Does fundamental-frequency discrimination measure virtual pitch discrimination?

Christophe Micheyl,^{a)} Kristin Divis, David M. Wroblewski, and Andrew J. Oxenham
*Auditory Perception and Cognition Laboratory, Department of Psychology, University of Minnesota,
Minneapolis, Minnesota 55455*

(Received 4 November 2009; revised 16 July 2010; accepted 19 July 2010)

Studies of pitch perception often involve measuring difference limens for complex tones (DLCs) that differ in fundamental frequency (F0). These measures are thought to reflect F0 discrimination and to provide an indirect measure of subjective pitch strength. However, in many situations discrimination may be based on cues other than the pitch or the F0, such as differences in the frequencies of individual components or timbre (brightness). Here, DLCs were measured for harmonic and inharmonic tones under various conditions, including a randomized or fixed lowest harmonic number, with and without feedback. The inharmonic tones were produced by shifting the frequencies of all harmonics upwards by 6.25%, 12.5%, or 25% of F0. It was hypothesized that, if DLCs reflect residue-pitch discrimination, these frequency-shifted tones, which produced a weaker and more ambiguous pitch than would yield larger DLCs than the harmonic tones. However, if DLCs reflect comparisons of component pitches, or timbre, they should not be systematically influenced by frequency shifting. The results showed larger DLCs and more scattered pitch matches for inharmonic than for harmonic complexes, confirming that the inharmonic tones produced a less consistent pitch than the harmonic tones, and consistent with the idea that DLCs reflect F0 pitch discrimination. © 2010 Acoustical Society of America. [DOI: 10.1121/1.3478786]

PACS number(s): 43.66.Hg, 43.66.Fe [CJP]

Pages: 1930–1942

I. INTRODUCTION

Harmonic complex tones usually evoke a pitch that corresponds to the fundamental frequency (F0) of which their component frequencies are all integer multiples. This pitch is referred to variously as “residue pitch,” “virtual pitch,” “periodicity pitch,” “fundamental pitch,” or “complex pitch” (Schouten, 1940; de Boer, 1956; Schouten *et al.*, 1962; Terhardt, 1970; Cudahy, 1975). It plays a crucial role in the perception of music, where it defines musical notes, and where changes in pitch over time are used to convey melody (e.g., McDermott and Oxenham, 2008), and in speech perception, where it can convey information involving prosody, speaker identity and, in tone languages, lexical information (e.g., Cooper and Sorensen, 1981).

Over the past fifty years a large number of psychoacoustical experiments have been devoted to the perception of virtual pitch and its underlying mechanisms (for reviews, see: Moore, 2003; Plack *et al.*, 2005). Of most common interest has been the pitch and the pitch strength produced by certain combinations of pure tones or other periodic, or quasi-periodic sounds. Pitch is usually measured in “pitch matching” experiments, where listeners are asked to adjust the frequency of a tone (usually, a pure tone), so that its pitch matches that of the sound of interest (de Boer, 1956; Flanagan and Guttman, 1960; Schouten *et al.*, 1962; Patterson, 1973; Patterson and Wightman, 1976; Goldstein *et al.*, 1978; Moore *et al.*, 1992). Pitch *strength* can be measured directly in rating experiments (e.g., Fastl and Stoll, 1979), or in

matching experiments (e.g., Wiegand *et al.*, 1998). However, by far the most commonly used measure of pitch perception is the discrimination threshold. In a typical “pitch discrimination” experiment, listeners are presented with harmonic complex tones that differ in F0, and asked to indicate which tone had a higher (or lower, or just different) pitch. Thresholds measured in this task are traditionally referred to as “difference limens for F0” (DLF0s), or “pitch discrimination thresholds.” These thresholds are commonly assumed to measure the precision with which virtual pitch is represented in the auditory system (e.g., Goldstein, 1973; Srulovicz and Goldstein, 1983), and to be inversely related to pitch strength.

Faulkner (1985) challenged this assumption. He pointed out that when the tones being compared contain corresponding harmonics,¹ performance may be based on comparisons of the frequencies (or subjectively, pitches) of these harmonics, rather than on comparisons of F0 (or subjectively, fundamental pitch). Consistent with this, and with earlier results (Ritsma, 1963), Faulkner found that “DLF0s” were larger for tones that did not contain corresponding harmonics than for tones that did. Moreover, using a multiple-band statistical summation model (Green, 1958), he showed that the thresholds measured in conditions with corresponding harmonics could be accounted for quantitatively based on the discriminability of the individual component frequencies, as estimated from empirical measures of difference limens for frequency (DLFs) with pure tones. Based on these results, Faulkner concluded that “residue pitch discriminations are not used for pitch discrimination between sounds with corresponding components; rather, comparisons based on residue pitch are only employed where there are no common re-

^{a)}Author to whom correspondence should be addressed. Electronic mail: cmicheyl@umn.edu

solved components in the signal to be discriminated” and that “...experiments, in which signals with common components are compared, may not teach us anything about the perception of residue pitch.”

Moore and Glasberg (1990) questioned Faulkner’s conclusions on several grounds. First, they argued that the larger DLF0s for complex tones with no harmonics in common were due to the *timbre* differences between such tones interfering with listeners’ perception of *pitch*.² Consistent with this, they found that DLF0s for complex tones that had several low-rank harmonics in common (thereby allowing direct comparisons of component frequencies) were smaller when the tones also had common upper-rank harmonics (and thus, a similar timbre) than when they did not. Second, Moore and Glasberg measured difference limens for complex tones (DLCs),³ which were made inharmonic by shifting the frequencies of odd harmonics upward by $0.15 \times F_0$, and the frequencies of even harmonics downward by $0.15 \times F_0$. They reasoned that, since the residue pitch of inharmonic tones is ambiguous, if discrimination is based on residue pitch, DLCs should be larger for such tones than for harmonic tones, which have a well-defined pitch. On the other hand, if discrimination depends on component pitches, these inharmonic tones should yield the same DLCs as harmonic tones. The results showed larger DLCs for the inharmonic tones than for harmonic tones. Based on these results, Moore and Glasberg concluded that “DLCs for complex tones with common harmonics depend on residue pitch comparisons, rather than on comparisons of the pitches of the partials.”

Given the predominant role of DLC measurements in pitch-perception research, it seems worthwhile to examine these conclusions further. First, it is important to note that Moore and Glasberg’s first argument, involving the potentially distracting effects of timbre differences, does not necessarily address the question of how pitch judgments are made when no timbre differences are present. Second, timbre differences, such as those produced by changing the upper harmonic numbers present in a stimulus, could interfere with the ability of listeners to discriminate frequency differences between individual components as much as they interfere with listeners’ ability to discriminate residue pitch. Moore and Glasberg’s second argument, involving a comparison of DLCs for harmonic and inharmonic complex tones, provides potentially stronger evidence that DLCs involve residue pitch comparisons. However, the existing data are relatively sparse on this question. Of the three subjects tested by Moore and Glasberg (1990), only one showed consistently larger thresholds for inharmonic tones than for harmonic tones; one showed no clear differences, and the other showed a large difference (roughly a factor of 2) only for one of the four F_0 s tested (see Fig. 10 on p. 2174 in Moore and Glasberg’s article). In addition, the interpretation of these results is complicated somewhat by the fact that the inharmonic tone was produced by shifting the frequencies of odd harmonics upward and those of even harmonics downward. This resulted in an irregular pattern of frequency spacing. The greater proximity of some components in the inharmonic tone and/or the lack of spectral regularity could adversely affect listeners’ ability to accurately extract, compare, or combine the

pitches of individual components (Roberts and Bregman, 1991; Roberts and Bailey, 1996; Roberts, 1998). Finally, the complexes had equal-amplitude components (i.e., rectangular spectral envelopes), and the lowest harmonic number was not roved. Under such circumstances, discrimination could be based on the frequency of the lowest harmonic, or on some other aspect of the stimulus that co-varied with this frequency (e.g., the spectral center of gravity of the complex), without necessarily involving residue pitch. While such cues would have been present for both harmonic and inharmonic tones, it is unclear whether listeners’ reliance on them was the same in both conditions. It could be, for instance, that performance in the harmonic conditions was based on residue pitch, whereas performance in the inharmonic conditions was based on the frequency of the lowest component.

The aim of the current study was to further investigate the influence of inharmonicity on DLCs, with several differences from the most comparable earlier experiment of Moore and Glasberg (1990). First, to avoid potential issues of limited sample size, the experiments involved a minimum of seven listeners; the main experiment involved nine listeners. Second, the lowest harmonic number was randomized across stimulus presentation to prevent listeners from relying on the frequency of the lowest harmonic, or on a timbre (e.g., spectral-centroid) cue; in the main experiment, the roving of the lowest harmonic number was combined with interleaved tracking in such a way that the influence of lowest-frequency or timbre differences on DLCs could be estimated, and corrected for. Third, the components were shifted by the same amount within a given inharmonic complex, providing spectral regularity, and different size shifts were tested. Finally, in addition to discrimination thresholds, pitch matches were also collected using the same stimuli and same listeners to provide a more direct test of residue pitch accuracy.

