target voice, it would be very informative to know whether this priming benefit depended on (1) the masker type, (2) the size of $\Delta F0$, and (3) whether listeners were musically trained or not. In half of the conditions, a beeping tone was presented prior to the stimulus onset, which indicated the pitch of the target voice. Listeners were instructed to try hard to focus their attention on the voice spoken at this pitch. If the use of $\Delta F0$s could be maximized by attention toward the target $F0$, one would expect a larger masking release with the priming cue than without, and perhaps particularly for a large $\Delta F0$, and with speech maskers rather than buzzes. We did not have any strong expectation as to whether musicians or non-musicians would utilize this priming cue more efficiently. One could imagine that musicians would profit from their finer sense of pitch by being better able to utilize this priming cue and selectively attend to the target voice, perhaps specifically for a small $\Delta F0$ (more challenging). Alternatively, it might be equally possible that musicians already excel at switching their attention very quickly to the pitch of the target voice, and consequently the priming cue might actually benefit non-musicians more.

The beeping tone consisted of a 150-ms long harmonic complex at the target $F0$, with a standard speech-shaped spectral profile (different from the specific profile of the male speaker), which was presented four times separated by 150 ms of silence, and set at 65 dB SPL. The other half of the conditions were a direct replication of the diotic conditions in experiment 1 (filled symbols). Thus, there were 12 experimental conditions, resulting from 2 presentations (with or without priming) × 2 masker types (speech-modulated buzzes or 2-voice maskers) × 3 $\Delta F0$s (0, -2, and -8 semitones). All stimuli were presented diotically.

B. Results

A repeated-measures ANOVA was conducted to determine the influence of each factor on the SRTs shown in Fig. 3. The results are reported in Table I. Population had no main effect, nor did it interact in 2-, 3-, or 4-ways. The main effect of masker type reflected, as earlier, that SRTs were higher (by 8.5 dB on average) for speech maskers (right panels) than for buzzes (left panels), due to the large informational masking induced by speech maskers. The main effect of priming was significant overall, but it was detrimental. Also, it strongly interacted with masker type: indeed, the simple effect of priming was significant for 2-voice maskers [$F(1,22) = 49.5$, $p < 0.001$], but not for buzzes [$F(1,22) = 1.0$, $p = 0.318$]. Furthermore, the effect of priming for 2-voice maskers was itself restricted to the conditions of -2 and -8 semitones [$F(1,22) = 44.8$ and 30.3, $p < 0.001$ in both cases]; it was not significant for the 2-voice maskers at 0 semitones [$F(1,22) = 2.4$, $p = 0.136$], and not significant for the buzzes at any $\Delta F0$ size [$F(1,22) = 2.4$, $p = 0.139$].

Finally, there was a main effect of $\Delta F0$, which interacted with masker type. As in experiment 1, it was found that for buzzes, SRT decreased from 0 to -2 semitones ($p < 0.001$) but did not decrease further with -8 semitones ($p = 0.477$). In contrast, for 2-voice maskers, SRT decreased more gradually from 0 to -2 to -8 semitones ($p < 0.001$ in all three comparisons). In other words, the $\Delta F0$ benefit was steep with speech-modulated buzzes (benefit of 4.9 and 5.5 dB, respectively, at -2 and -8 semitones, on average over priming and

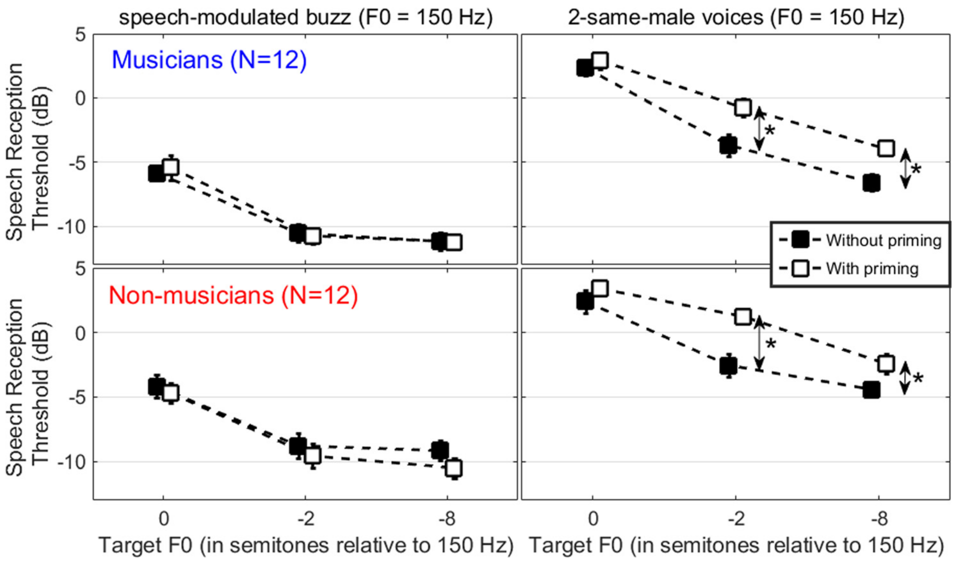

FIG. 3. (Color online) Mean SRTs measured in experiment 2, obtained by musicians (top panels) and non-musicians (bottom panels), for a monotonized target voice against speech-modulated buzzes (left panels) or two interfering voices (right panels), presented with and without a beeping tone that cued the pitch of the target voice. Here, asterisks indicate a significant effect of priming.

population), whereas it was more gradual for 2-voice maskers (4.2 and 7.1 dB, respectively, at −2 and −8 semitones, on average over priming and population).

C. Discussion

1. Detrimental effect of priming

Not only was there no benefit in providing listeners with the pitch of the voice they should attend to, but this manipulation was actually detrimental to performance in both populations. This result was certainly unexpected and remains quite puzzling. At first sight, one might think that the percept of the beeping tone was too similar to the percept of the speech-modulated buzzes. After all, both had buzz-like qualities (despite having a somewhat different timbre due to spectral envelope differences). This could have incited listeners to expect the pitch of the buzz to match the priming cue, despite our instructions. However, this interpretation does not hold because priming had in fact no effect for buzzes. Alternatively, one might think that the prime caused some sort of adaptation to the target $F0$, thus providing the masker with a "pop-out" benefit when it comes on, or that the prime acted as a distractor by being presented too close to the beginning of the target sentences. But once again, those interpretations do not hold on the basis that the impairment did not occur for buzzes; it occurred only against speech maskers, and only in the different-$F0$ conditions which leads us to think that this is to some extent related to $F0$ processing.

