Title

Roles of the target and masker fundamental frequencies in voice segregation

Journal
The Journal of the Acoustical Society of America, 136(3)

ISSN
0001-4966

Authors
Deroche, Mickael LD
Culling, John F
Chatterjee, Monita
et al.

Publication Date
2014-09-01

DOI
10.1121/1.4890649

Peer reviewed

Roles of the target and masker fundamental frequencies in voice segregation

Mickael L. D. Deroche[a)]
Department of Otolaryngology, Johns Hopkins University School of Medicine, 818 Ross Research Building, 720 Rutland Avenue, Baltimore, Maryland 21205

John F. Culling
School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT, United Kingdom

Monita Chatterjee
Auditory Prostheses and Perception Laboratory, Boys Town National Research Hospital, 555 N 30th Street, Omaha, Nebraska 68131

Charles J. Limb
Department of Otolaryngology, Johns Hopkins University School of Medicine, 818 Ross Research Building, 720 Rutland Avenue, Baltimore, Maryland 21205

(Received 7 January 2014; revised 4 July 2014; accepted 9 July 2014)

Intelligibility of a target voice improves when its fundamental frequency (F0) differs from that of a masking voice, but it remains unclear how this masking release (MR) depends on the two relative F0s. Three experiments measured speech reception thresholds (SRTs) for a target voice against different maskers. Experiment 1 evaluated the influence of target F0 itself. SRTs against white noise were elevated by at least 2 dB for a monotonized target voice compared with the unprocessed voice, but SRTs differed little for F0s between 50 and 150 Hz. In experiments 2 and 3, a MR occurred when there was a steady difference in F0 between the target voice and a stationary speech-shaped harmonic complex or a babble. However, this MR was considerably larger when the F0 of the masker was 11 semitones above the target F0 than when it was 11 semitones below. In contrast, for a fixed masker F0, the MR was similar whether the target F0 was above or below. The dependency of these MRs on the masker F0 suggests that a spectral mechanism such as glimpsing in between resolved masker partials may account for an important part of this phenomenon. © *2014 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4890649]

PACS number(s): 43.66.Dc, 43.71.Es, 43.71.Gv, 43.66.Hg [ELP] Pages: 1225–1236

I. INTRODUCTION

Against a competing voice, a target voice is better understood when spoken at a different fundamental frequency (F0) than when spoken at the same F0 (Brokx and Nooteboom, 1982; Bird and Darwin, 1998; Assmann, 1999; Drullman and Bronkhorst, 2004). Studying this phenomenon in a controlled manner often requires manipulating the F0 of speech sources. Experimenters can choose to fix the F0 of the masker and change that of the target, or fix the F0 of the target and change that of the masker. This choice has potential consequences for the effect of a difference in fundamental frequency (ΔF0).

The study by Brokx and Nooteboom (1982) was one of the earliest to investigate ΔF0 effects. In the first experiment, target sentences were processed by linear predictive coding analysis and resynthesis to have a fixed F0 of 100, 103, 106, 109, 120, or 200 Hz, and they were masked by competing sentences resynthesized at 100 Hz. In their method, therefore, the masker F0 was fixed and the target F0 was raised. Recognition of the target voice increased, when increasing the ΔF0. However, performance dropped for a ΔF0 of 12 semitones, which was taken as evidence that the two competing utterances could not be segregated when their harmonics coincided. This interpretation is very intuitive given the known phenomenon of perceptual fusion for octave differences (Broadbent and Ladefoged, 1957; Myers *et al.*, 1975) but it only holds provided that the target voice was equally intelligible irrespective of its F0 and this assumption might not be valid. The elevated F0 of the target voice could, itself, have resulted in reduced intelligibility. If so, the target resynthesized at 200 Hz might have been less intelligible, and performance would have dropped for a reason completely unrelated to the ΔF0. For clarity, the term *intrinsic intelligibility* will be used here to refer to the intelligibility of a voice regardless of any masking involved. Brokx and Nooteboom's second experiment examined more realistic differences in F0 range. They asked a male speaker to imitate female utterances (with a high voice pitch at about 220 Hz). When utterances were intentionally spoken on a monotone, listeners did not benefit much from differences in F0 range and the authors argued that it was because the male utterances were intrinsically less intelligible when spoken at a high F0. To put it otherwise, any benefit of ΔF0s would have been counteracted by a reduction in intrinsic intelligibility of the high-F0 target voice. If this argument can be made for sentences spoken deliberately at high F0s, there is no reason it cannot also be made for sentences resynthesized at high F0s.

[a)]Author to whom correspondence should be addressed. Electronic mail: mderoch2@jhmi.edu

The alternative method, fixing the target F0 and varying the masker F0, ensures that the target voice has identical intrinsic intelligibility for all experimental conditions. This is the choice made by Bird and Darwin (1998), who also used a single masking voice. They fixed the target F0 at 140 Hz and varied the masker, using either positive or negative shifts: 0, ±1, ±2, ±4, ±6, and ±8 semitones (in experiment 1). For the most extreme cases, their masker F0 was thus 88.2 or 222.2 Hz. They did not find any effect of the sign of ΔF0, meaning in their method that ΔF0 effects did not depend on the masker F0. This result is quite surprising as large changes in the F0 of a harmonic complex masker can result in very different opportunities for listeners to glimpse information about the target voice in between resolved masker partials. For instance, Deroche *et al*. (2013) measured speech reception thresholds (SRTs) for an unprocessed voice, naturally intonated, in the presence of harmonic complexes at 50-, 100-, and 200-Hz F0. For random-phase harmonic maskers (for which there is little contribution of additional factors such as the non-linear amplification of temporal dips in each masker's fundamental period), SRT decreased by about 3 dB for each doubling of the masker F0. If listeners are capable of glimpsing some target energy in between masker partials, the ΔF0 benefit measured at a fixed target F0 should be larger for a high masker F0 than for a low masker F0, because spectral dips broaden and deepen as F0 increases. On the other hand, Bird and Darwin (1998) found that ΔF0 effects relied strongly on the frequency region below 800 Hz. This result is rather consistent with the spectral glimpsing hypothesis. A similar result was also observed by Culling and Darwin (1993) where the ΔF0 benefit observed for the recognition of vowels was largely explained by a ΔF0 in the first formant region. Thus, the literature provides both consistent and inconsistent evidence for the idea that spectral glimpsing in between masker partials may contribute to ΔF0 effects.

This issue has important implications regarding the underlying mechanisms. There has been a long debate as to whether spectral or temporal mechanisms can best account for ΔF0 effects, which mirrors the "place vs time" debate in pitch perception (e.g., Houtsma and Goldstein, 1972, for a place model and Licklider, 1951, for a place-time model). Place models (such as a harmonic sieve) could easily account for a large dependency on the masker F0, but predicted performance is strongly constrained by resolution of the spectral analysis. Frequency selectivity of the peripheral auditory system as estimated by Glasberg and Moore (1990) seems not to be sufficiently fine for such models to predict the improvements observed in vowel-recognition experiments (Assmann and Summerfield, 1990). Consequently, temporal mechanisms such as auto-correlation of within-channel waveforms (Meddis and Hewitt, 1992) may be more likely to explain ΔF0 effects and although they may in principle be dependent on masker F0 (e.g., a time-domain comb-filter), this dependency has not been observed experimentally (Deroche *et al.*, 2014).

The present study investigated to what extent ΔF0 effects depend on the relative F0s of target and masker. The first experiment examined how the intrinsic intelligibility of the target voice varied with different F0 manipulations, ranging between 50 and 300 Hz. A noise background was used in order to avoid any influence of masker F0. Having found a range of F0s over which the intrinsic intelligibility of the target varied little, ΔF0s were created in experiments 2 and 3, by fixing the masker F0 and varying the target F0 above and below, or by fixing the target F0 and varying the masker F0 above and below. Stationary harmonic complexes with a speech-shaped spectral profile were used as maskers in experiment 2, while 400 simultaneous sentences were used to create babbles in experiment 3.

