The role of pitch and harmonic cancellation when listening to speech in harmonic background sounds

Daniel R. Guest, and Andrew J. Oxenham

Citation: The Journal of the Acoustical Society of America **145**, 3011 (2019); doi: 10.1121/1.5102169
View online: https://doi.org/10.1121/1.5102169
View Table of Contents: https://asa.scitation.org/toc/jas/145/5
Published by the Acoustical Society of America

The role of pitch and harmonic cancellation when listening to speech in harmonic background sounds

Daniel R. Guest[a] and Andrew J. Oxenham
Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455, USA

(Received 2 January 2019; revised 16 April 2019; accepted 19 April 2019; published online 21 May 2019)

Fundamental frequency differences ($\Delta F0$) between competing talkers aid in the perceptual segregation of the talkers ($\Delta F0$ benefit), but the underlying mechanisms remain incompletely understood. A model of $\Delta F0$ benefit based on harmonic cancellation proposes that a masker's periodicity can be used to cancel (i.e., filter out) its neural representation. Earlier work suggested that an octave $\Delta F0$ provided little benefit, an effect predicted by harmonic cancellation due to the shared periodicity of masker and target. Alternatively, this effect can be explained by spectral overlap between the harmonic components of the target and masker. To assess these competing explanations, speech intelligibility of a monotonized target talker, masked by a speech-shaped harmonic complex tone, was measured as a function of $\Delta F0$, masker spectrum (all harmonics or odd harmonics only), and masker temporal envelope (amplitude modulated or unmodulated). Removal of the masker's even harmonics when the target was one octave above the masker improved speech reception thresholds by about 5 dB. Because this manipulation eliminated spectral overlap between target and masker components but preserved shared periodicity, the finding is consistent with the explanation for the lack of $\Delta F0$ benefit at the octave based on spectral overlap, but not with the explanation based on harmonic cancellation. © *2019 Acoustical Society of America*. https://doi.org/10.1121/1.5102169

[VB] Pages: 3011–3023

I. INTRODUCTION

Pitch is a salient perceptual dimension of many common sounds and plays a key role in the perception of speech and music (Plack *et al.*, 2005; Oxenham, 2018). Voiced speech has a clear pitch determined by the fundamental frequency ($F0$) of vibration of the talker's vocal folds. Voice pitch has important suprasegmental functions, such as conveying emotion and emphasis (Frick, 1985), and is also a key indexical property of speech, conveying information about the talker's age and biological sex (Assmann *et al.*, 2006; Barreda and Assmann, 2018). Moreover, voice pitch helps listeners better understand speech in challenging listening conditions. A number of studies have demonstrated that as the $F0$ difference ($\Delta F0$) between competing talkers increases, perceptual segregation of the talkers becomes easier (Assmann and Summerfield, 1994; Assmann, 1998; Oxenham and Simonson, 2009; Deroche *et al.*, 2014a; Leclère *et al.*, 2017). Hereafter, we refer to such improvements as $\Delta F0$ benefit. Studies of $\Delta F0$ benefit have translational relevance because hearing-impaired (HI) listeners experience less $\Delta F0$ benefit than normal-hearing (NH) listeners, a deficit which is thought to play a role in the difficulties HI listeners experience in understanding speech in the presence of competing talkers (Summers and Leek, 1998; Oxenham, 2008). However, understanding the causes of this deficit and the best ways to resolve it is difficult because the mechanisms underlying $\Delta F0$ benefit are not yet completely understood in NH listeners.

One common approach to investigating $\Delta F0$ benefit is to use targets and maskers with monotone (i.e., non-time-varying) $F0$ contours and to measure speech intelligibility as a function of $\Delta F0$ between target and masker (de Cheveigné *et al.*, 1995; Assmann, 1998; Assmann and Paschall, 1998; Deroche *et al.*, 2014b,a; Leclère *et al.*, 2017; Madsen *et al.*, 2017). Although natural speech is not monotone, using monotone sounds to study $\Delta F0$ benefit is advantageous at least in part because it allows for findings to be interpreted in the context provided by the literature on pitch perception and segregation of monotone harmonic sounds (Assmann and Summerfield, 1990, 1994; Meddis and Hewitt, 1993; Micheyl *et al.*, 2006; Micheyl *et al.*, 2010; Micheyl and Oxenham, 2010; Wang *et al.*, 2012). The most extensively investigated case is that of two concurrent synthetic vowels (Assmann and Summerfield, 1990, 1994; Meddis and Hewitt, 1993; de Cheveigné *et al.*, 1995; de Cheveigné *et al.*, 1997; Assmann and Paschall, 1998).

Two factors likely play a key role in hearing out a monotone target from a monotone masker on the basis of $F0$ differences. First, $F0$ differences between two harmonic sounds create opportunities for listeners to glimpse target energy from target harmonics located in the spectral gaps between resolved masker harmonics. This phenomenon has been termed spectral glimpsing (Deroche and Culling, 2013; Deroche *et al.*, 2014a). Second, $F0$ differences enable the use of $F0$-guided segregation mechanisms to enhance the target and/or suppress the masker. Many models of such segregation mechanisms have been proposed, but no one model is yet definitively supported by behavioral or neurophysiological evidence, and a number of key questions remain unanswered (Assmann and Summerfield, 1990; de Cheveigné, 1993; Meddis and Hewitt, 1993; Cariani, 2003; Micheyl and Oxenham, 2010).

[a]Electronic mail: guest121@umn.edu

Most models of an *F0*-guided segregation mechanism share the general architecture of a phenomenological model of auditory peripheral processing followed by an *F*0 estimation stage and then a segregation stage. The auditory periphery stage typically includes a filterbank, representing the action of the basilar membrane, followed by a rectifying non-linearity, mimicking the function of the inner hair cells and auditory nerve. The *F*0 estimation stage then uses the output of the auditory periphery stage to derive estimates of the stimulus *F*0s. Finally, the *F*0 estimates are used to guide the segregation stage, which seeks to separate the competing sounds, suppress the masker, and/or enhance the target (Assmann and Summerfield, 1990; de Cheveigné, 1993; Meddis and Hewitt, 1993). Given its conceptual separability from the previous stages and relevance to the topic at hand, we will turn our focus primarily toward this last stage.

Assmann and Summerfield (1990) proposed a segregation mechanism that separated competing sounds by sampling the internal representation of the sound mixture according to an *F*0-guided rule which, in their "place-time" model, was based on sampling the autocorrelation functions (ACFs) of each channel (where a channel is a particular filter in an auditory filterbank) at the delays corresponding to the periods of each sound. Meddis and Hewitt (1993) proposed a segregation mechanism that, instead of separating the sounds by sampling from all of the channels, separated the sounds by decomposing the channels into two disjoint subsets. Specifically, the subset of channels with ACFs that produced the same dominant *F*0 estimate as the cross-channel pooled ACF (i.e., in which the ACF is computed first in each channel and then summed across channels) were assumed to predominantly represent one sound, while the complement of this subset was assumed to predominantly represent the competing sound. de Cheveigné (1993) addressed the problem in a somewhat different fashion, describing an algorithm that, rather than segregating the competing sounds, sought to isolate one while cancelling (i.e., eliminating or suppressing) the other. This "harmonic cancellation" mechanism filtered the sound mixture with a time-domain comb filter tuned to the *F*0 of the masker, a processing strategy that effectively eliminates the masker, while leaving the target mostly intact.

A target with an *F*0 one octave higher than the *F*0 of its masker poses an intriguing challenge for most models of *F*0-guided sound segregation. The first problem is that double *F*0 estimation is difficult in the presence of an octave $\Delta F0$ because such an *F*0 relationship produces ambiguous cues as to the presence of two different *F*0s (Micheyl and Oxenham, 2010). This problem is illustrated in Fig. 1, which shows the waveforms (top row), power spectra (middle row), and autocoincidence (AC) histograms for simulated auditory nerve fibers (ANFs) (lower row) in response to synthetic vowels (see the Appendix for details of the simulations). Each of the isolated vowels (the left and middle columns) produce clear cues in the AC histogram as to their *F*0s (i.e., they contain peaks at the periods of the two vowels). In contrast, in the AC histogram, the mixture of two vowels separated by an octave produces a clear peak at 10 ms (i.e., 100 Hz) but a notably less salient peak at 5 ms (i.e., 200 Hz). Thus, an *F*0-guided segregation mechanism might fail in the case of an octave $\Delta F0$ simply because it fails to detect the presence of two sounds or cannot clearly identify their *F*0s. It should be noted that this problem is not equally severe for all