There are several potential accounts for the $\Delta F0$ effect, and harmonic cancellation stands as one likely candidate (de Cheveigné, 1993; de Cheveigné *et al.*, 1995). It is thought that the auditory system is capable of filtering out the entire harmonic structure of a complex masker, consequently enhancing the TMR in conditions where target and masker differ in $F0$. For such a process to operate, however, the system needs a way to identify which $F0$ belongs to the target source, and which $F0$(s) belong(s) to the interfering source(s). To date, it is not known where in the brain or how this identification step occurs, but one may speculate that it interacts with the listener's knowledge and attention in the form of a top-down influence tuning the harmonic cancellation to the masker $F0$. A paradox arises when wondering how listeners could tune anything to the masker $F0$ while their attention is focused on tracking the target voice. This paradox is generally resolved by assuming that internal auditory processes must be tracking the pitch of competing voices at a pre-attentive level, but this paradox remains overall an open question. One possibility is that as listeners were asked to try hard to focus their attention on a particular pitch, perhaps this forced harmonic cancellation to tune to that particular $F0$, which would have canceled the target voice, leaving the 2-voice maskers relatively unaffected, thereby explaining the 2-dB reduction in masking release. No impairment would have occurred when all sources shared the same $F0$ since harmonic cancellation would have canceled all of them equally. In the case of non-speech maskers, one may imagine that other mechanisms—perhaps more automatic—engage in recognizing linguistic from non-linguistic units. So, the buzz $F0$ would be automatically identified without interference from the listener's attention, and subsequently fed to the harmonic cancellation stage. When thinking of realistic situations, however, the present result is in any case difficult to absorb. A form of priming to the target $F0$ occurs every time one is being called by one's name before hearing a message, and this certainly feels helpful, not detrimental. Therefore, this priming-induced impairment warrants further investigation in the future. For example, one exciting avenue would be to present the beeping tone at the masker $F0$ and ask listeners to ignore the voices spoken at this pitch. Although this seems quite a convoluted instruction to follow, perhaps this would turn the priming cue into an advantage rather than a drawback. At the very least, it would expand our understanding of the relationship between the harmonic cancellation stage and the listener's knowledge of the identity of competing $F0$s, which remains to this date largely unexplored.

2. The musician advantage effect

The two populations did not differ in this experiment. For speech-modulated buzzes, musicians and non-musicians obtained similar masking releases, about 5 dB for both −2 and −8 semitones (on average over priming conditions). For speech maskers, musicians obtained masking release of 4.8 and 7.9 dB, respectively, with −2 and −8 semitones $\Delta F0$ (on average over priming), while non-musicians obtained masking release of 3.6 and 6.3 dB, respectively, (on average over priming). Another finding confirmed by experiment 2 is that the $\Delta F0$ benefit arises for small $\Delta F0$s and asymptotes at 2 semitones for buzzes, whereas it is more gradual for competing voices. This observation is consistent with those by Deroche and Culling (2013), who measured SRTs for $\Delta F0$s of 0, +2, and +8 semitones (relative to 110 Hz) for the same masker types. For the speech-modulated buzzes, the masking releases were about 5 dB at both 2 and 8 semitones, whereas for the 2-voice maskers, the masking releases were 3 and 8 dB, respectively, at 2 and 8 semitones. The most likely interpretation for this distinct pattern of $\Delta F0$ benefit is that part of the masking release is of informational nature for speech maskers and requires a large $\Delta F0$ to start contributing (Darwin *et al.*, 2003). In contrast, the masking release obtained with buzzes may be largely energetic, and therefore similar to what has been observed for double-vowels experiments (Culling and Darwin, 1993; de Cheveigné *et al.*, 1997a; de Cheveigné *et al.*, 1997b). What the present results add to the literature is that musicianship does not change anything about this distinct pattern of masking release depending on masker type. This result is somewhat surprising considering that musicians should be less affected than non-musicians by informational masking (Oxenham *et al.*, 2003). One would therefore have expected musicians to display a more similar pattern of masking release in both masker types. However, this result is in line with the lack of musicianship effects in two other speech recognition studies (Ruggles *et al.*, 2014; Boebinger *et al.*, 2015) that used different maskers along the continuum between energetic and informational.

V. EXPERIMENT 3—UNCERTAINTY ABOUT A SMALL $\Delta F0$

A. Rationale and design

Since experiment 2 failed to help listeners optimize their use of $\Delta F0$s, we tried the opposite approach, i.e., hindering their use of $\Delta F0$s. Two factors were tested: (1) roving the masker $F0$, randomly across trials within a block, and (2) swapping the position of the target $F0$ above and below the masker $F0$, randomly across trials within a block (whereas the target $F0$ was always at or below the masker $F0$ in the first two experiments). The first factor focused on the role of masker consistency. For example, a process such as harmonic cancellation might benefit from fixing the value of masker $F0$ throughout an entire experiment, as opposed to having to adjust to a different masker $F0$ on every trial. The second factor focused on the role of target consistency, in absolute or relative terms. For example, if listeners preferentially tuned to the higher-pitch voice in a crowd, then the masking release might not be affected by roving the masker $F0$ as long as the target $F0$ is consistently $+n$ semitones above, but it would be affected by swapping the target $F0$, sometimes $+n$ semitones above, and sometimes $-n$ semitones below the masker $F0$, referred here as the $\pm n$ semitones condition. Thus, the present design resulted in 16 experimental conditions, resulting from 2 types of masker $F0$ roving [fixed at 125 Hz or variable over ten logarithmic steps between 100 and 150 Hz, as displayed in panel (c) of Fig. 1] × 2 masker types (speech-modulated buzzes or 2-voice maskers) × 4 $\Delta F0$s (0, −2, +2, ±2 semitones). All stimuli were presented diotically.

B. Results

A repeated-measures ANOVA was conducted with one between-subjects factor (population) and three within-subjects factors (masker type, masker roving, and $\Delta F0$) in order to determine the influence of each factor on the SRTs shown in Fig. 4. The results are reported in Table II.

There was a main effect of population reflecting that, on average across the three within-subjects factors, the mean SRT was 1.5 dB lower for musicians than for non-musicians. The main effect of masker type reflected that SRTs were overall elevated (by 8.5 dB) for 2-voice maskers compared with speech-modulated buzzes; a difference presumably due to informational masking for the most part. A modest interaction between these two factors indicated that musicians outperformed non-musicians by 2 dB with 2-voice maskers ($p = 0.003$), but not significantly with buzzes ($p = 0.095$).

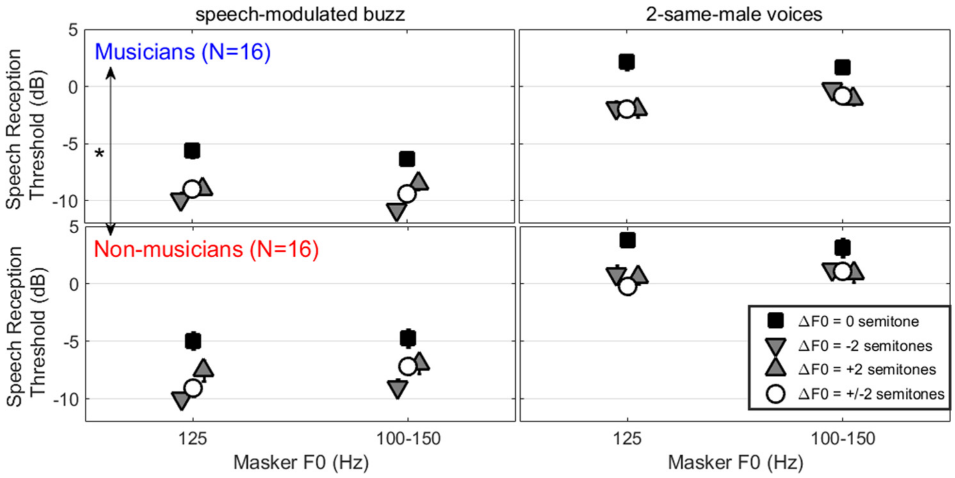

FIG. 4. (Color online) Mean SRTs measured in experiment 3, obtained by musicians (top panels) and non-musicians (bottom panels), for a monotonized target voice against speech-modulated buzzes (left panels) or two interfering voices (right panels). The masker $F0$ was either fixed at 125 Hz or varied randomly across trials between 100 and 150 Hz. Relative to the masker $F0$, the target $F0$ was either identical, 2 semitones above, below, or swapped across trials within a block. Here, the asterisk indicates the musician advantage effect.