II. GENERAL EXPERIMENTAL METHOD

A. Listeners

Fourteen listeners took part in experiment 1 and 18 listeners took part in experiments 2 and 3. They were between 20 and 45 years old and were paid for their participation. All listeners had pure tone thresholds less than 15 dB hearing level (HL) at frequencies between 0.25 and 8 kHz and English was their native language. Fourteen listeners completed all three experiments in the same order, within about 2 h, with breaks in between. Four listeners only performed experiments 2 and 3, within about 90 min.

B. Stimuli

A total of 19 blocks of 10 sentences were used for the target stimuli, covering seven, six, and six conditions for experiment 1, 2, and 3 respectively. In addition 20 other sentences were used for two practice blocks occurring at the beginning of the first experimental session. The same listener could thus participate in several experiments since different materials were used in each. All sentences, taken from the Harvard Sentence List (Rothauser *et al.*, 1969), have low predictability. The sentences were manipulated using the Praat PSOLA speech analysis and resynthesis package (Boersma and Weenink, 2013), which estimated the F0 contour for each sentence and resynthesized it on a steady value. The F0 of the manipulated voice varied over a much larger range than used in most studies in the literature, from 50 to 300 Hz. As a consequence, stimuli had to be filtered at the output of Praat to have them all equalized at 65 dB sound pressure level (SPL) while preserving their overall spectral shape (see Appendix A for a complete description). In regions of resolved partials, spectral peaks and dips necessarily varied as a function of F0, but in regions of unresolved partials, the excitation level remained identical regardless of the F0. Thus, energy from the manipulated voice was equally detectable in high frequency regions regardless of the F0, such that differences in excitation level between target and masker originated exclusively from the regions of resolved partials. All target sentences were at most 3-s long.

Each of the three experiments used a different masker: white noise, speech-shaped harmonic complexes, and 400-voice babbles, respectively, which are further detailed in the method section of each experiment.

C. Procedure

All listeners began the study with two practice runs using unprocessed sentences and white noise as a masker. They subsequently took part in experiment 1, 2, and 3 (or only experiment 2 and 3). Within each experiment, effects of order and materials were counterbalanced by presenting the target sentences in the same order while rotating the order of conditions for successive listeners. SRT was measured using a 1-up/1-down adaptive method. In this method, 10 target sentences are presented one after another, each one against the same masker. The target-to-masker ratio (TMR) is initially -32 dB and the first sentence is heard several times, each time with a 4-dB increase in TMR. Listeners attempt to type a transcript of the first sentence when they can hear two or three words. The correct transcript is then displayed on the screen, with five key words written in capitals, and the listener self-marks how many key words were obtained. Subsequent target sentences are presented only once and self-marked in a similar manner; the level of the target speech is decreased by 2 dB if the listener correctly identifies three or more of the five key words or else increased by 2 dB. Measurement of each SRT is taken as the mean TMR of the last eight trials.

D. Equipment

All experiments were performed at the Music Perception Laboratory of Johns Hopkins Hospital, and approved by an Institutional Review Board. Informed consent was obtained for all subjects. A user-interface was displayed on a touch-screen monitor, inside a sound-attenuating audiometric booth. Listeners used a keyboard to type their transcript. Signals were sampled at 44.1 kHz and 16-bit resolution, digitally mixed, D/A converted by a 24-bit Edirol UA-25 sound card and presented diotically over Sennheiser HD 280 headphones.

III. EXPERIMENT 1: F0 MANIPULATIONS

A. Rationale

To investigate ΔF0 effects in the laboratory, one needs to manipulate the harmonic structure of speech sources. It is likely, however, that the F0 manipulation of the target voice itself produces signal distortions that could result in this voice being less intelligible. Such variations in intelligibility would occur regardless of any masking by a competing source and are thus confounds for the masking effects of interest. The aim of the first experiment was to examine how much the intelligibility of a F0-manipulated voice varied as a function of its F0.

B. Method

There were seven experimental conditions. The target voice was either naturally intonated, i.e., unprocessed (with a F0 varying around a mean of 104 Hz), or monotonized with Praat at 50, 100, 150, 200, 250, and 300 Hz. This experiment was not concerned with masking but measurements of speech recognition in quiet might have been at ceiling, so a single broadband Gaussian white noise served as masker. Although maskers had a speech-shaped spectral profile in subsequent experiments, it was likely that listeners would rely more and more on low frequency regions when they were able to glimpse in between the partials of harmonic maskers. White noise, having intense energy in high-frequency regions, was chosen here rather than speech-shaped noise to extract the intrinsic intelligibility of F0-manipulated speech in conditions where performance was driven by relatively low-frequency regions. The masker was 3-s long with 30-ms onset and offset ramps, and presented at 65 dB SPL. Fourteen listeners resulted in two complete rotations of the conditions.

C. Results

Figure 1 presents the mean SRTs, averaged over the 14 listeners. Mauchly's test of sphericity indicated that the assumption of sphericity had been respected [$\chi^2(20) = 26.6$, $p = 0.161$]. An analysis of variance with one within-subject factor revealed a main effect of F0 [$F(6,78) = 15.7$, $p < 0.001$]. Out of all the possible pairwise comparisons, only some were of interest to address three questions. (1) Was the unprocessed voice more intelligible than any of the processed voices? Among all monotonized conditions, the lowest threshold was obtained at 100-Hz F0 and a single paired-samples t-test confirmed that SRT for the intonated condition was lower than for this 100-Hz condition [$t_{13} = -2.2$, $p = 0.048$]. (2) Was there an effect of F0 among monotonized voices? The analysis of variance was recalculated, excluding the intonated condition, and the effect of F0 was still significant [$F(5,65) = 9.9$, $p < 0.001$], and was further interrogated using pairwise comparisons with Bonferroni corrections (i.e., paired-samples t-tests at a significance level of 0.05/15). SRT for the voice monotonized at 300 Hz did not differ from that of the voice monotonized at 250 Hz [$p = 1.00$] or from that of the voice monotonized at 200 Hz [$p = 0.181$], but was higher than any other SRT [$p < 0.003$]. In other words, the monotonized voices became less intelligible as F0 increased and this was at least significant for the extreme value of 300-Hz F0. (3) Over what range of F0s was intelligibility of the monotonized voices roughly stable? This question is related to the probability of a type II error: The point here is to assess how confident we may be about the absence of an F0 effect. A power analysis was performed for

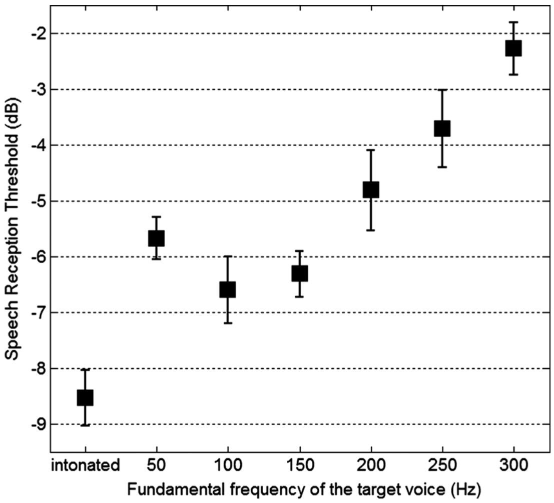

FIG. 1. Mean speech reception threshold measured for a target voice in a white noise background. The target voice was either naturally intonated or monotonized at F0s between 50 and 300 Hz. Error bars are ±1 standard error of the mean over 14 listeners.

each pairwise comparison between 50 and 200 Hz. Comparison between 100 and 150 Hz led to a very small effect size of 0.016 with an observed power of 0.071. Within this 50-Hz range, we can be fairly confident that monotonized voices would be equally intelligible. Comparisons of 100, 150, and 200 Hz relative to 50 Hz led to effect sizes of 0.123, 0.096, 0.148, with observed powers of 0.239, 0.194, 0.286, respectively. These are still small effects but intelligibility might vary a little. Finally, comparisons of 100 and 150 Hz relative to 200 Hz led to effect sizes of 0.249 and 0.267, with observed powers of 0.484, 0.522. In other words, as F0 increases beyond 150 Hz, it becomes more and more dangerous to conclude of an absence of effect, i.e., that intrinsic intelligibility would be stable.