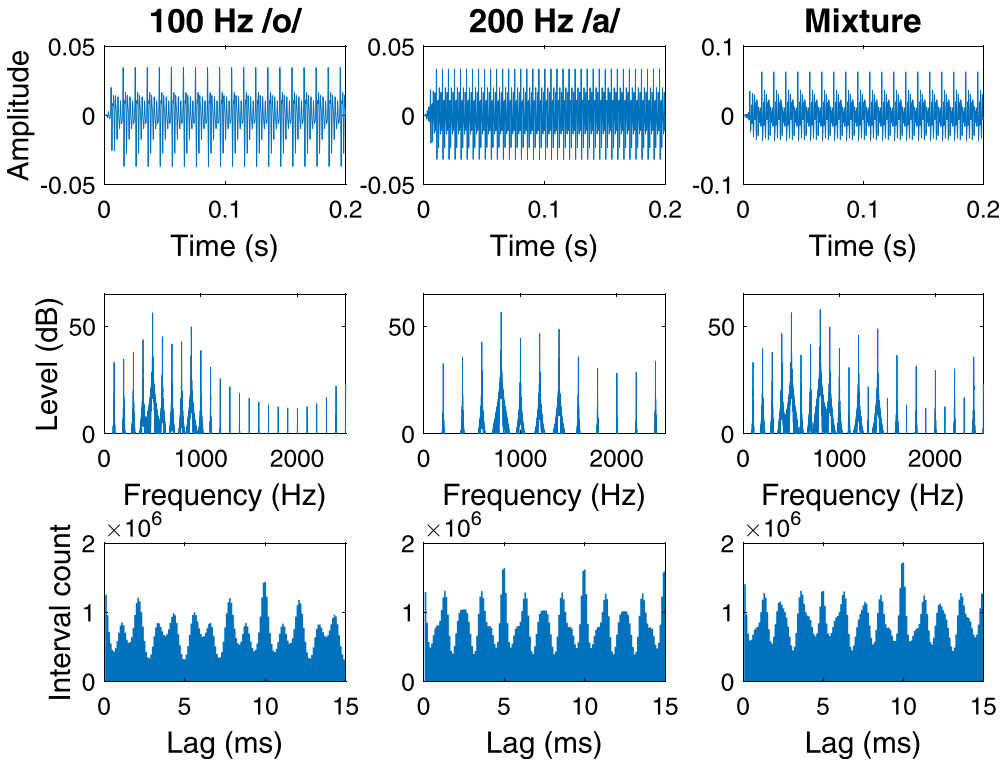

FIG. 1. (Color online) Time domain waveforms (top), power spectra (middle), and pooled AC histograms of simulated ANF responses for synthetic vowels. The first two columns show an /o/ with a 100 Hz *F*0 and an /a/ with a 200 Hz *F*0, respectively, while the last column shows a mixture of the two vowels.

mechanisms, as not all of the aforementioned segregation mechanisms require two $F0$ estimates. For example, neither the channel assignment mechanism proposed by Meddis and Hewitt (1993) nor the harmonic cancellation mechanism proposed by de Cheveigné (1993) require the estimation of both $F0$s. However, even if this issue can be solved and the correct $F0$ estimate(s) can be made, the segregation stage itself may still pose problems. Consider, for example, the case of harmonic cancellation. Even with the correct $F0$ estimates, cancelling the masker via a neural comb filter tuned to the masker $F0$ would also cancel the target because all the harmonics of the target are also harmonics of the masker. This phenomenon can be seen in Fig. 2, which shows that applying a cancellation filter to eliminate a masker with an $F0$ one octave below the target $F0$ eliminates both the target and masker, whereas the same mechanism suppresses the masker periodicity, while leaving the target periodicity intact when the target and masker are separated by a non-octave interval. It seems reasonable, then, to predict that the octave $\Delta F0$ may be challenging for human listeners as well, if humans use a similar mechanism.

Human performance in segregating two talkers separated by an octave $\Delta F0$ has been previously investigated by Brokx and Nooteboom (1982). In one experiment, they presented listeners with target sentences resynthesized via linear predictive coding (LPC) analysis with monotone $F0$s in intervals of 0, 1, 2, 3, 4, or 12 semitones (ST) above 100 Hz. The masker was composed of a continuous stream of the target talker's speech, which was resynthesized with a monotone 100 Hz $F0$. The authors found that the rate of errors in reporting the target speech decreased monotonically with increasing $\Delta F0$, except at 12 ST, where performance was similar to performance at about 1 ST $\Delta F0$. This finding is broadly consistent with the qualitative predictions of the cancellation model above, that an octave $\Delta F0$ should provide little $\Delta F0$ benefit.

However, the multitude of alternative explanations for the findings of Brokx and Nooteboom (1982) limit our ability to draw strong conclusions from it with respect to models of $\Delta F0$ benefit. First, distortions introduced by the synthesis algorithm may have reduced the inherent intelligibility of the target talker and thus offset any potential $\Delta F0$ benefit. Deroche *et al.* (2014a) discussed this issue and demonstrated that speech from a male talker, resynthesized via the pitch-synchronous overlap-add technique (PSOLA; Moulines and Charpentier, 1990) with various monotonized $F0$ contours, produced higher (poorer) speech reception thresholds (SRTs) against a white noise background when the $F0$ was outside of a natural male $F0$ range than within, providing support for this explanation. Second, fixed $F0$ differences of an octave may have produced some perceptual fusion of the two voices because the two voices share a pitch chroma (Huron, 1991). Third, because every harmonic of the target was also a harmonic of the masker, limited opportunities for spectral glimpses in the masker may have reduced speech intelligibility. Thus, although behavioral data for the octave $\Delta F0$ exist, they are of little value for informing models of $\Delta F0$ benefit.

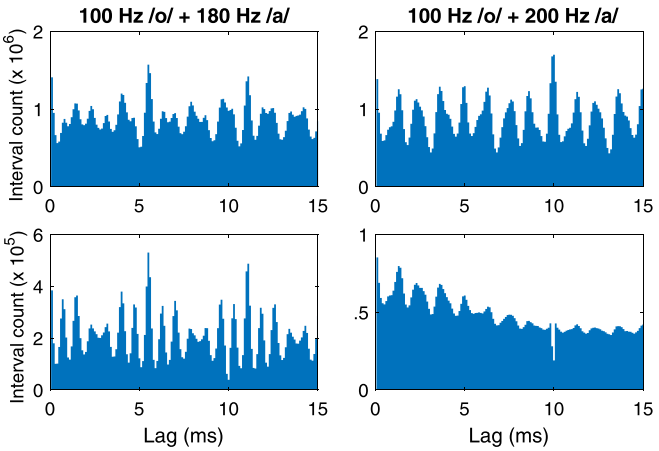

FIG. 2. (Color online) Pooled AC histograms of stimulated ANF responses for mixtures of synthetic vowels. The top row shows unprocessed responses, while the bottom row shows responses processed with a cancellation filter tuned to 100 Hz (de Cheveigné, 1993). The left column shows responses to a mixture of a 100 Hz /o/ and a 180 Hz /a/, while the right column shows responses to a mixture of a 100 Hz /o/ and a 200 Hz /a/.

To address these limitations and attempt to rule out some of these competing explanations, we conducted two experiments measuring SRTs for a target talker in the presence of various maskers. First, to assess the extent to which the Brokx and Nooteboom (1982) findings could be explained by reductions in target intelligibility due to $F0$ manipulations, we first performed an experiment that measured SRTs of the target stimuli against white noise, in a similar fashion to Deroche *et al.* (2014a) (Experiment 1). We also used STRAIGHT (Kawahara, 1997; Kawahara *et al.*, 1999), a high-quality speech manipulation program, to manipulate the target speech in an effort to minimize the reductions in target intelligibility due to $F0$ manipulations. Second, we measured SRTs for the same target talker in the presence of harmonic complex tones (HCTs) as a function of ΔF0 (Experiment 2). In that experiment, we also included a condition wherein spectral overlap was eliminated between target and masker at the octave $\Delta F0$ (by removing the even harmonics of the lower-$F0$ stimulus), while preserving their shared periodicity in an effort to test explanations based on spectral overlap. Additionally, we compared performance with and without broadband temporal modulation of the masker to determine whether a listener's ability to benefit from $\Delta F0$ interacted in any way with their ability to benefit from temporal modulations.

II. GENERAL METHODS

A. Stimuli

1. Targets

The target stimuli were recordings of sentences from the IEEE/Harvard sentence lists (Rothauser *et al.*, 1969) made at the University of Minnesota by a single male talker. The geometric mean $F0$ of the talker across all sentences, estimated via STRAIGHT, was 90 Hz. The recordings were made at a sampling rate of 22.05 kHz but were resampled to 20 kHz for processing and playback. Lists 1 and 2 (20 sentences) were used as practice stimuli while lists 3–66 (640 sentences) were used as test stimuli.

All signal processing was conducted in MATLAB (The MathWorks, Natick, MA). First, the targets were analyzed with STRAIGHT (Kawahara, 1997; Kawahara *et al.*, 1999). Then, the targets were resynthesized by STRAIGHT with their natural *F*0 contour as well as with monotone *F*0 contours set to 80, 95.14, 160, and 190.27 Hz (0, 3, 12, and 15 ST above 80 Hz, respectively). Next, the monotone 80-Hz targets were further processed to have two types of spectral structure (here, spectral structure refers to specific patterns of component levels). For the first type, SS-all, the targets were synthesized with all of their harmonics. For the second type, SS-odd, the voiced and unvoiced portions of the targets were synthesized separately, and the even harmonics of the voiced portion were removed with a zero-phase comb filter tuned to 2*F*0 before summation with the unvoiced portions. Thus, the voiced portions of speech in the SS-odd targets contained only odd harmonics, while the unvoiced portions of the SS-odd targets remained identical to those of the SS-all targets. The relative root mean squared (rms) levels of the voiced and unvoiced portions were maintained at the sentence level across all *F*0s and spectral structures.

After the targets were synthesized, they were passed through 1024-tap finite impulse response (FIR) filters designed to match the unresolved portions of their excitation patterns (EPs, i.e., simulated outputs of auditory filters as a function of filter center frequency; Glasberg and Moore, 1990; Moore and Glasberg, 1987) at a given rms level (Deroche *et al.*, 2014a; Leclère *et al.*, 2017). Specifically, the unresolved portions of the EPs of each sentence were matched to those of the sentence's corresponding 80 Hz SS-all version. Here, "unresolved portion" refers to frequencies above 2000 Hz. This cutoff was chosen to be just above the frequency of the 10th harmonic of the highest *F*0 used in this experiment (i.e., approximately beyond the limits of peripheral resolvability for all of our stimuli; Bernstein and Oxenham, 2003; Houtsma and Smurzynski, 1990; Shackleton and Carlyon, 1994) to ensure that potential segregation cues based on resolved harmonics were preserved at this stage of signal processing. The effect of this processing can be seen in Fig. 3, which shows the average EP of each stimulus type.