TABLE II. Statistics for experiments 3 and 4, using a mixed between/within ANOVA. Population consisted of musicians or non-musicians. Masker type was either speech-modulated buzzes or 2-voice maskers. The masker $F0$ was either fixed at 125 Hz, or roved pseudo-randomly across trials within a block among ten discrete values between 100 and 150 Hz. The $\Delta F0$ was 0, −2, +2, or ±2 semitones in experiment 3; it was 0, −8, +8, or ±8 semitones in experiment 4. Mauchly's tests of sphericity were never significant in experiment 3 [$\chi2(5) < 5.2$, $p > 0.389$]. In experiment 4, Mauchly's tests of sphericity were not significant [$\chi2(5) < 5.6$, $p > 0.344$], except for the interaction masker type by $\Delta F0$ [$\chi2(5) = 18.3$, $p = 0.003$] whose degree of freedom was adjusted with Greenhouse-Geisser correction. *indicates p values less than 0.05.

Factors	Experiment 3	Experiment 4
population	$F(1,30) = 6.8$ $p = 0.014$*	$F(1,30) = 11.4$ $p = 0.002$*
masker type	$F(1,30) = 1747.5$ $p < 0.001$*	$F(1,30) = 1370.6$ $p < 0.001$*
masker $F0$	$F(1,30) = 5.5$ 0.026*	$F(1,30) = 5.6$ $p = 0.025$*
$\Delta F0$	$F(3,90) = 87.7$ $p < 0.001$*	$F(3,90) = 131.9$ $p < 0.001$*
population × masker type	$F(1,30) = 4.2$ $p = 0.050$*	$F(1,30) = 4.6$ $p = 0.041$*
population × masker $F0$	$F(1,30) = 1.4$ $p = 0.253$	$F(1,30) = 1.3$ $p = 0.266$
population × $\Delta F0$	$F(3,90) = 0.4$ $p = 0.765$	$F(3,90) = 0.5$ $p = 0.657$
masker type × masker $F0$	$F(1,30) = 0.7$ $p = 0.413$	$F(1,30) = 5.6$ $p = 0.025$*
masker type × $\Delta F0$	$F(3,90) = 9.2$ $p < 0.001$*	$F(2.2,65.0) = 38.6$ $p < 0.001$*
masker $F0$ × $\Delta F0$	$F(3,90) = 4.3$ $p = 0.007$*	$F(3,90) = 0.4$ $p = 0.778$
population × masker type × masker $F0$	$F(1,30) = 5.7$ $p = 0.024$*	$F(1,30) = 0.5$ $p = 0.466$
population × masker type × $\Delta F0$	$F(3,90) = 0.3$ $p = 0.848$	$F(2.2,65.0) = 0.6$ $p = 0.578$
population × masker $F0$ × $\Delta F0$	$F(3,90) = 1.1$ $p = 0.359$	$F(3,90) = 2.2$ $p = 0.095$
masker type × masker $F0$ × $\Delta F0$	$F(3,90) = 0.9$ $p = 0.457$	$F(3,90) = 0.9$ $p = 0.461$
4-way	$F(3,90) = 0.9$ $p = 0.462$	$F(3,90) = 0.8$ $p = 0.471$

There was also a complex interaction between population, masker type, and masker roving. Examining the simple effect of population for each factorial combination of the other two factors, we found that musicians outperformed non-musicians in the presence of 2-voice maskers with fixed or variable $F0$ (by 2.2 dB, $p = 0.004$; and by 1.7 dB, $p = 0.008$, respectively). Musicians also outperformed non-musicians in the presence of buzzes with variable $F0$ (by 1.8 dB, $p = 0.022$), but not for buzzes with fixed $F0$ ($p = 0.482$). In other words, the musician advantage effect was observed overall, except for the simplest cases of buzzes consistently presented at 125 Hz.

Regardless of population, there was an effect of $\Delta F0$, as expected. The mean SRT at 0-semitones $\Delta F0$ (black squares) was higher than the mean SRT at −2, +2, or ±2 semitones ($p < 0.001$ in every case), i.e., a masking release was observed whenever a $\Delta F0$ was available. The amount of masking release depended on the masker type and the $\Delta F0$. For buzzes, −2 semitones provided more benefit than +2 or ±2 semitones ($p < 0.001$ in both cases) which did not differ between them ($p = 0.179$). For 2-voice maskers, all three benefits were similar ($p > 0.471$).

Roving the masker $F0$ across trials elevated SRTs overall, but this factor interacted with $\Delta F0$: it had no effect at 0 semitones (in fact, slightly lowering SRTs) whereas it tended to elevate SRTs when a $\Delta F0$ was available. To clarify these trends, SRTs for the three conditions of $\Delta F0$ were subtracted from SRTs at 0-semitones $\Delta F0$ to extract the $\Delta F0$ benefit, in each of the three cases of −2, +2, and ±2 semitones, displayed in Fig. 5. A similar ANOVA was conducted to determine the influence of each factor on the $\Delta F0$ benefits. Varying the masker $F0$ across trials reduced the masking release by 1.1 dB on average, and this effect was similar for −2, +2, or ±2-semitones $\Delta F0$, regardless of masker type. Also, note that this additional ANOVA did not reveal a main effect of population, nor did it interact in 2-, 3-, or 4-ways. At least for small $\Delta F0$s such as 2 semitones, musicians did not exploit $\Delta F0$s more efficiently than non-musicians.

C. Discussion

1. The musician advantage effect

Musicians obtained lower SRTs than non-musicians, and this difference was particularly pronounced in speech-on-speech masking conditions. Our main hypothesis was that this could be due to musicians being better able to exploit $\Delta F0$s between competing voices than non-musicians. The present data did not support this interpretation: when examining the $\Delta F0$ benefits, both groups obtained similar amounts of masking release and population did not interact with any other factors (Fig. 5).

Another way to strengthen the idea that the musician advantage effect is not due to a better ability to exploit $\Delta F0$s is to examine SRTs for the 0-semitones conditions. Non-musicians obtained SRTs of −5 and +3.5 dB for buzzes and 2-voice maskers, respectively. In comparison, musicians obtained SRTs of −6 and +2 dB, respectively. To a small degree, the musician advantage effect already exists in the absence of $\Delta F0$. Statistical analysis of this subset of data falls just short of significance ($p = 0.052$; the same test was significant in experiment 4, although it was not significant in experiments 1 and 2). Nonetheless, this reinforces the idea that musicians tend to outperform non-musicians for reasons that are unrelated to $F0$-segregation.