II. EXPERIMENT 1

A. Methods

1. Stimuli

The stimuli were harmonic and inharmonic complex tones. The harmonic tones consisted of harmonics 2–7 or 3–8 of a reference F_0 , $F_{0\text{ref}}$, that was nominally equal to 400 Hz. The 400-Hz F_0 falls roughly in the middle of the musical pitch range, being close to the F_0 of A above middle C (440 Hz), which is commonly used as a standard by musicians. The actual F_0 s of the tones presented on a trial were equal to $F_{0\text{ref}}(1 + \Delta F_0/100)^{-1/2}$ for the higher- F_0 tone, and $F_{0\text{ref}}(1 - \Delta F_0/100)^{-1/2}$ for the lower- F_0 tone, so that the geometric mean of the two F_0 s presented on a trial was equal to $F_{0\text{ref}}$. The F_0 difference, ΔF_0 (expressed in % of the lower F_0), was varied adaptively during the course of a block of trials, as explained in the following section. On each trial, $F_{0\text{ref}}$ (in Hz), was drawn at random from a 20%-wide uniform probability distribution centered on the nominal F_0 (400 Hz). This across-trial “roving” of F_0 was introduced to discourage listeners from forming a representation of the standard in long-term memory, and then comparing incoming tones against this fixed representation rather than against each

other. The inharmonic tones were obtained by shifting all components in the harmonic tone upward by ΔF Hz. In this experiment, the frequency shift, ΔF , was equal to 25% of $F_{0\text{ref}}$ on each trial.

The tones were 400 ms in duration each, including 20-ms on and off (raised-cosine) ramps. The starting phases of the harmonics were randomized independently over a 0–360° range with uniform distribution, and the tones had a level of 50 dB SPL per component, corresponding to an overall level of 57.8 dB SPL. The tones were presented in pink noise with a spectrum level of 20 dB SPL/Hz at 800 Hz. This level was selected to mask any potential distortion products generated by the complexes. The noise started 400 ms before the first tone on a trial and ended 400 ms after the offset of the second tone on a trial.

2. Procedure

DLCs were measured using a two-interval, two-alternative forced-choice (2I-2AFC) paradigm and a transformed two-down, one-up procedure that tracks the 70.7%-correct point on the psychometric function (Levitt, 1971). On each trial, two tones were presented, which differed in F_0 as indicated above. The higher- F_0 tone was played either first or second, with equal probability. The task of the listener was to indicate which of the two observation intervals contained the higher- F_0 tone. The two observation intervals were marked on the computer screen. Listeners gave their responses by pressing “1” or “2” on a computer keyboard. At the beginning of an adaptive “run” (block of trials), the ΔF_0 was set to 4% of $F_{0\text{ref}}$. It was increased following each incorrect response and decreased following two consecutive correct responses. The amount by which ΔF_0 was increased or decreased was equal to a factor of 4 until the first reversal in the direction of the change in the tracking variable from “up” to “down.” This factor was reduced to 2 after the second up-to-down reversal, and to $\sqrt{2}$ after the third up-to-down reversal. If the adaptive procedure called for a ΔF_0 greater than 4% or less than 0.001%, the value of ΔF_0 was set to the corresponding limit and the procedure continued. A provision was made in the program, such that if this happened more than four times within a run, the run would have been automatically terminated, and no threshold would have been recorded for that run; however, in practice, this never happened. The adaptive tracking procedure stopped after a total of four reversals with the smallest step size were obtained. Threshold was computed as the geometric mean of the ΔF_0 values at the last four reversals.

Two adaptive tracks were randomly interleaved on each run. On one track, the lower- F_0 complex consisted of harmonics 2–7 while the higher- F_0 complex consisted of harmonics 3–8. The complex with a lowest harmonic number, N , of 3 was perceived as having a brighter timbre than the complex with an N of 2. On such tracks, therefore, timbre was consistent with pitch: the lower-pitch sound had a duller timbre, while the higher-pitch sound had a brighter timbre. In the other track, it was the opposite, and timbre was inconsistent with pitch. These two types of tracks are referred to as “consistent” and “inconsistent” tracks, respectively. Listeners were instructed that they should focus on the pitch of the

sounds, and try to ignore timbre as much as possible. Thresholds were computed separately for the two tracks, based on the ΔF_0 values on the last four turn-points of the adaptive staircase procedure within each track. Therefore, two threshold estimates were obtained for each adaptive run; one for each track. For the inharmonic tones, the same procedure was used, except that the frequencies of the components were shifted upward by a constant amount (ΔF) in Hz, as explained earlier.

Listeners performed between two and four blocks of this experiment. Each block involved six threshold measurements (three for each of the two tracks described above) in each condition. Two stimulus conditions were tested: harmonic tones and inharmonic tones. In addition, two “feedback” conditions were tested. In the “feedback” condition, a message (“correct” or “wrong”) was displayed on the computer screen following each response. In the “no-feedback” condition, no such message was displayed. The rationale for including a no-feedback condition was that we were interested in determining whether, and to what extent, listeners’ ability to follow selectively residue-pitch changes and ignore irrelevant timbre changes in the harmonic and shifted conditions was dependent upon on the provision of trial-wise feedback.

The four resulting combinations of conditions (harmonic or inharmonic combined or not with feedback) were tested in separate blocks of trials in completely randomized order. The rationale for randomizing the testing order of the different conditions was to avoid a systematic influence of presentation order on the results.

3. Listeners

Listeners were selected based on their results in a preliminary test during which they performed a minimum of three runs of the interleaved tracking procedure with feedback, using the same harmonic complex tones as in the main experiment.⁴ The objective of this preliminary test was to exclude from the study listeners who were unable to perform the basic discrimination task (with roving of the lowest harmonic number) sufficiently well, for their data to be usable. Listeners who produced DLCs larger than 2% on at least one of the two adaptive tracks—indicating that they were not very accurate discriminators, or had trouble ignoring the timbre differences caused by roving the lowest harmonic number—were not retained for further participation. This was the case for five out of the sixteen listeners who took part in the screening test. Of the remaining eleven listeners, nine (aged 19–27) completed the experiment. All had normal hearing, defined as pure-tone thresholds of 15 dB HL or less at octave frequencies between 500 and 8000 Hz. The listeners were 5 females and 4 males with a range of 7–17 years of musical training (mean=10.6). Four played the piano. The others listed their main instrument as one of the following: voice, bass clarinet, saxophone, or percussion. During testing, listeners were seated inside a double-walled sound-attenuating booth (IAC). Participants provided written informed consent prior to inclusion in the study and were paid for their participation.

FIG. 1. Mean DLCs in Experiment 1. The open bars show thresholds estimated based on trials in which the F0 and the lowest harmonic number moved in the same direction (i.e., consistent tracks). The shaded bars show thresholds estimated based on trials in which the F0 and the lowest harmonic number moved in opposite directions (i.e., inconsistent tracks). Thresholds are expressed as a percentage of the frequency spacing between consecutive harmonics. For the harmonic tones, this spacing was equal to the F0 of the lower-F0 complex. For the frequency-shifted (inharmonic) tones, it was equal to the F0 of the lower-F0 complex before the constant upward frequency shift was applied. The error bars indicate geometric standard errors above the geometric mean. To facilitate reading, the values of the mean DLCs are also shown on each bar.

4. Apparatus

A Madsen Conera™ Diagnostic Audiometer (GN Otopneumetrics, A/S) was used for pure-tone audiometry. During the experiments proper, stimulus presentation and response collection were controlled using the AFC software package (Stefan Ewert, Universität Oldenburg) under Matlab (The MathWorks, Inc.). The stimuli were generated digitally and played out via a soundcard (LynxStudio L22) with 24-bit resolution and a sampling frequency of 32 kHz. They were presented monaurally to the listener via Sennheiser HD 580 headphones.

B. Results and discussion

Figure 1 shows the (geometric) mean DLCs across the nine listeners who took part in this experiment. A three-way (harmonicity $\times$ track $\times$ feedback) repeated-measure analysis of variance (RMANOVA) on these log-transformed DLCs showed no significant main effect of harmonicity [$F(1,8) = 0.02$, $p = .902$], but it showed a significant main effect of both track (i.e., a difference between “consistent” and “inconsistent” tracks) [$F(1,8) = 6.95$, $p = 0.030$] and feedback [$F(1,8) = 6.52$, $p = 0.034$], as well as a significant harmonicity $\times$ track interaction [$F(1,8) = 7.59$, $p = 0.025$]. This interaction reflects the observation that while DLCs for the harmonic tones were roughly the same on the two tracks, DLCs for the shifted tones were considerably smaller (by more than an order of magnitude on average) for the consistent tracks than for the inconsistent tracks. A two-way (feedback $\times$ track) RMANOVA on the DLCs measured with just the inharmonic tones showed a significant main effect of track [$F(1,8) = 9.32$, $p = 0.016$], with no significant effect of feedback [$F(1,8) = 3.47$, $p = 0.099$], or interaction [$F(1,8) = 1.99$, $p = 0.196$]. The same analysis on the DLCs measured using harmonic tones showed no significant main effect of

track [$F(1,8) = 1.63$, $p = 0.238$], no significant effect of feedback [$F(1,8) = 1.15$, $p = 0.316$], and no significant interaction [$F(1,8) = 0.83$, $p = 0.389$].

