Supplementary Text

Materials and Methods

Participants.

All experiments were approved by the Committee on the use of Humans as Experimental Subjects at the Massachusetts Institute of Technology, and were conducted with the informed consent of the participants. In all experiments except for Experiment 5 (which contained only two trials per participant), we excluded poorly performing participants using hypothesis-neutral performance criteria. In Experiments 1-4, we selected exclusion criteria a priori based on our expectations of what would constitute good performance. For Experiment 6, we conducted a pilot experiment in the lab, and selected online participants who performed comparably to good in-lab participants.

Musicianship: Across all experiments, musicians were defined as individuals with five or more years of self-reported musical training and/or active practice/performance. Non-musicians had four or fewer years of self-reported musical training and/or active practice/performance.

Experiment 1: 16 participants were recruited for Experiment 1. 4 of these participants had an average performance across all conditions of less than 60% correct and their data were removed. The remaining participants (N=12, 5 female, mean age = 32.5 years, S.D. = 19.0 years) included 6 non-musicians and 6 musicians (24.5 years of training and active performance, S.D. = 16.0).

Experiment 2:
34 participants were recruited for Experiment 2. 2 of these did not finish the experiment, and their data were not analyzed. Of the remaining 32 participants, 7 scored less than 60% correct across all conditions and were removed. The remaining participants (N=25, 12 female, mean age= 31.4 years, S.D.=14.7 years) included 10 non-musicians and 15 musicians (mean=16 years of musical training and active performance, S.D.=12.2 years).

Experiment 3: 35 participants were recruited for Experiment 3. 3 of these did not complete the experiment or were removed for non-compliance. An additional 9 participants were removed from analysis because their average threshold across conditions (using the first run of each condition, allowing unbiased threshold estimates from the last three runs of each condition for the remaining participants) was greater than 3% (approximately half a semitone). The remaining participants (N=23, 9 female, mean age=36.0 years, S.D.=16.6 years) included 11 non-musicians and 12 musicians (mean=17.1 years of musical training and active, S.D.=13.5)

Experiment 4: 77 participants were recruited online for Experiment 4. 39 participants were removed from analysis because their average threshold across the first run of all conditions was greater than 3% (approximately half a semitone). The high number of excluded participants is most likely due to the tedious nature of this experiment (because of the forced intertrial interval in some of the conditions, which made it easy to lose focus). The remaining participants (N=38, 16 female, mean age=37.6 years, S.D.=11.4 years) included 27 non-musicians and 11 musicians (mean=8.5 years of musical training and active, S.D.=2.7).

Experiment 5: Before analysis, we removed the participants who completed the demographic survey in either less than 20 seconds, which we believed made it unlikely that those participants read and paid attention to all the questions, or greater than 3 minutes. 1 and 17 participants were excluded via these timing criteria, respectively. After these exclusions, 1150 people completed Experiment 5 (680 female, mean age = 35.0 years, S.D. = 11.6 years). 371 reported some form

of musical training, and 290 of those reported greater than four years of training (mean=14.4 years, S.D.=9.2 years).

Experiment 6: 450 participants were recruited for the online component of Experiment 6. We sought to obtain mean performance levels comparable with those of compliant and attentive participants run in the lab. To this end, we excluded participants whose average performance across conditions fell below a cutoff. We ran 10 participants in the lab to establish this cutoff. We used the average threshold from the best two-thirds of these in-lab participants (7 of the 10) across all 5 of the no-delay conditions as the cutoff for inclusion in the online experiment. This yielded a cutoff value of 2.18%, and 286 of the online participants were excluded from analysis because their average threshold (across all 5 of the no-delay conditions, using the first run for each of these conditions) was greater than this cutoff. We used only the first run to determine inclusion, and subsequently analyzed only the remaining three runs, to avoid selection bias in the threshold estimates. The remaining set of 164 participants (73 female, mean age=35.3 years, S.D.=9.1 years) included 73 who reported greater than four years of musical training (mean=11.0 years, S.D.=6.9 years).

Audio Presentation: In-Lab.
In all experiments, a MacMini computer running Psychtoolbox for MATLAB (1) was used to play sound waveforms. Sounds were presented to participants at 70 dB SPL over Sennheiser HD280 headphones (circumaural) in a soundproof booth (Industrial Acoustics). Sound levels were calibrated with headphones coupled to an artificial ear, with a microphone at the position of the eardrum. Participants logged their responses via keyboard press.

Audio Presentation: Online.
We used the crowdsourcing platform provided by Amazon Mechanical Turk to run experiments that necessitated large numbers of participants (Experiments 5 and 6), or when in-person data collection was not possible due to the COVID-19 virus (Experiment 4). Each participant in these studies used a calibration sound to set a comfortable level, and then had to pass a 'headphone check' experiment that helped ensure they were wearing headphones or earphones as instructed (2) before they could complete the full experiment. The experimental stimuli were set to 15.5 dB below the level of the calibration sound, to ensure that stimuli were never uncomfortably loud. Participants logged their responses by clicking buttons on their computer monitors using their mouse.

Feedback.
Feedback (correct/incorrect) was given after each trial for all tasks except for the two test trials for about half of the participants of Experiment 5 (see below).