2. Uncertainty about the masker $F0$

Roving the masker $F0$ from trial to trial did reduce the masking releases to a small degree. One could in principle attribute this effect either to uncertainty about the masker $F0$, or uncertainty about the target $F0$ since it was positioned relative to the masker $F0$. However, there was overall no effect of randomly swapping the target $F0$ above or below the masker $F0$ within a block (the ±2 semitones condition), even when the latter was also roved between 100 and 150 Hz across trials. Thus, listeners coped very well with uncertainty about the target $F0$. Therefore, it appears that the small reduction in masking release due to roving of the competing

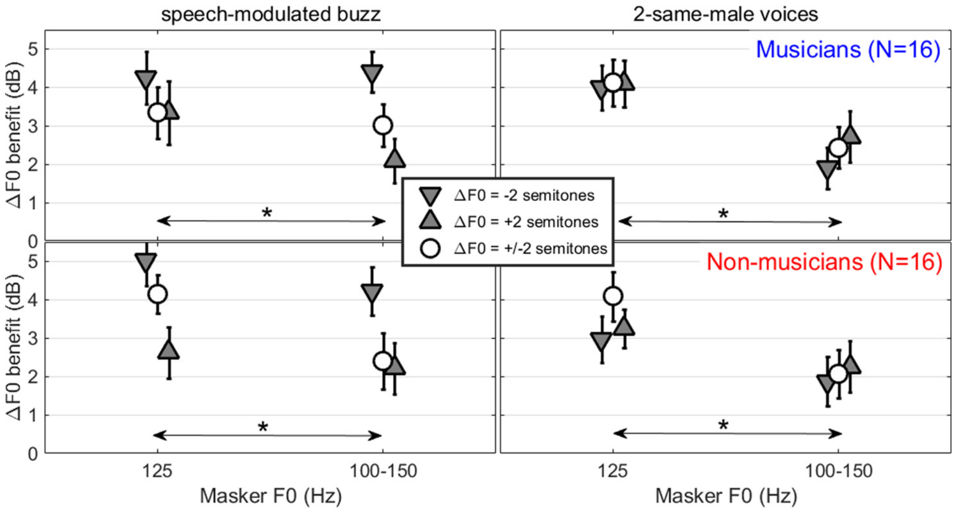

FIG. 5. (Color online) Mean Δ$F0$ benefits (extracted from the difference in SRTs relative to the 0-semitones baseline) measured in experiment 3. Here, asterisks indicate a significant effect of roving the masker $F0$ across trials which resulted in a small reduction in masking release, while swapping the target $F0$ above or below the masker $F0$ had no effect.

$F0$s within a block was primarily caused by uncertainty about the masker $F0$.

3. *Spectral glimpsing*

The amount of masking release was roughly similar whether the Δ$F0$ was −2, +2, or ±2 semitones, in the presence of 2-voice maskers. In contrast, for buzzes, a Δ$F0$ of −2 semitones (downward triangles) provided a little more benefit than the other two cases. This is presumably due to spectral glimpsing effects: for a given masker $F0$, there are spectral dips in resolved regions of the masker where target energy can potentially be glimpsed, and there is in principle more chance for this to happen when the target $F0$ is low (and therefore fills in the spectral dips) than when it is high. The reason why this does not happen for 2-voice maskers may be that it is difficult to glimpse spectrally in a temporally-modulated harmonic masker (Deroche *et al.*, 2014b; Leclère *et al.*, 2017). Note that these effects were not the focus of the present study: the conditions of −2 and +2 semitones were primarily used to obtain baselines from which to compare the ±2 semitones condition. However, the ±2 semitones condition always led to SRTs in between the fixed conditions of +2 and −2 semitones, revealing no effect of interest and emphasizing the idea that listeners perform just as well with unexpected target $F0$s.

VI. EXPERIMENT 4—UNCERTAINTY ABOUT A LARGE Δ*F0*

A. Rationale and design

As shown in experiments 1 and 2, the amount of masking release is gradual as a function of the Δ$F0$ size, in the presence of interfering voices. The reason for this continuing improvement could relate to $F0$-streaming or the fact that listeners are able to use a large Δ$F0$ to avoid confusing which voice they should attend to (Darwin *et al.*, 2003). We reasoned that a design identical to experiment 3 but with 8-semitones Δ$F0$ might lead to different results, because the informational masking release provided by large Δ$F0$s may be more prone to these uncertainty effects. Thus, experiment 4 consisted of 16 experimental conditions, resulting from 2 types of masker $F0$ roving (fixed at 125 Hz or variable over 100–150 Hz) × 2 masker types (speech-modulated buzzes or 2-voice maskers) × 4 Δ$F0$s (0, −8, +8, ±8 semitones). All stimuli were presented diotically.

B. Results

A repeated-measures ANOVA was conducted with one between-subjects factor (population) and three within-subjects factors (masker type, masker roving, and Δ$F0$) in order to determine the influence of each factor on the SRTs shown in Fig. 6. The results are reported in Table II.

There was a main effect of population: the mean SRT was overall 1.7 dB lower for musicians than non-musicians. Population also interacted with masker type, as musicians outperformed non-musicians by 1.3 dB for buzzes ($p = 0.027$), and even more, by 2.1 dB, for speech maskers ($p < 0.001$). In essence, this is very similar to what was observed in experiment 3, except that, here, the difference between the two populations reached significance for buzzes overall. But once again, the musician advantage effect was more easily observable in speech-on-speech masking conditions than in speech-on-buzz masking conditions.

Roving the masker $F0$ across trials elevated SRTs slightly, but this factor interacted with masker type: it elevated SRTs by 0.7 dB for buzzes ($p = 0.007$), while it did not have any effect for the 2-voice maskers ($p = 0.835$).

As expected, there was an effect of Δ$F0$. The mean SRT at 0-semitones Δ$F0$ (black squares) was higher than the mean SRT at −8, +8, or ±8 semitones ($p < 0.001$), i.e., a masking release was observed whenever a Δ$F0$ was available. This benefit depended on the masker type since it was 2.6 dB larger (on average) with speech maskers than with buzzes, a difference that did not occur in experiment 3. Presumably, this was due to the additional informational component made available since the Δ$F0$ was large enough, here 8 semitones. To illustrate the size of these masking releases, SRTs for the three conditions of Δ$F0$ were subtracted from SRTs at 0-semitones Δ$F0$. A similar ANOVA was conducted to determine the influence of each factor on the Δ$F0$ benefits shown in Fig. 7. For buzzes, a Δ$F0$ of −8 semitones (downward triangles in the left panels) provided a larger benefit than Δ$F0$s

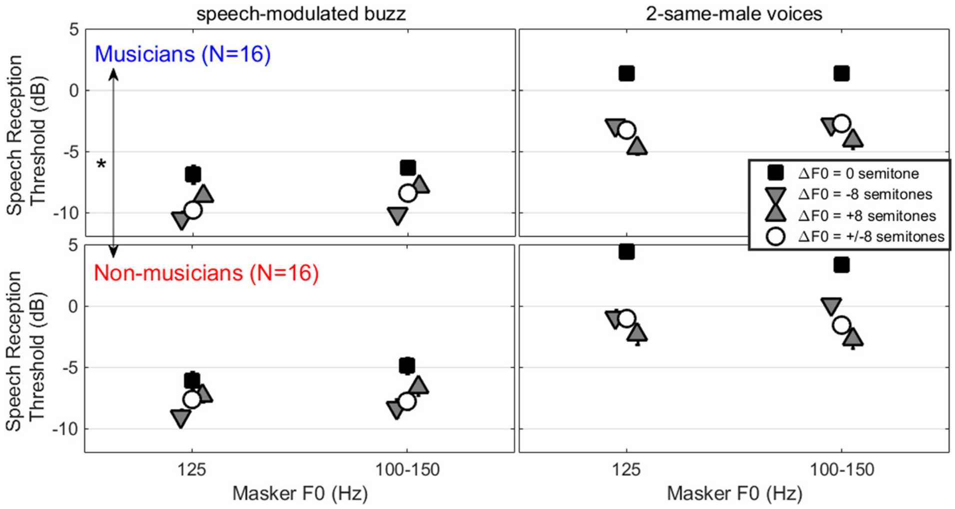

FIG. 6. (Color online) Mean SRTs measured in experiment 4, obtained by musicians (top panels) and non-musicians (bottom panels), for a monotonized target voice against speech-modulated buzzes (left panels) or two interfering voices (right panels). The masker $F0$ was either fixed at 125 Hz or varied randomly across trials between 100 and 150 Hz. Relative to the masker $F0$, the target $F0$ was either identical, 8 semitones above, below, or swapped across trials within a block. Here, the asterisk indicates the musician advantage effect.

of +8 or ±8 semitones ($p < 0.001$ in both cases) while the latter two did not differ ($p = 0.066$). In contrast, for 2-voice maskers, a $\Delta F0$ of +8 semitones (upward triangles in the right panels) provided a larger benefit than $\Delta F0$s of −8 or ±8 semitones ($p < 0.004$) while the latter two did not differ ($p = 0.242$).