D. Discussion

The most striking result of this experiment is that resynthesizing a voice at a high F0 results in a substantial loss of intelligibility, as evidenced by the fact that SRT for the 300-Hz voice was 4 dB higher than for the 100-Hz voice. F0 manipulations thus produce degradations of the original speech signal that hinder its intrinsic intelligibility. One form of degradation may be that formants are less well defined by the sparsely distributed harmonics when F0 is set to a high value. As a consequence, vowels may be confused. It is also interesting that among the monotonized conditions, SRT was lowest for a F0 of 100 Hz, which is closest to the average 104-Hz F0 of the original recordings. So, it may be that a voice also loses intrinsic intelligibility when its F0 mismatches the natural resonances of the speaker's vocal tract, as it may be for a male speaker with a voice pitch at 300 Hz. To limit the artificial quality of F0-manipulated speech, and thus preserve intrinsic intelligibility, the position of formants should be adjusted according to the F0 (Kawahara *et al.*, 1999). Straight, another speech analysis and resynthesis package, might provide a better way to manipulate speech while limiting degradations in intrinsic intelligibility. However, our preliminary use of Straight to monotonize sentences between 50 and 300 Hz also led to differences in excitation level in unresolved regions (as it was the case for Praat), because the reduction in spectral density is not sufficiently compensated by increases in the amplitude of resolved partials. As far as this design was concerned, a particular care would also have to be paid to equalize masking in high-frequency regions (as described in Appendix A).

The unprocessed intonated voice was more intelligible, by about 2 dB, even when compared with the voice monotonized at the most reasonable value of 100 Hz. Thus, it appears that a voice with a flat F0 contour is always less intelligible than a voice with a natural F0 track. Natural variations of F0 are an essential component of prosody and it has been well documented that prosody helps speech intelligibility (Collier and Hart, 1975; Darwin 1975; Cutler *et al.*, 1997). Although intensity and duration cues convey some prosodic cues, F0 variations are also important, especially at rates of 2–4 Hz suggesting that accents at the syllable rate are critical (Binns and Culling, 2007). Furthermore, speech manipulated with an inverted F0 contour is even less intelligible than monotonized speech (Binns and Culling, 2007), so the poorer intelligibility of monotonized voices here could not only be due to an absence of F0 variation reducing the overall contribution of prosody, but also to unnatural intonation (here flat) mismatching other prosodic cues. Although keeping the variations of F0 as intact as possible would be desirable for purposes of intrinsic intelligibility, it is not desirable for investigations of ΔF0 effects because mechanisms such as glimpsing in between masker partials or periodicity-based mechanisms may be disrupted by these F0 variations (see Sec. VI D).

The main objective of this experiment was to find a range of F0s over which intelligibility of a monotonized voice would not vary greatly. The results of the power analysis showed that the F0 (for this male talker) should vary within a range narrower than 50 Hz between 100 and 150 Hz to really ensure that intelligibility would be identical. This "equal-intelligibility" criterion was too stringent to examine masking-related effects of target F0 and masker F0 in subsequent experiments. A range of F0s between 50 and 178.2 Hz seemed a reasonable compromise. On one hand, this 22-semitones range enabled the use of large ΔF0s of ±11 semitones in experiments 2 and 3. On the other hand, according to the SRTs observed in Fig. 1, variations in intrinsic intelligibility between these values were expected to be limited to about 1 dB.

IV. EXPERIMENT 2: BUZZ MASKER

A. Rationale

In the presence of harmonic complexes, listeners may be able to glimpse in between masker partials, but there cannot be any opportunity to glimpse when the target and masker share a common F0 throughout. If spectral glimpsing plays a role at all in the observed ΔF0 effects, one should find that the masking release (MR) obtained by a given ΔF0 is larger when the masker F0 is above the target F0 than when it is below by an equal amount.

B. Method

Experiment 2 used harmonic complexes with a speech-shaped spectral profile, referred to as *buzzes*. Three broadband harmonic complexes, 3-s long with 30-ms onset and offset ramps, with partials in random phase, based on F0s of 50, 94.4, and 178.2 Hz, were passed through a linear-phase finite impulse response filter designed to match the excitation pattern of average speech. The average speech was based on a larger corpus of 400 sentences spoken by the same male talker. Like the target stimuli, all buzz maskers were equalized at 65 dB SPL and their excitation level in unresolved regions was the same irrespective of F0 (see Appendix B for a complete description).

There were six experimental conditions. There were two conditions of same-F0, in which the target voice and the buzz masker were either both at 50 Hz, or both at 178.2 Hz. There were four conditions of different-F0: for a target F0 fixed at 94.4 Hz, the masker F0 was 50 Hz (11 semitones below) or 178.2 Hz (11 semitones above); for a masker F0 fixed at 94.4 Hz, the target F0 was 50 or 178.2 Hz. With 18 listeners, there were three complete rotations of the conditions.

C. Results

The left panel of Fig. 2 presents the mean SRTs over the 18 listeners. The effect of F0 concerned the target, the masker, or both, so it was tested via three paired-samples t-tests with Bonferroni corrections. When target and masker shared the same F0 (black squares), it did not matter that F0 was 50 or 178.2 Hz [$t_{17} = -0.8$, $p = 1.00$]. When the masker F0 was fixed at 94.4 Hz (gray triangles), it did not matter whether the target F0 was 50 or 178.2 Hz [$t_{17} = -1.9$, $p = 0.222$]. However, when the target F0 was fixed at 94.4 Hz, SRT was much lower for a masker F0 at 178.2 than at 50 Hz [$t_{17} = 7.2$, $p < 0.001$].

In addition, since the range of 50–178.2 Hz exceeded slightly the confident range for equal intrinsic intelligibility observed in experiment 1, a power analysis was performed on the same-F0 conditions (black squares). The effect size was small, 0.188, with an observed power of 0.118, thus suggesting that intrinsic intelligibility did not vary greatly within the range chosen in experiment 2.

D. Discussion

When the target voice was fixed at an F0 of 94.4 Hz, a F0 which did not coincide with the harmonic structure of the masker, SRT decreased considerably by increasing the masker F0 (white circles). The most likely interpretation for this masker F0 effect is that listeners are able to glimpse important speech cues in regions of spectrally resolved masker partials. In the particular case (which happens only in the laboratory) where the target partials coincide exactly with the masker partials throughout the entire sentence duration, there is nothing to glimpse within the masker spectral dips, irrespective of how deep these dips may be, which is why there is no F0 effect in the same-F0 conditions (and also because intrinsic intelligibility was stable).

The point of varying the target F0 (gray triangles) was to examine whether glimpsing could also depend to some extent on the target F0. One might indeed think that there are better cues to be glimpsed when more target partials accumulate in between masker partials, i.e., at low target F0. The effect of target F0 did however not reach significance, suggesting that glimpsing does not depend strongly on it. To understand this result, it is useful to refer to an energy-detection model (Sec. VI B). This model suggests that the role played by spectral glimpsing is still modest for a masker F0 at 94.4 Hz. As a consequence, it does not make much of a difference whether energy from a low-F0 target is available in auditory filters centered at the masker spectral dips or whether energy from a high-F0 target is available in auditory filters centered at the target spectral peaks.

As a conclusion, the MR provided by a ΔF0 was larger when the F0 of the buzz was 11 semitones above the target than below. For harmonic complexes with F0s in the human voice range, a substantial part of the MR caused by ΔF0s might then be due to the listeners' ability to glimpse in between partials.