Fisher LSD tests comparing thresholds measured with and without feedback, which were performed separately for harmonic and shifted tones, and for each track, failed to show any statistically significant difference. Thus, it appears that the main effect of feedback in the overall RMANOVA was not driven by a particular condition. Surprisingly, on average across tracks and harmonicity conditions, DLCs measured with feedback (0.34%) were slightly larger than DLCs measured without feedback (0.28%). We have no explanation for this counter-intuitive finding. However, the important point is that very similar effects of harmonicity and track were observed regardless of whether feedback was provided or not: in both cases, DLCs were significantly larger on inconsistent than on consistent tracks for inharmonic tones, whereas for harmonic tones, DLCs measured on consistent and inconsistent tracks were similar.

The finding of larger DLCs on inconsistent than on consistent tracks for inharmonic tones indicates that listeners were inclined to choose the observation interval that contained the complex with the higher lowest-harmonic number more often than the other interval. A likely explanation for this effect is that these inharmonic tones evoked more ambiguous (multiple), less salient, and more variable residue pitches than the harmonic tones. This presumably made it difficult for listeners to focus on residue pitch, and increased their reliance on other cues such as brightness or, more generally, on cues related to the frequency of the lowest harmonic or spectral center of gravity, but unrelated to residue pitch. However, it is clear that listeners’ responses were not based *entirely* on such cues. Otherwise, thresholds would have been close to the largest $\Delta F0$ that was allowed by the tracking procedure (4%) on inconsistent tracks and close to the smallest allowed $\Delta F0$ (0.001%) on consistent tracks. Instead, thresholds were only around 1.35% on average for inconsistent tracks and around 0.07% on average for consistent tracks. Thus, listeners had smaller (i.e., better) thresholds on inconsistent tracks than they should have if they had been relying solely on timbre or lowest-harmonic cues, even in conditions where no trial-by-trial feedback was provided. This indicates that listeners were relying to some extent on cues unrelated to the frequency of the lowest harmonic—either residue pitch or, perhaps, comparisons of individual corresponding components higher than the lowest (e.g., comparing the fourth harmonic in one tone with the fourth harmonic in the other tone).

All listeners received feedback in half the conditions tested, and it is possible that their performance in the absence of feedback was still influenced by the feedback received in earlier runs. It may therefore be that the effect of timbre on responses could have been greater in some listeners if they had not been provided with any feedback at all. Based on earlier results, showing that the influence of timbre on pitch judgments varies greatly between individuals (Smooenburg, 1970; Laguiton *et al.*, 1998; Schneider *et al.*, 2005; Seither-Preisler *et al.*, 2007), actual effects in the complete absence of feedback would likely have been subject dependent.

FIG. 2. Mean “unbiased” DLCs computed based on the results of consistent and inconsistent tracks in Experiment 1. The error bars indicate geometric standard errors above the geometric mean. To facilitate reading, the values of the mean DLCs are also shown on each bar.

To remove the influence of “timbre” bias on DLCs, we used an analysis inspired by previous articles (Oxenham and Buus, 2000; Micheyl and Oxenham, 2004). This analysis combines the thresholds measured on inconsistent and consistent tracks to compute “unbiased” thresholds, which are estimates of the thresholds that would have been measured, had harmonic number not had any influence on the listeners’ responses. The details are described in Appendix A. Here, we simply give the equation that was used to compute the unbiased thresholds:

$$DLC_u = 100 \left[\sqrt{\left(1 + \frac{DLC^+}{100}\right) \left(1 + \frac{DLC^-}{100}\right)} - 1 \right]. \quad (1)$$

In this equation, DLC^+ and DLC^- denote the DLCs (in %) measured on consistent and inconsistent tracks, respectively, and DLC_u is the “unbiased” DLC (in %). This equation was applied to each pair of measured DLCs within a given run, and the results were averaged (geometrically) within and across listeners.

The resulting mean unbiased DLCs, which are shown in Fig. 2, were larger by a factor of about 2, on average, for the inharmonic complexes than for the harmonic complexes. The statistical significance of this effect was confirmed using a two-way (harmonicity $\times$ feedback) RMANOVA on the log-transformed unbiased DLCs. This showed a significant main effect of harmonicity [$F(1,8)=36.68$, $p<0.0005$]. There was no significant main effect of feedback [$F(1,8)=3.30$, $p=0.107$], or interaction [$F(1,8)=1.95$, $p=0.200$].

The finding that, after removing the influence of timbre, DLCs were significantly larger for inharmonic than for harmonic tones is qualitatively consistent with the earlier finding of Moore and Glasberg (1990). Quantitatively, the effect of inharmonicity observed in the current study (a factor of 2.2 on average) is somewhat larger than that observed in the earlier study (a factor of 1.6 on average). This could be due to inter-individual differences, or to methodological differences, such as the degree and direction of frequency shift: $\pm 15\%$ of F_0 for alternating components in Moore and Glasberg vs. $+25\%$ of F_0 here. The reason why we used a frequency shift of 25% of F_0 is that we reasoned that this value corresponded to the largest possible deviation from harmonicity that can be obtained by frequency shifting. In addition, the randomization of the lowest harmonic number in our

experiment may have led to a greater reliance on residue pitch than in the conditions tested by Moore and Glasberg, where no roving was undertaken.

In summary, the results of experiment 1 show that DLCs based on harmonic tones are lower than those based on inharmonic tones, by about a factor of 2. Thus, in situations where the lowest frequency is not a reliable cue, listeners seem to rely on a residue-pitch cue that is stronger for harmonic than for inharmonic tone complexes. This is true, despite the fact that we used regularly spaced inharmonic components produced by coherent shifting, which have been shown to produce many effects that are similar to those found with harmonic stimuli, particularly in the area of perceptual organization (e.g., Roberts and Brunstrom, 1998, 2001).

One potential limitation of our conclusion is the fact that the DLC estimates for the inharmonic tones were based on a combination of thresholds from consistent and inconsistent tracks, according to the Eq. (1). Had we simply computed geometric means of the DLCs from the two types of track, we would have obtained a rather different result: nearly identical thresholds for inharmonic complexes (geometric mean across feedback and no-feedback conditions=0.33) and inharmonic complexes (geometric mean across feedback and no-feedback conditions=0.34). Most other studies that have roved the lowest harmonic number of a complex tone (e.g., Houtsma and Smurzynski, 1990; Bernstein and Oxenham, 2003; Oxenham *et al.*, 2009) have roved randomly within a single track, making our comparisons with earlier data dependent on the assumption that Eq. (1) is a valid way of estimating DLCs in this context. The aim of experiment 2 was to test whether our initial results could be reproduced using a more standard single-track method. It is important to point out that there is no guarantee *a priori* that a single-track procedure with randomly interleaved “consistent” and “inconsistent” trials will return valid estimates of DLCs, when bias is present. The outcome depends on how exactly the bias influences the listener’s responses, and on the parameters of the tracking procedure. However, we reasoned that, if larger DLCs for the inharmonic complex tones than for the harmonic complex tones were also observed using a single-track procedure, this would provide some further assurance that this difference was not an artifact of our calculation of unbiased thresholds.

III. EXPERIMENT 2

A. Methods

The stimuli and procedure were similar to those of experiment 1. The only difference was that this experiment used a single-track procedure with feedback after each trial. Whether the first or second interval had the higher harmonic numbers was selected at random with equal *a priori* probability for each trial. As in Experiment 1, the tone with the higher F_0 (before shifting, when shifting was applied), was also assigned randomly to the first or second interval, with equal probability. Six threshold estimates were obtained in each of two test conditions (harmonic and inharmonic), for

each listener. Six of the nine listeners who participated in Experiment 1 were tested in the current experiment.

B. Results and discussion

The (geometric) mean DLCs obtained in this experiment were 0.40% for the harmonic tones, and 0.94% for the inharmonic tones. (The corresponding geometric standard errors were approximately equal to 12% and 20% of the cited geometric means, respectively). Thus, DLCs were about two times larger for the inharmonic tones than for the harmonic tones. This difference was statistically significant [paired t test, $t(5)=2.83$, $p=0.037$], and was not significantly different from that measured in the same six listeners in Experiment 1, as indicated by the lack of significant interaction in a two-way (experiment $\times$ harmonicity) RMANOVA [$F(1,5)=0.50$, $p=0.510$]. The results show that the finding of larger DLCs for inharmonic tones than for harmonic tones in the previous experiment was not just a by-product of the use of two interleaved tracks, or an artifact of the way in which unbiased DLCs were computed based on the results from the two tracks.