C. Discussion

1. The musician advantage effect

As in experiment 3, musicians obtained lower SRTs than non-musicians, and this difference was particularly pronounced in speech-on-speech masking conditions. Our hypothesis was that this could be due to the fact that musicians exploited $\Delta F0$s better than non-musicians, but the present data did not support this interpretation: when examining the $\Delta F0$ benefits (Fig. 7), both groups obtained similar amounts of masking release.

As before, it is useful to examine SRTs for the 0-semitones conditions (black squares of Fig. 6). On average, non-musicians obtained SRTs about −5.5 and +4 dB for buzzes and 2-voice maskers, respectively. In comparison, musicians obtained SRTs of about −6.5 and +1.5 dB for buzzes and 2-voice maskers. In other words, the musician advantage effect already exists in the absence of $\Delta F0$. Statistical analysis of this subset of data found the population difference to be significant ($p = 0.002$). Musicians obtained lower SRTs regardless of their use of $\Delta F0$s.

2. Limited effect of uncertainty

In this experiment, masker $F0$ roving did not have the same effect as it had in experiment 3: it did not interact with $\Delta F0$, meaning that it had no influence on the masking releases shown in Fig. 7. Instead it interacted with masker type, having overall a detrimental influence for buzzes but not for speech maskers. It is possible that uncertainty about the masker $F0$ could have been resolved by the listener's directed attention, provided that 8 semitones had represented a more salient cue than in experiment 3. But given that this effect was weak and inconsistent between experiments 3 and 4, it was presumably not very meaningful. Also, there was no more evidence for an impairment caused by the random swapping of the target $F0$ above and below the masker $F0$ (±8 semitones). The only puzzling fact was that the size of the masking releases was overall low compared to those obtained in experiments 1 or 2: 2–4 dB for buzzes, and 4–6 dB for 2-voice maskers (Fig. 7). It remains unclear whether this is due to a slightly lower masker $F0$ (125 Hz as

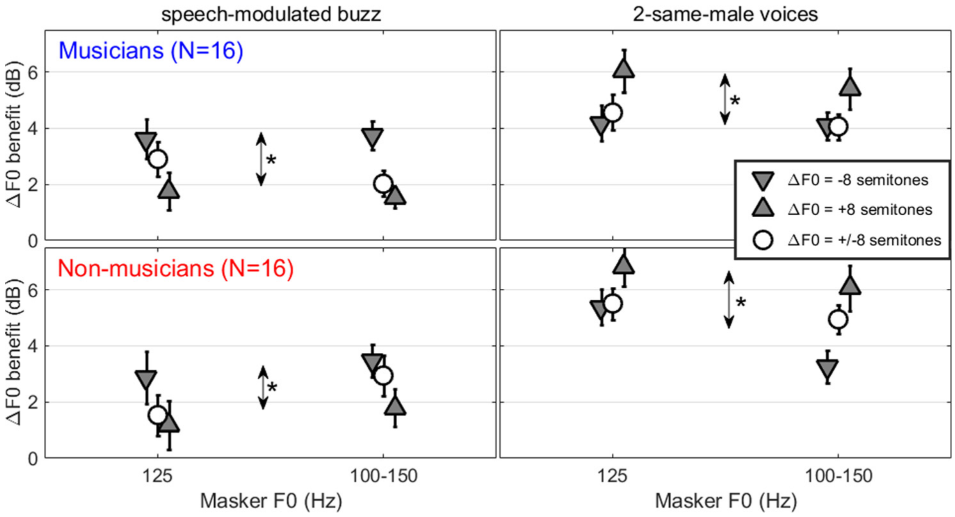

FIG. 7. (Color online) Mean $\Delta F0$ benefits (extracted from the difference in SRTs relative to the 0-semitones baseline) measured in experiment 4. Roving the masker $F0$ across trials or swapping the target $F0$ above or below the masker $F0$ had no effect. Here, asterisks indicate an opposite effect of the sign of $\Delta F0$ depending on masker type.

opposed to 150 Hz) leaving fewer opportunities for spectral glimpsing (discussed in Sec. VI C 3), or whether this could be due to uncertainty effects that would have to some degree generalized across conditions, reducing all masking releases.

3. Spectral glimpsing and informational masking release

As in experiment 3, it is likely that the differences in masking release for -8 semitones (downward triangles in the left panels of Fig. 7) and $+8$ semitones (upward triangles in the left panels of Fig. 7) against buzzes are due to spectral glimpsing effects. When the target $F0$ is low, listeners may be able to glimpse important cues in between resolved partials of the masker. When the target $F0$ is high, its energy may not ideally fall within the masker spectral dips; consequently spectral glimpsing may be of little use in this case. What is interesting is that while the same aspects of energetic masking applied to speech maskers—and that spectral glimpsing should have played a minimal role for a $\Delta F0$ of $+8$ semitones—this condition actually led to the largest benefits (upward triangles in the right panels of Fig. 7). One interpretation is that a large part of the $\Delta F0$ benefit observed against interfering voices is of informational nature, and it is perhaps easier to track the higher-pitch voice in a crowd. Note that Assmann (1999) observed this effect for a situation using two voices (i.e., 1 masking talker), although it was not the case across all sizes of $\Delta F0$s tested in his study.

VII. GENERAL DISCUSSION

A. The musician advantage effect

This study examined the hypothesis that the musician advantage effect was due to a better ability to segregate voices on the basis of $\Delta F0$s. In experiments 1 and 2, musicians did not significantly outperform non-musicians but they did in experiments 3 and 4. Considering that the group of musicians largely overlapped across the four experiments, and that the method and stimuli were similar, this lack of replicability highlights that the musician advantage effect is not a robust phenomenon, consistent with the literature on this topic. However, in both experiments 3 and 4, the musician advantage effect was larger in the presence of 2-same-male voices than in the presence of speech-modulated buzzes. This interaction between musicianship and the nature of masking (energetic versus informational) is consistent with the idea that the phenomenon is more likely to be observed in situations that involve a lot of informational masking (Oxenham *et al.*, 2003). Furthermore, the $\Delta F0$ benefits obtained by musicians were never significantly larger than those obtained by non-musicians, and this was true in all four experiments and whether the maskers were speech-modulated buzzes or 2-same-male voices. Therefore, overall, the present results did not support the proposed hypothesis.