2. Maskers

For Experiment 1, the maskers were samples of white Gaussian noise, which were freshly generated on each trial. All the maskers in Experiment 1 were presented at an rms level of 70 dB sound pressure level (SPL). White Gaussian noise was selected over speech-shaped noise in order to both replicate the stimulus design of Deroche *et al.* (2014a) and to emphasize the lower harmonics of the stimuli (which did not have EPs matched across conditions) over the higher harmonics (which had EPs matched across conditions). The lower (peripherally resolved) harmonics were emphasized here because it was thought they would likely play a more important role in *F*0-guided segregation than the higher harmonics (Bird and Darwin, 1997).

For Experiment 2, the maskers were speech-shaped HCTs, freshly synthesized on each trial. The maskers were synthesized with the same *F*0s and spectral structures as described above for the targets. For SS-all maskers, pure tones up to the Nyquist frequency were added in random phase at harmonic frequencies of the *F*0. Then, the tones were passed through an FIR filter designed to match their EPs to the average EP of the targets with the corresponding *F*0. For SS-odd maskers, the same process was followed, but additionally the stimuli were filtered with a comb filter designed to remove the even harmonics. Using this procedure, the EPs of the SS-odd maskers closely matched the average EPs of the targets with the corresponding spectral

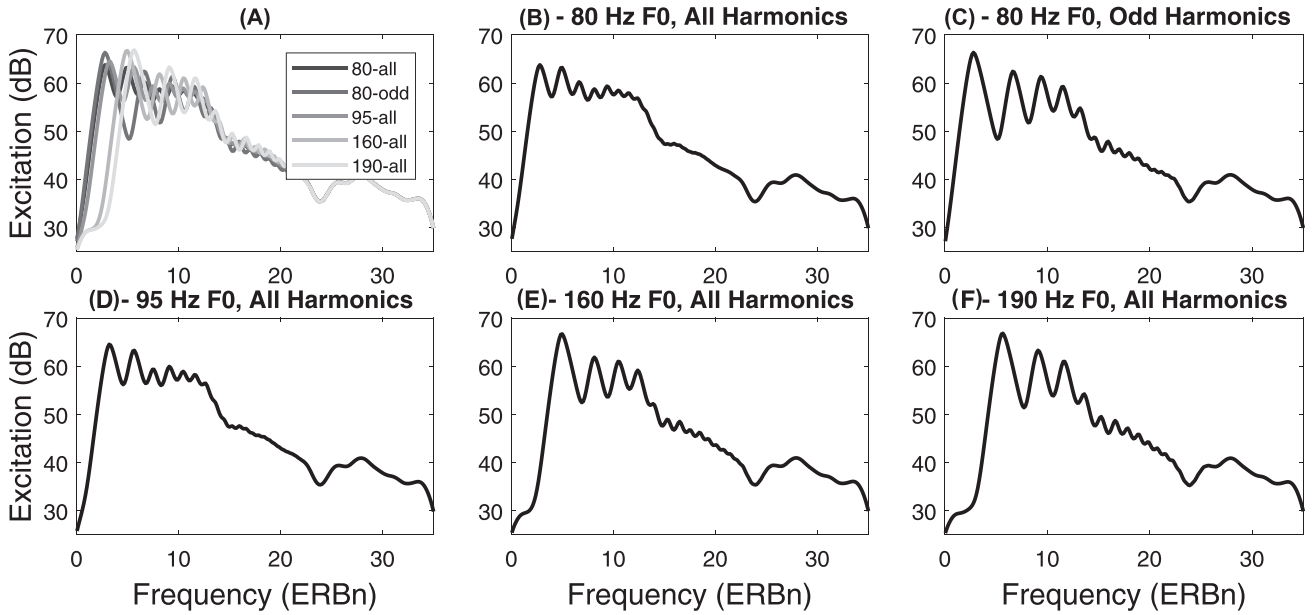

FIG. 3. Average excitation patterns for each speech stimulus type plotted on a frequency scale of estimated number of equivalent rectangular bandwidths (ERBn) of the auditory filters (Glasberg and Moore, 1990). Panels (B), (D), (E), and (F) show SS-all stimuli (with 80, 95, 160, and 190 Hz F0s, respectively), while Panel (C) shows SS-odd 80 Hz stimuli. Panel (A) shows all five stimulus types superimposed, with contrast indicating stimulus type.

structure and $F0$. All of the maskers in Experiment 2 were presented at an rms level of 70 dB SPL.

In some conditions of Experiment 2, the maskers were amplitude modulated by broadband temporal envelopes extracted from speech. To extract these envelopes, the 80 Hz SS-all target sentences were concatenated and then gaps of silence between sentences greater than 50 ms in length were removed. The concatenated speech was full-wave rectified and then zero-phase filtered by a 4th-order lowpass Butterworth filter with a cutoff frequency of 40 Hz. A random segment of this envelope was extracted and applied to the masker on each trial in these conditions.

3. Combined stimuli

The masker began and ended with 100-ms raised-cosine ramps and was presented for 750 ms before and after the target. The target level was varied adaptively, so the combined level of the target and masker varied throughout the experiment. In Experiment 1 the target-to-masker ratio (TMR) was not permitted to exceed 10 dB, while in Experiment 2 the TMR was not permitted to exceed 4 dB. These values were set based on the range of TMRs observed in pilot testing. There were only seven trials in Experiment 1 in which listeners were tested at the ceiling TMR value and then remained at that TMR value for the next trial (because they reported fewer than three keywords correctly), and there were no such trials in Experiment 2, so it was assumed that these ceiling values had little impact on the measured SRTs.

B. Procedure

SRTs were measured using a one-up one-down adaptive procedure adapted from Deroche *et al.* (2014a). In each run, ten sentences were presented in sequence. The TMR began at −28 dB on the first sentence, and listeners were given the opportunity to repeat the first sentence multiple times, with each repeat increasing the TMR by 4 dB. Once they could hear approximately half the sentence, they typed what they heard. A transcript of the sentence with capitalized keywords was then displayed and listeners indicated how many keywords they correctly identified. For the subsequent sentences, listeners only had one opportunity to hear the sentence before making their responses. After each trial, the TMR was increased by 2 dB if the listener reported correctly identifying at least three of the five target keywords and decreased by 2 dB if the listener reported correctly identifying two or fewer target keywords. The SRT of each run was defined as the average of the TMRs of the last eight sentences.

Before beginning the experiment, the listeners were instructed first verbally and then again in written form about the procedure. They were instructed to grade their answers based on whether or not they contained the same sounds as the transcript if spoken aloud and not on the basis of spelling. The listeners were provided with a small set of examples demonstrating correct grading procedures. In both experiments, the listeners completed two runs of practice before data collection began. The parameters in each round of practice (e.g., $\Delta F0$) were randomly selected from the range of possible parameters within the experiment for each listener. Listeners were encouraged to take breaks after every six runs, and individual sessions never lasted more than 2 h.

All stimuli were presented to listeners diotically over HD650 headphones (Sennheiser, Old Lyme, CT) via a Lynx E22 sound card (Lynx Studio Technologies, Costa Mesa, CA) in sound attenuating booths. Listeners completed the experiment via a graphical user interface generated via custom MATLAB scripts.

C. Listeners

Twenty listeners participated in Experiment 1 (15 female, 5 male) and 20 different listeners participated in Experiment 2 (14 female, 6 male). All listeners were between 18 and 38 years of age, were native speakers of American English, and had pure tone thresholds no greater than 20 dB hearing level (HL) at octave frequencies from 250 Hz to 8 kHz. Participants in Experiment 1 were recruited through a University of Minnesota Department of Psychology research participation pool and were compensated with their choice of extra course credit or $10 per hour. Participants in Experiment 2 were recruited from an in-house participant database and compensated with $10 per hour. Experiment 1 took approximately 1.5 h to complete, while Experiment 2 took approximately 6 h to complete. All participants provided written informed consent prior to participating, and all protocols were approved by the Institutional Review Board of the University of Minnesota.

III. EXPERIMENT 1: SPEECH PERCEPTION IN NOISE

A. Design and rationale

Experiment 1 examined the intelligibility of the target talker in white background noise as a function of the target talker's $F0$, spectral structure, and intonation. Seven conditions were tested. In the first condition, denoted INT, the targets were the unprocessed speech stimuli with their natural intonation intact. In the second condition, denoted INT-PROC, the targets were the speech stimuli analyzed and resynthesized with STRAIGHT but without any modifications to $F0$. Because we expected the INT stimuli to be more intelligible than the monotone stimuli (described below), the INT-PROC condition was included to help determine what portion of this difference in intelligibility could be attributed to distortions produced by STRAIGHT and what portion could be attributed to the flattening of the $F0$ contour. In the next four conditions, denoted 80-all, 95-all, 160-all, and 190-all, the targets were processed speech stimuli with SS-all spectral structure (i.e., their voiced speech contained all of its harmonics) and monotone $F0$s of 80, 95, 160, and 190 Hz, respectively. In the final condition, denoted 80-odd, the targets were processed speech stimuli with SS-odd spectral structure (i.e., their voiced speech contained only odd harmonics) and a monotone $F0$ of 80 Hz. Two runs per participant were completed for each condition, yielding a total 14 runs (seven conditions x two runs per condition). Lists were randomly assigned to runs for each participant.