V. EXPERIMENT 3: BABBLE MASKER

A. Rationale

In more common situations of conversation, ΔF0 effects take place with masking voices, not stationary buzzes. However, the case of masking voices is very complicated with many additional factors at play. (1) First, masking voices have a glottal excitation. There are short temporal dips within the fundamental period of within-channel temporal envelopes which may allow listeners a better TMR (Summers and Leek, 1998). Whether listeners utilize these extremely short temporal dips depends on several factors, including outer hair cell function, masker level, masker F0, masker spectral profile, and perhaps even the role of the

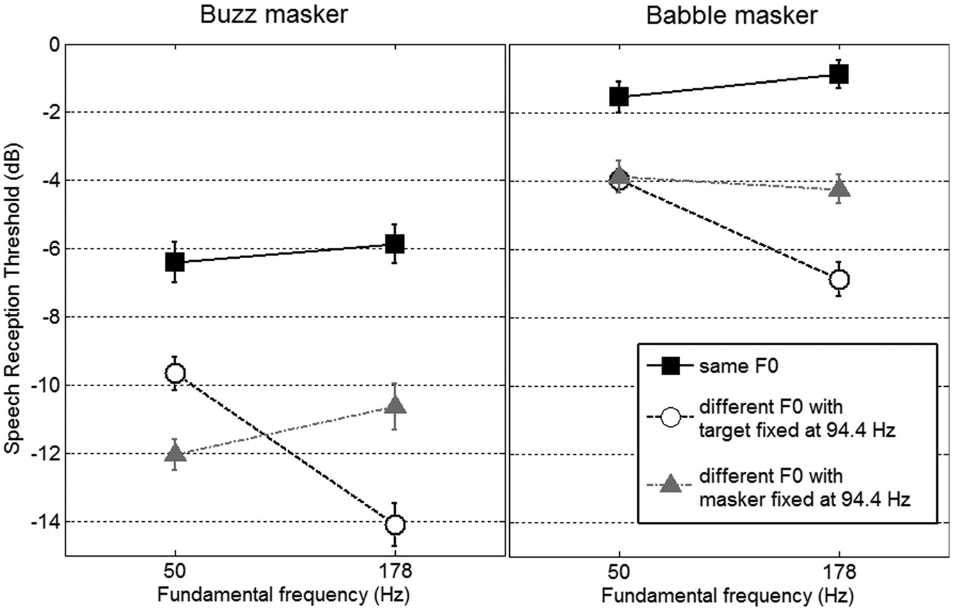

FIG. 2. Mean speech reception threshold measured for a voice against a buzz (left panel) or against a babble (right panel). Target and masker either shared the same F0, at 50 and 178.2 Hz, or had F0s that were ±11 semitones apart in all possible configurations. Error bars are ±1 standard error of the mean over 18 listeners.

media olivocochlear reflex. These factors could easily interfere with ΔF0 effects. (2) Listeners may also "listen in the dips" of the broadband temporal envelope of the masking sentence. Perhaps, the more listeners can glimpse temporally, the less they need to glimpse spectrally. So this ability could potentially interfere with the spectral glimpsing effects. (3) Even when target and masking sentences are filtered in different frequency bands, such that energetic masking is largely absent, large amounts of masking, referred to as informational masking, occur for speech-on-speech configurations that do not occur for speech-on-noise configurations (Kidd *et al.*, 2005). F0 can be used as a cue for sequential segregation (Darwin *et al.*, 2003; Drullman and Bronkhorst, 2004). It may be that the mechanisms of sequential segregation dominate over the mechanisms of simultaneous segregation when the task involves a large amount of informational masking. (4) The slow modulations of speech (less than 10 Hz) are essential to articulation (Houtgast and Steeneken, 1985). Although "listening in the dips" can generally provide MR in temporally fluctuating maskers, these slow temporal modulations would also mask those of the target voice and this would be very detrimental to target intelligibility. Competing voices would therefore produce a substantial amount of modulation masking (Dau *et al.*, 1999). It is currently unclear how this factor could interfere with ΔF0 effects, but it is certainly not impossible. The aim of this last experiment, which was more generally the aim of the entire study, was to approach the case of masking voices while limiting as much as possible the role of these additional factors to see whether spectral glimpsing effects such as those observed with buzzes could occur with maskers made of real speech material. Babbles were constructed from the entire material of 400 sentences, in which each sentence was monotonized at the same F0. The temporal overlap between the 400 sentences was controlled to produce a pseudo-stationary masker with shallow modulation of within-channel envelopes at the F0 rate, and informational masking was expected to be largely absent since each of the 400 concurrent sentences was 26 dB lower than the target sentence (for the final babble to be at 65 dB SPL). The resulting babbles had a buzz-like quality with the exception of sibilant high-frequency sounds produced by the addition of aperiodic cues in the 400 sentences.

B. Method

Babbles were made of 400 concurrent sentences, the same ones that served for the buzz creation. For each of the three F0s, 50, 94.4, and 178.2 Hz, each sentence was monotonized using the Praat package and filtered in exactly the same way as the target sentences. Two more complications arose when adding the 400 monotonized sentences together. First, to monotonize a given sentence, Praat uses a glottal pulse source. This means that all sentences at the output of Praat have a synchronized excitation, irrespective of their temporal envelope. Adding all sentences together would result in the harmonic partials of the babble being roughly in phase despite each sentence having different onsets; consequently, there would be some temporal dips in the strongly modulated within-channel envelopes of such babble. For a F0 of 50 Hz, it is likely that listeners can listen in these 20-ms temporal dips (see, for instance, Deroche *et al.*, 2013). In real life however, simultaneous talkers do not have synchronized excitation, so these phase effects were undesirable. To eliminate them, a random delay was applied to every sentence when adding the 400 sentences together, but the same set of delays was applied for the three F0s. The delay could be as long as the difference in duration between any two sentences. The resulting babble displayed no obvious modulation at the F0, nor on a longer timescale, caused, for instance, by the first syllable always occurring 200–300 ms after onset. Second, adding the 400 sentences together resulted in constructive interference in some spectral regions and destructive interference in others. Each of the three babbles was therefore filtered to have the excitation pattern of the average over the 400 monotonized sentences. This final filtering produced excitation patterns that were essentially identical to those of the buzz at the same F0. The design of experiment 3 was otherwise identical to that of experiment 2. Eighteen listeners resulted again in three complete rotations of the conditions.

C. Results

The right panel of Fig. 2 presents the mean SRTs over the 18 listeners. The effect of F0 concerned again the target, the masker, or both, and was tested via three paired-samples t-tests with Bonferroni corrections. When target and masker shared the same F0, SRT was a little higher at 178.2 than at 50 Hz [$t_{17} = -3.4$, $p = 0.010$]. When the masker F0 was fixed at 94.4 Hz, it did not matter whether the target F0 was at 50 or 178.2 Hz [$t_{17} = 1.0$, $p = 1.00$]. When the target F0 was fixed at 94.4 Hz, SRT was lower for a masker F0 at 178 than at 50 Hz [$t_{17} = 7.2$, $p < 0.001$].

A power analysis was performed on the same-F0 conditions (black squares) and revealed an effect size of 0.806 with an observed power of 0.897. Contrary to experiment 2, there was an effect of increasing both F0s from 50 to 178.2 Hz.

D. Discussion

When the target voice shared the same harmonic structure as the babble, SRT was a little elevated for a F0 at 178.2 Hz as opposed to 50 Hz (black squares). Since this effect was largely absent in experiment 2 (as evidenced by the power analysis) using the exact same range, it is presumably not related to changes in intrinsic intelligibility, but rather related to the masker type. A particular care had been paid to generate the babbles with the same excitation patterns as buzzes. Furthermore, the random delay applied to each of the 400 sentences constituting the babble considerably reduced its within-channel envelope modulations. Although passing the babbles through a simulation of auditory filters with realistic phase responses did not reveal any obvious temporal dip across many center frequencies, it is difficult to exclude the possibility that some residual modulations had remained when many partials of a low F0 interacted within a filter centered at relatively high frequencies. Portions of lower masker intensity within these residual modulations may have allowed listeners a better TMR, which might perhaps explain why recognition of the target voice was a little better for the babble at 50 Hz than for the babble at 178.2 Hz.

Except for this F0 effect in the same-F0 conditions, the general pattern of results was consistent with experiment 2. SRTs were lower in the presence of the ΔF0; there was no effect of target F0 when the babble F0 was fixed; but there was an effect of the babble F0 when the target F0 was fixed. The MR provided by the ΔF0 was overall more modest with babble than with buzz, but increased substantially with the masker F0 in both cases.