As in any study comparing a control population with a population of interest, there is the possibility that some subjects in the population of interest did not possess enough of the traits of interest. In our study, this translates into whether the subjects selected to be part of the musicians group were sufficiently "good" musicians. Unfortunately, we did not measure $F0$ discrimination thresholds, and thus we cannot be sure that the musicians involved in this study had lower $F0$ thresholds than the non-musicians. This was only expected on the basis of the literature (see Sec. I A). Other studies did provide this information in combination with performance in speech intelligibility tasks. For example, Boebinger *et al.* (2015) confirmed that musicians had lower thresholds for pure tone discrimination (at 1 kHz) than non-musicians, but failed to show a musician advantage in SRT against four different types of masker. Similarly, Ruggles *et al.* (2014) confirmed that musicians had lower $F0$ thresholds than non-musicians (8 versus 30 cents) but failed to observe a musician advantage in SRT. Furthermore, they did not observe any correlation between $F0$ thresholds and clinical measures of speech-in-noise (QuickSIN and Adaptive HINT). Therefore, the idea that pitch sensitivity leads to better intelligibility of masked speech should at the very least be regarded with caution. But more to the point, it is somewhat dubious to judge musicianship on the basis of pitch sensitivity. As shown in experiments 3 and 4, the musician advantage effect can be observed even in the absence of $\Delta F0$s. So there must be other explanations for this phenomenon. Differences in high-level processes are obvious candidates, either in the form of better executive functions (Bialystok and DePape, 2009; Zuk *et al.*, 2014), better auditory attention (Strait *et al.*, 2010), larger working memory capacity (Besson *et al.*, 2011), or better cognitive abilities since Boebinger *et al.* (2015) found that non-verbal IQ predicted SRT in noise whereas musicianship did not.

One important aspect to bear in mind is that the fine sensitivity to pitch, exhibited by highly trained musicians, concerns very subtle $\Delta F0$s. For example, Micheyl *et al.* (2006) showed that piano players reached $F0$ thresholds down to 1.9 cents whereas players of strings and winds reached thresholds down to 1.5 cents. Although these differences supported the hypothesis that self-tuning one's instrument sharpens pitch sensitivity, they are incredibly small differences for a factor that is learned over decades of life. As for non-musicians (excluding amusics), at worst they might obtain thresholds just over 1 semitone (7% of $F0$ in Fig. 2 of Micheyl *et al.*, 2006). So, it is possible that our choice of 2-semitones $\Delta F0$ might not have been small enough to tax musicians' potential advantage. Our choice was simply guided by the scale of $\Delta F0$s that has been used in the literature on the cocktail-party situation. For double-vowels, very small $\Delta F0$s have been used, but there is some debate as to whether those effects are genuinely related to $F0$-segregation or perhaps compounded by waveform interactions. Indeed, Assmann and Summerfield (1990, 1994) and Culling and Darwin (1994) had shown that the beating between unresolved partials could result in specific segments where the spectral envelope of one vowel was more easily detectable. de Cheveigné (1999) later disputed this view by showing that the size of the benefits due to waveform interactions was (on average over vowel pairs) relatively small compared to the benefits of just 7 cents $\Delta F0$. In any case, for running speech, the spectral envelope is constantly changing over time, so listeners have little opportunity to use these waveform interactions to glimpse a particular vowel. Consequently, $\Delta F0$ effects have rarely been tested below 1 or 2 semitones for speech (Brokx and Nooteboom, 1982; Bird and Darwin, 1998; Darwin *et al.*, 2003). So once again, it seems possible that musicians could make a better use than non-musicians of very small $\Delta F0$s, but one might question the generalization of such an advantage to realistic situations. As demonstrated by Leclère *et al.* (2017), most of the benefit in SRT is obtained by instantaneous $\Delta F0$s (even with a mean $\Delta F0$ of 0 semitones). Those instantaneous $\Delta F0$s result from the different intonation patterns of the competing harmonic sources, and are as large as several semitones even with same-gender speakers. In other words, it may not matter to have an advantage over a scale of cents because in realistic situations $\Delta F0$s exist within a scale of semitones, and are therefore accessible to anyone, musicians or not.

Another point that is worth mentioning is that non-musicians are not devoid of training in the task of understanding a voice in a cocktail-party. If musicians could acquire benefits from parsing concurrently presented instruments, why would non-musicians not have acquired benefits from parsing concurrently presented voices? This is a somewhat provocative statement since musicians have been estimated to spend as much as 10 000 h by age 21 practicing their instrument (Ericsson *et al.*, 1993). It is hard to estimate how many hours anyone by age 21 would have spent in noisy/multitalker environments but at the very least, one should acknowledge that situations requiring humans to recognize speech in noisy backgrounds (whatever the noise is) are extremely common. Sustaining conversations in adverse signal-to-noise ratios could potentially act as a form of auditory training, which may perhaps be reflected by differences between children and adults in those exact situations (for example, illustrated by the incredible amount of informational masking induced by contralateral distractors for children, observed in Wightman *et al.*, 2010). In other words, normal-hearing adults are perhaps all experts at $F0$-segregation of voices, but musicianship might help with higher-level processes engaged in those situations accounting for the musician advantage phenomenon in some of them.

B. Predictability and uncertainty

Besides musicianship, the present study also examined a number of aspects related to the use of $\Delta F0$s. In experiment 1, $\Delta F0$s were shown to provide a benefit even in the complete absence of energetic masking. In some sense, one could say that $F0$ acted as a purely informational cue. Since informational masking is known to vary considerably across listeners (Neff and Dethlefs, 1995; Oh and Lufti, 1998; Lufti *et al.*, 2003) and between musicians and non-musicians (Oxenham *et al.*, 2003), this result encouraged us to test whether it was possible to experimentally modulate the strength of this informational component by providing predictable or unpredictable conditions in which those $\Delta F0$s could be used. There are several examples of such effects for speech intelligibility.

Collin and Lavandier (2013) wondered whether listeners could anticipate the locations of temporal dips in a speech-modulated (envelope of 1 voice) speech-shaped noise to glimpse some information about the target voice, a phenomenon often referred to as "listening-in-the-dips." They compared constant maskers (i.e., same modulated noise within a block) to freshly generated maskers (which changed from trial to trial) and found about a 1.5-dB larger benefit in the case of constant maskers. This benefit is directly caused by the predictability of dips in a masker which help listeners make a stronger use of them. Freyman *et al.* (2004) primed listeners with the target sentence except the last word and observed a release from speech-on-speech masking on the identification of the last word (which was not primed). Thus, experience or knowledge of the target voice can facilitate its intelligibility. Johnsrude *et al.* (2013) used the coordinate response measure and asked listeners to report coordinates from strangers' voices, either masked by other strangers' voices or masked by their spouse's voice. Listeners were better at ignoring their spouse, suggesting that experience or knowledge of the masking voice (which is supposedly outside of the focus of attention) can also facilitate intelligibility of the target voice. Allen *et al.* (2011) observed a release from speech-on-speech masking when presenting the masking voice at a predictable rather than unpredictable spatial location. As a whole, this body of data confirms that consistency of certain characteristics of voices can help speech recognition.