Experiment 1 was designed to assess the extent to which the signal processing applied to the speech stimuli affected their intelligibility, independent of the presence of an $F0$

difference between target and masker (the key independent variable in Experiment 2). Additionally, it sought to replicate the basic findings of Deroche *et al.* (2014a) while using STRAIGHT instead of Praat PSOLA (Boersma and Weenink, 2019) to manipulate the speech stimuli. Based on prior research (Assmann and Nearey, 2008; Deroche *et al.*, 2014a), we first hypothesized that *F*0s closer to the natural average *F*0 of our talker (approximately 90 Hz) would result in lower (better) SRTs. That is, manipulating the *F*0 of the talker away from its natural range should reduce the intelligibility of the speech. We also hypothesized that the 80-odd condition would result in poorer speech intelligibility than the 80-all condition. Such a hypothesis is reasonable, given the unnatural quality of the 80-odd speech and its sparser sampling of the spectral envelope. Finally, we hypothesized that the intonated speech would be more intelligible than any of the monotone speech. This hypothesis was motivated by the well-established finding that monotone speech is less intelligible than speech with natural *F*0 variations (Miller *et al.*, 2010; Deroche *et al.*, 2014a; Madsen *et al.*, 2017).

B. Analysis

1. Data preprocessing

Two pre-defined criteria were used to screen data before data analysis. First, means and standard deviations for each condition were calculated (data for each listener were averaged across the listener's two runs before calculation of these statistics). Then, any individual threshold was excluded as an outlier if it was more than three standard deviations from the mean in that condition. Four runs from three listeners were excluded based on this criterion. Second, a listener's data were excluded if the listener failed spotchecks on their self-grading performance. To perform spotchecks, 30 responses were randomly selected from all of the responses of each listener. If their reported number of correct keywords deviated from the true number of correct keywords by more than 15%, their data were excluded. One listener failed this criterion. Hence, only data from 19 participants were analyzed and presented below.

2. Statistical model and tests

A linear mixed-effects model was used to analyze the results from Experiment 1. The only fixed effect was condition, while random effects included random intercepts and slopes for listener and list (i.e., a maximal random effects structure given our data; Barr *et al.*, 2012). The model was fit in the R programming language using the lme4 package via penalized maximum likelihood estimation (Bates *et al.*, 2015). After fitting the model but before proceeding with analysis, diagnostic checks by visual inspection of a QQ plot of the standardized residuals and a plot of residuals versus fitted values were made to ensure satisfactory normality and independence of the residuals.

The model was analyzed in two ways. First, the significance of condition was assessed using F tests in a type III analysis of variance (ANOVA). The ANOVA employed the Satterthwaite approximation for degrees of freedom and was implemented using the lmerTest package in R (Kuznetsova *et al.*, 2017). Second, each of our hypotheses was analyzed by a Wald χ^2 linear contrast test. The contrast tests were implemented using the phia package in R (De Rosario-Martinez, 2015). All of the statistical tests in Experiment 1 were jointly corrected using the Holm-Bonferroni method (Abdi, 2010) and the corrected p values are reported below. A criterion of $\alpha = 0.05$ was used to assess statistical significance.

C. Results

The results of Experiment 1 are plotted in Fig. 4. The one-way ANOVA revealed a significant effect of condition [$F(6,11) = 23.45$, $p < 0.001$]. Pairwise linear contrasts between the 80-all condition and each of the other SS-all conditions (i.e., 95-all, 160-all, 190-all) revealed no significant difference between 80-all and 95-all (estimated mean difference $= 0.70$ dB, $\chi^2_1 = 0.97$, $p = 0.32$) or between 80-all and 160-all (estimated mean difference $= 1.54$ dB, $\chi^2_1 = 4.68$, $p = 0.091$), but revealed a significant difference between 80-all and 190-all (estimated mean difference $= 1.87$ dB, $\chi^2_1 = 9.41$, $p = 0.011$). In summary, while our analysis confirmed our first hypothesis that shifting the *F*0 of the talker away from its natural range would elevate SRTs, this change was only significant for the highest *F*0 we tested. A linear contrast between the 80-all and 80-odd conditions revealed a significant difference between the two conditions (estimated mean difference $= 2.39$ dB, $\chi^2_1 = 13.02$, $p = 0.0018$), confirming our hypothesis that the 80-odd condition would be more difficult than the 80-all condition. Finally, a linear contrast between the INT condition and an average of the 80-all and 95-all conditions (the two monotone conditions with *F*0s close to the average *F*0 of the talker's natural speech) revealed a significant benefit of natural speech over monotone processed speech (estimated mean difference $= 2.65$ dB, $\chi^2_1 = 18.83$, $p < 0.001$). However, we cannot confidently attribute this effect strictly to the elimination of intonation, as opposed to artifacts produced by STRAIGHT, because *post hoc* tests revealed that while the INT and INT-PROC conditions were not significantly different

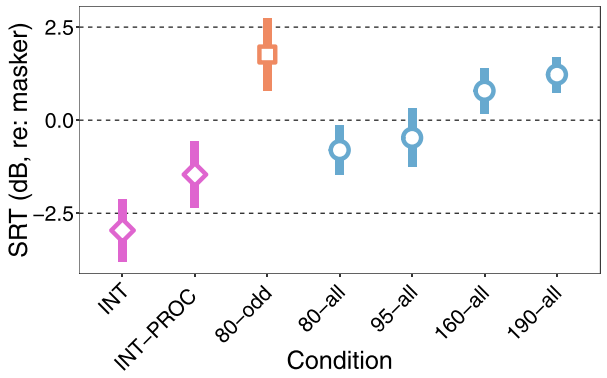

FIG. 4. (Color online) SRTs versus condition for Experiment 1. The SRTs for each listener's two runs per condition were averaged before averaging data across listeners. Error bars indicate ± 1.96 standard error of the mean. Data from SS-all conditions is shown in blue circles, data from SS-odd conditions is shown in orange squares, and data from intonated conditions is shown in purple diamonds.

(estimated mean difference = 1.22 dB, $\chi_1^2 = 3.65$, $p = 0.11$), the INT-PROC and an average of the 80-all and 95-all conditions were also not significantly different (estimated mean difference = 1.43 dB, $\chi_1^2 = 5.18$, $p = 0.091$).

D. Discussion

Our findings are broadly consistent with previous literature. In particular, Deroche *et al.* (2014a) reported comparable findings for a similar experiment that also used a processed male talker and a white-noise masker. They found SRTs were about 1.5 dB higher when the monotone target *F*0 was shifted one octave upward (from 100 to 200 Hz), and we likewise found SRTs were about 1.5 dB higher (although this change was not significant after correction for multiple comparisons) when the monotone target *F*0 was shifted one octave upward (from 80 to 160 Hz). Deroche *et al.* also found that unprocessed speech had SRTs that were about 2 dB better than artificially monotonized speech, while we found that SRTs were about 2.5 dB better for natural speech than monotonized speech. The lack of significant differences between the INT-PROC condition and the 80-all condition or between the INT-PROC condition and the INT condition in our experiment suggests that this effect of processing may have not been due solely to *F*0 manipulations but also to distortions introduced by the speech manipulation software. However, the contrast between the INT-PROC condition and INT condition cannot provide insight into additional processing artifacts that are likely introduced when changes are made to the *F*0 contour, so further research will be needed to address this issue more completely.

One novel finding from Experiment 1 is that removing the even harmonics reduced the talker's intelligibility. As previously mentioned, this could have been a consequence of the unnatural timbre created by this manipulation. In addition, it could have been partially caused by the sparser sampling of the talker's spectral envelope in this condition. The main value of these findings is in providing context to the results of Experiment 2 presented below. In Experiment 2, in cases where the target *F*0 was varied and the masker *F*0 was fixed, Δ*F*0 was confounded with target *F*0. But based on the results of Experiment 1 and under the simplifying assumption that the speech-in-noise and speech-in-tone psychometric functions are reasonably similar, we can estimate that the impact of target *F*0 on SRTs in Experiment 2 should not be much greater than about 2 dB.

IV. EXPERIMENT 2: SPEECH PERCEPTION IN MODULATED AND UNMODULATED HARMONIC TONES

A. Design and rationale

In Experiment 2, four independent variables of interest were manipulated in a fully factorial design. The first variable was the absolute *F*0 difference between target and masker (Δ*F*0), which had four levels (0, 3, 12, and 15 ST). The second variable was target *F*0 range, i.e., whether the target was assigned to the lower *F*0 (always 80 Hz) or the higher *F*0. Target *F*0 range therefore had two levels (Target Low and Target High). The third variable was spectral structure, which had two levels (SS-all, SS-odd). The spectral structure manipulation was always applied to the sound assigned to the lower *F*0. In the text and figures below, spectral structure and target *F*0 range are labeled jointly by indicating first which sound had the fixed lower *F*0 and second what that sound's spectral structure was (e.g., Target-All means the target *F*0 was 80 Hz while the masker *F*0 varied and the target had all of its harmonics; Masker-Odd means the masker *F*0 was 80 Hz while the target *F*0 varied and the masker only had its odd harmonics). The fourth variable was masker modulation, which had two levels (unmodulated and modulated with a broadband speech envelope, as described in the general methods). Two runs per participant were completed for each possible combination of variables, yielding a total of 64 runs (4 Δ*F*0 × 2 spectral structures × 2 target *F*0 ranges × 2 masker modulations × 2 runs per condition). Lists were randomly assigned to runs for each participant.

