EXPLORING THE DIMENSIONS OF HUMAN SENSITIVITY TO MUSICAL TONALITY

Chris Fagan, Charles E. Wright, Charles Chubb
Department of Cognitive Sciences, UC Irvine

Corresponding author:
Charles Chubb
Department of Cognitive Sciences
UC Irvine
Irvine, CA 92697-5100
cfchubb@uci.edu
949-824-1481

Abstract

This study investigates strategies available to listeners for classifying tone scrambles: rapid, randomly-ordered sequences of pure tones drawn from the thirteen notes from G_5 to G_6. In a given condition, the listener strives to classify each scramble in accordance with a specified feedback rule. For each of three conditions ("structured-random," "major-minor," and "squeezed-spread"), we measure the impact exerted on judgments by each of the thirteen different notes, given that the note was assigned to (a) the last tone in the scramble, (b) a non-terminal scramble tone when the last tone is a tonic, and (c) a non-terminal tone when the last tone is not a tonic. We also measure sensitivity to tonal information at different sequence locations. Results: (1) impact functions differ markedly in the major-minor versus the structured-random tasks; (2) although superficially similar to the major-minor task, the squeezed-spread task is much more difficult; (3) listeners show heightened sensitivity to the last tone of the scramble; (4) especially in the major-minor task, the note of the last tone affects judgments differently than the note of a non-terminal tone; and (5) in the major-minor task, sensitivity to information from all tones is enhanced if the scramble ends with a tonic.

1 Introduction

1.1 Tonality, macroharmony, and centricity

This paper focuses on tonality perception. However, as observed by Krumhansl, 2004, "the term tonality takes on diverse meanings within different musical periods, cultures, and theoretical traditions. Its meaning also depends on the disciplinary approach taken to understanding tonality" (p 253). It is therefore important that we start by circumscribing carefully the particular meaning we give this term. First, in this study, we restrict our attention to music that can be constructed from the notes of the (equal-tempered) chromatic scale used in western music. Second, our focus is on the perceptual properties that emerge from music due to the relative frequencies with which different notes occur in the music. In other words, we focus here on qualities that derive from the histogram (i.e., the first-order statistics) characterizing the notes in the music.

There are at least two general ways in which we anticipate that the histogram of a piece is likely to influence the perceptual qualities the music produces. First, the histogram is likely to select a specific subset of the set of all possible notes, thereby establishing the "machroharmony" of the music, to use Tymoczko's (2011) term. Loosely speaking, we can think of the machroharmony of the music as the set of tones to which it assigns non-zero probabilities. Second, the histogram may or may not emphasize (i.e., assign a substantially higher probability to) a single note above all others. If it does, then this note is likely to be established as the tonal center, or "tonic," of the music. A great deal of effort has been focused on understanding the features of music as it unfolds that control which tone listeners perceive to be the tonic at any instant in time (e.g., Brown, 1988, Cohen, 1991; Oram & Cuddy, 1995; Smith & Schmuckler, 2004). Tonally centric music seems able to evoke a range of qualities distinct from those evoked by tonally acentric music, and many of these qualities seem to be related to the special status of the tonic. For reasons that remain obscure, listeners tend to experience the tonic as the "goal" of tonally centric music, the note to which the music needs to return in order to be able to rest with stable finality.

One can use the same machroharmony and yet produce various different auditory qualities by establishing different notes as the tonic. For example, different musical modes may all share the same machroharmony (the eight notes of a given major scale) but differ in the notes they take as their tonic. In particular, the C ionian mode (the major mode) has a machroharmony comprising the seven notes, C, D, E, F, G, A, and B. The same is true of the A aeolian mode (the natural minor mode). However, in the C ionian mode, the C is established as the tonic whereas in the

A Aeolian mode the A is the tonic. For many listeners, this difference in choice of tonic leads to strikingly distinct auditory qualities.

Indeed, much of the underlying motivation for the work reported here originates in the observation that for many if not all listeners, music in the major mode sounds “happy” whereas music in the minor mode sounds “sad.” Why it is that these mood words seem very naturally to apply to the qualities evoked by these two sorts of music has baffled musicians, composers and philosophers for centuries; although it is not hard to see how a picture or a poem might be happy or sad by virtue of the circumstances to which it refers, a melody refers to nothing outside itself, yet the “mood” of a melody seems to strike many listeners with stunning immediacy. (For an overview of the main theories that have been offered to account for this effect, see Crowder, 1984.) It should be noted, however, that although substantial research supports the claim that the major mode sounds happy and the minor mode sounds sad (Crowder, 1984,1985a,b; Dalla Bella, Peretz, Rousseau, & Gosselin, 2001; Gagnon & Peretz, 2003; Gerardi & Gerken, 1995; Henlein, 1928 (as reanalyzed by Crowder, 1984); Hevner, 1935; Kastner & Crowder, 1990; Whissell & Whissell, 2000), a number of studies document that the difference between the major *versus* minor modes is not as vividly distinct for many listeners as one might expect given its importance in western music theory (e.g., Blechner, 1977; Crowder, 1985a; Halpern, 1984; Halpern, Bartlett & Dowling, 1998; Leaver & Halpern, 2004).

1.2 Feedback-free versus feedback-driven tasks

Naturally enough, much of the research in music cognition has tended to use stimuli closely related in form to the music produced by various cultures, and listeners are asked to rate various qualities of these stimuli. For example, in the probe-tone technique pioneered by Krumhansl & Shepard (1979) and Krumhansl & Kessler (1982), a sequence of tones or chords is presented so as to establish a melodic context characteristic of western music. Then a final probe-tone is presented, and listeners are asked to rate the goodness of the fit of the probe-tone within the context established by the leading sequence. In the years since the probe-tone technique was introduced, this method has dominated research in tonality perception and has produced many important insights. The goal of such experiments is to discover what factors control which tones fit well for the listener into which contexts. Note that in the probe-tone task, there is no right answer to the question being posed; the listener him/herself is the final judge of how well a tone fits into a given context. Therefore it is not appropriate to give feedback in this task.

The distinction between feedback-free *versus* feedback-driven tasks is important because these two general classes of tasks provide very different sorts of information. In any task, the participant must base his/her responses across trials on some internally computed, decision statistic extracted from the stimulus. Typically feedback-free tasks are required if the goal is to analyze the factors that imbue music with some particular quality X for the listener. In the case of $X =$ sadness, for example, we may be convinced from our own experience that it makes sense to use the word “sad” to characterize some music and not other music. This gives us reason to suppose that, like ourselves, other listeners have access to an internal statistic whose value naturally and spontaneously reflects the degree of music’s sadness. If our goal is to understand the physical features of the music that control the value of this particular, internally available statistic, we must rely on the participant as the expert. We cannot give feedback because we do not know the right answer: it is precisely what makes music sad for the listener that we seek to discover. In essence, what we learn in a feedback-free experiment of this sort is the functional meaning the participant assigns to the quality X as it applies to music.