Experiment 1: Discriminating instrument notes with intervening silence
Procedure: Participants heard two instrument notes per trial, separated by varying amounts of silence (0, 5, and 10 seconds) and judged whether the second note was higher or lower than the first note. Participants heard 30 trials per condition, and all conditions were intermixed. The first stimulus for a trial began one second after the response was entered for the previous trial, such that there was at least a 1-second gap between successive trials.

Stimuli: Instrument notes were derived from the RWC Instrument database, which contains recordings of chromatic scales played on different instruments (3). We used recordings of baritone saxophone, cello, ukulele, pipe organ and oboe, chosen to cover a wide range of timbres.

Instrument tones were manipulated using the STRAIGHT analysis and synthesis method (4-6). STRAIGHT is normally used to decompose speech into excitation and vocal tract filtering, but can also decompose a recording of an instrument into an excitation signal and a spectrotemporal filter. If the voiced part of the excitation is modeled sinusoidally, one can alter the frequencies of individual harmonics, and then recombine them with the unaltered instrument body filtering to generate inharmonic notes. This manipulation (6) leaves the spectral shape of the instrument largely intact. Previous studies with speech suggest that the intelligibility of inharmonic speech is comparable to that of harmonic speech (7). The frequency jitters for inharmonic instruments were chosen in the same way as the jitters for the inharmonic synthetic tones used in Experiments 2-6 (described below). The same pattern of jitter was used for both notes in a trial. STRAIGHT was also used to frequency-shift the instrument notes to create pairs of notes that differed in f0 by a specific amount. Audio was sampled at 16,000 Hz. All notes were 400 ms in duration and were windowed by 20 ms half-Hanning windows.

Each trial consisted of two notes. The second note differed from the first by .25 or .5 semitone. To generate individual trials, the f0 of the first note of each trial was randomly selected from a uniform distribution over the notes in a Western classical chromatic scale between 196 and 392 Hz (G3 to G4). A recording of this note, from an instrument selected from the set of 5 that were used (baritone saxophone, cello, ukulele, pipe organ and oboe), was chosen as the source for the first note in the trial (instruments were counterbalanced across conditions). If the second note in the trial was higher, the note 1 semitone above was used to generate the second note (the note 1 semitone lower was used if the second note of the trial was lower). The two notes were analyzed and modified using the STRAIGHT analysis and synthesis method (4-6); the notes were f0-flattened to remove any vibrato, shifted to ensure that the f0 differences would be exactly the intended f0 difference apart, and resynthesized with harmonic or inharmonic excitation. Some instruments, such as the ukulele, have slightly inharmonic spectra. These slight inharmonicities were removed for the Harmonic conditions due to the resynthesis.

Stimuli for Experiments 2-6.

Experiments 2, 3, 4, 5.1, and 6 used the same types of tones. The stimuli for Experiment 5.2 are described below. Synthetic complex tones were generated with exponentially decaying temporal envelopes (decay constant of 4 s^{-1}) to which onset and offset ramps were applied (20 ms half-Hanning window). The sampling rate was 16,000 Hz for Experiment 2, and 48,000 Hz for all others. Prior to bandpass filtering, tones included all harmonics up to the Nyquist limit, in sine phase, and were always 400 ms in duration.

In order to make notes inharmonic, the frequency of each harmonic, excluding the fundamental, was perturbed (jittered) by an amount chosen randomly from a uniform distribution, $U(-.5, .5)$. This jitter value was chosen to maximally perturb f0 (lesser jitter values did not fully remove peaks in the autocorrelation at the period of the original f0 (8)). Jitter values were multiplied by the f0 of the tone, and added to the frequency of the respective harmonic. For example, if the f0 was 200 Hz and a jitter value of -0.39 was selected for the second harmonic; its frequency would be set to 322 Hz. To minimize salient differences in beating, jitter values were constrained (via rejection sampling) such that adjacent harmonics were always separated by at least 30 Hz. The same jitter pattern was applied to every note of the stimulus for a given trial, such that the spectral pattern shifted coherently up or down, even in the absence of an f0. Except for the Inharmonic-Fixed condition of Experiment 3, where one random jitter pattern was used for entire blocks of the experiment, a new jitter pattern was chosen for each trial.

Each complex tone was band-pass filtered in the frequency domain with a Gaussian transfer function (in log frequency) centered at 2,500 Hz with a standard deviation of half an octave. This filter served to ensure that participants could not perform the tasks using changes in the spectral envelope, and also to minimize timbral differences between notes in the Interleaved Harmonic condition. The filter parameters were chosen to ensure that the f0 was attenuated (to eliminate variation in a spectral edge at the f0) while preserving audibility of resolved harmonics (harmonics below the 10^{th}, approximately). The combination of the filter and the masking noise (described below) rendered the frequency component at the f0 inaudible.

To ensure that differences in performance for Harmonic and Inharmonic conditions could not be mediated by distortion products, we added masking noise to these bandpass filtered notes. We low pass filtered pink noise using a sigmoidal (logistic) transfer function in the frequency domain. The sigmoid had an inflection point at the third harmonic of the highest of the two notes on a trial, and a maximum slope yielding 40 dB of gain or attenuation per octave. We scaled the noise so that the noise power in a gammatone filter (one ERB_N in bandwidth (9), implemented as in (10)) centered at the f0 was 10 dB lower than the mean power of the three harmonics of the highest note of the trial that were closest to the 2,500 Hz peak (and thus had greatest magnitude) of the Gaussian spectral envelope (8). This noise power is sufficient to mask distortion products at the f0 (11, 12). This filtered and scaled pink noise was added to each note, and did not continue through the silence in 'delay' conditions. Noise has been reported to facilitate the perception of the f0 of a set of harmonics (13, 14) in contexts where the harmonic frequencies are embedded in relatively high levels of noise. Because the noise in our stimuli was focused at the f0 rather than the higher harmonics that composed our tones, it seems less likely to have produced such a benefit, but we never specifically manipulated it to assess its effect.