Three predictions were made regarding Experiment 2. First and most importantly, we expected that the pattern of results for speech segregation at the octave Δ*F*0 would be inconsistent with an explanation based on harmonic cancellation. Our key *a priori* hypothesis to this effect was that when the target had an *F*0 one octave above the masker *F*0, removing the even harmonics from the masker would improve SRTs (i.e., Masker-Odd would produce better thresholds than Masker-All at the octave Δ*F*0). This manipulation eliminated spectral overlap between the voiced portions of the target and masker but did not alter their shared periodicity. Thus, this hypothesis was consistent with an explanation of the findings of Brokx and Nooteboom (1982) based on spectral glimpsing, but not consistent with an explanation based on harmonic cancellation. Examination of this hypothesis was supplemented by a number of *post hoc* comparisons in order to ascertain whether the broader pattern of results also favored spectral glimpsing over harmonic cancellation. It should be emphasized at this point, however, that our experimental design focused on resolving the question of harmonic cancellation versus spectral glimpsing at the octave Δ*F*0 specifically and cannot resolve this issue at other Δ*F0s* more generally. Second, in cases where the target and masker had all of their harmonics, Δ*F*0 benefit would generally be larger when the masker *F*0 was higher than the target *F*0 than vice versa (i.e., Δ*F*0 benefit would be larger in Target-All than Masker-All). This prediction was consistent with the hypothesis that maskers with higher *F*0s offer better opportunities for spectral glimpsing than maskers with lower *F*0s (Deroche *et al.*, 2014a). Third, broadband temporal envelope modulation of the masker would improve intelligibility of the target talker. This hypothesis was motivated by previous findings that speech envelope modulations imposed on a stationary masker can improve speech intelligibility (Peters *et al.*, 1998; Qin and Oxenham, 2003; Leclère *et al.*, 2017). More generally, the modulated masker conditions were included to determine whether any interactions were present between a listener's ability to benefit from spectral dips and their ability to benefit from temporal dips, but *a priori* only a main effect of masker modulation was anticipated.

B. Analysis

The procedures used in Experiment 1 for data preprocessing were repeated here. Two outlier runs from 2 subjects were excluded from data analysis based on these procedures.

1. Statistical model and tests

A linear mixed-effects model was used to analyze the results from Experiment 2. In the present model, the fixed effects included main effects of $\Delta F0$, spectral structure, target $F0$ range, and masker modulation, as well as all possible interactions. Random effects included random intercepts for listener and list as well as random listener slopes. The same procedures used to fit the model for Experiment 1 were used for Experiment 2.

The model was analyzed in three ways. First, the significance of the main effects and interactions were assessed using F tests in a type III ANOVA. The ANOVA employed the Satterthwaite approximation for degrees of freedom. Second, each of our hypotheses was analyzed by a Wald χ^2 linear contrast test. Finally, *post hoc* linear contrast tests were employed as needed to aid interpretation of the significant main effects and interactions. For all contrast tests presented below, it can be assumed that when a model term is not explicitly mentioned it was averaged for that test. The same R packages used to conduct tests in Experiment 1 were used to conduct these tests. All of the statistical tests in Experiment 2 were jointly corrected using the Holm-Bonferroni method.

C. Results

The means and 95% confidence intervals for each condition in Experiment 2 are shown in Figs. 5 (unmodulated masker) and 6 (modulated masker). The means and 95% confidence intervals for $\Delta F0$ benefit are shown in Fig. 7,

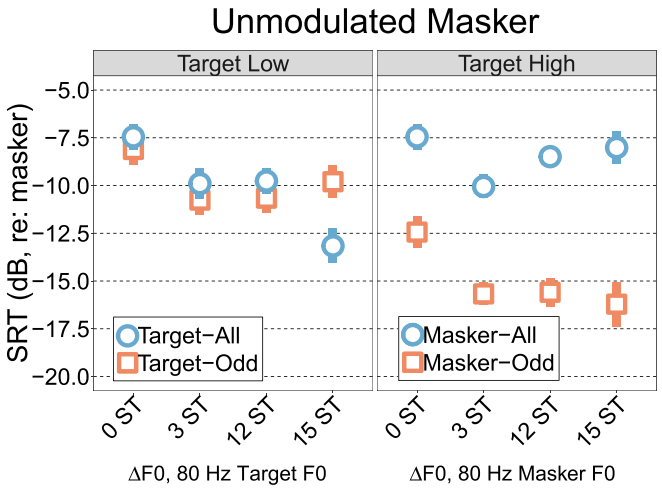

FIG. 5. (Color online) SRTs vs $\Delta F0$ for the unmodulated masker. The SRTs for each listener's two runs were averaged before averaging data across listeners. Error bars indicate ± 1.96 standard error of the mean. Data from Target Low conditions are shown in the left figure, while data from Target High conditions are shown in the right figure. Data from SS-all conditions are shown in blue circles while data from SS-odd conditions are shown in orange squares. For this figure, data from the 0 ST Masker-All and 0 ST Target-All conditions were averaged because the task and stimuli were identical in these two conditions.

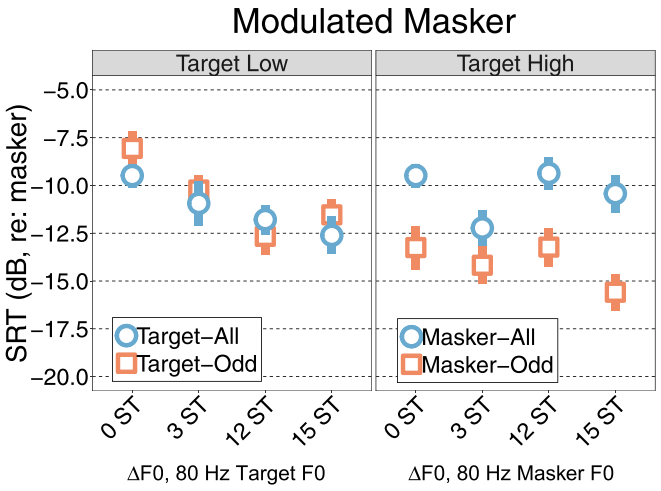

FIG. 6. (Color online) SRTs vs $\Delta F0$ for the modulated masker. The SRTs for each listener's two runs were averaged before averaging data across listeners. Error bars indicate ± 1.96 standard error of the mean. Data from Target Low conditions are shown in the left figure while data from Target High conditions are shown in the right figure. Data from SS-all conditions are shown in blue circles while data from SS-odd conditions are shown in orange squares. For this figure, data from the 0 ST Masker-All and 0 ST Target-All conditions were averaged because the task and stimuli were identical in these two conditions.

while the means and 95% confidence intervals for SS-odd benefit (i.e., improvement in SRTs from removing the even harmonics of the lower $F0$ sound) at each level of $\Delta F0$ are shown in Fig. 8. The ANOVA revealed significant main effects of spectral structure [$F(1, 92) = 161.00$, $p < 0.001$], $\Delta F0$ [$F(3, 80) = 37.38$, $p < 0.001$], target $F0$ range [$F(1, 107) = 79.51$, $p < 0.001$] and masker modulation [$F(1, 40) = 16.60$, $p = 0.0041$]. The interaction between $\Delta F0$ and target $F0$ range was significant [$F(3, 1043) = 18.59$, $p < 0.001$], as were the interactions between target $F0$ range and spectral structure [$F(1, 1040) = 230.44$, $p < 0.001$] and between $\Delta F0$ and spectral structure [$F(3, 1050) = 4.67$, $p = 0.048$]. No other model terms reached significance.

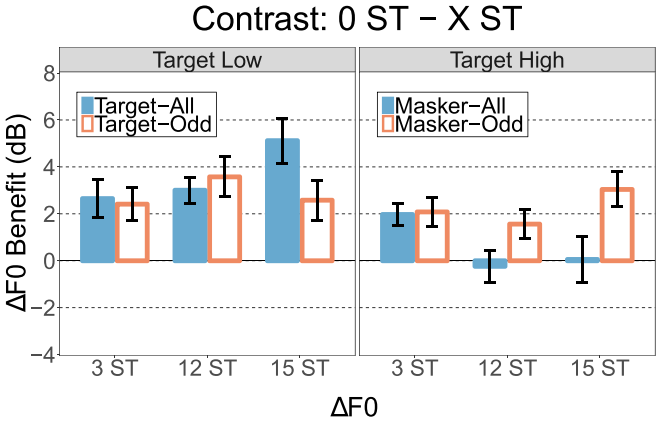

FIG. 7. (Color online) $\Delta F0$ benefit versus $\Delta F0$. $\Delta F0$ benefit was calculated as the difference between average SRTs in the 0 ST $\Delta F0$ and the given $\Delta F0$. Data shown here were averaged first across masker modulation and then across listeners. Error bars indicate ± 1.96 standard error of the mean. The left panel shows data from the Target Low conditions, while the right panel shows data from the Target High conditions. Data from SS-all conditions are shown in filled blue bars while data from SS-odd conditions are shown in unfilled orange bars.

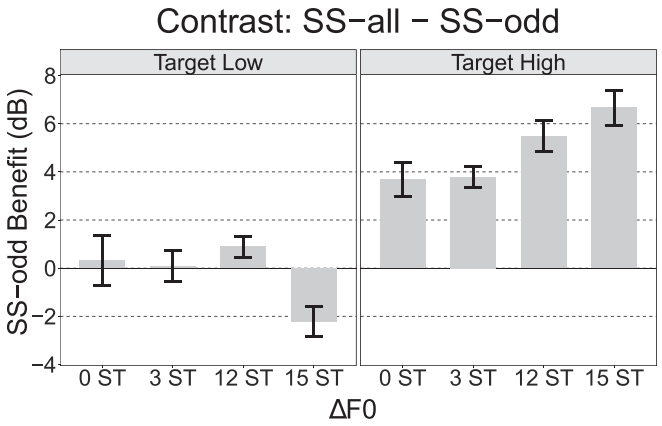

FIG. 8. SS-odd benefit versus $\Delta F0$. SS-odd benefit was calculated as the difference between average SRTs in the SS-all and SS-odd conditions. Data shown here were averaged first across masker modulation and then across listeners. Error bars indicate ± 1.96 standard error of the mean. The left panel shows data from the Target Low conditions while the right panel shows data from the Target High conditions.