By contrast, the decision statistics used in feedback-driven tasks (e.g., tasks in which the correct response is revealed to the participant at the end of each trial) cannot be assumed to reflect any such easily named musical qualities. The cost of providing feedback is that we allow the participant to use any strategy that works to optimize performance; therefore we can have no guarantee that the decision statistic the participant adopts will reflect any particular musical quality we might name. The benefit of providing feedback is that it may enable the participant (with enough practice) to optimize his/her performance in making the required judgment. If, after sufficient training, the participant really has optimized his/her performance, then by analyzing the dependence of the participant's responses on the features of the stimulus, we can gain insight into the processing constraints imposed by the functional architecture of the participant's perceptual system.

In contrast to much of the previous research in music cognition, in this project we use feedback-driven tasks. As suggested by the foregoing discussion, our purpose is not to understand how the physical features of music produce any particular, nameable qualities for listeners. Instead, our aim is to begin the process of cataloguing the dimensions of sensitivity listeners have to variations in tonality (where tonality is understood as the probability distribution characterizing the relative frequencies with which different notes enter the music). The stimuli we use are designed specifically for this purpose.

1.3 Tone scrambles

These stimuli are called *tone scrambles.* A tone scramble η is a rapid sequence of briefly windowed, pure tones presented in random order. In the studies reported here, the tones are drawn from the set of thirteen notes from G_5 to G_6 in the standard tuning (A 440) equally-tempered scale and are presented at a rate of around 15/sec. The *histogram* of η is the probability distribution that gives, for each note $n \in \{G_5, A\flat, \cdots, G_6\}$, the proportion of notes in the scramble that are assigned note n. When our focus is on the temporal location of a tone in a scramble, we shall refer to the tone as a "pip." We will write N_{pips} for the number of pips in the scramble and $\eta(t)$ for the note assigned to the t^{th} pip of a scramble for $t = 1, 2 \cdots, N_{pips}$. The scrambles used in all of the studies described here have $N_{pips} = 26$ (except in the case of one listener in one condition in which $N_{pips} = 27$).

On each trial, the listener hears a single scramble and attempts to classify the scramble according to a particular rule (determined by the experimental condition) which depends only on the relative frequencies with which different notes occur in the scramble. For example, in one of our conditions (in which G is established as the tonic by virtue of being the most commonly occurring tone in each scramble), the listener strives to judge whether the sum total of B's and E's (the third and sixth degrees of the major diatonic scale with tonic G) in the scramble is greater than the sum total of B♭'s and E♭'s (the third and sixth degrees of the minor diatonic scale with tonic G).

1.4 The analogy between tonality perception and color perception

The seemingly spontaneous nature of the qualities evoked by variations in tonality has parallels to the perception of color.

The visual world is awash in color, with a virtually unlimited number of distinct hues of varying saturation and brightness immediately apparent in any scene, yet all of these colors are derived from three classes of receptors sensitive to different wavelengths of light, the *S*-, *M*- and *L*-cone classes, which define three distinct dimensions of sensitivity to wavelength. Note, however, that this aspect of our experience of color–that it results from exactly three dimensions of sensitivity–is by no means directly obvious from our visual experience. It required careful empirical investigation to establish this fact (e.g., Maxwell, 1855).

Our working hypothesis is that just as human vision has three classes of cones that are differentially sensitive to the wavelengths present in a light, human audition has a small number of mechanisms that are differentially sensitive to the different notes occurring in our scrambles. Let $f_k : \{G_5, A\flat, \cdots, G_6\} \rightarrow \Re$ be the function that gives the sensitivity of the k^{th} tonality mechanism to different notes. For our purposes, the important implication of the color analogy is that a listener should only be able to discriminate two scrambles if the difference ϕ between their histograms has a non-zero correlation with one or more of the sensitivity functions f_k.

Although we hypothesize that perception of the tonal properties of scrambles is analogous to color perception, several differences between scrambles *versus* lights complicate this picture. First, in tonally centric music, what seems to matter most in determining the quality of the music is not (as one might infer from the analogy to color) the note histogram *per se*, but rather the histogram of scale degrees relative to the tonic established by the music. For example, there is a close qualitative similarity of music in the aeolian mode (the natural minor) irrespective of the tonic used to center the music. In the current experiments, we will sidestep this issue by using a fixed tonic across all experimental conditions. Second, a light is instantaneously analyzed by the retina; by contrast, a scramble unfolds in time, raising the possibility that some pips (e.g., the first and/or last) may influence the activation in a given tonal mechanism more strongly than others. Third, tonally centric melodies have a strong tendency to end on the tonics they establish. Such resolution confers a sense of stability to the ending that is likely to be lacking if the last note is, for example, the seventh degree of the scale. This raises the possibility that the very last pip of a scramble may exert not only a stronger influence on the activation of a given tonality mechanism but also a pattern of influence that differs from that exerted by non-terminal pips. As will become clear in section 3, the model we use strives to preserve the logic suggested by the analogy to color while taking these complexities into account.

2 Methods

2.1 Pips

Each pip in a scramble is a pure tone, 65 ms in duration, comprising 3250 samples presented at 50,000 samples per second. The envelope of each pip is ramped to prevent clicking between pips using the following raised cosine windowing function:

$$W(x) = \begin{cases} \frac{1}{2}\big(1 - cos(\frac{\pi x}{1125})\big) & 1 \leq x \leq 1125 \\ 1 & 1126 \leq x \leq 2125 \\ \frac{1}{2}\big(1 - cos(\frac{\pi(x-3250)}{1125})\big) & 2126 \leq x \leq 3250 \end{cases} \quad (1)$$

2.2 Scramble classification tasks

We call any function V assigning non-negative integers to the set $\{G_5, A\flat, \cdots, G_6\}$ a *note-count vector* provided $\sum_{n=G_5}^{G_6} V(n) = N_{pips}$. Note that if η has note-count vector V, then its histogram is equal to $\frac{V}{N_{pips}}$.