Experiment 2: Discriminating synthetic tones with intervening silence
Procedure: The procedure was identical to that for Experiment 1, except stimuli were synthetic tones.

Stimuli: Each trial consisted of two notes, described above in **Stimuli for Experiments 2-6**. The second tone differed from the first by .25 or .5 semitones. The first note of each trial was randomly selected from a uniform distribution on a logarithmic scale spanning 200 to 400 Hz. Tones were either Harmonic, Inharmonic, or Interleaved Harmonic. Interleaved Harmonic notes were synthesized by removing harmonics [1,4, 5, 8, 9, etc.] in one note, and harmonics [2, 3, 6, 7, etc.] in the other note. They were otherwise identical to the Harmonic tones (identical bandpass filter in the frequency domain, as well as noise to mask a distortion product at the fundamental). This manipulation was intended to isolate f0-based pitch, as it removes the note-to-note spectral correspondence between harmonics.

Experiment 3: Discriminating synthetic tones with a consistent inharmonic spectrum
Procedure: Participants heard two notes per trial, separated by varying amounts of silence (0, 1 and 3 seconds) and were asked whether the second note was higher or lower than the first note. Unlike Experiments 1 and 2, which used the method of constant stimuli, participants completed 2-up-1-down adaptive threshold measurements for each condition. Each run ended after 10 reversals. For the first 4 reversals, the f0 changed by a factor of 2, and for subsequent reversals by a factor of $\sqrt{2}$. Each adaptive run was initialized at an f0 difference of 1 semitone (approximately 6%), and the maximum f0 difference was limited to 16 semitones. The adaptive procedure continued if participants reached this 16 semitone limit; if they continued to get trials

incorrect the f0 difference remained at the 16 semitone limit, and if they got two in a row right, the f0 difference would decrease from 16 semitones by a factor of 2 or $\sqrt{2}$ depending on how many reversals had already occurred. In practice participants who hit this limit repeatedly were removed before analysis due to our exclusion criteria. Thresholds were estimated by taking the geometric mean of the final 6 reversals. The first stimulus for a trial began one second after the response was entered for the previous trial, such that there was at least a 1-second gap between successive trials.

Stimuli: Each trial consisted of two notes, described above in **Stimuli for Experiments 2-6**. The first note of each trial was randomly selected from a uniform distribution on a logarithmic scale spanning 200 to 400Hz. Tones were either Harmonic, Inharmonic, or Inharmonic-Fixed, separated into three blocks, the order of which was counterbalanced across participants. Participants completed 12 adaptive thresholds within each block (3 delay conditions x 4 runs per condition). For the Inharmonic-Fixed block, a random jitter pattern was chosen at the beginning of the block and used for every trial within the entire block.

Experiment 4: Discriminating synthetic tones with a longer intertrial interval

Procedure: The procedure for Experiment 4 was identical to that for Experiment 3, except that the experiment was run on Amazon Mechanical Turk, with different time intervals between trials. Participants completed two sets of adaptive threshold measurements. In the first, trials were initiated by the participant, and could begin as soon as they entered the response for the previous trial and clicked a button to start the next trial. Four adaptive threshold measurements per condition were taken in this way (Harmonic and Inharmonic stimuli both with and without a 3-second delay). In the second set, a mandatory 4-second pause was inserted between each trial, which could be initiated by the participant once the pause had elapsed. Four threshold measurements for the same conditions were taken in this way, and thresholds were estimated by taking the geometric mean of the final 6 reversals. The two sets of measurements were randomly intermixed.

To perform adaptive procedures online, stimuli were pre-generated (instead of being generated in real-time as was done for the in-lab studies). For each condition and possible f0 difference, the stimuli were drawn randomly from a set of 20 pre-generated trials (varying in the f0 of the first note, and in the jitter pattern for the Inharmonic trials).

Stimuli: Stimuli were identical to those for the Harmonic and Inharmonic conditions in Experiment 3.

Experiment 5: One-shot discrimination with longer intervening delay

Procedure. Participants were recruited using the Amazon Mechanical Turk crowdsourcing platform. In the main experiment, each participant completed two trials – one each of two trial types. In the first type of trial, they heard two consecutive notes and were asked whether the second note was higher or lower than the first. Notes always differed by a semitone. For the second type of trial, participants heard the first note, then were directed to a demographic survey, then were presented with the second note, and then were asked to respond. Before the presentation of the first note participants were told that they would be subsequently asked whether a second note heard after the survey was higher or lower in pitch than the note heard before the survey. Each participant heard the same stimulus type (Harmonic, Inharmonic, or Interleaved Harmonic) for each of the two trials. The order of the trials (with and without the intervening

survey) was randomized across participants. For the trials with the demographic survey, we collected time stamps of when participants heard the first and second note in order to calculate the delay between notes. Because we thought feedback might affect the results, we ran two versions of the experiment. In the first version, participants received no feedback but were allowed to listen to the stimuli twice if they wished. In the second version there was feedback and participants heard each stimulus only once. 592 of the 1150 participants completed the first version of the experiment; the remaining 558 participants completed the second version. We found that there was no significant difference between the two versions of the experiment in any of the conditions. We thus combined data across the two versions.