Finally, SRTs were overall much higher in experiment 3 than in experiment 2. A combination of 400 voices, all set at 26 dB lower than the level of the target sentences, is rather unlikely to have produced informational masking (Simpson and Cooke, 2005). There was also no silence or pause to listen to. This, however, does not imply that there was no temporal modulation. The babble contained sibilant high-frequency sounds covering a large range of random modulation rates (in the same way that there is a full range of modulation rates in stationary noise). Modulations at rates around 10 Hz, which the buzz lacked, could have masked the slow modulations of the target temporal envelope, essential to its articulation (Houtgast and Steeneken, 1985). This large elevation of thresholds may therefore be attributed to modulation masking occurring in high frequency channels.

VI. GENERAL DISCUSSION

When investigating ΔF0 effects, the present study showed that the choice of experimental design, namely, fixing the target F0 and varying the masker F0, or vice versa, is an important one to consider. Both target F0 and masker F0 may influence the measured SRTs and these effects are reviewed below.

A. Variations in intrinsic intelligibility

Experiment 1 showed that resynthesizing a voice with an arbitrary F0 contour can result in loss of intrinsic intelligibility. Because formants are less well defined at high F0s or because some unexpectedly low or high F0s mismatch the natural resonances of a speaker's vocal tract, a voice manipulated experimentally can lose intrinsic intelligibility. Although different materials and a different speech synthesis was used in Brokx and Nooteboom (1982), it is possible that the reduction in performance occurring for a ΔF0 of 12 semitones was partly due to the target voice being less intelligible when monotonized at 200 Hz. It would be interesting to reexamine how much performance really drops for octave differences when controlling for intrinsic intelligibility. The use of Straight might alleviate some of the signal distortions that occur when manipulating F0 (Kawahara *et al.*, 1999).

B. Do effects of ΔF0 depend on their sign?

To further examine the influence of masker F0 on the MR provided by the 11-semitones ΔF0, an analysis of variance was performed by subtracting SRTs for the same-F0 condition (black squares) from the different-F0 condition in which the *target* was fixed (white circles), with two within-subject factors (masker F0 × masker type). As illustrated in the left panel of Fig. 3, there was a main effect of masker F0 [F(1,17) = 105.5, $p < 0.001$]. The MR was overall larger at high masker F0 (7.1 dB on average) than at low masker F0 (2.6 dB on average). The main effect of masker type was at the significance level [F(1,17) = 4.4, $p = 0.050$], indicating that the MR tended to be larger with a buzz masker (5.8 dB on average) than with a babble masker (4.0 dB on average). The interaction was not significant [F(1,17) = 1.3, $p = 0.263$].

A similar analysis was conducted for the conditions in which the *masker* was fixed, with two within-subject factors (target F0 × masker type). As illustrated in the right panel of Fig. 3, there was again a main effect of masker type [F(1,17) = 12.9, $p = 0.002$], indicating that the MR was larger with a buzz masker (5.2 dB on average) than with a babble masker (2.8 dB on average), but there was no main

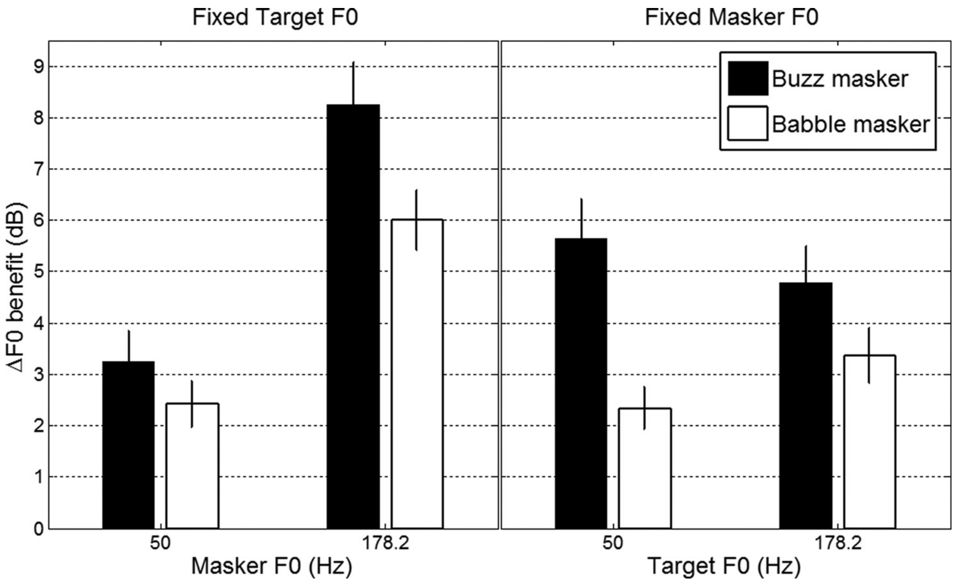

FIG. 3. Masking releases obtained from the 11-semitones ΔF0, evaluated as a function of masker F0 (left panel) or target F0 (right panel) for the two masker types. Error bars are ±1 standard error of the mean over 18 listeners.

effect of target F0 [F(1,17) = 0.02, $p = 0.893$]. There was also no interaction [F(1,17) = 2.6, $p = 0.126$].

The main results show a strong dependency of ΔF0 effects on the masker F0, but no dependency on the target F0. The interpretation that seems most likely is that listeners are capable of glimpsing speech information in between resolved partials of the masker, since this masker F0 effect disappeared when the harmonic structure of the target matched that of the masker. But if so, why did the target F0 have no effect? A low-F0 target should in principle "fill in" the masker spectral dips more than a high-F0 target. A simple energy-detection model further explored the spectral glimpsing hypothesis and attempted to resolve this apparent paradox. Being purely spectral, this account was examined using excitation patterns, computed from rounded-exponential filters, equally spaced on an ERB-scale with level dependency (Glasberg and Moore, 1990).

Figure 4 shows the excitation patterns of the buzz masker and the target voice averaged over the 400 sentences of the entire material, both at 65 dB SPL, in each of the six experimental conditions. Note that the figures would be essentially the same with babble rather than buzz as their excitation patterns were identical. The absolute thresholds are represented with dashed gray lines, at a hearing level of 0 dB. In each panel, the target energy is detectable in auditory filters where its excitation level is both above absolute threshold and above the excitation level of the masker. It is then possible to integrate this area to provide an estimate of how detectable the target is at this TMR of 0 dB.

An increase in the F0 of a harmonic complex (be it the target or the masker) results both in a reduction in spectral density and in an increase in the level per partial. A key point to bear in mind is that these two factors vary differently as a function of F0: dips deepen a lot more than peaks grow. Using flat-spectrum harmonic complexes, this difference is very striking (Deroche *et al.*, 2014), but the same is true for speech-shaped harmonic complexes, as it can be seen by comparing the two top panels of Fig. 4. Furthermore, the variation in the peaks/dips ratio is not linear as F0 increases. At a F0 of 50 Hz, peaks and dips are equivalent in size. At a F0 of 94.4 Hz, dips are only a little more pronounced than peaks. At a F0 of 178.2 Hz, dips are substantially more pronounced than peaks. This peaks/dips ratio at each F0 is the basis for the differential role of masker F0 and target F0 observed in the present study.

In the same-F0 conditions (top panels), the excitation patterns of target and masker coincide very well (as it was intended through the stimuli generation). Detectability of the target energy is minimal in both cases. In the different-F0 conditions where the target F0 was fixed at 94.4 Hz (middle panels), target energy is barely detectable above the excitation level of the low-F0 masker in auditory filters centered at the target peaks (because the size of spectral peaks has increased a little from 50- to 94.4-Hz F0). In contrast, target energy is largely available in auditory filters centered at the spectral dips of the high-F0 masker (because the size of spectral dips has increased considerably from 50 to 178.2 Hz). In the different-F0 conditions where the masker F0 was fixed at 94.4 Hz (bottom panels), energy of the low-F0 target is primarily detectable at the masker spectral dips whereas energy of the high-F0 target is primarily detectable at the target spectral peaks, but overall the integration area is equal between the two conditions. This result essentially arises because the size of spectral dips differs little from the size of spectral peaks at a masker F0 of 94.4 Hz. Consequently, the potential benefit of spectral glimpsing for low-F0 targets is not sufficient to offset the detectability of intense partials of high-F0 targets.