The experiment reported here uses three task conditions. In each of these, the listener is asked on a given trial to classify a tone scramble η into one of two classes. In each condition, there are two base note-count vectors, B_1 and B_2. On a given trial, the listener hears a scramble η with some note-count vector V and is required to classify the scramble according to whether V is closer in Euclidean distance to B_1 *versus* B_2, that is, according to whether

$$|V - B_1| = \sqrt{\sum_{n=G_5}^{G_6} \big(V(n) - B_1(n)\big)^2} < \sqrt{\sum_{n=G_5}^{G_6} \big(V(n) - B_2(n)\big)^2} = |V - B_2|. \tag{2}$$

Let

$$D(n) = B_1(n) - B_2(n) \qquad \text{and} \qquad A(n) = \frac{B_1(n) + B_2(n)}{2} \qquad \text{for } n = G_5, A\flat, ..., G_6. \tag{3}$$

Then a small amount of algebra shows that for random notes X_V and X_A with probability distributions $\frac{V}{N_{pips}}$ and $\frac{A}{N_{pips}}$ respectively, V is closer to B_1 than to B_2 in Euclidean distance if and only if

$$\frac{1}{N_{pips}} \sum_{t=1}^{N_{pips}} D\big(\eta(t)\big) = E[D(X_V)] > E[D(X_A)]. \tag{4}$$

Thus, an ideal strategy for making the required judgment is to compute a statistic that takes the average value of D applied to the notes of all pips in the scramble and compares the result to the criterion $E[D(X_A)]$. For this reason we call D the *target discrimination function* for a given one of the three tasks we ask our participants to try to perform.

2.3 Conditions

There were three parts of the experiment, all with the same protocol but using different pairs of base note-count vectors. Thus, each part of the experiment required the participant to attempt to classify scrambles based on a different sort of tonal quality, i.e., to try to achieve a different target discrimination function. The base note-count vectors and target discrimination functions for the three tasks are given in Table 1.

Please insert Table 1 around here

In the Structured-Random condition (see the first part of Table 1), B_1 and B_2 differed in tonal "structuredness", with B_1 emphasizing the tonic, the fourth, the fifth and the octave of the diatonic scale and B_2 assigning equal numbers of all thirteen notes. Each of the base note-count vectors used in this part of the experiment was neutral in mode, seeming neither major nor minor. However, the two base note-count vectors yielded scrambles that sounded quite distinct.

In the Major-Minor condition (see the second part of Table 1), B_1 and B_2 differed in mode, with B_1 emphasizing major thirds and major sixths *versus* minor thirds and minor sixths and B_2 emphasizing minor thirds and minor sixths *versus* major thirds and major sixths.

In the Squeezed-Spread condition (see the third part of Table 1), B_1 and B_2 were superficially similar to the base note-count vectors of condition 2; however, each of B_1 and B_2 was ambiguous in mode. B_1 emphasized major thirds and minor sixths *versus* minor thirds and major sixths; B_2 emphasized minor thirds and major sixths *versus* major thirds and minor sixths. Our aim in this condition was to produce note-count vectors comparable in their physical difference to the base note-count vectors of condition 2 but unlikely to be discriminable using a decision statistic sensitive to the difference between the major *versus* minor modes. We anticipated that listeners would be much less sensitive to this difference than to the difference in condition 2, despite the superficial similarity of this judgment to the condition 2 judgment.

In each condition, various different note-count vectors V were used to generate scrambles. In particular, for a condition in which the base note-count vectors were B_1 and B_2, with D and A given by Eq. 3, we made sure that the set Θ of note-count vectors we used satisfied the following conditions:

1. $B_1 \in \Theta$ and $B_2 \in \Theta$.
2. The set of functions $\{\phi = V - A | V \in \Theta\}$, spanned the set of all functions f for which $\sum_{n=G_5}^{G_6} f(n) = 0$. (This condition is crucial in order to insure that the resulting data allow us to estimate the relative sensitivity with which the listener's decision statistic is influenced by different notes in $\{G_5, A\flat, \cdots, G_6\}$.)
3. There was always a correct response: i.e., V was always closer in Euclidean distance to one of B_1 *versus* B_2.
4. For every note-count vector $V \in \Theta$, there existed a complementary note-count vector $V' \in \Theta$ such that $V + V' = A$.

In different conditions, slightly different algorithms were used to produce the set Θ of note-count vectors used to generate stimuli. To take one example, in the structured-random condition, the set Θ contained B_1, B_2 (the two base note-count vectors themselves), and in addition, the forty-four note-count vectors generated by taking $B_1 + Q_k, B_1 - Q_k, B_2 + Q_k$, and $B_2 - Q_k$, where Q_k is function given by the k^{th} row of Table 2.

Please insert Table 2 around here

The 11 rows of Table 2 are orthogonal functions. [1] It is easy to check that this method yields a set Θ of 46 note-count vectors satisfying the conditions above.

An experimental block included one trial per each note-count vector $V \in \Theta$. On the trial in which the note-count vector was V, the scramble contained $V(n)$ pips with note n for each $n \in \{G_5, A\flat, \cdots, G_6\}$, and these N_{pips} notes were randomly sequenced.

2.4 Procedure

2.4.1 Listeners

Six listeners took part in the experiment; we will refer to them as listeners 1, 2, 3, 4, 5 and 6. Listeners 1,2 and 3 participated in Condition 1 (the "structured-random" task); all six listeners participated in Condition 2 (the "major-minor" task), and listeners 1 and 2 participated in Condition 3 (the "squeezed-spread" task). All listeners had normal hearing by self-report. Listeners 1 - 5 had experience with playing music, but none had any significant formal training. None of these listeners had perfect pitch. By contrast, Listener 6 had extensive training as a concert pianist as well as perfect pitch.

[1] Specifically, the rows Table of Table 2 are generated by appending a zero to each of the 11 Hadamard functions of length 12 that sum to 0.

2.4.2 Equipment

Listeners were seated in a quiet laboratory room. Stimuli were generated in Matlab on a PC running Windows XP and were delivered over Sennheiser HD 265 headphones at approximately 70dB (precise control of volume is not critical to the design of the experiment, so listeners were free to adjust the volume as necessary for comfort).

2.4.3 Trials and blocks

Each trial consisted of one scramble followed by a prompt to identify it as one of two types, as determined by the condition. For instance, in condition 2, the "major" response was indicated by the "1" key and minor by the "2" key. After the response, the listener was presented with the text "Correct" or "Incorrect" on the screen. Finally, an additional key press was required to begin each trial, allowing listeners to pause at their discretion during sessions. In most conditions, those listeners who participated ran 50 blocks, each of 46 trials (including one trial per each of the note-count vectors used in that condition per block). The exceptions were that Listener 1 ran 83, Listener 2 ran 110, Listener 3 ran 100 and Listener 6 ran 165 blocks in condition 2; Listener 2 ran 40 blocks in condition 1; Listener 1 ran 31 blocks in condition 3. In each case, prior to data collection, the listener went through several practice blocks.

3 Modeling

3.1 The data

For each trial, the sequence of 26 notes presented on that trial was stored along with the listener's response. We write $\eta_j(t)$ for the note assigned to the t^{th} pip of the j^{th} scramble presented to the listener in this condition, and $R(j)$ for the listener's response to this scramble.