From this standpoint, it seemed probable that $F0$ cue usage in the present tasks would have provided just another example where predictability would have enhanced the masking release and uncertainty would have hindered it, but this was not the case. First, experiment 2 failed to maximize the use of $\Delta F0$s, as in fact priming listeners to the pitch of the target voice created a reduction in masking release which only occurred against speech maskers and only in the presence of $\Delta F0$s, but for both musicians and non-musicians. Second, experiments 3 and 4 did not show much impairment (<1 dB) caused by uncertainty about the competing $F0$s, regardless of masker type, regardless of whether the $\Delta F0$ was 2 or 8 semitones, and regardless of musicianship. As a conclusion, it is not trivial from an experimenter's perspective to modulate the strength of $\Delta F0$ effects, and this seems a surprising trait for a cue that can be purely informational.

ACKNOWLEDGMENTS

This research was supported by a NSERC grant awarded to V.L.G., and NIH Grant Nos. R01 DC004786, R01 DC004786-08S1, and R21 DC011905 to M.C. We are grateful to all participants for their time and effort.

APPENDIX

The method used in this study for SRT measurement relies on self-scoring: in each trial, subjects typed the words they could identify, and then pressed enter which made the correct response appear on the screen, with five keywords written in capitals. They were instructed to compare their transcript to the actual answer and count the number of correct keywords and report it (between 0 and 5). This report drove the increase or decrease in TMR presented on the subsequent trial. Thus, the self-scoring had the potential to bias the measurement. We wanted to verify the accuracy of those reports in order to examine (1) whether it depended on the experimental condition (i.e., the within-subjects factors), and (2) whether it differed between musicians and non-musicians. To this aim, we screened all log files that contained the transcripts along with the correct answer and the respective number of words reported, and spotted the number of mistakes made by subjects in counting, disregarding obvious typos, verb tense, singular/plural errors, and words with the same root (e.g., friend/friendly).

Across all four experiments, the average number of errors was 0.57 per block and per subject. In other words, subjects miscounted about once every 17 sentences. The distribution of these errors depended on the number of words reported: 7.8% at 0 words, 14.9% at 1 word, 20.3% at 2 words, 22.7% at 3 words, 22.1% at 4 words, and 12.2% at 5 words. This distribution can be better appreciated when bearing in mind that the adaptive staircase attempts to target a level of 50% performance. As the staircase progresses, more trials are presented when performance is around 2 or 3 words correct than where performance is at 0 or 5 words correct. Consequently, errors are themselves more concentrated around this mid-level of performance, being slightly skewed toward a higher number of words because subjects were more often overestimating the number of words (57.6% of errors) than underestimating it (42.4%). Taking a closer look at the size of these errors: 54.0% of errors were by +1 (i.e., subjects reporting one more word than they actually got); 39.5% of errors were by −1 (i.e., subjects reporting one less word than they actually got); 3.0% of errors were by +2; 1.3% of errors were by −2; and the remaining 2.2% of errors were by ±3, 4, or 5 and appeared to be operational mistakes (subjects typing enter to move to the next sentence, which by default used the number of words which had been recorded on the previous sentence). Therefore, in a very large majority (93.5%) of cases, subjects simply miscounted plus/minus one word, and in the remaining 6.5%, there was no obvious intention or sign that subjects consistently attempted to overestimate their performance. So, we conclude that those were "honest" mistakes.

More importantly, out of all these errors, most of them had no impact on the staircases, because the adaptive method gave the same output for 0, 1, or 2 words reported (i.e., the same increase in TMR) and 3, 4, or 5 words reported (i.e., the same decrease in TMR). The only mistakes that would have affected the SRT measurement were the ones crossing the 50% point. Across all experiments, there were 133 cases where subjects reported 3, 4, or 5 words while they should have reported 2 or less, and there were 102 cases where subjects reported 0, 1, or 2 words while they should have reported 3 or more. These more problematic errors represented 25.4% of all errors. Note, as mentioned above, that 25.4% is a little more than 20% and this is because more trials were concentrated around the 50% point of the psychometric functions. Also, the ratio of overestimation/underestimation is 56.6% to 43.4% which is very similar to the ratio observed pooling all errors together. To summarize, there was on average 1 mistake every 69 sentences that had the potential to affect the SRT measurement. Considering that our experiments contained 120 or 160 sentences, this now amounts to only 2 instances of those errors per experiment and per subject. Furthermore, if we consider that underestimation errors eventually canceled out overestimation errors, either within the same block or by averaging across subjects, then there are only 13% of those problematic errors left that really did bias the SRT measurement. This amounts to 1 occurrence every 523 sentences.

In conclusion, we are confident that these errors (in reporting the number of words identified) would have had a negligible influence on the SRTs measured. Nonetheless, we passed the number of errors through the same between/within ANOVA employed for each experiment. There was never a main effect of population [$F(1,22) < 0.1$, $p = 0.928$ in experiment 1; $F(1,22) < 0.1$, $p = 0.787$ in experiment 2; $F(1,30) = 0.2$, $p = 0.658$ in experiment 3; and $F(1,30) = 1.2$, $p = 0.287$ in experiment 4]. Thus, there was no basis to support the idea that musicians would have overestimated their performance. As for the within-subject factors, there was also very little to report. Experiments 2 and 3 revealed no main effects or interactions. There was a main effect of masker type in both experiment 1 and 4 [$F(1,22) = 5.2$, $p = 0.033$ and $F(1,30) = 7.6$, $p = 0.010$] suggesting that subjects miscounted words more often for speech-modulated buzzes than for 2-voice maskers (but this effect was completely absent in experiments 2 and 3). Finally, there were two 3-way interactions that occurred in experiment 4, but they were inconsistent with the patterns of experiment 3, despite having a similar design. As a consequence, those were presumably spurious interactions which were unlikely to have any meaning. In general, errors were tight to the lexical/syntactical structure used among sentences. The counterbalancing (a full rotation between speech material and experimental conditions, across subjects) enabled those errors to be equally allocated across conditions. This is why within-subject factors are unlikely to systematically influence the errors in number of words reported.

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

Assmann, P. F. (**1999**). "Fundamental frequency and the intelligibility of competing voices," in *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, CA (August 1–7), pp. 179–182.

Assmann, P. F., and Summerfield, Q. (**1990**). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. **88**, 680–697.

Assmann, P. F., and Summerfield, Q. (**1994**). "The contribution of waveform interactions to the perception of concurrent vowels," J. Acoust. Soc. Am. **95**, 471–484.

Baskent, D., and Gaudrain, E. (**2016**). "Musician advantage for speech-on-speech perception," J. Acoust. Soc. Am. **139**, EL51–EL56.

Besson, M., Chobert, J., and Marie, C. (**2011**). "Transfer of training between music and speech: Common processing, attention, and memory," Front. Psychol. **2**, 94.

Bialystok, E., and DePape, A. M. (**2009**). "Musical expertise, bilingualism, and executive functioning," J. Exp. Psychol. Hum. Percept. Perform. **35**, 565–574.

Bidelman, G. M., Krishnan, A., and Gandour, J. T. (**2011**). "Enhanced brainstem encoding predicts musicians' perceptual advantages with pitch," Eur. J. Neurosci. **33**, 530–538.

Bilsen, F. A. (**1976**). "Pronounced binaural pitch phenomenon," J. Acoust. Soc. Am. **59**, 467–468.

Binns, C., and Culling, J. F. (**2007**). "The role of fundamental frequency contours in the perception of speech against interfering speech," J. Acoust. Soc. Am. **122**, 1765–1776.

Bird, J., and Darwin, C. J. (**1998**). "Effects of a difference in fundamental frequency in separating two sentences," in *Psychophysical and Physiological Advances in Hearing*, edited by A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis (Whurr, London), pp. 263–269.