The mean DLCs obtained in this experiment were not significantly different from the mean unbiased DLCs that were computed by pooling data from consistent and inconsistent tracks in Experiment 1 (0.43% for harmonic tones, and 0.92% for inharmonic tones, on average across feedback and no-feedback conditions). This good correspondence provides some assurance that the assumptions on which our calculation of unbiased DLCs in Experiment 1 was based are reasonable. Although the two procedures yield similar estimates of DLC, the dual-track procedure that was used in Experiment 1 provides more information than the single-track procedure, because it allows the extent to which the listener's responses were influenced by timbre or lowest-harmonic number cues to be assessed, both on the level of individual listeners and on the level of the group data. Future studies could use this technique to estimate the relative weight assigned to F0 and spectrum by individual listeners in the absence of feedback in a more precise way than has been done in previous studies (e.g., Laguitton *et al.*, 1998; Schneider *et al.*, 2005; Seither-Preisler *et al.*, 2007).

Taken together, the results from experiments 1 and 2 suggest that DLCs are influenced by the harmonicity of the components in the complex. However, it could be argued that our paradigm, which involved roving the lowest harmonic number and providing feedback based on the F0, may have forced listeners to use residue pitch cues more than would be normally done in situations where there are no timbral or spectral differences between the comparison tones. Experiment 3 addressed this issue by repeating experiment 1 but without roving the lowest harmonic number within each trial, and without feedback.

IV. EXPERIMENT 3

A. Rationale and methods

In this experiment, DLCs were again measured for harmonic and inharmonic tones, but this time, the lowest harmonic number, N , was the same in the two observation in-

tervals on a given trial. This number (2 or 3) was randomized across trials with equal probability. Therefore, listeners could perform the task by basing their responses solely on the frequency of the lowest harmonic, or on cues that co-varied with this frequency, within each trial. We reasoned that if listeners' responses were based solely on these cues, DLCs should not be significantly larger for the inharmonic tones than for the harmonic tones. On the other hand, if listeners were still relying on residue pitch, a difference in DLCs should again be observed. Seven of the listeners who participated in Experiment 1 took part in this experiment, of whom six also took part in Experiment 2. For these six, experiments 2 and 3 were run in counterbalanced order across listeners, after they had completed Experiment 1. In this experiment, no feedback was provided to the listeners.

B. Results and discussion

The mean unbiased DLCs that were obtained in this experiment were 0.32% for the harmonic tones and 0.48% for the inharmonic tones (compared to 0.46% and 1.01%, respectively for the same seven listeners in Experiment 1). The difference between the two conditions was statistically significant [paired t -test, $t(6)=2.63$, $p=0.039$]. However, this difference, corresponding to a factor of 1.5, was smaller than that observed in the same seven listeners in Experiment 1 (a factor of 2.2). This difference in the effect of inharmonicity between the two experiments was confirmed by a significant experiment $\times$ harmonicity interaction in a two-way RMANOVA [$F(1,6)=10.63$, $p=0.017$]. The RMANOVA also showed a significant main effect of *experiment* [$F(1,6)=37.82$, $p=0.001$], consistent with the observation that DLCs were generally smaller in the current experiment than in Experiment 1.

A possible reason for the finding that DLCs were smaller overall in this experiment than in Experiment 1 is that the changes in timbre introduced by the random roving of the lowest harmonic number in Experiment 1 had a distracting influence (Moore and Glasberg, 1990). Another plausible explanation is that, when such roving was not used, listeners could take advantage of cues related to the frequency of the lowest harmonic. This reliance on cues other than residue-pitch could explain why frequency-shifting had a smaller effect in this experiment than in Experiment 1. From this point of view, the results of this study support the practice of roving the lowest harmonic number in studies of residue-pitch discrimination. These results are consistent with the notion that, when this is not done, listeners' performance and thresholds may be significantly influenced by cues unrelated to residue pitch. However, the fact that DLCs were significantly larger for inharmonic than for harmonic tones even in the absence of roving indicates that residue-pitch still plays a role even in such conditions.

V. EXPERIMENT 4

A. Rationale and methods

The aim of this experiment was to test whether the effects of inharmonicity on DLCs were also present for smaller shift sizes and to determine the relationship between shift

FIG. 3. Mean unbiased DLCs as a function of frequency-shift size. The error bars show geometric standard errors above the geometric mean. To facilitate reading, the values of the mean DLCs are also shown on each bar.

size and the effects of inharmonicity on DLCs. The DLCs were measured for complexes shifted by 12.5% and 6.25% of the reference F0. Other than this, the same stimuli and procedure as in Experiment 1 were used. Visual feedback was always provided after each trial. Seven of the nine listeners from Experiment 1 participated in this experiment. Three of these listeners had completed Experiments 2 and 3 before they took part in the current experiment.

B. Results

The (geometric) mean unbiased DLCs obtained in this experiment are shown in Fig. 3. To facilitate comparisons with Experiment 1, the mean unbiased DLCs across the same seven listeners in Experiment 1 are also shown for 0 and 25% shifts. Figure 3 shows that DLCs increased with increasing frequency shift. This was confirmed by a significant main effect of shift size [$F(3,18)=14.13$, $p=0.002$], and a significant linear contrast [$F(1,6)=20.36$, $p=0.004$], consistent with a roughly linear increase in DLCs as a function of shift size. Overall the data are consistent with the view that DLCs are adversely influenced by inharmonicity, and that they worsen as the degree of inharmonicity increases.

Our shifts were extended only to 25% because further increases in the shift would have led to increasingly harmonic tones, up to a 50% shift, at which point all the components would have been odd harmonics of an F0 corresponding to $F_{0,\text{ref}}/2$. Thus, with shifts greater than 25% we might expect DLCs to again decrease, reaching a minimum again at a 50% shift. Although not tested here, results consistent with this prediction were presented by Bernstein and Oxenham (2008), who found that for F0s between about 150 and 300 Hz (which in their experiments were high enough to yield resolved harmonics) DLCs for odd-harmonic complexes were not significantly different from DLCs for full harmonic complexes measured under the same conditions in the same listeners (see Fig. 2 on p. 1657 in the cited article).

VI. EXPERIMENT 5

A. Rationale

The aim of this experiment was to obtain information on the perceived pitch(es) and pitch accuracy (or pitch strength) of the stimuli used in Experiments 1–4. Listeners adjusted the frequency of a pure tone so that its pitch matched that of a harmonic or inharmonic complex tone, which was similar to those used in Experiments 1–4. We reasoned that the perception of multiple pitches and/or reduced pitch strength for inharmonic tones would be reflected in more scattered pitch matches and/or multi-modal pitch-match histograms than for harmonic tones.

B. Stimuli, procedure, and listeners

Each trial of the pitch-matching procedure involved the successive presentation of two sounds. The first, referred to as the “reference” stimulus, was a harmonic or inharmonic complex tone similar to those used in Experiment 1, i.e., a 400-ms complex tone (including 20-ms raised-cosine onset and offset ramps) consisting of harmonics 2–7 or 3–8 of a 400 Hz nominal F0, the frequencies of which could all be shifted by a constant amount—0% (harmonic), 6.25%, 12.5%, or 25% of the F0, which at the nominal F0 of 400 Hz corresponded to shifts of 0, 25, 50, and 100 Hz. At the beginning of each matching procedure, the actual F0 of that reference complex was selected at random from a 20%-wide rectangular distribution centered on the nominal F0 (400 Hz).

The reference stimulus was followed after a 500-ms gap by a “comparison” 400-ms pure tone (also including 20 ms raised-cosine ramps) with an adjustable frequency. At the beginning of each matching procedure, the frequency of the comparison tone was selected randomly from a two-octave-wide distribution centered on 400 Hz, with 1-semitone steps. Listeners were instructed to adjust the frequency of the comparison tone so that its pitch matched that of the reference tone as precisely as possible. They could increase or decrease the frequency of the comparison tone at will, using six virtual buttons on the computer screen: two large buttons for ± 4 -semitone (about 26%) steps, two medium-sized buttons for ± 1 -semitone (about 6%) steps, and two small buttons for $\pm 1/4$ -semitone (about 1.45%) steps. A “hear again” button was also available. After each button press, the reference and comparison stimuli were regenerated with the appropriate parameters. The different conditions (harmonic and inharmonic reference) were tested in random order.

Five of the listeners who took part in Experiments 1 and 4 completed the current experiment. Each listener performed 90 matches in each of the four test conditions (the harmonic condition, and three different inharmonic conditions corresponding to ΔF s of 6.25, 12.5, and 25% of the F0). For each match, the deviation (in semitones) was computed between the frequency of the comparison tone at the end of the match, and the F0 of the reference tone before any shift (hereafter referred to as the “original” F0—i.e., the F0 before the shift was imposed). Pitch-match histograms were produced by

FIG. 4. Pitch match histograms. The different panels correspond to different frequency-shift sizes (ΔF), from 0% of F_0 at the top (which corresponds to the harmonic case), to 25% of F_0 at the bottom. The dotted vertical line indicates the center of the histogram, which corresponds to the original F_0 , defined as the F_0 of the complex tone before the application of the frequency shift, ΔF . Vertical dashed lines indicate 1 octave below and above the original F_0 . Panels in the left-hand column correspond to conditions in which the rank of the lowest harmonic (or component, inharmonic cases), N , was equal to 2. Panels in the right-hand column correspond to conditions in which N was equal to 3.

counting the number of matches that fell within 1/4-semitone bins from 20 semitones below to 20 semitones above the original F_0 .