3.2 The general form of the model

We will assume that on the j^{th} trial of a given condition, the listener extracts a statistic Γ_j from the scramble presented on that trial. We further assume that this statistic is degraded by adding a standard normal random variable X_j (with X_j's independent across trials). Then for some criterion $Crit$ (that the listener is assumed to select so as to optimize performance), we assume that the listener responds "2" if

$$\Gamma_j - X_j > Crit \tag{5}$$

and otherwise responds "1." The main issue we will need to resolve is how best to define Γ_j to illuminate the processes our listeners are using to make their decisions. As described in Secs. 3.4 and 3.5, we find it useful to explore two different models of Γ_j, the "last-pip-special" model and the "resolution-interaction" model. Before describing these models, we briefly sketch the Bayesian fitting method we use to derive parameter estimates for both models.

3.3 Fitting methods

For each participant in each condition, Markov-chain Monte-Carlo simulation is used to derive a sample from the posterior density characterizing the joint distribution of the model parameters. Each parameter was assigned a prior density uniform on the interval $(-10, 10)$, which proved to be many times larger than the 95% credible intervals ultimately observed. In an initial simulation stage, multiple, random starting points were used, and a total sample of size 150,000 samples was collected. Then the median sample vector m from this set was taken as a starting point for a search process that obtained the maximum likelihood parameter vector $MaxLV$ in the neighborhood of m. Finally, a new MCMC process of length 40,000 was run using $MaxLV$ as the starting point. In all cases, the last 20,000 samples provided what appeared to be a stable estimate of the posterior density.

3.4 The "last-pip-special" model of Γ_j

As we observed in Sec. 1.4, a scramble unfolds in time. This leads us to anticipate that Γ_j may be more sensitively influenced by the notes at some pip locations than others. In addition, tonally centric music has a strong tendency to resolve to its tonic. This suggests that notes occurring on the last pip of a scramble may exert a pattern of influence that differs from that exerted by non-terminal pips. A rudimentary model that suffices to detect such effects if they exist is given by Eq. 6:

$$\Gamma_j = \sum_{t=1}^{N_{pips}-1} W(t)F(\eta_j(t)) + W_{last}F_{last}(\eta_j(N_{pips})), \tag{6}$$

where F and F_{last} both map $\{G_5, A\flat, \cdots, G_6\}$ into $\Re$, sum to zero, and have norm 1. In addition, we require $W_{last} \in \Re$ to be positive (otherwise, the sign of F_{last} can reverse with that of W_{last}); and we require the sum of the weights $W(1), \cdots, W(N_{pips}-1)$ to be positive (otherwise, the sign of F can reverse with the sign of W).

In the last-pip-special model,

1. the function F gives the relative impacts exerted on the listener's decision statistic Γ_j by the different notes $\{G_5, A\flat, \cdots, G_6\}$ when they occur on non-terminal pips of the scramble;
2. for $t = 1, 2, \cdots, N_{pips} - 1$, $W(t)$ gives the sensitivity with which Γ_j is influenced by notes occurring on pip t;
3. F_{last} gives the relative impacts exerted on Γ_j by the different notes when they occur on the last pip of the scramble, and
4. W_{last} gives the sensitivity with which Γ_j is influenced by notes occurring on the last pip of the scramble.

We did in fact fit this "last-pip-special" model to the data of all of our listeners. We present the results for Listener 1 in the major-minor task (Fig. 1) to motivate our primary model.

Please insert Fig. 1 around here

It should be noted that Listener 1 was the best performer (out of six listeners) in the major-minor task. The panel on the left of Fig. 1 gives the estimated pip-note impact functions F and F_{last} (of Eq. 6) derived under the last-pip-special model. Notes of the scale are arranged on the horizontal axis in accordance with the circle of fifths in this figure and in all other figures showing model fits because the pip-note impact functions achieved by our listeners take simpler, smoother forms when plotted on the circle of fifths than when plotted as a function of notes ordered by pitch height. Error bars are 95% Bayesian credible intervals. The thin line marked by square tokens gives the target discrimination function for the judgment. The thick black line shows F, and the dashed line shows F_{last}. The figure on the right shows $W(t)$ for pips $t = 1, 2, \cdots, N_{pips}$, where we have set $W(N_{pips}) = W_{last}$.

Note first that for this listener, the normalized impact function F that gives the relative impacts exerted on his judgments by notes occurring on non-terminal pips of the scramble matches the target function (marked by squares) fairly well (although there are significant deviations). Next note that for this listener the last pip does indeed exert dramatically stronger influence than any other pip on the decision. This difference in influence is strikingly abrupt: $W(t)$ is essentially flat across pips $t = 1, 2, \cdots, 25$ with a value near 0.7 and then jumps up to around 2.7 on the last pip. Perhaps more interestingly, the note-dependent pattern of influence exerted by the last pip (F_{last}) differs strongly from the pattern of influence exerted by the non-terminal pips in the scramble (F).

It is an implication of Eq. 4 that an ideal performer of this task should weight information from all pips equally and should use the same pip-note impact function on the last pip as he/she does on every other pip. In particular, Listener 1 would have done much better to distribute his pip sensitivity function W more evenly across all pips instead of loading so much weight on the last pip. In addition, Listener 1 would also have done better to use his non-terminal pip-note function F on the last pip instead of the function F_{last} that he actually used.

These observations raise a puzzle: Why did Listener 1 allow the last pip to exert the powerful and counterproductive influence that it did? Given that Listener 1 significantly outperformed the other five listeners in this task, we speculated that the peculiar influence exerted by the last pip on his judgments might actually be a feature of the optimal strategy available to human listeners. Perhaps it is only by allowing this last pip to exert the influence it does that Listener 1 is able to maximize his sensitivity to the information available from the other pips in the tone scramble; in essence, we suspected, the peculiar jolt delivered by the last pip to Listener 1's judgments is the price he pays to boost his sensitivity to non-terminal pips.

What sort of computation might one expect to produce effects of this sort, released as they are by the last pip of the scramble? In struggling to imagine an answer to this question, we were led to think about sentence processing. In extracting the meaning of a sentence, as the sentence unfolds, one builds a structure (a parse-tree) in which the roles of components of the sentence are determined. Only when the entire sentence has been parsed are roles of all components finally fixed. Thus, in an important sense, the impact exerted on the meaning of the sentence by any given component may not be firmly resolved until the sentence has ended.

Fig. 1 suggests that Listener 1 is unable to achieve the optimal strategy defined by Eq. 4. Suppose that instead the most sensitive device the listener has available for purposes of making the required judgments operates like a parser, extracting melodic meaning from the scramble by building a structure in which the roles of all tones are fixed only after the last pip has been heard. In this case,

1. one might expect the last-pip-special model of Eq. 6 to yield strong, and not necessarily interpretable, effects of the last pip on performance.

2. notes occurring at pips early in the scramble should interact with the note assigned to the last pip in influencing judgments.