Before completing the two main experiment trials and the survey, participants completed 10 practice trials without a delay, with the same type of tone that they would hear in the two main experiment trials (i.e. if a participant would hear inharmonic stimuli in the two experiment trials, their 10 practice trials would contain inharmonic stimuli). Participants received feedback on each of these practice trials. The stimulus difference in all trials was 1 semitone.

Stimuli. Stimuli for Experiment 5 were identical to those used in Experiments 2-4 (Harmonic, Inharmonic and Interleaved Harmonic conditions).

Experiment 6: Individual differences in tone discrimination

Procedure: Both in lab and on Mechanical Turk, participants completed 2-up-1-down adaptive threshold measurements. The instructions were to judge whether the second note was higher or lower than the first note. The adaptive procedure was identical to that used in the first set of threshold measurements of Experiment 4 (each trial could be initiated by the participant as soon as they had entered their response for the previous trial). For the in-lab participants, we used the best three runs from each condition to set the inclusion criteria for online studies. For online participants, the first run of each condition was used to determine inclusion, and the final three runs of each condition were used for analysis. The order of the adaptive runs was randomized for each participant. There were 4 runs for each of the 10 conditions, for a total of 40 adaptive runs, randomized in order for each participant. Thresholds were estimated as the geometric mean of the final 6 reversals. Participants received feedback after each trial.

Stimuli: Participants in lab and online were tested on 5 different types of stimuli, presented either with no delay or a 3-second delay between tones (10 conditions). The five types of stimuli were as follows:

(1) Harmonic, (2) Inharmonic, and (3) Interleaved Harmonic, all identical to the same conditions in Experiment 2. (4) Pure Tones: We used the 4^{th} harmonic of the f0 (f0*4) so that the stimuli would overlap in frequency with the complex tones used in other conditions, which were filtered so that first audible harmonic was generally the 3^{rd} or 4^{th} harmonic. Low pass masking noise was omitted from the Pure Tone condition given that distortion products were not a concern. Given the similarity in mean thresholds between the Pure Tone and Harmonic conditions, and the high correlation between them across participants, the absence of noise in this condition does not appear to have influenced the results. (5) Random Harmonic: For each note, two harmonics were randomly chosen from harmonics 1 to 4, 5 to 8, etc. By chance, some harmonics could be present in both notes. This manipulation was intended to be intermediate between the Harmonic and Interleaved Harmonic conditions.

In all conditions, the f0 of the initial tone for each trial was chosen randomly from a uniform distribution on a logarithmic scale spanning 200-400 Hz.

As in Experiment 4, stimuli were pre-generated to enable online threshold measurements. For each condition and possible f0 difference, the stimuli were drawn randomly from a set of 20 pre-generated trials (varying in the f0 of the first note, and in the jitter pattern for the inharmonic conditions).

Sample Sizes.
Experiments 1 and 2: A power analysis of pilot data for Experiments 1 and 2 showed an effect size of d=1.25 for the difference between harmonic and inharmonic conditions at 5 seconds. We thus aimed to run at least 6 musicians and 6 non-musicians to be able to analyze the two groups separately and have an 80% chance of detecting the harmonic advantage at a $p<.05$ significance level (using paired t-tests). This number of participants also left us well-powered to observe an interaction between harmonicity and delay for both groups (8 participants were needed to have an 80% of detecting an interaction with an effect size of that seen in pilot data, $\eta_p^2 = .2$). Power analyses for Experiments 1-4 and Experiment 6 (in lab baseline) used G*Power (15). Experiment 2 was run in combination with other experiments (not described here) that were not as well powered and required more data, hence the additional participants.

Experiments 3 and 4: We performed power analyses for Experiments 3 and 4 using a pilot experiment with 17 participants where we measured thresholds either with or without a 3-second delay. The pilot experiment used the same method and analysis as Experiments 3 and 4, but without the 1-second-delay and Fixed-Jitter condition of Experiment 3 or the intertrial delays of Experiment 4. The effect size from this pilot experiment for the Harmonic-Inharmonic difference at the 3-second delay was d=1.49. Based on the rough intuition that the effect of the Inharmonic-Fixed manipulation or the intertrial delay might produce an effect approximately half this size, we sought to be 80% likely to detect an effect half as big as that observed in our pilot data, at a $p<.05$ significance (using a two-sided Wilcoxon signed-rank test). This yielded a target sample size of 17 participants. We did not plan to recruit equal numbers of musicians and non-musicians due to the similarity between groups in Experiments 1-2.

Experiment 5: For Experiment 5 we performed a power analysis by bootstrapping using a pilot version of the experiment (similar to the current version but without practice trials). For each possible sample size we computed the bootstrap distribution of the difference in performance between the Harmonic and Inharmonic conditions with a delay, as well as a null distribution obtained with conditions permuted across participants. We sought the sample size yielding an 80% chance of seeing a significant Harmonic-Inharmonic difference (i.e. where 95% of the bootstrap samples showed a difference exceeding the 97.5^{th} percentile of the null distribution), yielding a target sample size of 178 participants in each condition. We made no attempt to recruit equal numbers of musicians and non-musicians, as we did not plan to analyze those groups separately given the similar results across groups observed in the experiments we ran prior to this.