C. Results and discussion

Figure 4 shows the mean pitch-match histograms that were obtained after pooling data across the five listeners who took part in this experiment. Pitch-match histograms for complexes in which the lowest harmonic number, N , was equal to 2 (prior to the application of any frequency shift) are shown on the left; pitch-match histograms for complexes in which N was equal to 3 are shown on the right. No salient and systematic differences were apparent between these two conditions. Therefore, the measures indicated below were computed after pooling the data across the $N=2$ and $N=3$ conditions. For harmonic tones (top panel), most (91%) of the matches fell within ± 1.5 semitones of the original F_0 . Matches that fell outside of this range usually fell close to 1 octave above the original F_0 (marked by dashed vertical lines).

As the size of the frequency shift (ΔF) increased from 0 (harmonic condition, top panel) to 25% of the F_0 (bottom panel), three effects were observed. First, the largest histogram peak, which at the 0% ΔF was centered on the original F_0 (indicated by the dotted line in the center of these histograms) shifted slightly to the right of the original F_0 , and

was about 1 semitone above the original F_0 at the 25% ΔF . This increase in perceived pitch with increasing ΔF was expected based on the results of earlier studies, which found that shifting the frequency of all components in a harmonic complex tone produces an increase in the perceived pitch—an effect traditionally referred to as the “pitch shift of the residue” (Schouten, 1940; de Boer, 1956; Patterson, 1973; Patterson and Wightman, 1976). These earlier results indicate that the pitch shifts are usually well modeled as linear functions of the frequency shift (at least, for relatively small frequency shifts), with a slope typically between 0.1 and 0.3; the exact slope value depends on stimulus parameters, and varies somewhat across listeners and studies. The present data, which indicate a pitch shift of about 1 semitone (6% of the nominal F_0 , or 24 Hz at 400 Hz) for a ΔF of 25% yield a slope of 0.24. This is well within the range of slopes observed in earlier studies. In fact, under stimulus conditions directly comparable to those used in the current study—namely, harmonic complexes with an original F_0 of 400 Hz and a lowest harmonic number equal to 2 or 3 prior to shifting—Patterson and Wightman’s (1976) pitch-matching data in two listeners show pitch shifts corresponding to a factor of about 1.06 (i.e., 1 semitone) for a ΔF equal to 25% of F_0 (see the lower panel in their Figs. 1 and 2).

The second effect apparent in Fig. 4 relates to the fact that, as ΔF increased, smaller peaks in the histograms appeared at frequencies that did not correspond precisely to the original F_0 (which, for the shifted complexes, was equal to the frequency spacing between consecutive harmonics), or to octaves of it. In particular, with the ΔF corresponding to 25% of the F_0 , a peak also emerged about 1 semitone below the original F_0 . A smaller peak located about 13 semitones below the original F_0 was also observed at that ΔF . This matches close to the sub-octave may be related to the fact that, for 25% ΔF , the frequencies of components in the complex tone were approximately equal to a set of odd harmonics of $F_0/2$, as suggested by Patterson (1973).

The third effect apparent in these histograms is the increasing dispersion of the matches with increasing ΔF . This is indicated by wider and shorter distributions going from the top panel to the bottom panel in Fig. 4. These effects confirm that the inharmonic tones used in the discrimination tasks of Experiments 1–4 had a less salient (as reflected in the smaller histogram peaks) and more ambiguous (as indicated by the multiple peaks) pitch than the harmonic tones. This outcome is qualitatively consistent with the findings of significantly larger DLCs for inharmonic than for harmonic tones in these preceding experiments. A more quantitative approach to relating histograms from pitch matches to DLCs is described in Appendix B.

VII. GENERAL DISCUSSION

Four experiments found smaller DLCs for harmonic tones than for inharmonic tones produced by shifting the frequencies of harmonics by up to 25% of F_0 . The effect was larger when the lowest harmonic number present was roved across intervals (about a factor of 2) than when it was not (a factor of 1.5), but was statistically significant in all cases.

The final experiment, involving pitch matches, showed that the deterioration in DLCs with increasing inharmonicity was paralleled by increasing variability in the pitch matches, and the appearance of multiple peaks in the pitch-match histograms. Overall the results are consistent with the idea, implicit in many studies, that F0 discrimination involves discrimination of the residue pitch (Moore and Glasberg, 1990), rather than discrimination of the pitches (or frequencies) of individual components (Faulkner, 1985).

Importantly, the present results indicate that DLCs reflect residue pitch discrimination even when the complexes being compared contain several corresponding harmonics: in three of the four experiments in which DLCs were measured, five out of the six components in the two complex tones that were presented on a trial had the same nominal ranks, and proximal frequencies; in Experiment 3, all of the six components in the two complexes being compared had corresponding ranks; yet DLCs were still significantly larger for inharmonic tones than for harmonic tones. This conclusion comes with one qualification, however. In Experiment 3, in which the lowest harmonic number was not randomized within trials, the difference in DLCs between harmonic and inharmonic conditions was smaller than that observed in Experiments 1 and 2, in which the lowest harmonic number was roved. This indicates that in psychoacoustic studies in which the lowest harmonic number is not randomized within trials, DLCs may be more heavily influenced by cues related to the lowest harmonic number, and thus less indicative of residue-pitch discrimination ability, than in studies in which the lowest harmonic is randomized within trials. However, this observation may be specific to complex tones that consist of equal-level components, as were used here. Cues related to the frequency of the lowest harmonic number are less likely to play a role in experiments in which the complexes are passed through bandpass filters with relatively shallow slopes (e.g., Carlyon and Shackleton, 1994; Micheyl and Oxenham, 2004), as this likely attenuates or eliminates edge-shift effects and spectral-centroid shifts as the F0 changes.

The finding that the difference in DLCs between harmonic and inharmonic tones was larger when the lowest harmonic number was roved across intervals than when it was not suggests that perceived differences in timbre interfere more readily with the discrimination of residue pitch when the pitch is ambiguous (being either weak, or multiple). Our method of interleaving two independent tracks, with either consistent or inconsistent changes in F0 and spectrum, provides a way of assessing the effects of timbre on judgments both at a group and an individual level. This in turn provides a method to evaluate individual differences in the dominance of “spectral” vs. “periodicity” pitch in a more nuanced way than has been possible in earlier studies (Laguitton *et al.*, 1998; Schneider *et al.*, 2005; Seither-Preisler *et al.*, 2007).

These findings, which provide clear evidence that DLCs for complex tones with corresponding components effectively measure residue-pitch discrimination, have direct practical implications for pitch-perception researchers, as they validate DLF0s as a measure of residue-pitch discriminability and pitch strength. Moreover, the findings confirm and quantify the influence of randomizing the lowest harmonic

number across intervals (within trials), and they indicate that although perhaps not absolutely necessary, when applied properly, such randomization reduces the influence of timbre or other cues related to changes in the frequency of the lowest harmonic—but unrelated to residue-pitch perception. Evidently, the roving technique must be adequate to ensure that the level of performance targeted or measured in an experiment cannot be accounted for based on residual cues related to the frequency of the lowest harmonic (see Oxenham *et al.*, 2009 for a discussion of the merits and limitations of different randomization techniques). The roving technique that was used in the current study, in combination with the dual-tracking procedure and the method for calculating unbiased thresholds, provides a useful approach for the measurement of difference limens for residue pitch, while simultaneously assessing the influence of cues related to the frequency of the lowest harmonic, including timbre (brightness) differences.

Despite the practical implications of the current findings, they do not provide strong evidence for or against existing place-based or temporal theories of pitch perception. Both spectral pattern-matching models (e.g., Wightman, 1973) and temporal autocorrelation models (e.g., Meddis and Hewitt, 1991) produce representations of pitch that are more ambiguous (i.e., contain broader peaks and/or more spurious peaks) in response to inharmonic tones than in response to perfectly harmonic tones. Thus, while detailed investigations using computer implementations of these models might reveal that some of them are better able than others to reproduce precisely the results of this study, it seems likely that such conclusions would depend on detailed assumptions concerning the calculation of discrimination-threshold estimates and pitch matches based on the computed pitch representations.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Institutes of Health (NIDCD R01 DC 05216). Thanks go to Celia Wolk Gershenson and Herb Pick for organizing the REU Program in the Behavioral and Cognitive Sciences, which made it possible for the second author to become involved in the research project, and prompted us to carry out the study. We are grateful to Brian Roberts, Chris Plack, and an anonymous reviewer for many constructive suggestions.

APPENDIX A: DERIVING “UNBIASED” THRESHOLDS FROM INTERLEAVED TRACKS

In this Appendix, the mathematical model underlying the calculation of “unbiased” DLCs [Eq. (1) in the main text] is explained. The intuition on which this model is based is that, all other things being equal, an increase in the frequency of the lowest component biases the listener toward judging the second complex as “higher” than the first along some internal dimension. We model this by assuming that the decision variable in the “pitch discrimination” tasks is a linear combination of two variables, one representing a perceived change in residue pitch, the other representing a perceived change in timbre.