We began examination of the model by performing a number of contrast tests to investigate $\Delta F0$ benefit in conditions that were comparable to previous literature (i.e., where both sounds had all of their harmonics). We first performed contrast tests between 0 ST $\Delta F0$ and all other levels of $\Delta F0$ for cases in which the masker $F0$ was fixed at 80 Hz (i.e., in the Masker-All conditions). These tests revealed that increasing $\Delta F0$ from 0 to 3 ST (estimated mean difference $= 1.57$ dB, $\chi^2_1 = 10.19$, $p = 0.025$) improved SRTs, whereas increasing $\Delta F0$ from 0 to 12 ST (estimated mean difference $= -0.21$ dB, $\chi^2_1 = 0.19$, $p = 1.00$) or from 0 to 15 ST (estimated mean difference $= 0.62$ dB, $\chi^2_1 = 1.51$, $p = 1.00$) did not produce significant changes in SRTs. Next, we compared these effects with the corresponding effects when instead the target $F0$ was fixed at 80 Hz (i.e., in the Target-All conditions). In contrast to what was observed in Masker-All conditions, increasing $\Delta F0$ in the Target-All conditions from 0 to 3 ST (estimated mean difference $= 1.90$ dB, $\chi^2_1 = 15.85$, $p = 0.0014$), from 0 to 12 ST (estimated mean difference $= 1.87$ dB, $\chi^2_1 = 14.03$, $p = 0.0036$) and from 0 to 15 ST (estimated mean difference $= 4.81$ dB, $\chi^2_1 = 86.61$, $p < 0.001$) all improved SRTs. This pattern of results is generally consistent with prior literature on $\Delta F0$ benefit. In particular, the failure to find a significant 12 ST $\Delta F0$ in the Masker-All condition resembles the finding of Brokx and Nooteboom (1982) that a target $F0$ one octave above the masker $F0$ produced little $\Delta F0$ benefit. Also, a significant $\Delta F0$ benefit for a target $F0$ 3 ST above the masker $F0$ in a speech-shaped tonal masker has been reported previously (Leclère *et al.*, 2017).

Next, interaction contrast tests of the $\Delta F0$ benefit observed in Masker-All versus Target-All conditions confirmed that $\Delta F0$ benefit was significantly larger in Target-All than in Masker-All at 12 ST (estimated mean difference $= 2.08$ dB, $\chi^2_1 = 9.29$, $p = 0.039$) and 15 ST (estimated mean difference $= 4.19$ dB, $\chi^2_1 = 35.20$, $p < 0.001$) but not at 3 ST (estimated mean difference $= 0.33$ dB, $\chi^2_1 = 0.25$, $p = 1.00$). In other words, the benefit of a given $\Delta F0$ depended on the direction of the $\Delta F0$, with cases where the masker $F0$ was higher than the target $F0$ generally producing larger a $\Delta F0$ benefit than vice versa. Further interaction contrasts confirmed that this sign effect was larger for larger $\Delta F0$s: specifically, it was significantly larger for the 15 ST $\Delta F0$ than the 12 ST $\Delta F0$ (estimated mean difference $= 2.11$ dB, $\chi^2_1 = 8.74$, $p = 0.048$). Deroche *et al.* (2014a), using a $\Delta F0$ of 11 ST, reported a similar phenomenon, finding that $\Delta F0$ benefit was larger in cases where the masker $F0$ was higher than the target $F0$ than vice versa. However, at the octave $\Delta F0$, this sign effect offers us no insight into the relative roles of harmonic cancellation and spectral glimpsing. At the Masker-All octave $\Delta F0$ (i.e., when the target $F0$ is one octave above masker $F0$) spectral glimpsing performs poorly because every target harmonic is also a masker harmonic, while at the Target-All octave $\Delta F0$ (i.e., when the masker $F0$ is one octave about target $F0$) spectral glimpsing performs better because there is one target harmonic to glimpse between each masker harmonic. Harmonic cancellation performs similarly: at the Masker-All octave $\Delta F0$ a cancellation filter tuned to the masker $F0$ would eliminate every harmonic of the target, while at the Target-All octave $\Delta F0$ a cancellation filter tuned to the masker $F0$ would eliminate only every other harmonic of the target.

Next, we examined the effects of removing the even harmonics of the target (i.e., Target-All vs Target-Odd). Averaged across levels of $\Delta F0$, removing the harmonics of the target had no significant effect (estimated mean difference $= -0.45$ dB, $\chi^2_1 = 3.34$, $p = 0.68$). The difference between Target-All and Target-Odd at the 15 ST $\Delta F0$ with the unmodulated masker constituted the largest effect of removing the even harmonics of the target, but even this difference was not significant after correction for multiple comparisons (estimated mean difference $= -1.81$, $\chi^2_1 = 6.62$, $p = 0.14$). One might have expected removing the even harmonics of the target to have had an effect at the 15 ST $\Delta F0$ because, at this $\Delta F0$, many of the harmonics of the target which were furthest from masker harmonics (e.g., 4th harmonic, 6th harmonic, 8th harmonic) were the ones that were removed from the target. In contrast, at other $\Delta F0$s, the harmonics which were removed from the target by the Target-Odd manipulation were usually either overlapping with masker harmonics (between 0 and 12 ST $\Delta F0$) or often very close to masker harmonics (3 ST $\Delta F0$). Again, however, such a finding would not have helped us distinguish between spectral glimpsing and harmonic cancellation because the harmonics at the 15 ST $\Delta F0$, which were most readily glimpsed, would also have been the harmonics least distorted by a harmonic cancellation filter tuned to the masker $F0$, so both models would predict that removing the target's even harmonics at this $\Delta F0$ should be detrimental.

In an effort to determine which explanation could best account for the lack of a Masker-All octave $\Delta F0$ benefit, we examined the effects of removing the even harmonics of the masker (i.e., Masker-All vs Masker-Odd). In contrast to the effects of removing the even harmonics of the target, removing the even harmonics of the masker significantly improved SRTs at all $\Delta F0$s: 0 ST (estimated mean difference $= 3.91$ dB, $\chi^2_1 = 68.27$, $p < 0.001$), 3 ST (estimated mean difference $= 4.65$ dB, $\chi^2_1 = 92.14$, $p < 0.001$), 12 ST (estimated mean difference $= 5.39$ dB, $\chi^2_1 = 125.54$, $p < 0.001$), and 15 ST

(estimated mean difference = 5.31 dB, $\chi_1^2 = 118.48$, $p < 0.001$). The significant benefit of Masker-Odd over Masker-All at 12 ST $\Delta F0$ confirmed our hypothesis that removing the even harmonics of the masker when the target $F0$ was one octave above the masker $F0$ would improve speech intelligibility. In principle (and as discussed more thoroughly below in the Discussion section), this result is consistent with an explanation for the lack of a $\Delta F0$ benefit at the Masker-All octave $\Delta F0$ on the basis of spectral glimpsing but not on the basis of harmonic cancellation. The benefit of Masker-Odd over Masker-All at the 0 ST $\Delta F0$ has analogous implications. Interestingly, the benefit of the removing the masker's even harmonics appeared to grow larger at higher ΔF0s (see Fig. 7), but an interaction contrast of this effect at 0 ST $\Delta F0$ versus 15 ST $\Delta F0$ did not reveal significant differences (estimated mean difference = 1.40 dB, $\chi_1^2 = 4.42$, $p = 0.43$).

A contrast test between the 0 ST $\Delta F0$ Masker-Odd and 12 ST $\Delta F0$ Target-All conditions was also performed, as it provided an interesting test of spectral glimpsing. In this contrast, the target was the same in both conditions (80 Hz $F0$, with all its harmonics), while in the former condition the masker had the odd harmonics of an 80 Hz $F0$ and in the latter condition the masker had the even harmonics 80 Hz $F0$ (equivalently, all the harmonics of 160 Hz $F0$). In that sense, one might expect spectral glimpsing to produce similar results in both conditions because there were the same number of "gaps" in both cases and only their positions differed. However, this test confirmed that these two conditions differed significantly (estimated mean difference = 2.64 dB, $\chi_1^2 = 25.29$, $p < 0.001$). Examining their excitation patterns in Fig. 3 reveals that most of the spectral glimpsing opportunities in the 12 ST $\Delta F0$ Target-All condition were at the first target harmonic, whereas in 0 ST $\Delta F0$ Masker-Odd condition fairly deep spectral glimpses were distributed across the lower-order even harmonics of the target. This phenomenon may help explain the significant difference between these conditions.

Finally, the main effect of masker modulation confirmed that masker modulation affected SRTs, and the contrast test between modulated and unmodulated masker conditions revealed that masker modulation resulted in a small but significant improvement in SRTs (estimated mean difference = 0.97 dB, $\chi_1^2 = 16.60$, $p < 0.001$). The lack of interactions between masker modulation and other fixed effects in the model suggests that listeners were able to take advantage of temporal modulations in the masker without losing the ability to take advantage of other segregation cues like $\Delta F0$. This confirmed our hypothesis that masker modulation would produce an improvement in SRTs but have no interaction with other manipulations.

D. Discussion

The key novel finding of Experiment 2 was that listeners can exploit spectral glimpses in a harmonic complex tone introduced by removing that tone's even harmonics, even when the target and the tone share a pitch chroma. Our other findings were broadly consistent with the previous literature, and in particular the findings of Brokx and Nooteboom (1982). Of all the conditions tested in Experiment 2, the Masker-All conditions were perhaps most analogous to those of Experiment 1 in Brokx and Nooteboom (1982). In their experiment, Brokx and Nooteboom observed a $\Delta F0$ benefit when the target $F0$ was 3 ST above the masker $F0$ but little or no $\Delta F0$ benefit when the target $F0$ was 12 ST above the masker $F0$. Similarly, as can be seen in the right-hand panel of Fig. 7, we observed a significant $\Delta F0$ benefit when the target $F0$ was 3 ST above the masker $F0$ but no $\Delta F0$ benefit when the target $F0$ was 12 ST above the masker $F0$. Thus, despite using an HCT masker instead of a speech masker, we replicated this key finding of Brokx and Nooteboom.