Boebinger, D., Evans, S., Rosen, S., Lima, C. F., Manly, T., and Scott, S. K. (**2015**). "Musicians and nonmusicians are equally adept at perceiving masked speech," J. Acoust. Soc. Am. **137**, 378–387.

Boersma, P., and Weenink, D. (**2013**). Praat: Doing phonetics by computer [computer program]. Version 5.3.85, http://www.praat.org/ (Last viewed September 8, 2014).

Brattico, E., Naatanen, R., and Tervaniemi, M. (**2002**). "Context effects on pitch perception in musicians and nonmusicians: Evidence from event-related-potential recordings," Music Percept. **19**, 199–222.

Brattico, E., Pallesen, K. J., Varyagina, O., Bailey, C., Anourova, I., Jarvenpaa, M., Eerola, T., and Tervaniemi, M. (**2009**). "Neural discrimination of nonprototypical chords in music experts and laymen: An MEG study," J. Cogn. Neurosci. **21**, 2230–2244.

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

Carcagno, S., and Plack, C. J. (**2011**). "Subcortical plasticity following perceptual learning in a pitch discrimination task," J. Assoc. Res. Otolaryngol. **12**, 89–100.

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

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

Corrigall, K. A., Schellenberg, E. G., and Misura, N. M. (**2013**). "Music training, cognition, and personality," Front. Psychol. **4**, 222.

Cramer, E. M., and Huggins, W. H. (**1958**). "Creation of pitch through binaural interaction," J. Acoust. Soc. Am. **30**, 413–417.

Culling, J. F., and Colburn, H. S. (**2000**). "Binaural sluggishness in the perception of tone sequences and speech in noise," J. Acoust. Soc. Am. **107**, 517–527.

Culling, J. F., and Darwin, C. J. (**1993**). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by f0," J. Acoust. Soc. Am. **93**, 3454–3467.

Culling, J. F., and Darwin, C. J. (**1994**). "Perceptual and computational separation of simultaneous vowels: Cues arising from low frequency beating," J. Acoust. Soc. Am. **95**, 1559–1569.

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

de Cheveigné, A. (**1993**). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am. **93**, 3271–3290.

de Cheveigné, A. (**1999**). "Waveform interactions and the segregation of concurrent vowels," J. Acoust. Soc. Am. **106**, 2959–2972.

de Cheveigné, A., Kawahara, H., Tsuzaki, M., and Aikawa, K. (**1997a**). "Concurrent vowel segregation. I. Effects of relative amplitude and F0 difference," J. Acoust. Soc. Am. **101**, 2839–2847.

de Cheveigné, A., McAdams, S., Laroche, J., and Rosenberg, M. (**1995**). "Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement," J. Acoust. Soc. Am. **97**, 3736–3748.

de Cheveigné, A., McAdams, S., and Marin, C. (**1997b**). "Concurrent vowel segregation. II. Effects of phase, harmonicity and task," J. Acoust. Soc. Am. **101**, 2848–2856.

Deroche, M. L. D., and Culling, J. F. (**2013**). "Voice segregation by difference in fundamental frequency: Effect of masker type," J. Acoust. Soc. Am. **134**, EL465–EL470.

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., Lavandier, M., and Gracco, V. L. (**2017**). "Reverberation limits the release from informational masking obtained in the harmonic and binaural domains," Attn., Percept., Psychophys. **79**, 363–379.

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.

Ericsson, K. A., Krampe, R. T., and Tesch-Römer, C. (**1993**). "The role of deliberate practice in the acquisition of expert performance," Psychol. Rev. **100**, 363–406.

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

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

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

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.

Kishon-Rabin, L., Amir, O., Vexler, Y., and Zaltz, Y. (**2001**). "Pitch discrimination: Are professional musicians better than non-musicians?," J. Basic Clin. Physiol. Pharmacol. **12**, 125–144.

Koelsch, S., Schroger, E., and Tervaniemi, M. (**1999**). "Superior preattentive auditory processing in musicians," Neuroreport **10**, 1309–1313.

Leclère, T., Lavandier, M., and Deroche, M. L. D. (**2017**). "The intelligibility of speech in a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location," Hear. Res. **350**, 1–10.

Lutfi, R. A., Kistler, D. J., Oh, E. L., Wightman, F. L., and Callahan, M. R. (**2003**). "One factor underlies individual differences in auditory informational masking within and across age groups," Percept. Psychophys. **65**, 396–406.

Micheyl, C., Delhommeau, K., Perrot, X., and Oxenham, A. J. (**2006**). "Influence of musical and psychoacoustical training on pitch discrimination," Hear. Res. **219**, 36–47.

Miller, S. E., Schlauch, R. S., and Watson P. J. (**2010**). "The effects of fundamental frequency contour manipulations on speech intelligibility in background noise," J. Acoust. Soc. Am. **128**, 435–443.

Musacchia, G., Sams, M., Skoe, E., and Kraus, N. (**2007**). "Musicians have enhanced subcortical auditory and audiovisual processing of speech and music," Proc. Natl. Acad. Sci. U.S.A. **104**, 15894–15898.

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

Oh, E. L., and Lufti, R. A. (**1998**). "Nonmonotonicity of informational masking," J. Acoust. Soc. Am. **104**, 3489–3499.

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

Pantev, C., Elbert, T., Ross, B., Eulitz, C., and Terhardt, E. (**1996**). "Binaural fusion and the representation of virtual pitch in the human auditory cortex," Hear. Res. **100**, 164–170.

Parbery-Clark, A., Skoe, E., and Kraus, N. (**2009a**). "Musical experience limits the degradative effects of background noise on the neural processing of sound," J. Neurosci. **29**, 14100–14107.

Parbery-Clark, A., Skoe, E., Lam, C., and Kraus, N. (**2009b**). "Musician enhancement for speech-in-noise," Ear. Hear. **30**, 653–661.

Plomp, R., and Mimpen, A. M. (**1979**). "Improving the reliability of testing the speech-reception threshold for sentences," Audiology **18**, 43–52.

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.

Ruggles, D. R., Freyman, R. L., and Oxenham, A. J. (**2014**). "Influence of musical training on understanding voiced and whispered speech in noise," PLoS One **9**, e86980.

Shaw, W. A., Newman, E. B., and Hirsh, I. J. (**1947**). "The difference between monaural and binaural thresholds," J. Exp. Psychol. **37**, 229–242.

Spiegel, M. F., and Watson, C. S. (**1984**). "Performance on frequency-discrimination tasks by musicians and nonmusicians," J. Acoust. Soc. Am. **76**, 1690–1695.

Strait, D. L., Kraus, N., Parbery-Clark, A., and Ashley, R. (**2010**). "Musical experience shapes top-down auditory mechanisms: Evidence from masking and auditory attention performance," Hear. Res. **261**, 22–29.

Wightman, F. L., Kistler, D. J., and O'Bryan, A. (**2010**). "Individual differences and age effects in a dichotic informational masking paradigm," J. Acoust. Soc. Am. **128**, 270–279.

Zendel, B. R., and Alain, C. (**2012**). "Musicians experience less age-related decline in central auditory processing," Psychol. Aging **27**, 410–417.

Zuk, J., Benjamin, C., Kenyon, A., and Gaab, N. (**2014**). "Behavioral and neural correlates of executive functioning in musicians and non-musicians," PLoS One **9**(6), e99868.