First, we consider the case of harmonic tones. For this case, the general expression for the decision variable is,

$$D = w_p \ln\left(\frac{F0_h}{F0_l}\right) + w_t \ln\left(\frac{F_h}{F_l}\right). \quad (\text{A1})$$

In this equation, w_p and w_t denote the weights assigned to the subjective pitch (F0) and timbre (lowest frequency) changes, respectively, $F0_l$ and $F0_h$ denote the lower and higher F0s, and F_l and F_h denote the lower and higher frequencies of the lowest component in the complex. Our choice of using logarithms of F0 (or frequency) ratios, rather than arithmetic F0 (or frequency) differences, reflects the view that constant F0 (or frequency) steps in octaves (or semitones) correspond to approximately equal changes in perceived pitch (or brightness). The second additive term on the right-hand side of Eq. (A1) quantifies the “timbre bias.” In the absence of timbre bias (i.e., when $w_t=0$), the decision depends solely on residue pitch. We add the constraint that $w_p + w_t = 1$.

We seek to find the F0 difference that would correspond to 70.7% correct (a d' of 0.77 in our 2-interval, 2-alternative forced-choice task), *in the absence of any timbre bias*. We denote this F0 difference as DLC_u , where the subscript, u , stands for “unbiased.” Unfortunately, we cannot measure this quantity directly. What we have access to are thresholds measured under conditions where some timbre bias was almost certainly present, although we do not know how large the bias was. Fortunately, we have measured thresholds from two adaptive tracks: “consistent” tracks, in which the F0 and the lowest component frequency moved in the same direction, and “inconsistent” tracks, in which these two variables moved in opposite directions. Hereafter, these two tracks are indicated using superscripts: + for consistent tracks, and – for inconsistent tracks. The thresholds measured on these two tracks are denoted DLC^+ and DLC^- . These thresholds are related to the corresponding F0 ratios by the following equations,

$$DLC^+ = 100 \left(\frac{f0_h^+}{f0_l^+} - 1 \right) \quad (\text{A2})$$

and

$$DLC^- = 100 \left(\frac{f0_h^-}{f0_l^-} - 1 \right) \quad (\text{A3})$$

where the lowercase $f0$'s are the F0s of the lower- and higher-F0 tones at threshold on the two considered tracks.

From Eq. (A1), the decision variables for the two tracks are

$$\begin{aligned} D^+ &= w_p \ln\left(\frac{F0_h^+}{F0_l^+}\right) + w_t \ln\left(\frac{3F0_h^+}{2F0_l^+}\right) \\ &= \ln\left(\frac{F0_h^+}{F0_l^+}\right) + w_t \ln\left(\frac{3}{2}\right) \end{aligned} \quad (\text{A4})$$

and

$$\begin{aligned} D^- &= w_p \ln\left(\frac{F0_h^-}{F0_l^-}\right) + w_t \ln\left(\frac{2F0_h^-}{3F0_l^-}\right) \\ &= \ln\left(\frac{F0_h^-}{F0_l^-}\right) - w_t \ln\left(\frac{3}{2}\right) \end{aligned} \quad (\text{A5})$$

These equations can be rewritten in terms of DLC^+ and DLC^- , as follows:

$$d^+ = \ln\left(\frac{DLC^+}{100} + 1\right) + w_t \ln\left(\frac{3}{2}\right), \quad (\text{A6})$$

$$d^- = \ln\left(\frac{DLC^-}{100} + 1\right) - w_t \ln\left(\frac{3}{2}\right), \quad (\text{A7})$$

where the lowercase d 's refer to specific values of the decision variable, determined by the measured DLCs.

In the framework of signal detection theory (Green and Swets, 1966), the decision variables, D^+ and D^- , correspond to the differences between the means of the two probability density functions, conditioned on the two stimulus alternatives (i.e., higher-F0 in first interval and higher-F0 in second interval) of the 2I-2AFC task. By the definition of the index of sensitivity, d' , these differences are related to d' by a proportionality constant. It follows that, unless the magnitude of the internal noise differed between the two tracks, at threshold (corresponding to a fixed d'), d^+ and d^- must be equal. We define a new variable, $d \equiv d^+ = d^-$. As a result, we are left with a solvable system of two equations with two unknowns:

$$\begin{aligned} d &= \ln\left(\frac{DLC^+}{100} + 1\right) + w_t \ln\left(\frac{3}{2}\right), \\ d &= \ln\left(\frac{DLC^-}{100} + 1\right) - w_t \ln\left(\frac{3}{2}\right). \end{aligned} \quad (\text{A8})$$

It is straightforward to solve this system for d . The answer is,

$$d = \frac{1}{2} \left[\ln\left(\frac{DLC^-}{100} + 1\right) + \ln\left(\frac{DLC^+}{100} + 1\right) \right]. \quad (\text{A9})$$

We can now come back to our original problem of finding DLC_u . In our model, DLC_u is related to d by the following equation,

$$d = \ln\left(\frac{DLC_u}{100} + 1\right), \quad (\text{A10})$$

which is just a particular case of Eq. (1), obtained by setting w_t to 0. Inserting Eq. (A9) into Eq. (A10), rearranging, and simplifying, we finally obtain

$$DLC_u = 100 \left[\sqrt{\left(\frac{DLC^-}{100} + 1\right)\left(\frac{DLC^+}{100} + 1\right)} - 1 \right], \quad (\text{A11})$$

which is equivalent to Eq. (1) in the main text.

We now consider the case of shifted (inharmonic) complexes. For such stimuli, the residue pitch is not equal to F0. Rather, it is approximately equal to F0 plus a shift, which is proportional to the frequency shift applied to all components in the complex (de Boer, 1956; Schouten et al., 1962; Patter-

son, 1973; Patterson and Wightman, 1976). Thus, for this case, the equation for the decision variable is

$$D_s = w_p \ln \left(\frac{F0_h + \alpha \Delta f}{F0_l + \alpha \Delta f} \right) + w_t \ln \left(\frac{F_h}{F_l} \right), \quad (\text{A12})$$

where the subscript, s , stands for “shifted,” and the new variable, α , is the proportionality constant relating the physical frequency shift (Δf) to the perceived residue-pitch shift. In addition, the lowest-component frequencies in the lower- and higher-F0 complexes are now shifted. Thus, for “consistent” tracks, we have

$$D_s^+ = w_p \ln \left(\frac{F0_h^+ + \alpha \Delta f}{F0_l^+ + \alpha \Delta f} \right) + w_t \ln \left(\frac{3F0_h^+ + \Delta f}{2F0_l^+ + \Delta f} \right), \quad (\text{A13})$$

and for “inconsistent” tracks, we have

$$D_s^- = w_p \ln \left(\frac{F0_h^- + \alpha \Delta f}{F0_l^- + \alpha \Delta f} \right) + w_t \ln \left(\frac{2F0_h^- + \Delta f}{3F0_l^- + \Delta f} \right). \quad (\text{A14})$$

Due to the presence of additive terms in the ratios between the parentheses, the solution of this system of equations is more involved than for the harmonic case, and it is not reproduced here. In particular, unlike that for the harmonic case, the solution cannot be expressed in terms of DLC^+ and DLC^- only; it depends, also, on $F0_l^-$ and $F0_l^+$ (or $F0_h^-$ and $F0_h^+$), and α . If $F0_{\text{ref}}$ (the geometric mean of the two F0s presented on a trial) is constant and identical for both tracks, $F0_l^-$ and $F0_l^+$ (or $F0_h^-$ and $F0_h^+$) can be determined based on the measured DLC^+ and DLC^- . However, in the current experiments, $F0_{\text{ref}}$ varied randomly across trials, making it impossible to compute precisely the value of the unbiased DLC (DLC_u) using Eqs. (A13) and (A14). The value of α can be estimated from pitch-matching data. Unfortunately, not all listeners took part in the pitch-matching experiment (Experiment 5). Thus, for four of the listeners who took part in Experiment 1, we did not have values of α .

To assess the extent to which using Eq. (A11) instead of the more complicated solution based on Eqs. (A13) and (A14) influenced the results, we measured the error between unbiased thresholds computed using Eq. (A11) and the corresponding (true) values. We found that, for realistic settings of the variables (e.g., $\Delta f = 100$ Hz; $\alpha = 0.3$; and $F0_h^-$ and $F0_h^+$ values chosen to yield DLC^+ and DLC^- approximately equal to those measured with inharmonic tones in Experiment 1), the unbiased threshold estimated using Eq. (A11) was only about 7% lower than the true value. Note that this outcome was obtained using the largest Δf tested in the experiments (25% of the F0), and with α set to 0.3, which is larger than the mean value estimated from the results of Experiment 5 (0.24); thus, it is likely that the 7% error between the estimated DLC_u and the true DLC_u in these calculations provides an upper bound. In addition, because the estimated DLC_u was lower than the true DLC_u , this error cannot explain our findings of larger DLC_u values for shifted complexes than for harmonic complexes. Accordingly, Eq. (A11) (i.e., Eq. (1) in the main text) was used to analyze the data of both conditions.