The single strongest rule constraining melodic structure is the requirement that melodies end on the tonic. For most listeners, melodies that end on a tonic achieve a stable sense of finality that is lacking in melodies that don't. It thus seems likely that the strongest interactions we are likely to find (if any exist at all) will be conditioned on whether or not the scramble being processed ends on a tonic.

3.5 The "resolution-interaction" model of Γ_j

In order to investigate this possibility, we elaborate the model of Eq. 6 so that the influence exerted by any non-terminal pip in a scramble depends not only on the location of that pip and the note assigned to it but also on whether or not the scramble ends on a tonic. Specifically, we assume in this model that

$$\Gamma_j = \begin{cases} \sum_{t=1}^{N_{pips}-1} W_{tonic}(t) F_{tonic}(\eta_j(t)) + W_{last} F_{last}(\eta_j(N_{pips})) & \text{if } \eta(N_{pips}) = G_5 \text{ or } G_6 \\ \sum_{t=1}^{N_{pips}-1} W_{other}(t) F_{other}(\eta_j(t)) + W_{last} F_{last}(\eta_j(N_{pips})) & \text{otherwise.} \end{cases} \quad (7)$$

This model includes separate sensitivity functions for non-terminal pips when the scramble ends on a tonic (W_{tonic}) *versus* on some note other than the tonic (W_{other}), and separate non-terminal pip-note impact functions for scrambles that end on a tonic (F_{tonic}) *versus* on some note other than a tonic (F_{other}).

In order to keep the number of parameters in this model manageable, we approximate W_{tonic} (the sensitivity function for non-terminal pips when the scramble ends on a tonic) and W_{other} (the sensitivity function for non-terminal pips when the scramble does not end on a tonic) by third order polynomials. In addition, to insure that parameter values are uniquely determined, we impose the following constraints:

1. Each of W_{last} and the means of W_{tonic} and W_{other} are required to be greater than 0. (Otherwise, the signs of F_{last}, F_{tonic} and F_{other} and can reverse with those of W_{last}, W_{tonic} and W_{other} respectively).

2. The means of F_{tonic} and F_{other} are required to be 0. (Otherwise, the means of F_{tonic} and F_{other} can trade off with $Crit$).

3. The norms of F_{tonic}, F_{other} and F_{last} are required to be 1. (Otherwise, the norms of F_{tonic} and F_{other} can trade off the mean values of W_{tonic} and W_{other} and the norm of F_{last} can trade off with W_{last}).

This model has a total of 43 free parameters: 4 for each of W_{tonic} and W_{other}, 1 for W_{last}, 11 for each of F_{tonic}, F_{other} and F_{last}, and 1 for $Crit$.

There are two distinct ways in which the impact of non-terminal pips might depend on whether or not the scramble ends on a tonic. On the one hand, the relative impacts exerted by different notes occurring on non-terminal pips may depend on whether the scramble resolves to the tonic or not. In the context of the resolution-interaction model of Eq. 7, such a result would show up as a significant difference between F_{tonic} *versus* F_{other}. On the other hand, it might simply be that resolution to the tonic adds a quality of well-formed-ness to a tone scramble that makes its tonal quality more legible in a sense akin to that in which a grammatical sentence is easier to read. If so, then we might expect performance to be better for scrambles ending in tonics *versus* other notes. In the context of the resolution-interaction model of Eq. 7, such a result would be signaled by finding that the mean value of W_{tonic} is significantly greater than the mean value of W_{other}.

4 Results

In broad strokes, listeners performed (1) well in the structured-random classification task, (2) well in the major-minor classification task, but (3) poorly in the spread/squeezed classification task. This point is made clearly by Fig. 2.

Please insert Fig. 2 around here

For each listener and each of the three classification tasks, this figure plots the sensitivity-per-pip of the decision statistic achieved by the listener to the target function for the task. [2] For this plot, we use the fits of the last-pip-special model (Eq. 6) to estimate overall sensitivity-per-pip. This model yields pip-note impact functions F and F_{last} giving the relative impacts exerted on the listener's judgments by notes occurring at non-terminal pips (F) and by notes occurring on the last pip (F_{last}). The model also yields the function W that gives the sensitivity of the listener's decision statistic to notes occurring at each of the pips $t = 1, 2, \cdots, N_{pips}$. For D the target discrimination function for the task, we estimate sensitivity-per-pip by

$$\text{sensitivity-per-pip} = \frac{D}{N_{pips}|D|} \bullet \left(\left(\sum_{t=1}^{N_{pips}-1} W(t) \right) F + W(N_{pips}) F_{last} \right) \tag{8}$$

where the $\bullet$ indicates inner product: sensitivity-per-pip is the average value, across all N_{pips} pips t, of the correlation of D with the pip-note impact function for pip t scaled by $W(t)$, the participant's sensitivity to information from pip t. The plotted values of sensitivity-per-pip correspond roughly to

The reader will note that Listeners 1 and 2 achieve much higher sensitivity-per-pip for the structured-random and major-minor tasks than they do for the squeezed-spread task. Listener #2's sensitivity-per-pip in the squeezed-spread task is near zero. Although Listener #1 achieves slightly higher sensitivity-per-pip in this task, his level is still low in comparison to the levels he achieves in the structured-random and major-minor tasks. It should also be noted that Listener #1 spent time between blocks of experimental trials listening to enriched exemplars of the two sorts of scrambles so as to refresh his sense of the difference he was striving to discriminate whereas listener #2 did not. Another point to note: one might have expected Listener 6, a highly trained concert pianist with absolute pitch, to have an advantage in the major-minor task. Despite extensive practice in this task, however, her sensitivity-per-pip is no higher than those achieved by any of the other five listeners and is substantially lower than the sensitivity-per-pip achieved by Listeners 1 and 2.

4.1 Fits of the resolution-interaction model

The fits of the resolution-interaction model of Eq. 7 are given in Figs. 3 for the structured-random task, 4 and 5 for the major-minor task and 6 for the spread/squeezed task. The plotting conventions in each figure are the same. In each of these figures, the panels on the left show the estimated pip-note impact functions F_{tonic} (dark gray markers), F_{other} (light gray markers) and F_{last} (dashed line, black markers) derived under the resolution-interaction model of Eq. 7. Notes of the scale follow the circle of fifths along the horizontal axis. All three functions are constrained to sum to zero, and all three are normalized: i.e., for each, the sum of squared values is 1. The small, open squares

[2] The reader may wonder why it is necessary to introduce the statistic sensitivity-per-pip instead of just plotting percent correct in Fig. 2. The main reason for this is that no particular effort was made to insure that the conditions tested in different tasks were equally difficult in the sense of affording equal levels of information for purposes of making the required classifications. The statistic sensitivity-per-pip compensates for such differences in difficulty across tasks. In addition, whereas it is easy to derive credible intervals for sensitivity-per-pip, it is difficult to estimate reasonable confidence intervals for percent correct in a given condition since any such condition includes a variety of stimuli differing in difficulty.

mark the target discrimination function for the given task. Panels in the center plot $F_{other} - F_{tonic}$, and panels on the right plot the estimated pip-location sensitivity functions W_{tonic} (thick line) and W_{other} (thin line) for $t = 1, 2, \cdots, N_{pips}$, where $W_{other}(N_{pips}) = W_{tonic}(N_{pips}) = W_{last}$.