To provide more compelling evidence from a modeling perspective, a similar figure as Fig. 4 can be generated for different TMRs. Figure 5 shows how the integration area progressively decreases as TMR decreases in each of these six experimental conditions. In the same-F0 conditions (top panels), the integration area reflects the overall shift between the two excitation patterns almost identical in shape. As expected, the integration area decreases in a linear way until it has completely disappeared once the target excitation level

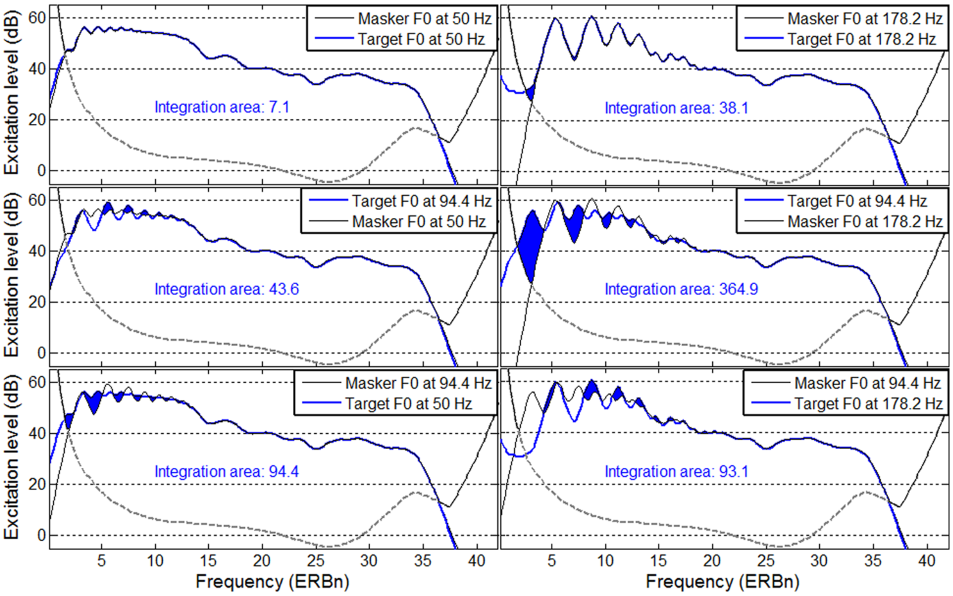

FIG. 4. (Color online) Excitation patterns of target and masker, at a TMR of 0 dB, in each of the six experimental conditions used in experiments 2 and 3. Energy from the target voice is particularly detectable at the spectral dips of the high-F0 masker (middle right panel), accounting for the large effect of masker F0 observed experimentally.

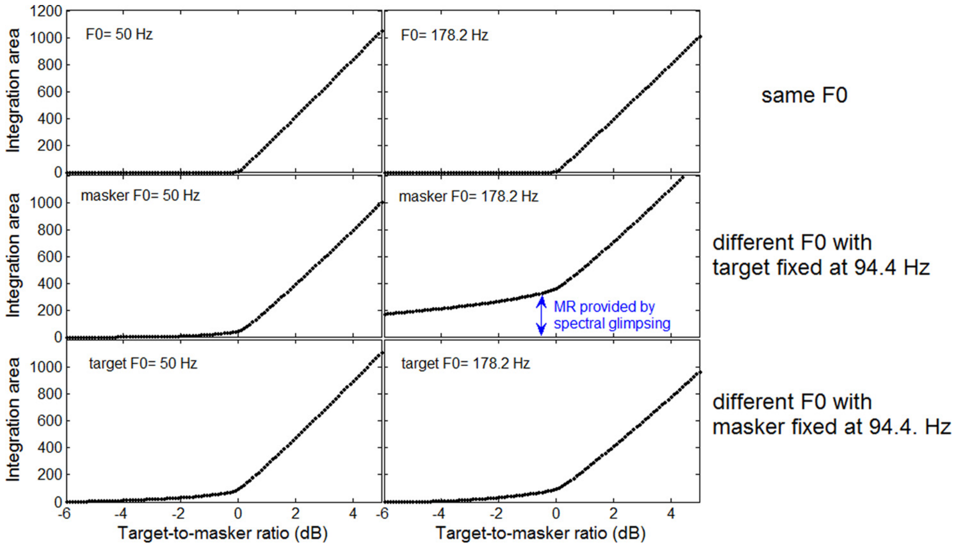

FIG. 5. (Color online) Variation of the integration area, reflecting the detectability of target energy in a given masking situation, as a function of TMR. For a high masker F0 different from that of the target (middle right panel), target energy remains available at spectral dips in between resolved masker partials, enabling the target to be detectable at negative TMR.

is below that of the masker. In the different-F0 conditions where the target is fixed at 94.4 Hz, the pattern of the integration area with TMR for a low-F0 masker (left middle panel) is similar to the same-F0 conditions. This is an important prediction since it suggests that any ΔF0 benefit occurring at such a low masker F0 (such as the 2–3 dB effect shown in the left panel of Fig. 3) would have to be accounted for by another mechanism, e.g. a periodicity-based mechanism. In contrast, the integration area for a high-F0 masker (middle right panel) takes on a distinct, shallower, slope at negative TMR because target energy remains largely available in auditory filters centered close to the masker spectral dips. Finally, for a masker F0 fixed at 94.4 Hz (bottom panels), the two conditions lead to similar patterns in which the integration area is not quite nil but nonetheless very small at negative TMR. This confirms again that, at a masker F0 of 94.4 Hz, spectral dips are not sufficiently larger than spectral peaks to see a substantial glimpsing-related difference between these two conditions.

The present model generally supported the proposed hypothesis that spectral glimpsing plays an important role in ΔF0 effects, accounting for the large dependency on the masker F0. However, the model also cast a slightly different conclusion on the effect of target F0. The reason why no effect of target F0 was observed in the present study might have been due to the fact that these two conditions were tested for a masker F0 at 94.4 Hz, for which spectral glimpsing may not play a considerable role. Had this design shifted upwards by one octave, testing the effect of target F0 between 100 and 356.4 Hz at a fixed masker F0 of 188.8 Hz, it is possible that an effect of target F0 would be observed, with a low-F0 target leading to better SRT than a high-F0 target. Note however that such a F0 range would pose some intrinsic intelligibility issues at least for a male talker. Regardless of the choice of the F0 range, the effect of the masker F0 would still be larger than the effect of target F0 in these designs. Therefore, a safe conclusion is that effects of ΔF0 may generally depend on their sign, being larger for negative ΔF0s (i.e., masker above target) than for positive ΔF0s. This dependency is stronger in designs where the target F0 is fixed than in designs where the masker F0 is fixed, and in both cases would depend on how much spectral glimpsing is involved in the task.