Experiment 6: For Experiment 6, we performed a power analysis by bootstrapping pilot data (an earlier version of Experiment 6 with slightly different stimuli). For each of a set of sample sizes we computed bootstrap distributions of the interaction term (difference of differences between the conditions being compared, Harmonic/Inharmonic and Interleaved Harmonic), as well as null distributions obtained by permuting conditions across participants. We found that a sample size of 154 yielded a 90% chance of seeing the interaction present in our pilot data at a $p<.05$ significance level. We ran more than this number of participants to allow performance-based

exclusion. As with Experiments 3-5, we made no attempt to recruit equal numbers of musicians and non-musicians.

Statistics.

For Experiments 1, 2, and 5 we calculated percent correct for each condition. For Experiments 1 and 2, data were evaluated for normality with Lilliefors' composite goodness-of-fit test. Data for Experiment 1 passed Lilliefors' test, and so significance was evaluated using paired t-tests and repeated-measures ANOVAs. We used mixed-model ANOVAs to examine the effects of musicianship (to compare within- and between-group effects). Data for Experiment 2 were non-normal due to ceiling effects in some conditions, and so significance was evaluated with the same non-parametric tests used for the threshold experiments (described below).

For Experiment 5, the significance of the differences between conditions and the significance of interactions were calculated via bootstrap (10,000 samples). To calculate the significance of the interaction between conditions, we first calculated the interaction (the difference of differences in means with and without a delay). For instance, for the Harmonic and Inharmonic conditions this term is as follows:

$$(\mu_{Harmonic}^{no-delay} - \mu_{Harmonic}^{delay}) - (\mu_{Inharmonic}^{no-delay} - \mu_{Inharmonic}^{delay})$$

Then, we approximated a null distribution for this interaction, permuting conditions across participants and recalculating the difference of differences 10,000 times. To determine statistical significance we compared the actual value of the interaction to this null distribution.

Data distributions were non-normal (skewed) for threshold experiments (Experiments 3,4 and 6) as well as for Experiment 2, so non-parametric tests were used for all comparisons. Wilcoxon signed-rank tests were used for pairwise comparisons between dependent samples (for example, two conditions for the same participant group). To compare performance across multiple conditions or across musicianship we used F statistics for repeated-measures ANOVAs (for within group effects) and mixed-model ANOVAs (to compare within and between group effects). However, because data were non-normal, we evaluated the significance of the F statistic with approximate permutation tests, randomizing the assignment of the data points across the conditions being tested 10,000 times, and comparing the F statistic to this distribution.

We used Spearman's rank correlations to examine individual differences in Experiment 6. Correlations were corrected for the reliability of the threshold measurements using the Spearman correction for attenuation (16). We used standardized Cronbach's alpha as a measure of reliability (17, 18). This entailed calculating the Spearman correlation between pairs of the 3 analyzed adaptive track thresholds for each condition, averaging these three correlations, and applying the Spearman-Brown correction to estimate the reliability of the mean of the three adaptive threshold measurements. Standard errors for correlations were estimated by bootstrapping the correlations 10,000 times. To calculate the significance of the interaction between conditions, we first calculated the interaction (the difference of differences). For instance, for the Harmonic and Inharmonic conditions, this term was:

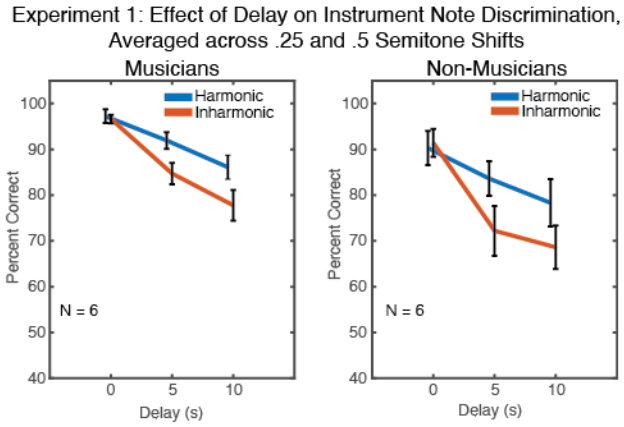
$$(r_{s_{Harmonic}}^{no-delay} - r_{s_{Harmonic}}^{delay}) - (r_{s_{Inharmonic}}^{no-delay} - r_{s_{Inharmonic}}^{delay})$$

Then, we approximated a null distribution for this interaction, permuting conditions across participants and recalculating the difference of differences 10,000 times. To determine statistical significance we compared the actual value of the interaction to this null distribution.