The main questions the resolution-interaction model was introduced to address are:

1. Do the notes of the chromatic scale exert different patterns of influence if a tone scramble ends on a tonic than they do if it ends on some other note?

2. Is the listener more sensitive to information at non-terminal pips of tone scrambles ending on tonics *versus* other notes?

The first question receives an ambiguous answer from the data. Some of the differences $F_{other} - F_{tonic}$ plotted in the middle panels of Figs. 3, 4, 5 and 6 seem to deviate significantly from zero at certain points suggesting an affirmative answer to question 1. However, in no instance are these deviations from 0 very large; moreover, the patterns of differences produced by different listeners seem largely idiosyncratic. Likelihood ratio tests (e.g., Mood, Graybill & Boes, 1974) were used to test the null hypothesis that $F_{tonic} = F_{other}$ for each participant in each condition. The results are given in Table 3.

Please insert Table 3 around here

In only one condition was the null hypothesis decisively rejected. This was for Listener 3 in the structured-random task. (It should be noted, however, that the p-value of the likelihood ratio test for this listener in the major-minor task (the only other task this listener performed) was also rather low, 0.020.) We conclude that non-terminal pip-notes show only weak, unsystematic tendencies to exert different patterns of influence on trials in which the tone scramble ends on a tonic than they do on trials in which the scramble does not end on a tonic. [3]

The second question receives a more interesting answer. Figure 7 addresses the question of whether listeners are more sensitive to information at non-terminal pips of tone scrambles ending on tonics *versus* other notes. This figure plots the average (across pips $t = 1, 2, \cdots, N_{pips} - 1$) A_{tonic} of $W_{tonic}(t)$ (black markers) and the average A_{other} of $W_{other}(t)$ (gray markers) for all listeners is all tasks. Error bars give 95% Bayesian credible intervals. Table 4 gives the results of likelihood ratio tests of the null hypothesis that $A_{tonic} = A_{other}$ for all listeners in all tasks.

Please insert Table 4 around here

A_{tonic} and A_{other} are equal within measurement error for Listeners 1, 2 and 3 in the structured-random task and also for Listeners 1 and 2 in the squeezed-spread task. A different picture emerges, however, in the major-minor task. Here we find that A_{tonic} differs significantly from A_{other} for each of Listeners 1, 3, and 6 (Bonferroni-corrected p-values do not reach significance for listeners 2, 4 and 5), and in each of these three cases, $A_{tonic} > A_{other}$ suggesting that in the major-minor task these listeners do indeed have higher sensitivity to the signals provided by non-terminal pips when the tone scramble ends on a tonic than they do when it ends on some other note.

Next we consider the patterns of results specific to each of the three different tasks.

[3]Some readers might note that there are three of these comparisons with p values less than 0.05. However, this is a multiple comparison situation, so a Bonferroni correction is appropriate. Thus to maintain the experiment-wise alpha at 0.05, the per-comparison alpha needs to be 0.0045.

4.2 Structured/Random Task

Results for Listeners 1, 2 and 3 in the structured-random task are plotted in Fig. 3.

Please insert Fig. 3 around here

Listeners varied in their ability to perform the required judgment. Consistent with sensitivities-per-pip plotted in Fig. 2, Listener 1 had the best performance (87.9% correct), followed by Listener 2 with 79.1% performance, and Listener 3 had the worst performance (71.7%). Note that although all pips were equally informative for this task, all three listeners show higher sensitivity for the last pip than for non-terminal pips.

For all three listeners, the three pip-note impact functions deviate significantly from the target discrimination function. This deviation is most striking for pip-note C, the fourth (subdominant) degree of both the major and minor diatonic scales with tonic G. Although listeners received trial-by-trial feedback that gave this note substantial positive weight, none of listeners were able to achieve pip-note impact functions (either F_{tonic}, F_{other} or F_{last}) that weighted this note much higher than 0. By contrast, the listeners were able achieve heightened sensitivity to occurrences of pip-note D, the fifth (dominant) degree of the scale. Another interesting deviation of the pip-note impact functions of all listeners from the target discrimination function is seen in the relative impact exerted on judgments by high *versus* low tonics (G_6's *versus* G_5's). For all three listeners, high tonics are more effective than low tonics in promoting a “structured” response. Also, for all three listeners, $D\sharp$'s (minor sixths) occurring in non-terminal pips were more effective at promoting a “random” response than expected from the target discrimination function.

4.3 Major-minor task

Results for the five untrained listeners are shown in Fig. 4.

Please insert Fig. 4 around here

Results for the sixth listener, a highly trained concert pianist with perfect pitch, are shown in Fig. 5.

Please insert Fig. 5 around here

In this task, consistent with Fig. 2, Listener 1 achieved 82% correct; Listener 2 achieved 77% correct; Listener 3 achieved 68% correct; Listener 4 achieved 73% correct; Listener 5 achieved 70% correct; and Listener 6 achieved 69% correct.

For all except Listener 2, the last pip exerts markedly higher impact than the other pips. For Listeners 1, 3, 4 and 5, the relative impacts of different notes occurring on the last pip (given by F_{last}) are strikingly different than the impacts exerted by those same notes in non-terminal pips (given by F_{tonic} and F_{other}). By contrast, for Listeners 2 and 6 F_{last} does not seem to diverge strikingly in form similar in form from F_{tonic} and F_{other}. For Listeners 1, 2, 3, 4, and 5, the relative impact exerted by the major 6^{th} (E) when it occurred on the last pip was lower than it was on non-terminal pips. Listeners 1, 2, 4 and 5 showed a similar effect for the major 3^{rd} (B). Also, for each of Listeners 1 - 5, high tonics (G_6's) exerted impact that was much more strongly major when they occurred on the last pip than they did when they occurred on non-terminal pips. Interestingly, except in the case of participant 1, this effect did not occur for the low tonics (G_5's).

The results for Listener 6 differ from those for the other listeners in interesting ways. The three pip-note impact functions F_{last}, F_{tonic} and F_{other} are similar in form for this listener. What is especially striking is that for this listener, $G\sharp$'s exert a very strong minor influence whenever they occur. Indeed, they are much more effective in promoting a "minor" response than are minor sixths ($D\sharp$'s) and slightly but significantly more effective than minor thirds ($A\sharp$'s). No other Listener showed such an effect. Finally, we note that although Listener 6 was a highly trained musician with perfect pitch, she did not perform especially well at this task. Her non-terminal pip-note impact functions do not correlate especially strongly with the target function, nor are the average values of her pip sensitivity functions especially high.