C. ΔF0 effects in more natural situations

It is unclear how much spectral glimpsing is involved in speech-on-speech situations. In the presence of a single masking utterance, Bird and Darwin (1998) did not find any effect of the ΔF0 sign, even though the masker F0 range varied between 88 and 222 Hz. Assmann (1999) also used ΔF0s with both positive and negative signs, because he let listeners report words belonging to either source. With ΔF0s of 0, 1, 2, 6, and 8 semitones, intelligibility was similar for the high-F0 and the low-F0 voice. Thus, with a single masking voice, it does not seem to matter whether the masker F0 is high or low. Those observations radically contrast with the present results (both experimental and modeling). One possible explanation is that listeners obtain some MR by combining spectral glimpsing and temporal glimpsing from the modulations of the temporal envelope of a single masking voice. A large contribution of temporal glimpsing could thus diminish the relative contribution of spectral glimpsing to the MR. An even stronger interaction between these two sources of MR could occur if the modulations of the temporal envelope somehow hindered the listener's ability to glimpse in between partials, in which case listeners might listen in temporal dips or listen in spectral dips, but not both. If that is the case, spectral glimpsing would provide substantial MR for stationary maskers, but little MR for temporally fluctuating maskers. Another possible explanation comes from the fact that masking voices have a glottal excitation, resulting in highly modulated waveforms in unresolved regions. Additional MRs are provided by the non-linear amplification of the basilar membrane, enhancing the detectability of speech energy at temporal dips in the fundamental period of stationary harmonic maskers (Summers and Leek, 1998). Note that this mechanism may be thought of as a form of temporal glimpsing, but in contrast to the first explanation, this mechanism relates to extremely short temporal dips (e.g., 5 or 10 ms), where listeners can clearly not attempt to "listen in the dips." The MRs provided by spectral glimpsing are restricted to the region of resolved partials, whereas the MRs provided by the basilar membrane compression are restricted to the region of unresolved partials. Depending on the magnitude of the MR provided by each source and whether or not listeners can understand speech at 50% intelligibility without requiring information from the other region, one source of MR may diminish the relative contribution of the other.

For naturally intonated target voices, spectral glimpsing may perhaps contribute even more than for monotonized voices. Target partials from a fixed harmonic structure are not often located right at the masker spectral dips; whereas a dynamic F0 contour may fill in spectral dips to their full extent. Glimpsing is presumably more useful when the cues being glimpsed are richer. A dynamic F0 contour would reveal spectral details about formants and formant transitions that may be very useful to glimpse. For a naturally intonated masking voice on the other hand, listeners could face difficulties in glimpsing in spectral dips that are constantly changing over time. Furthermore, masker F0 modulation could fill in the dips given a limited temporal resolution (Sec. VI D).

Finally, periodicity in the masking voices is likely to provide an additional source of MR, but may be restricted to the voiced portions. In this study, the MR was overall smaller for a babble than for a buzz masker, including in the conditions for which spectral glimpsing played no role (at a 50-Hz F0 masker). The most likely interpretation is that aperiodic cues in the 400 masking sentences degraded periodicity in relatively high-frequency channels, which reduced the overall benefit, compared to the buzz whose periodicity was intact across the entire spectrum. Deroche *et al.* (2014) disentangled the role of periodicity from that of spectral dips in harmonic and inharmonic complexes and concluded that both may indeed contribute to the MR but behave as two independent mechanisms. Most relevant to the present study, they found that the MR attributed to masker periodicity did not depend much on masker F0. The dependency of ΔF0 effects on masker F0 observed here may therefore relate exclusively to account based on spectral glimpsing.

D. Effect of F0 modulation and reverberation on the masker

Deroche and Culling (2011) were interested in examining whether speech recognition in the presence of a buzz masker depended entirely on the masker harmonicity, or whether it could also be influenced by the target harmonicity. They disrupted harmonicity by applying a sinusoidal modulation of F0 as well as simulated reverberation to each of the competing sources. F0 modulation on the masker alone resulted in a small elevation of SRT which became much larger in reverberant conditions. These results supported the idea that masker harmonicity was the critical factor. In addition, the results were similar whether the masker partials were in sine or in random phase, and no impairment was observed for a buzz in reverberant conditions as long as its F0 remained steady, while reverberation flattened any strong modulations of the within-channel temporal envelopes. Because their results were not influenced by phase effects, Deroche and Culling argued that F0 segregation relied heavily on low-order partials of the masker and that autocorrelation-based models provided a plausible account for these ΔF0 effects. In the light of the present results, it is likely that the impairments generated by F0 modulation and reverberation on the masker were also partly caused by the filling of the spectral dips in the buzz maskers. Given limited temporal resolution, F0 modulation would fill in the spectral dips in between resolved partials and this effect would be exacerbated by reverberation. Periodicity-based mechanisms may certainly be an important account of ΔF0 effects, but at least for buzz maskers, part of the MR might be attributed to spectral glimpsing, a process that may technically not require harmonicity, but simply dips in a spectral template.

VII. CONCLUSION

Three experiments measured SRTs for a target voice masked by white noise, buzz maskers, and babble maskers, respectively. In experiment 1, F0 manipulations always resulted in some loss of intrinsic intelligibility of the target voice compared with the unprocessed, naturally intonated, voice. These impairments were however worse beyond a F0 of 150 Hz. One should therefore bear in mind that ΔF0 effects can be confounded by effects intrinsic to the experimental manipulations of speech sources. In experiment 2 and 3, using buzzes or babbles as maskers, the benefit of an 11-semitones ΔF0 was similar whether the target F0 was above or below the masker F0 fixed at 94.4 Hz, but larger when the masker F0 was above the target fixed at 94.4 Hz than below. The strong dependency of ΔF0 effects on the masker F0 provides in general more support for a contribution from place models.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health Grant Nos. R01 DC004786, R01 DC004786-08S1, and R21 DC011905 to M.C.

APPENDIX A

The Praat PSOLA package (Boersma and Weenink, 2013) is a useful tool to examine effects of ΔF0 between competing voices, because it lets the user resynthesize a given speech sample with a completely arbitrary F0 contour. In the present study, manipulations focused on the very simple case where F0 is fixed throughout the entire duration, but any arbitrary manipulation is possible. Some F0 manipulations, however, introduce substantial variations in root-mean-square (RMS) level. As illustrated in the left panel of Fig. 6, monotonizing the F0 contour of a sentence in our speech material (spoken by a male voice) at 50 Hz resulted in 4.3-dB decrease in RMS level. In contrast, monotonizing it at 300 Hz resulted in a 1.5-dB increase in RMS level. With a F0 range between 50 and 300 Hz, the RMS level could thus vary by as much as 6 dB. One cannot simply equalize the RMS level by multiplying the waveforms by a factor

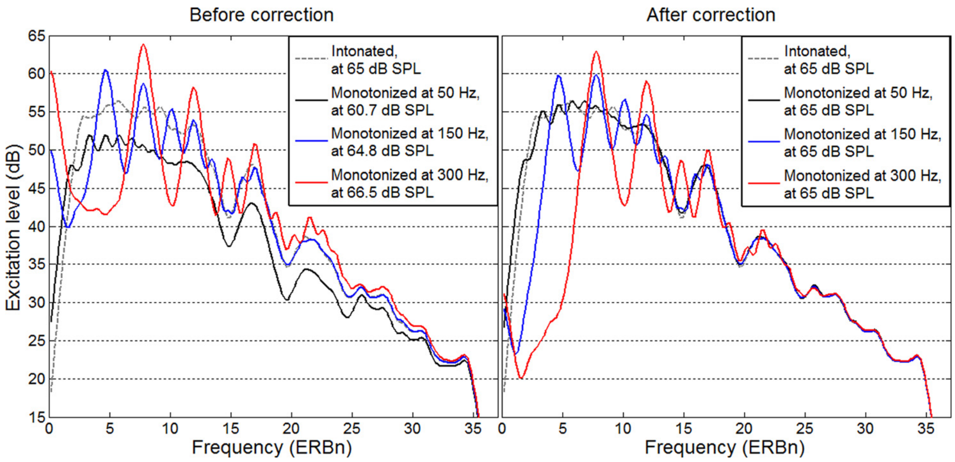

FIG. 6. (Color online) Details of the spectral correction applied to one sentence manipulated by the Praat package to hold a fixed F0 pattern throughout. The filtering ensured that speech stimuli, irrespective of F0, were equalized to a RMS level of 65 dB SPL and their excitation level in regions of unresolved partials was within 0.2 dB.

because this would result in speech stimuli having different excitation levels throughout the spectrum, depending on F0. It is obviously not possible to equate the excitation level in resolved regions because spectral peaks and dips will differ in size and in position as a function of F0, but it is possible to equate the excitation level in unresolved regions. Furthermore, the differences in excitation pattern between F0 manipulations of a given sentence are more complex than a spectral tilt as described in Appendix B. Therefore, a method was adopted to correct for the Praat-induced differences in excitation patterns. The excitation patterns of the monotonized and intonated stimuli were first smoothed in resolved regions, for center frequencies between 0.8 × F0 and the unresolvability cutoff. The smoothing was performed using three passes of rectangular window that shrank as center frequency increased (because the lower the center frequency, the more smoothing was needed, to eliminate the peaks and dips caused by the presence of resolved partials). The smoothed excitation patterns revealed more clearly the spectral regions where partials received too much or too little gain. The monotonized stimulus was then passed through a finite-impulse-response filter with 1024 coefficients, whose frequency response was the difference between the smoothed excitation patterns of the intonated and monotonized stimuli. The right panel of Fig. 6 illustrated this spectral correction for the stimuli monotonized at 50, 150, 300 Hz. Irrespective of the F0 at which stimuli were monotonized, the excitation level in unresolved regions was within 0.2 dB after correction.