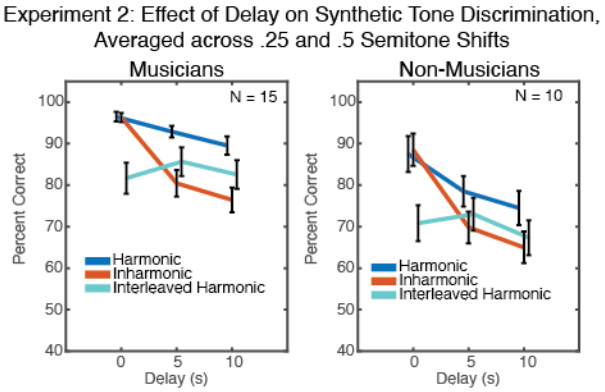
References

1. M. Kleiner, D. Brainard, D. Pelli, What's new in Psychtoolbox-3? (2007).
2. K. J. P. Woods, M. H. Siegel, J. Traer, J. H. McDermott, Headphone screening to facilitate web-based auditory experiments. *Atten. Percept. Psychophys.* **79**, 2064-2072 (2017).
3. M. Goto, H. Hashiguchi, T. Nishimura, R. Oka (2003) RWC Music Database: Music Genre Database and Musical Instrument Sound Database. in *The 4th International Conference on Music Information Retrieval (ISMIR 2003)*, pp 229-230.
4. H. Kawahara, STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology* **27**, 349-353 (2006).
5. H. Kawahara, M. Morise, TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. *SADHANA* **36**, 713-722 (2011).
6. J. H. McDermott, D. P. W. Ellis, H. Kawahara (2012) Inharmonic speech: A tool for the study of speech perception and separation. in *Proceedings of SAPA-SCALE* (Portland, OR), pp 114-117.
7. S. Popham, D. Boebinger, D. P. Ellis, H. Kawahara, J. H. McDermott, Inharmonic speech reveals the role of harmonicity in the cocktail party problem. *Nat. Commun.* **9**, 2122 (2018).
8. M. J. McPherson, J. H. McDermott, Diversity in pitch perception revealed by task dependence. *Nat. Hum. Behav.* **2**, 52-66 (2018).
9. B. R. Glasberg, B. C. J. Moore, Derivation of auditory filter shapes from notched-noise data. *Hear. Res.* **47**, 103-138 (1990).
10. M. Slaney (1998) Auditory toolbox version 2. in *Interval Research Corporation Technical Report* (Interval Research Corporation).
11. D. Pressnitzer, R. D. Patterson, "Distortion products and the perceived pitch of harmonic complex tones" in Physiological and Psychophysical Bases of Auditory Function, D. J. Breebaart, Ed. (Shaker Publishing, 2001), pp. 97-104.
12. S. Norman-Haignere, J. H. McDermott, Distortion products in auditory fMRI research: Measurements and solutions. *NeuroImage* **129**, 401-413 (2016).
13. T. Houtgast, Subharmonic pitches of a pure tone at low S/N ratio. *J. Acoust. Soc. Am.* **60**, 405-409 (1976).
14. J. W. Hall, R. W. Peters, Pitch for nonsimultaneous successive harmonics in quiet and noise. *J. Acoust. Soc. Am.* **69**, 509-513 (1981).
15. F. Faul, E. Erdfelder, A.-G. Lang, A. Buchner, G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175-191 (2007).
16. C. Spearman, The proof and measurement of association between two things. *Am. J. Psychol.* **15**, 72-101 (1904).
17. L. J. Cronbach, Coefficient alpha and the internal structure of tests. *Psychometrika* **16**, 297-334 (1951).
18. C. F. Falk, V. Savalei, The relationship between unstandardized and standardized alpha, true reliability, and the underlying measurement model. *J. Pers. Assess.* **93**, 445-453 (2011).

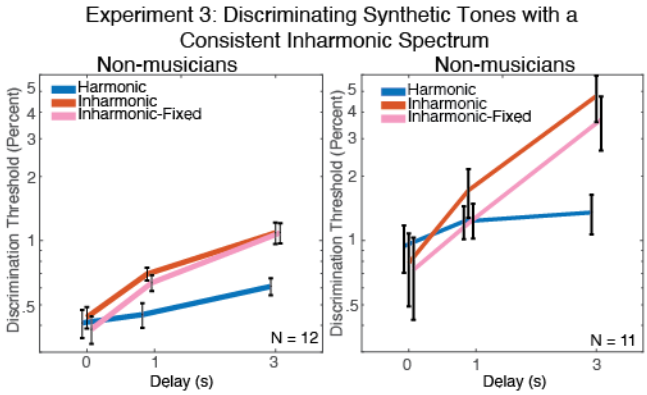
Supplementary Figures



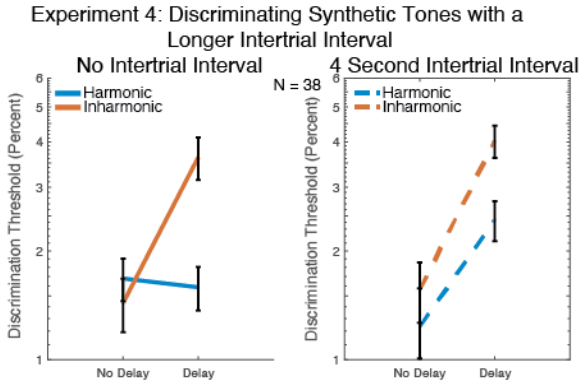
Supplementary Figure 1. Results from Experiment 1, plotted separately for musicians and non-musicians. Results are averaged across the two difficulty levels (.25 and .5 semitones) to maximize power. Error bars show standard error of the mean. There was no interaction between musicianship, harmonicity and delay length ($F(2,20)=0.58$, $p=.57$, $\eta_p^2=.06$), and the interaction between delay and harmonicity was significant in non-musicians alone ($F(2,10)=6.48$, $p=.02$, $\eta_p^2=.56$)).