4.4 Squeezed/Spread Task

The fits of the resolution-interaction model for participants 1 and 2 in the squeezed-spread task are shown in Fig. 6.

Please insert Fig. 6 around here

As suggested by Fig. 2, Listeners 1 and 2 were both poor at discriminating the two sequence types, achieving 60% and 52% performance respectively. The slightly above-chance performance achieved by listener 1 in this task may be partly attributable to the fact that this listener spent time between blocks of experimental trials listening to enriched exemplars of the two sorts of scrambles so as to refresh his sense of the difference he was striving to discriminate.

5 Discussion

A cursory glance at the model fits for different listeners reveals dramatic individual differences. These differences may be due to genuine functional differences in the auditory systems of our listeners. Alternatively, these differences may reflect different strategies in processing the stimuli. Although our intent in using feedback-driven tasks was to enable our listeners to optimize their performance, it is impossible to know how well any given listener has achieved this goal. Especially in light of the large variations in skill demonstrated by different listeners, it seems likely that some of our listeners might well discover different and more effective strategies with further practice, strategies that might produce substantial changes in the fits provided by our models. Our interpretation of the data must be tempered by these considerations.

Our analysis has revealed that the decision statistics achieved by our listeners are complicated in several important respects:

1. The last pip of a tone scramble often exerts much greater influence on these decision statistics than do the non-terminal pips;

2. In many cases, the pip-note impact function F_{last} reflecting the influences exerted by notes assigned to the last pip of a tone scramble is strikingly different in form from the impact functions F_{tonic} and F_{other} characterizing the influences exerted notes assigned to non-terminal pips;

3. In some cases, the note assigned to the last pip interacts with notes assigned to non-terminal pips of the scramble in determining the value of the decision statistic. Specifically,
   (a) for three listeners in the major-minor task, sensitivity to information from non-terminal pips was significantly higher if the tone scramble ended with a tonic than if it ended with some other note.
   (b) for one listener in the structured-random task, the relative sensitivity to different notes occurring at non-terminal pips was significantly different if the scramble ended on a tonic than if it ended in a non-tonic.

We should not let these complexities overshadow the central finding of our experiments, however: Our listeners were able to extract decision statistics that were

1. tuned to specific scramble histogram properties, and
2. sensitively influenced by information from all pips of the tone scramble.

Neither of the models we fit (last-pip-special or resolution-interaction) provides much insight into the strengths with which the notes assigned to different pips of the scramble may interact to control the listener's responses. The last-pip-special model includes no interaction terms at all; the influence exerted by each pip depends only on its location and the note that has been assigned to it. The resolution-interaction model enables us to examine a specific and limited form of interaction: whether notes assigned to non-terminal pips exert different patterns and/or strengths of influence depending on whether or not the last pip of the scramble is a tonic.

It is possible that notes assigned to different pips interact in determining our listeners' responses in ways we have not thought to investigate. However, the fact that for most listeners in most tasks, (see, e.g., Listener 1 in both the structured-random and major-minor tasks), the pip-location sensitivity functions W in the last-pip-special model, and W_{tonic} and W_{other} in the resolution-interaction model, are flat and elevated well above zero across all pips $t = 1, 2, \cdots, N_{pips} - 1$ shows that irrespective of what higher order interactions may be operating to determine our listeners' responses, our models are capturing strong and systematic first-order effects.

It should also be noted that in the last-pip-special model, the pattern of influence exerted by different notes $n \in \{G_5, A\flat, \cdots, G_6\}$ is constant and given by $F(n)$ across all non-terminal pips $t = 1, 2, \cdots, N_{pips} - 1$. (The note-dependent pattern of influence is allowed to differ only on the last pip.) In the resolution-interaction model, the pattern of influence exerted by different notes n is also constant across all non-terminal pips; however, in this model, the pattern of influence exerted by different notes n is allowed to take one form ($F_{tonic}(n)$) if the last pip is assigned a tonic vs. another form ($F_{other}(n)$) if the last pip is assigned some note other than a tonic. As our results show, for most listeners in most conditions, the fit of this model supports the idea that $F_{tonic} = F_{other}$.

We thus conclude that irrespective of what additional, undisclosed interactions may be operating to affect our listeners' judgments, our findings implicate decision processes in which the notes assigned to non-terminal pips in a scramble operate individually, forcefully and each with the same note-specific pattern of influence to control responding.

We submit that these decision processes point to the existence of tonality mechanisms that can operate rapidly to capture information pip-by-pip from the unfolding tone scramble. We assume that a given listener has some number $N_{tonality}$ of tonality mechanisms, the k^{th} of which is characterized by a sensitivity function $f_k : \{G_5, A\flat, \cdots, G_6\} \rightarrow \Re$ that gives the activation produced in that mechanism by the different notes used in our scrambles. We further assume that the only pip-note impact functions our listeners can achieve with high sensitivity are linear combinations of the sensitivity functions $f_1, f_2, \cdots, f_{N_{tonality}}$, which we can assume without loss of generality are linearly independent. Under this assumption, the dimensionality of the space of pip-note impact functions our listeners can achieve with high sensitivity is equal to $N_{tonality}$. Because any pip-note impact function is constrained to sum to 0, the space of all pip-note impact functions has dimensionality 12.

Thus, our basic assumptions imply that $N_{tonality} \geq 1$ and $N_{tonality} \leq 12$. Our experimental results enable stronger conclusions. Specifically, they provide evidence that in fact

1. $N_{tonality}$ is strictly greater than 1, and also that
2. $N_{tonality}$ is strictly less than 12.

The primary evidence that $N_{tonality} > 1$ comes from considering the performance of listeners 1 and 2 in the structured-random and major-minor tasks. These two listeners achieved decision statistics in these different tasks that were tuned to distinctly different histogram properties. This can be seen by visually comparing the plots in the left-hand panels of Fig. 3 with the corresponding plots in Fig. 4. The non-dashed lines in the upper two, left-hand panels of Fig. 3 tend to be high for the tonics (G_5 and G_6) and the dominant (D) and low for all other notes. By contrast, the non-dashed lines in upper two, left-hand panels of Fig. 4 tend to be high for the major sixth and major third (E and B), low for the minor sixth and minor third ($D\sharp$ and $A\sharp$) and near 0 for all other notes. These observations suggest that for listeners 1 and 2, the non-terminal pip-note sensitivity functions for the structured-random task should correlate weakly with the non-terminal pip-note sensitivity functions for the major-minor task. Table 5 confirms this expectation.