As previously noted, there are competing potential explanations for failing to find a significant $\Delta F0$ benefit when the target $F0$ was one octave above the masker $F0$. First, it could be explained by reduced intelligibility of the target talker as the $F0$ is shifted away from its natural range. In other words, a change of 0 dB in SRT relative to the 0 ST condition might actually be interpreted as evidence of $\Delta F0$ benefit because if there truly were no $\Delta F0$ benefit one might expect a reduction in intelligibility relative to the 0 ST condition. Our design cannot in principle rule out this possibility; however, the small difference between the intelligibility of the target talker with an 80 Hz $F0$ versus with a 160 Hz $F0$ observed in Experiment 1 (approximately 1.5 dB change in SRT, see Fig. 4) suggests that if this explanation holds then the effect size of a $\Delta F0$ benefit in this case scenario ought to be less than 1–2 dB. This may not be an insignificant benefit, but in comparison to other $\Delta F0$ benefit effects observed in this experiment (e.g., the ∼2 dB improvement going from 0 ST to 3 ST $\Delta F0$ or the ∼5 dB improvement going from 0 ST to 15 ST $\Delta F0$ in Target-All), it could be considered fairly small.

Second, it could be explained by perceptual fusion between the target and masker due to their shared pitch chroma at the octave (Huron, 1991). However, as was revealed by our *post hoc* analyses, a significant octave $\Delta F0$ benefit *was* observed in Target Low conditions (see Fig. 7). This suggests that the octave $\Delta F0$ alone did not promote perceptual fusion of the target and masker. Furthermore, other powerful segregation cues were present (e.g., onset asynchrony), which likely prevented perceptual fusion of the two sounds.

Third, it could be explained by the failure of a cancellation mechanism due to the shared periodicity of the target and masker at this $\Delta F0$. The improvement in performance in Masker-Odd versus Masker-All at the octave $\Delta F0$, however, casts doubt on this explanation. The Masker-Odd condition eliminated spectral overlap between the voiced portions of the target and masker by removing the even harmonics of the lower-F0 sound but preserved their shared periodicity. If the lack of an octave $\Delta F0$ benefit in the Masker-All conditions could be attributed *entirely* to the shared periodicity of the target and masker interfering with a cancellation mechanism, this manipulation should have had little to no effect on SRTs. However, we found that this manipulation resulted in a large improvement in speech intelligibility (see Fig. 7). Thus, our results provide no evidence in favor of an explanation based on a *unique* harmonic cancellation mechanism

(de Cheveigné, 1993). At the same time, our results offer no evidence against a harmonic cancellation mechanism being employed at other $\Delta F0$s, or against a harmonic cancellation mechanism which is combined with, or applied after, a spectral glimpsing mechanism.

Fourth, it could be explained by spectral glimpsing. That is, when the target $F0$ is one octave above the masker $F0$ minimal opportunities are present to spectrally glimpse the target between masker harmonics, resulting in poor SRTs. We favor this explanation, as it accounts well not only for the lack of a Masker-All octave $\Delta F0$ benefit, but also for a number of other important features of our data. For instance, the beneficial effect of removing the masker's even harmonics when the target $F0$ was one octave above the masker $F0$ is consistent with an explanation based on spectral glimpsing because this manipulation introduced opportunities for spectral glimpsing in the masker (compare the corresponding panels of the stimuli EPs in Fig. 3). Similarly, this explanation is consistent with the finding that an octave $\Delta F0$ was beneficial in the Target-All condition but not in the Masker-All condition. In the Target-All octave $\Delta F0$ condition, one target harmonic was present between each masker harmonic (i.e., within each opportunity for a spectral glimpse in the masker), whereas in the Masker-All octave $\Delta F0$ condition, each target harmonic coincided with a masker harmonic. Hence, at the octave $\Delta F0$, more opportunities for spectral glimpses were available in Masker-All than in Target-All. The finding that a 15 ST $\Delta F0$ produced better SRTs when the target $F0$ was below the masker $F0$ than vice versa is also consistent with an explanation based on spectral glimpsing. In our stimuli EPs plotted in Fig. 3, it can be clearly seen that the higher $F0$ stimuli have deeper inter-peak dips than the lower $F0$ stimuli, suggesting that a higher $F0$ masker would generally offer more opportunities for spectral glimpsing than a lower $F0$ masker (Deroche *et al.*, 2014b,a).

As discussed above, modulating the masker with a speech envelope produced a small improvement in SRTs. While previous studies that imposed speech envelopes on noise maskers have shown similar benefits (Peters *et al.*, 1998; Qin and Oxenham, 2003; Freyman *et al.*, 2012), Leclère *et al.* (2017) imposed speech envelopes on a tonal masker and found improved SRTs only when the masker was intonated, and not when the masker was monotone. The origin of this discrepancy is unclear. One possibility involves differences in how the masker envelopes were extracted in the two studies, while another possibility is that the true effect size is nonzero but small enough that most studies on the scale of ours or that of Leclère *et al.* would have low power and would thus detect the effect unreliably. More generally, other types of envelope modulations imposed on tonal maskers have been shown to have no impact or even a detrimental impact on speech intelligibility, a finding which is thought to be related to the lack of inherent envelope fluctuations in tonal maskers (Stone *et al.*, 2012; Oxenham and Kreft, 2014).

V. GENERAL DISCUSSION

The present experiments replicated a number of key findings in the study of $\Delta F0$ benefit but also, via a novel manipulation of the spectral structure of the masker, provided evidence that listeners are readily able to segregate a target and masker separated by an octave $\Delta F0$, provided that spectral overlap between target and masker is reduced. We argued that spectral glimpsing provided a parsimonious qualitative account of the key features of our data and that, in the case of the lack of a $\Delta F0$ benefit when the target $F0$ was one octave above the masker $F0$, our results were inconsistent with an explanation based on a unique harmonic cancellation mechanism. However, this should not be interpreted as implying that harmonic cancellation is not used by the auditory system at other $\Delta F0$s or in general. Instead, it merely shows that a model of $\Delta F0$ benefit based on harmonic cancellation that does not explicitly incorporate frequency selectivity cannot account for our data at the octave $\Delta F0$. This issue was anticipated by de Cheveigné (1993), who wrote: "we do not mean to imply that peripheral frequency analysis plays no role, or only a minor role, in harmonic sound separation." Although an extensive examination of how to integrate frequency selectivity into a model of $\Delta F0$ benefit based on harmonic cancellation is beyond the scope of this paper, one simple possibility might be to selectively apply the cancellation filter to the outputs of auditory filters that are dominated by the masker [i.e., little representation of the target periodicity is present, or the signal-to-noise ratio (SNR) is poor]. Thus, the outputs of auditory filters with a good SNR would be left unaffected while the SNR at outputs of auditory filters with unfavorable SNRs before processing might be improved by cancellation.

As was discussed in the introduction, HI listeners experience less $\Delta F0$ benefit than NH listeners (Summers and Leek, 1998). This is also true for cochlear-implant (CI) listeners (Stickney *et al.*, 2004) and for NH listeners listening to vocoded stimuli (Qin and Oxenham, 2005). These findings have generally been attributed to the poor spectral resolution caused by broadened auditory filters in the case of HI listeners and a low number of analysis bands in the case of CI/vocoder listeners. Diminished spectral resolution could also impair $\Delta F0$ benefit because it reduces the ability of listeners to take advantage of spectral glimpses. Given the variety of situations in which spectral glimpsing appears to have played an important role in sound segregation in the present experiments, this explanation seems tenable. Diminished spectral resolution could also impair $\Delta F0$ benefit by compromising an $F0$-guided segregation mechanism. The $F0$ estimation stage of such a mechanism might be compromised by reducing access to spectrally resolved low-order harmonics which play an important role in pitch perception (Oxenham, 2018), while the segregation stage of the mechanism might be compromised by reducing the extent to which the competing sounds can be decomposed into separate spectral channels which can be operated on independently. The present experiments, other than by providing evidence against cancellation being the only segregation mechanism in question, provide little insight into this possibility. However, despite continued uncertainties, the multiple ways in which diminished spectral resolution might limit $\Delta F0$ benefit suggest that implementing strategies to improve spectral resolution for listeners with impaired hearing ought to be a key objective for the development of auditory prostheses and

interventions. For example, in the case of CIs, multipolar stimulation strategies may improve spectral resolution and help restore access to $\Delta F0$ benefit (Berenstein *et al.*, 2008; Smith *et al.*, 2013), although recent studies suggest that the fine spectral resolution required to restore $F0$ perception may be beyond current CIs, even with novel stimulation strategies (Mehta and Oxenham, 2017).

We conclude by proposing a few experiments that could strengthen the present results and shed further light on the nature of $\Delta F0$ benefit. First, an extension of the experiment to include a speech masker could help elucidate the role of $\Delta F0$ benefit when listening to more realistic maskers. Second, a replication of the experiment with a female talker with her $F0$ shifted down, rather than a male talker with his $F0$ shifted up, could help determine more clearly to what extent our results were confounded by variations in talker intelligibility as a function of $F0$ manipulations. Finally, exploring the relative impact on target intelligibility of different types of masker harmonicity manipulations could help further clarify the role of a harmonic cancellation mechanism in $\Delta F0$ benefit.

ACKNOWLEDGMENTS

This work was supported by the following funding sources: UMN College of Liberal Arts Graduate Fellowship awarded to D.R.G., UMN Department of Psychology Summer Graduate Fellowship awarded to D.R.G., NIH R01 DC005216 awarded to A.J.O, and NSF NRT-UtB1734815.