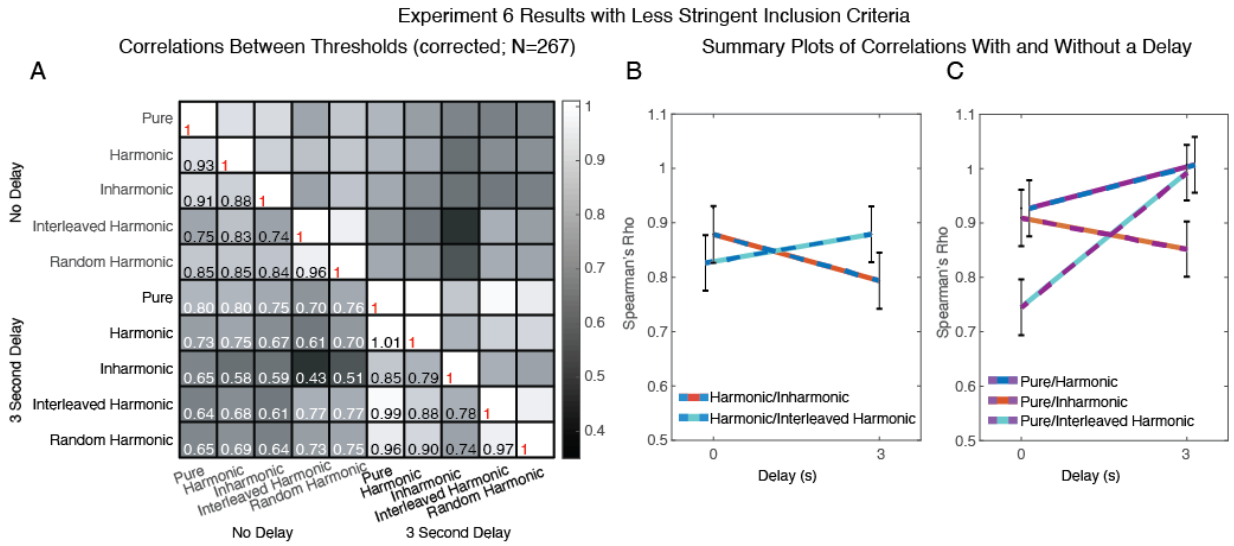
Supplementary Figure 2. Results from Experiment 2, plotted separately for musicians and non-musicians. Results are averaged across the two difficulty levels (.25 and .5 semitones) to maximize power. Error bars show standard error of the mean. As in Experiment 1, the effects were qualitatively similar for musicians and non-musicians. Although there was a significant main effect of musicianship ($F(1,23)=10.28$, $p<.001$, $\eta_p^2=.99$), the interaction between the effects of delay and harmonicity was significant in both musicians ($F(2,28)=20.44$, $p<.001$, $\eta_p^2=.59$) and non-musicians ($F(2,18)=11.99$, $p<.001$, $\eta_p^2=.57$), and there was no interaction between musicianship, stimulus type (Harmonic, Inharmonic, Interleaved Harmonic), and delay length ($F(4,92)=0.19$, $p=.98$, $\eta_p^2=.01$).



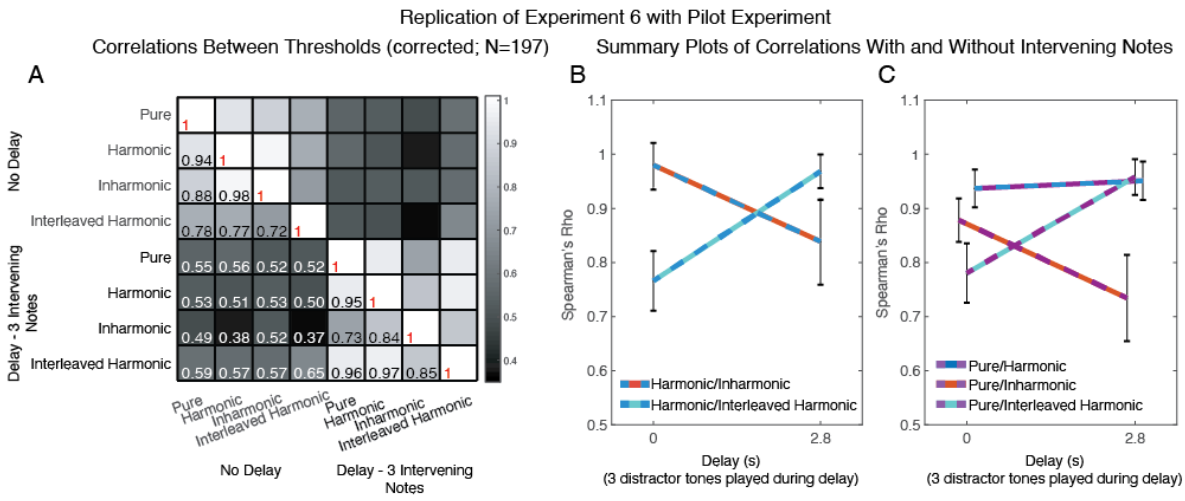
Supplementary Figure 3. Results from Experiment 3, plotted separately for musicians and non-musicians. Error bars show within-subject standard error of the mean. We again observed significant interactions between the effects of delay and harmonicity in both musicians ($F(4,44)=3.85$, $p=.009$, $\eta_p^2=.26$) and non-musicians ($F(4,40)=3.04$, $p=.028$, $\eta_p^2=.23$), and no interaction between musicianship, stimulus type (Harmonic, Inharmonic, Inharmonic-Fixed), and delay length ($F(4,84)=1.05$, $p=.07$, $\eta_p^2=.05$).