APPENDIX B: QUANTIFYING THE RELATIONSHIP BETWEEN PITCH MATCHES AND DISCRIMINATION THRESHOLDS

In an attempt to go beyond qualitative comparisons, the pitch-matching data were used to derive quantitative predictions of DLCs as a function of the frequency shift, ΔF . How can discrimination thresholds be predicted from matching data? Intuitively, larger discrimination thresholds reflect noisier internal representations of the stimulus, and should correspond to more scattered matches, i.e., broader histogram peaks. A commonly used measure of scatter is the standard deviation (SD) of the matches. However, when the distribution of matches exhibits multiple modes, the SD does not necessarily provide a good measure of discriminability. Therefore, it seems important to take into account the shape of the distribution of matches. The approach used here involved drawing pairs of samples at random from the empirical distribution of pitch matches.⁵ These samples can be thought of as “virtual pitch matches.” A large number (10,000) of such pairs were obtained. In each pair, one sample was drawn from the distribution of pitch matches corresponding to one stimulus condition (e.g., harmonic with a lowest harmonic rank of 2); the other sample was drawn from the distribution of pitch matches corresponding to the corresponding stimulus condition, with the other lowest-harmonic rank (in this example, harmonic with a lowest harmonic rank of 3). This was done to simulate the testing conditions of Experiment 1. For half of the pairs, the sample corresponding to one lowest-component condition was multiplied by $(1 + \Delta F0)^{1/2}$, while the other sample was divided by $(1 + \Delta F0)^{1/2}$; for the other half of the samples, the opposite was done. This simulates the presentation of complexes differing in F0 by $\Delta F0\%$. The value of $\Delta F0$ was varied adaptively, so that the computed proportion of “correct responses” across all pairs was equal to 0.7, which corresponds approximately to the proportion correct targeted in the discrimination experiments. The proportion of correct responses was computed as the number of trials on which the sample that was multiplied by $(1 + \Delta F0)^{1/2}$ was larger than the other sample from the same trial.

Using this approach, we obtained the following predictions for discrimination thresholds based on the pitch-match data: 1.8% for harmonic tones; 3.4% for inharmonic tones with a ΔF of 6.25% of F0 (i.e., 25 Hz); 18% for inharmonic tones with a ΔF of 12.5% of F0 (i.e., 50 Hz); and 30% for inharmonic tones with a ΔF of 25% of F0 (i.e., 100 Hz). As for the measured DLCs, these predicted thresholds increase with increasing ΔF . However, they are generally much larger than the mean unbiased DLCs that we measured in the human listeners in these same stimulus conditions (0.44%, 0.52%, 0.83%, and 0.95%, respectively). We can think of two possible lines of explanation for this outcome. The first is that human listeners are more sophisticated, and less sub-optimal, than the model that we used to predict DLCs based on pitch matches. The model is based on the assumption that listeners compare pitch sensations, which are represented mathematically as random variables drawn independently from distributions of pitch matches, and select the observation interval associated with the larger value as containing

the higher-pitch stimulus. This very simple decision strategy (which is traditionally referred to in the psychophysical literature as the “differencing” strategy) does not presume any knowledge concerning the statistical distribution of pitches and pitch shifts on the part of the observer—for instance, knowledge of the fact that small pitch shifts were more likely than very large ones. It is conceivable that the human listeners acquired such knowledge through repeated exposure to the stimuli early on in the experiments. A more sophisticated decision strategy, which takes into account such knowledge, could yield smaller thresholds. This begs the question: what thresholds could be achieved by an optimal observer?

Using a classic result in estimation theory (Edgeworth, 1908), the smallest discrimination threshold (corresponding to a given d'), $\theta_{\min}$, that can possibly be achieved based on observations, X , drawn from a known probability distribution, $f(X, \psi)$, where ψ is the parameter of interest, is

$$\theta_{\min} = \frac{d'}{\sqrt{I_{\psi}}}. \quad (\text{B1})$$

In this equation, I_{ψ} represents the Fisher information, a measure of the amount of statistical information about an underlying parameter, ψ , which is carried by a random variable, X . In the current application, the relevant probability distributions are distribution of pitch matches, and the underlying parameter is F0. In this context, the Fisher information for a given distribution of pitch matches, $f(X, \psi)$, can be computed as

$$I_{\psi} = E \left[\left(\frac{\partial}{\partial \psi} \ln f(X; \psi) \right)^2 \right], \quad (\text{B2})$$

where $E[\cdot]$ is the mathematical expectation operator, and $\partial/\partial\psi$ denotes partial derivation with respect to ψ . We evaluated this expression numerically for the different pitch-match distributions measured in Experiment 5, and plugged the results into Eq. (B1), with d' set to 0.7702, to obtain minimum-threshold estimates. These estimates represent the smallest F0-discrimination thresholds that can be achieved by making optimal use (in a maximum-likelihood sense) of all of the statistical information contained in the distributions of pitch matches. Thus, if these distributions faithfully represent the variability in the listeners' pitch percepts, the values computed with this Fisher-information approach provide a lower bound on pitch-discrimination thresholds.

The resulting values were 1.45%, 1.65%, 1.87%, and 1.86% for the harmonic, 6.25%-shift, 12.5%-shift, and 25%-shift conditions, respectively. These values are smaller than those obtained earlier using the “differencing” model, especially for large frequency shifts. This is not surprising, because for multi-modal distributions, the “differencing” model is sub-optimal. More interestingly, the value for the harmonic condition (1.45%) is substantially larger (by a factor of 3.3) than the mean unbiased DLC that was measured in human listeners in the corresponding condition of Experiment 4 (0.44%). This indicates that, even assuming an optimal observer, it is impossible to account for the low DLCs measured in human listeners under the two assumptions that a) these DLCs reflect comparisons between pitch percepts, and

b) the distributions of these pitch percepts are faithfully reflected in the measured distributions of pitch matches.

This conclusion leads us to consider a second possible explanation for the observed discrepancy between the pitch-match distributions and the predicted DLCs, which is that the distributions of pitch matches measured in this study (and perhaps, those measured in prior studies as well) do not faithfully reflect the accuracy of listeners' residue-pitch percepts. A possible reason why the pitch-match distributions may not have faithfully reflected the precision of the listeners' pitch percepts in this study is that, whereas the DLC measurements involved comparisons between two complex tones, the matches were between a complex tone and a pure tone. As suggested in the main text, timbre differences may interfere with residue-pitch comparisons. If so, the large timbre difference between the reference complex tone and the comparison pure tone in the pitch-matching experiment may have made it more difficult for listeners to accurately compare pitches than in the discrimination experiments. This greater difficulty could result in more variable pitch matches than expected based on the precision of the listeners' pitch percepts in the discrimination experiments, which was reflected in the DLCs.

Additional work is needed to clarify the relationship between pitch-matching data and pitch-discrimination thresholds. The two approaches described in this Appendix, i.e., the differencing model and the Fisher information, provide two mathematically principled approaches for quantifying this relationship might prove useful in future work.

¹“Corresponding harmonics” refers to harmonics that have the same rank, and/or have proximal frequencies. In previous publications, expressions such as “common harmonics” (Faulkner, 1985), or “overlapping harmonics” (Moore and Glasberg, 1990) have also been used.

²Timbre differences can interfere with the perception of pitch differences in various ways. First, listeners could confuse pitch with brightness, that dimension of timbre along which tones can be ordered from low to high—similar to pitch (Chuang and Wang, 1978; Hesse, 1982; Singh, 1989). Second, differences in spectral region can induce differences in pitch between complex tones having identical F0s (van den Brink, 1977; Ohgushi, 1978). Third, salient changes in timbre may have a distracting influence, and attract listeners' attention away from (less salient) changes in pitch.

³Strictly speaking, inharmonic complex tones do not have an F0. Therefore, the abbreviation “DLF0” does not apply to such stimuli. Consequently, in this article, following Moore and Glasberg (1990), we use the more neutral abbreviation “DLC” to refer to discrimination thresholds measured with inharmonic complex tones. The abbreviation is also used whenever it is not clear, *a priori*, whether the thresholds that they refer to truly reflect discrimination of F0, as opposed to discrimination of individual component frequencies or other aspects of the complex tones (e.g., spectral centroid).

⁴All but one of the listeners who were eventually included in the study performed no more than three runs of the adaptive procedure during this initial test. The remaining listener performed an additional block of three runs during the preliminary test, after he mentioned not being fully attentive during one of the first three runs. Although it might be argued that this screening test, which used only harmonic tones, could induce a practice advantage for the harmonic condition in the main experiment, we think that this was not the case because the thresholds measured during this short screening test were not significantly larger than the thresholds that were measured subsequently in the same listeners, under the same test conditions, in the main experiment. We believe that the absence of significant practice-related improvements in thresholds in this group of listeners stems from the fact that only listeners who had very low thresholds in the first place were included into the study.