APPENDIX B

To create harmonic complexes with a speech-shaped spectral profile (referred to as buzzes), complex signals with equal-amplitude partials are passed through a finite-impulse-response filter designed to match the average long-term excitation pattern of speech. For F0s up to about 150 Hz, the RMS level at the output of such a filter changes little (by less than 0.1 dB). However, for larger F0s, the RMS level is noticeably smaller because the reduction in spectral density is not sufficiently compensated by increases in the amplitude of low-order partials. The left panel of Fig. 7 illustrates that a harmonic complex based on a F0 of 300 Hz has a RMS level reduced by 1.2 dB compared with the level of average speech to which it is spectrally shaped. Equalizing the RMS level at 65 dB SPL by multiplying the signal would shift the excitation pattern upwards and consequently, the excitation level in regions of unresolved partials would change depending on F0. To prevent this effect and have excitation level equated in unresolved regions (in the event that listeners rely

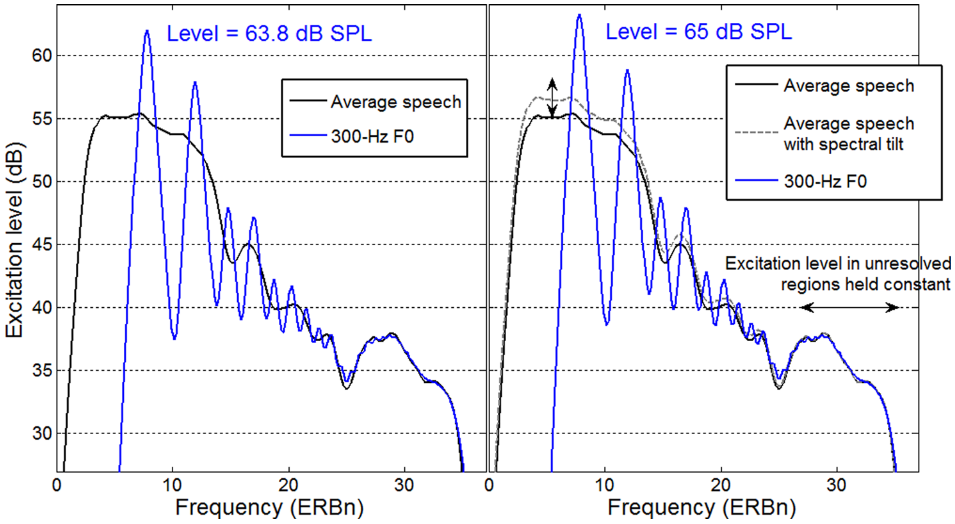

FIG. 7. (Color online) Details of the way harmonic complexes with a speech-shaped spectral profile were generated. The excitation level in unresolved regions was held constant, irrespective of F0, and a spectral tilt was added to the excitation pattern of average speech to increase the amplitude of low-order partials so as to compensate for the reduction in spectral density at high F0s.

relatively more on high-frequency regions when target and masker share a common F0, because they cannot glimpse in low-frequency regions), a spectral tilt was added to the excitation pattern of average speech. This spectral tilt was simply a gain that linearly decreased for center frequencies up to the unresolvability cutoff (as defined in Shackleton and Carlyon, 1994). In practice, an algorithm was written which progressively increased the amount of spectral tilt in the excitation pattern of average speech (illustrated by the dotted line in the right panel of Fig. 7), and the RMS level of the filtered complexes was calculated. This algorithm ended when the RMS level reached 65 dB SPL. Harmonic complexes filtered accordingly had higher amplitudes in their low-order partials, but the excitation level in unresolved regions was the same, irrespective of their F0. In the present study, buzzes based on F0s of 50, 94.4, and 178.2 Hz were created using this filtering process; they were all equalized at 65 dB SPL and their excitation level above 2.8 kHz (24 ERBn) was within 0.05 dB.

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

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.

Boersma, P., and Weenink, D. (**2013**). Praat: doing phonetics by computer. Computer program. Version 5.3.57. http://www.praat.org/ (Last viewed 10/23/2013).

Broadbent, D. E., and Ladefoged, P. (**1957**). "On the fusion of sounds reaching different sense organs," J. Acoust. Soc. Am. **29**, 708–710.

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

Collier, R., and Hart, J.'t (**1975**). "The role of intonation in speech perception," in *Structure and Process in Speech Perception*, edited by A. Cohen and S. G. Nooteboom (Springer Verlag, Heidelberg), pp. 107–121.

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.

Cutler, A., Dahan, D., and van Donselaar, W. (**1997**). "Prosody in the comprehension of spoken language: A literature review," Lang. Speech **40**, 141–201.

Darwin, C. J. (**1975**). "On the dynamic use of prosody in speech perception," in *Structure and Process in Speech Perception*, edited by A. Cohen and S. G. Nooteboom (Springer Verlag, Heidelberg), pp. 178–193.

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.

Dau, T., Verhey, J., and Kohlrausch, A. (**1999**). "Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers," J. Acoust. Soc. Am. **106**, 2752–2760.

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

Deroche, M. L. D., Culling, J. F., and Chatterjee, M. (**2013**). "Phase effects in masking by harmonic complexes: Speech recognition," Hear. Res. **306**, 54–62.

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

Drullman, R., and Bronkhorst, A. (**2004**). "Speech perception and talker segregation: Effects of level, pitch, and tactile support with multiple simultaneous talkers," J. Acoust. Soc. Am. **116**, 3090–3098.

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.

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

Houtsma, A., and Goldstein, J. (**1972**). "The central origin of the pitch of complex tones: Evidence from musical interval recognition," J. Acoust. Soc. Am. **51**, 520–529.

Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (**1999**). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Commun. **27**, 187–207. http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_e.html (Last viewed 04/07/2014).

Kidd, G., Mason, C., Brughera, A., and Hartmann, W. M. (**2005**). "The role of reverberation in release from masking due to spatial separation of sources for speech identification," Acta Acust. Acust. **91**, 526–535.

Licklider, J. C. R. (**1951**). "A duplex theory of pitch perception," Experientia **7**, 128–134.

Meddis, R., and Hewitt, M. J. (**1992**). "Modeling the identification of concurrent vowels with different fundamental frequencies," J. Acoust. Soc. Am. **91**, 233–245.

Myers, T. F., Zhukova, M. G., Chistovich, L. A., and Mushnikov, V. N. (**1975**). "Auditory segmentation and the method of dichotic stimulation," in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. A. A. Tatham (Elsevier, New York), pp. 243–273.

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.

Shackleton, T. M., and Carlyon, R. P. (**1994**). "The role of resolved and unresolved harmonics in pitch perception and frequency modulation," J. Acoust. Soc. Am. **95**, 3529–3540.

Simpson, S., and Cooke, M. P. (**2005**). "Consonant identification in *N*-talker babble is a non-monotonic function of *N*," J. Acoust. Soc. Am. **118**, 2775–2778.

Summers, V., and Leek, M. R. (**1998**). "Masking of tones and speech by Schroeder-phase harmonic complexes in normally hearing and hearing-impaired listeners," Hear. Res. **118**, 139–150.