Supplementary Figure 4: Results from Experiment 4, measuring discrimination of synthetic tones with and without a delay between notes, and with and without a longer intertrial interval. Results from trials with (right) and without (left) an added 4 second delay between trials are plotted separately. Error bars show within-subject standard error of the mean. The interaction between within-trial delay (0 vs. 3 seconds) and stimulus type (Harmonic vs. Inharmonic) was present both with and without the longer intertrial interval (with: $F(1,37)=4.92$, $p=.03$, $\eta_p^2=.12$; without: $F(1,37)=12.34$, $p=.001$, $\eta_p^2=.25$).



Supplementary Figure 5. Individual differences results with less stringent inclusion criteria. Instead of including only those participants who performed as well as in-lab participants, we excluded participants if their average threshold across all 5 conditions on the first run of the no-delay trials was greater than 5% (just under 1 semitone). This excluded 183 of 450 participants, leaving 267 participants (136 female, mean age=35.8 years, S.D.=9.5 years). 94 of these participants reported greater than four years of musical training (mean=10.8, S.D.=6.6 years). (A) Matrix of the correlation between thresholds for all pairs of conditions. Correlations are Spearman's rho, corrected for the reliability of the threshold measurements (i.e., corrected for attenuation). (B) Comparison between Harmonic/Inharmonic and Harmonic/Interleaved Harmonic correlations, with and without a delay. The interaction between Harmonic/Inharmonic and Harmonic/Interleaved Harmonic correlations remained significant even with this more lenient inclusion criteria (difference of differences between correlations with and without a delay = 0.14, p=.044). Error bars in b and c show standard error of the mean, calculated via bootstrap. (C) Comparison between Harmonic/Pure, Inharmonic/Pure, and Interleaved Harmonic/Pure correlations, with and without a delay. The interaction between the Inharmonic/Pure and Interleaved Harmonic/Pure correlations likewise remained significant with the less stringent inclusion criteria (difference of differences between correlations with and without a delay = 0.30, p<.001).



Supplementary Figure 6. Replication of individual differences results with pilot experiment. Results of Experiment 6 were replicated in two pilot experiments, the data from which are combined in this figure. Both pilot experiments were run online using 2-down-1-up adaptive procedures. Participants discriminated tones identical to those used in Experiment 6, except for the Random Harmonic conditions, which we thus omitted from this figure. Instead of having a 3-second silent delay between test tones, both pilot experiments presented three intervening distractor notes between the test tones. In these conditions, participants heard the first test tone, a 200ms pause, three back-to-back 800ms notes, a 200ms pause, and then the second test tone (yielding a total delay between the two test tones of 2.8 seconds). For two of the four adaptive runs, intervening notes were harmonic, and for the other two runs they were inharmonic (intervening tones were generated in the same way as the main test tones). The runs for all stimulus conditions were randomly ordered throughout the experiment. 310 participants completed the first pilot experiment, in which the intervening notes were chosen randomly from a 7-semitone distribution surrounding the first note (loosely modeled after the method used in Semal & Demany, 1990). 295 participants completed the second pilot experiment, in which the 3 intervening notes were chosen randomly from a uniform distribution spanning 178.2 Hz-449 Hz (200-400 Hz +/- 2 semitones). In the first pilot experiment, adaptive runs for tones without an inter-stimulus delay (and thus without intervening notes) were initialized at 1 semitone pitch difference, and adaptive tracks for intervening note conditions were initialized at a 2 semitone pitch difference. For the second pilot experiment, all adaptive tracks were initialized at a 2 semitone pitch difference. Because results from the two pilots were similar, we combined the data, and then used the same filtering procedure used in Experiment

6 – participants who performed worse than 2.18% across the first (of four) runs on conditions without intervening notes were removed from further analysis. This excluded 408 of the total 605 participants, leaving 197 participants (77 female, mean age=34.2 years, S.D.=9.8 years). 52 of these participants reported greater than four years of musical training (mean=10.9, S.D.=8.8 years). (A) Matrix of the correlation between thresholds for all pairs of conditions. Correlations are Spearman's rho, corrected for the reliability of the threshold measurements (i.e., corrected for attenuation). (B) Comparison between Harmonic/Inharmonic and Harmonic/Interleaved Harmonic threshold correlations, with and without a delay. The interaction between Harmonic/Inharmonic and Harmonic/Interleaved Harmonic correlations was significant in this pilot study (difference of differences between correlations with and without a delay = 0.34, p=.006), replicating the effect from Experiment 6. Error bars show standard error of the mean, calculated via bootstrap. (C) Comparison between Harmonic/Pure, Inharmonic/Pure, and Interleaved Harmonic/Pure correlations, with and without a delay. The interaction between the Inharmonic/Pure and Interleaved Harmonic/Pure condition was also significant (difference of differences between correlations with and without a delay = 0.32, p=.008), again replicating the effect seen in Experiment 6.