⁵For this purpose, the empirical distributions were obtained directly from

- the measured pitch matches, using a standard non-parametric density estimation technique known as the "Parzen window" technique, or "kernel smoothing" (Duda *et al.*, 2001). The kernel function was a standard normal with a width parameter corresponding to 1/32 semitones. The smoothed distribution functions were then used to compute empirical cumulative distributions. Using a table-lookup technique, the inverse of his distribution was applied to samples drawn from a uniform ([0;1]) distribution to produce samples having the desired distribution.
- Bernstein, J. G., and Oxenham, A. J. (2003). "Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number?" *J. Acoust. Soc. Am.* **113**, 3323–3334.
- Bernstein, J. G. and Oxenham, A. J. (2008). "Harmonic segregation through mistuning can improve fundamental frequency discrimination," *J. Acoust. Soc. Am.* **124**, 1653–1667.
- Carlyon, R. P., and Shackleton, T. M. (1994). "Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms?" *J. Acoust. Soc. Am.* **95**, 3541–3554.
- Chuang, C.-K., and Wang, W. S.-Y. (1978). "Psychophysical pitch biases related to vowel quality, intensity differences, and sequential order," *J. Acoust. Soc. Am.* **64**, 1004–1014.
- Cooper, W. E., and Sorensen, J. M. (1981). *Fundamental Frequency in Sentence Production* (Springer, Berlin), pp. 1–213.
- Cudahy, E. A. (1975). "Fundamental pitch discrimination for monotic and dichotic two-tone complexes," *J. Acoust. Soc. Am.* **57**(S1), S27.
- de Boer, E. (1956). "Pitch of inharmonic signals," *Nature (London)* **178**, 535–536.
- Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification* (Wiley, New York).
- Edgeworth, F. Y. (1908). "On the probable errors of frequency-constants," *J. R. Stat. Soc.* **71**, 499–512.
- Fastl, H., and Stoll, G. (1979). "Scaling of pitch strength," *Hear. Res.* **1**, 293–301.
- Faulkner, A. (1985). "Pitch discrimination of harmonic complex signals: Residue pitch or multiple component discriminations," *J. Acoust. Soc. Am.* **78**, 1993–2004.
- Flanagan, J. L., and Guttman, N. (1960). "Pitch of periodic pulses without fundamental component," *J. Acoust. Soc. Am.* **32**, 1319–1328.
- Goldstein, J. L. (1973). "An optimum processor theory for the central formation of the pitch of complex tones," *J. Acoust. Soc. Am.* **54**, 1496–1516.
- Goldstein, J. L., Gerson, A., Srulovicz, P., and Furst, M. (1978). "Verification of the optimal probabilistic basis of aural processing in pitch of complex tones," *J. Acoust. Soc. Am.* **63**, 486–497.
- Green, D. M. (1958). "Detection of multiple component signals in noise," *J. Acoust. Soc. Am.* **30**, 904–911.
- Green, D. M., and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics* (Krieger, New York).
- Hesse, H. P. (1982). "The judgement of musical intervals," in *Music, Mind, and Brain*, edited by M. Clynes (Plenum, New York), pp. 217–225.
- 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.
- Laguiton, V., Demany, L., Semal, C., and Liégeois-Chauvel, C. (1998). "Pitch perception: A difference between right- and left-handed listeners," *Neuropsychologia* **36**, 201–207.
- Levitt, H. (1971). "Transformed up-down methods in psychoacoustics," *J. Acoust. Soc. Am.* **49**, 467–477.
- McDermott, J. H., and Oxenham, A. J. (2008). "Music perception, pitch, and the auditory system," *Curr. Opin. Neurobiol.* **18**, 452–463.
- Meddis, R., and Hewitt, M. (1991). "Virtual pitch and phase sensitivity studied of a computer model of the auditory periphery. I: Pitch identification," *J. Acoust. Soc. Am.* **89**, 2866–2882.
- Micheyl, C., and Oxenham, A. J. (2004). "Sequential F0 comparisons between resolved and unresolved harmonics: No evidence for translation noise between two pitch mechanisms," *J. Acoust. Soc. Am.* **116**, 3038–3050.
- Moore, B. C. J. (2003). *An Introduction to the Psychology of Hearing* (Academic, London).
- Moore, B. C. J., and Glasberg, B. R. (1990). "Frequency discrimination of complex tones with overlapping and non-overlapping harmonics," *J. Acoust. Soc. Am.* **87**, 2163–2177.
- Moore, B. C. J., Glasberg, B. R., and Proctor, G. M. (1992). "Accuracy of pitch matching for pure tones and for complex tones with overlapping or non-overlapping harmonics," *J. Acoust. Soc. Am.* **91**, 3443–3450.
- Ohgushi, K. (1978). "On the role of spatial and temporal cues in the perception of the pitch of complex tones," *J. Acoust. Soc. Am.* **64**, 764–771.
- Oxenham, A. J., and Buus, S. (2000). "Level discrimination of sinusoids as a function of duration and level for fixed-level, roving-level, and across-frequency conditions," *J. Acoust. Soc. Am.* **107**, 1605–1614.
- Oxenham, A. J., Micheyl, C., and Keebler, M. V. (2009). "Can temporal fine structure represent the fundamental frequency of unresolved harmonics?" *J. Acoust. Soc. Am.* **125**, 2189–2199.
- Patterson, R. D. (1973). "The effects of relative phase and the number of components on residue pitch," *J. Acoust. Soc. Am.* **53**, 1565–1572.
- Patterson, R. D., and Wightman, F. L. (1976). "Residue pitch as a function of component spacing," *J. Acoust. Soc. Am.* **59**, 1450–1459.
- Plack, C. J., Oxenham, A. J., Popper, A. N., and Fay, R. R. (2005). *Pitch: Neural Coding and Perception* (Springer, New York).
- Ritsma, R. J. (1963). "On pitch discrimination of residue tones," *International Audiology* **2**, 34–37.
- Roberts, B. (1998). "Effects of spectral pattern on the perceptual salience of partials in harmonic and frequency-shifted complex tones: A performance measure," *J. Acoust. Soc. Am.* **103**, 3588–3596.
- Roberts, B., and Bailey, P. J. (1996). "Regularity of spectral pattern and its effects on the perceptual fusion of harmonics," *Percept. Psychophys.* **58**, 289–299.
- Roberts, B., and Bregman, A. S. (1991). "Effects of the pattern of spectral spacing on the perceptual fusion of harmonics," *J. Acoust. Soc. Am.* **90**, 3050–3060.
- Roberts, B., and Brunstrom, J. M. (1998). "Perceptual segregation and pitch shifts of mistuned components in harmonic complexes and in regular inharmonic complexes," *J. Acoust. Soc. Am.* **104**, 2326–2338.
- Roberts, B., and Brunstrom, J. M. (2001). "Perceptual fusion and fragmentation of complex tones made inharmonic by applying different degrees of frequency shift and spectral stretch," *J. Acoust. Soc. Am.* **110**, 2479–2490.
- Schneider, P., Sluming, V., Roberts, N., Scherg, M., Goebel, R., Specht, H. J., Dosch, H. G., Bleeck, S., Stippich, C., and Rupp, A. (2005). "Structural and functional asymmetry of lateral Heschl's gyrus reflects pitch perception preference," *Nat. Neurosci.* **8**, 1241–1247.
- Schouten, J. F. (1940). "The residue and the mechanism of hearing," *Proc. K. Ned. Akad. Wet.* **43**, 991–999.
- Schouten, J. F., Ritsma, R. J., and Cardozo, B. L. (1962). "Pitch of the residue," *J. Acoust. Soc. Am.* **34**, 1418–1424.
- Seither-Preisler, A., Johnson, L., Krumholz, K., Nobbe, A., Patterson, R., Seither, S., and Lütkenhöner, B. (2007). "Tone sequences with conflicting fundamental pitch and timbre changes are heard differently by musicians and nonmusicians," *J. Exp. Psychol. Hum. Percept. Perform.* **33**, 743–751.
- Singh, P. G. (1989). "Interaction of timbre and pitch in spectral discrimination tasks using complex tones," *J. Acoust. Soc. Am.* **86**(S1), S58.
- Smoorenburg, G. F. (1970). "Pitch perception of two-frequency stimuli," *J. Acoust. Soc. Am.* **48**, 924–941.
- Srulovicz, P., and Goldstein, J. L. (1983). "A central spectrum model: A synthesis of auditory-nerve timing and place cues in monaural communication of frequency spectrum," *J. Acoust. Soc. Am.* **73**, 1266–1276.
- Terhardt, E. (1970). "Frequency analysis and periodicity detection in the sensations of roughness and periodicity pitch," in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. F. Smoorenburg (Sijthoff, Leiden, The Netherlands), pp. 278–287.
- van den Brink, G. (1977). "Octave and fifth settings for pure tones and residue sounds," in *Psychophysics and Physiology of Hearing*, edited by E. F. Evans and J. P. Wilson (Academic, London), pp. 373–379.
- Wiegand, L., Patterson, R. D., Demany, L., and Carlyon, R. P. (1998). "Temporal dynamics of pitch strength in regular interval noises," *J. Acoust. Soc. Am.* **104**, 2307–2313.
- Wightman, F. L. (1973). "The pattern-transformation model of pitch," *J. Acoust. Soc. Am.* **54**, 407–416.