APPENDIX: ANF SIMULATIONS

The vowels used to generate Figs. 1 and 2 were synthesized via the implementation of the Klatt synthesizer (Klatt and Klatt, 1990) provided in the UR EAR toolbox (Bruce *et al.*, 2018). All voicing and source parameters were set to typical values and only $F0$, $F1$, $F2$, and $F3$ were manipulated. The formant frequencies were chosen to achieve desired vowel qualities according to averages published in Hillenbrand *et al.* (1995). Vowels synthesized with the 100 Hz $F0$ used formants appropriate for men while vowels synthesized with higher F0s used formants appropriate for women.

Once synthesized at 10 kHz, the vowels were resampled to 100 kHz and set to an rms level of 70 dB SPL. The vowels were then processed with the auditory nerve model of Bruce *et al.* (2018). 170 ANFs were simulated at 34 center frequencies (CFs) logarithmically spaced from 500 Hz to 16 kHz (5 ANFs per CF). A mixture of 20% low spontaneous rate, 20% medium spontaneous rate, and 60% low spontaneous rate fibers were simulated. The frequency tuning parameters used were those which corresponded to cat frequency tuning. For each vowel or vowel mixture, the responses to a 500 ms segment followed by 1 s of silence were simulated 100 times. The other model parameters were maintained at default values from the code provided by the authors. After the simulations were complete, where stated each spike train simulation was processed by the neural cancellation filter described in de Cheveigné (1993) with a time window of 0.1 ms. An autocoincidence (AC) histogram with a bin width of 0.1 ms was then computed for each spike train simulation according to the procedure described in Ruggero (1973). Finally, the AC histograms were pooled across repetitions, fibers, and CFs to generate pooled, or summary, AC histograms.

Abdi, H. (**2010**). "Holm's sequential Bonferroni procedure," in *Encyclopedia of Research Design*, edited by N. Salkind (Sage, Thousand Oaks, CA), pp. 573–577.

Assmann, P. F. (**1998**). "Fundamental frequency and the intelligibility of competing voices," in *Proceedings of the 14th International Congress of Phonetic Science*, August 1–7, San Francisco, CA, pp. 179–182.

Assmann, P. F., and Nearey, T. M. (**2008**). "Identification of frequency-shifted vowels," J. Acoust. Soc. Am. **124**, 3203–3212.

Assmann, P. F., Nearey, T. M., and Dembling, S. (**2006**). "Effects of frequency shifts on perceived naturalness and gender information in speech," in *Proceedings of the Ninth International Conference on Spoken Language Processing*, September 17–21, Pittsburg, PA, pp. 889–892.

Assmann, P. F., and Paschall, D. D. (**1998**). "Pitches of concurrent vowels," J. Acoust. Soc. Am. **102**, 1150–1160.

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.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (**2012**). "Random effects structure for confirmatory hypothesis testing: Keep it maximal," J. Mem. Lang. **68**, 255–278.

Barreda, S., and Assmann, P. F. (**2018**). "Modeling the perception of children's age from speech acoustics," J. Acoust. Soc. Am. **143**, EL361–EL366.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. **67**, 1–48.

Berenstein, C. K., Mens, L. H. M., Mulder, J. J. S., and Vanpoucke, F. J. (**2008**). "Current steering and current focusing in cochlear implants: Comparison of monopolar, tripolar, and virtual channel electrode configurations," Ear Hear. **29**, 250–260.

Bernstein, J. G. W., and Oxenham, A. J. (**2003**). "Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number?," J. Acoust. Soc. Am. **116**, 3323–3334.

Bird, J., and Darwin, C. J. (**1997**). "Effects of a difference in fundamental frequency in separating two speech messages," in *Psychophysics and Physiology of Hearing*, edited by A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis (Whurr Publishers Ltd., New York), pp. 263–270.

Boersma, P., and Weenink, D. (**2019**). "Praat: Doing phonetics by computer [computer program]," http://www.praat.org (Last viewed May 4, 2019).

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

Bruce, I. C., Erfani, Y., and Zilany, M. S. A. (**2018**). "A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites," Hear. Res. **360**, 40–54.

Cariani, P. (**2003**). "Recurrent timing nets for auditory scene analysis," in *Proceedings of the International Joint Conference on Neural Networks*, July 20–24, Portland, OR.

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., 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. M. H. (**1997**). "Concurrent vowel identification. 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**). "Roles of the target and masker fundamental frequencies in voice segregation," J. Acoust. Soc. Am. **136**, 1225–1236.
Deroche, M. L. D., Culling, J. F., Chatterjee, M., and Limb, C. J. (**2014b**). "Speech recognition against harmonic and inharmonic complexes: Spectral dips and periodicitby," J. Acoust. Soc. Am. **135**, 2873–2884.
De Rosario-Martinez, H. (**2015**). "phia: Post-hoc interaction analysis," https://cran.r-project.org/package=phia (Last viewed May 4, 2019).
Freyman, R. L., Griffin, A. M., and Oxenham, A. J. (**2012**). "Intelligibility of whispered speech in stationary and modulated noise maskers," J. Acoust. Soc. Am. **132**, 2514–2523.
Frick, R. W. (**1985**). "Communicating emotion: The role of prosodic features," Psychol. Bull. **97**, 412–429.
Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.
Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.
Houtsma, A. J. M., and Smurzynski, J. (**1990**). "Pitch identification and discrimination for complex tones with many harmonics," J. Acoust. Soc. Am. **87**, 304–310.
Huron, D. (**1991**). "Tonal consonance versus tonal fusion in polyphonic sonorities," Music Percept. **9**, 135–154.
Kawahara, H. (**1997**). "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited," in *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 21–24, Munich, Germany.
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," Speech Commun. **27**, 187–207.
Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857.
Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2017**). "lmerTest package: Tests in linear mixed effects models," J. Stat. Softw. **82**, 1–26.
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.
Madsen, S. M. K., Whiteford, K. L., and Oxenham, A. J. (**2017**). "Musicians do not benefit from differences in fundamental frequency when listening to speech in competing speech backgrounds," Sci. Rep. **7**, 12624.
Meddis, R., and Hewitt, M. J. (**1993**). "Modeling the identification of concurrent vowels with different fundamental frequencies," J. Acoust. Soc. Am. **91**, 233–245.
Mehta, A. H., and Oxenham, A. J. (**2017**). "Vocoder simulations explain complex pitch perception limitations experienced by cochlear implant users," J. Assoc. Res. Otolaryngol. **18**, 789–802.
Micheyl, C., Bernstein, J. G. W., and Oxenham, A. J. (**2006**). "Detection and F0 discrimination of harmonic complex tones in the presence of competing tones or noise," J. Acoust. Soc. Am. **120**, 1493–1505.
Micheyl, C., Keebler, M. V., and Oxenham, A. J. (**2010**). "Pitch perception for mixtures of spectrally overlapping harmonic complex tones," J. Acoust. Soc. Am. **128**, 257–269.
Micheyl, C., and Oxenham, A. J. (**2010**). "Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings," Hear. Res. **266**, 36–51.
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.
Moore, B. C. J., and Glasberg, B. R. (**1987**). "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns," Hear. Res. **28**, 209–225.
Moulines, E., and Charpentier, F. (**1990**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun. **9**, 453–467.
Oxenham, A. J. (**2008**). "Pitch perception and auditory stream segregation: Implications for hearing loss and cochlear implants," Trends Amplif. **12**, 316–331.
Oxenham, A. J. (**2018**). "How we hear: The perception and neural coding of sound," Annu. Rev. Psychol. **69**, 27–50.
Oxenham, A. J., and Kreft, H. A. (**2014**). "Speech perception in tones and noise via cochlear implants reveals influence of spectral resolution on temporal processing," Trends Hear. **18**, 1–14.
Oxenham, A. J., and Simonson, A. M. (**2009**). "Masking release for low- and high-pass filtered speech in the presence of noise and single-talker interference," J. Acoust. Soc. Am. **125**, 457–468.
Peters, R. W., Moore, B. C. J., and Baer, T. (**1998**). "Speech reception threshold in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," J. Acoust. Soc. Am. **103**, 577–587.
Plack, C. J., Oxenham, A. J., Fay, R. R., and Popper, A. N. (**2005**). *Pitch: Neural Coding and Perception* (Springer, New York).
Qin, M. K., and Oxenham, A. J. (**2003**). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," J. Acoust. Soc. Am. **114**, 446–454.
Qin, M. K., and Oxenham, A. J. (**2005**). "Effects of envelope-vocoder processing on F0 discrimination and concurrent-vowel identification," Ear Hear. **26**, 451–460.
Rothauser, E. H., Chapman, W. D., Guttmann, 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.
Ruggero, M. A. (**1973**). "Response to noise of auditory nerve fibers in the squirrel monkey," J. Neurophysiol. **36**, 569–587.
Shackleton, T. M., and Carlyon, R. P. (**1994**). "The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination," J. Acoust. Soc. Am. **95**, 3529–3540.
Smith, Z. M., Parkison, W. S., and Long, C. J. (**2013**). "Multipolar current focusing increasing spectral resolution in cochlear implants," in *Proceedings of the 35th IEEE International Conference on Engineering in Medicine and Biology Society*, July 3–7, Piscataway, NJ, pp. 2796–2799.
Stickney, G. S., Zeng, F.-G., Litovsky, R., and Assmann, P. (**2004**). "Cochlear implant speech recognition with speech maskers," J. Acoust. Soc. Am. **116**, 1081–1091.
Stone, M. A., Füllgrabe, C., and Moore, B. C. J. (**2012**). "Notionally steady background noise acts primarily as a modulation masker of speech," J. Acoust. Soc. Am. **132**, 317–326.
Summers, V., and Leek, M. R. (**1998**). "F0 processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss," J. Speech Lang. Hear. Res. **41**, 1294–1306.
Wang, J., Baer, T., Glasberg, B. R., Stone, M. A., Ye, D., and Moore, B. C. J. (**2012**). "Pitch perception of concurrent harmonic tones with overlapping spectra," J. Acoust. Soc. Am. **132**, 339–356.