Please insert Table 5 around here

For each of the three listeners who performed both the structured-random and major-minor tasks, this table gives the correlation of the non-terminal pip-note impact functions (F in Eq. 6) derived from the fits of the last-pip-special model to the data for both tasks. These two pip-note impact functions have low correlations for Listeners 1 and 2 (but not for Listener 3). We conclude that Listeners 1 and 2 can achieve at least two strongly distinct, non-terminal pip-note impact functions, implying that (for these two listeners) $N_{tonality} > 1$.

The strongest evidence that $N_{tonality} < 12$ comes from the fact that listeners 1 and 2, the most sensitive listeners in the structured-random and major-minor tasks, performed much less well in the squeezed-spread task. None of the pip-note impact functions (F_{tonic}, F_{other} or F_{last}) achieved by either listener was well-correlated with the squeezed-spread target discrimination function; moreover, the pip-location sensitivity functions (W_{tonic}, W_{other} or W_{last}) of listeners 1 and 2 were substantially lower than the pip-location sensitivity functions they achieved in other conditions. This suggests that a large component of the squeezed-spread target discrimination function is orthogonal to the space spanned by $f_1, f_2, \cdots, f_{N_{tonality}}$, implying that $N_{tonality} < 12$. We obtain additional evidence that $N_{tonality} < 12$ by considering the pip-note impact functions achieved by other listeners across the structured-random and major-minor tasks. The fact that we used trial-by-trial feedback makes it plausible to suppose that the pip-note sensitivity function achieved by a given listener in a given task was the linear combination of $f_1, f_2, \cdots, f_{N_{tonality}}$ whose correlation with the target discrimination function was maximal. If this is true, then any listener for whom $N_{tonality} = 12$ would always achieve a pip-note impact function that precisely matched the target discrimination function in any given task. On the contrary, however, the pip-note impact functions achieved by our listeners typically deviated (often quite strongly) from the target discrimination functions they were striving to achieve, again suggesting that $N_{tonality} < 12$.

5.1 The relation of the current results to the major and minor key profiles

Shepard (1982) and also Krumhansl & Kessler (1982) presented closely related geometric models that revolutionized our understanding of tonality perception. The model described by Krumhansl & Kessler (1982) is founded on two functions derived using the probe-tone technique, a "major key profile" reflecting the goodness with which different notes of the chromatic scale are perceived to fit into music in a major key and another "minor key profile" reflecting the goodness with which different notes are perceived to fit into music in a minor key. The major and minor key profiles are plotted (with the mean ratings for the tonic duplicated for G_5 and G_6), with the major key profile on top and the minor key profile on the bottom. The left panels plot the profiles as in Krumhansl & Kessler (1982), with notes arranged by pitch height; the right panels plot them with notes arranged in the circle of fifths (in line with the other figures in this paper).

Please insert Fig. 8 around here

How might the major and minor key profiles relate to the major-minor task? Suppose the listener has distinct major-sensitive and minor-sensitive tonality mechanisms, the major-sensitive mechanism tending to be highly activated by major musical contexts, and the minor-sensitive mechanism by minor musical contexts. Then a reasonable strategy for the listener to use in the major-minor task is to take the difference between the average activation produced by a given tone-scramble in the major-sensitive mechanism minus the average activation produced in the minor-sensitive mechanism and respond "major" if this difference is positive and "minor" if negative. Suppose in addition that the sensitivity functions characterizing these hypothetical major-sensitive and minor-sensitive mechanisms match the major and minor key profiles in form. Then we might expect the non-terminal impact functions achieved by our listeners in the major-minor task to resemble the function ProfileDiff defined in Eq. 9.

$$\text{ProfileDiff} = \text{MajorKeyProfile} - \text{MinorKeyProfile} \tag{9}$$

On the other hand the structured-random task requires the listener to assess whether a given scramble is closer to a G-centered diatonic scale (indeterminate whether major or minor) *versus* acentric (with all tones equally likely). A reasonable strategy for this task would be add together the average activations produced in the major-sensitive and minor-sensitive tonality mechanisms by a given scramble (thereby discarding information about majorness *versus* minorness but retaining information about structuredness) and compare this result to some criterion, responding "structured" if the summed activations exceeds the criterion and "random" otherwise. This suggests that the non-terminal impact functions achieved by our listeners in the structured-random task should resemble the function ProfileSum defined in Eq. 10.

$$\text{ProfileSum} = \text{MajorKeyProfile} + \text{MinorKeyProfile} \tag{10}$$

ProfileDiff and ProfileSum are plotted (dashed lines) in the upper and lower panels of Fig. 9 with means subtracted and normalized so as to make them comparable to the other impact functions plotted here. Also plotted in the upper panel of Fig. 9 is the mean non-terminal impact function (from the last-pip-special model) achieved by Listeners 1, 2, 3, 4 and 5 in the major-minor task. [4] In the lower panel, the mean non-terminal impact function achieved by Listeners 1, 2 and 3 in the structured-random task (solid line) is plotted with ProfileSum (dashed line).

Please insert Fig. 9 around here

Given the substantial differences in the individual impact functions achieved by different listeners in the major-minor task, the mean non-terminal, pip-note impact function from the major-minor task matches ProfileDiff tolerably well ($r = 0.82$). With the exception of an obvious difference at the low tonic (G_5), the mean impact function for the structured-random task also matches the function ProfileSum well ($r = 0.86$). Recall that although our listeners gave positive weight to the fifth (D) in the impact functions they achieved in the structured-random task, they were unable to give positive weight to the fourth (C) despite the qualitative similarity between these two degrees of the scale and despite the similar emphasis given these two notes in the target discrimination function used to give feedback in this task. Interestingly, this difference in sensitivity to the fifth *versus* the fourth is also seen in ProfileSum, which matches the mean impact function for the structured-random task very closely at C and D, giving higher weight to D.

Thus despite the great difference between the experimental paradigms used in the current paper versus the probe-tone method used by Krumhansl and Kessler (1982), we find reasonable convergence in the results obtained.

5.2 Some open questions

The experiments reported here are the first forays into a new experimental domain. It is therefore not surprising that they have raised more questions than they have answered. In hopes that others will take up the challenge of carrying this work forward, we close by listing some of the mysteries that remain unsolved.

1. We are used to thinking of listeners with perfect pitch as possessing generally heightened auditory powers compared to ordinary listeners. However, although Listener 6 is a highly trained concert pianist with perfect pitch, she performed relatively poorly in the major-minor task. This result raises a number of questions:
   (a) For Listener 6, $G\sharp$'s (the semitone above the tonic) exerted a strong "minor" influence on her judgments. What is the source of this strange effect? Will other trained musicians with perfect pitch show similar patterns?
   (b) The current finding raises the possibility that listeners with absolute pitch may actually be impaired in some auditory tasks in comparison to listeners without perfect pitch. Is this true? If so, on what tasks are they impaired? And what is the reason for this impairment?
2. The fourth and fifth degrees of the major and minor diatonic scales form qualitatively similar intervals with the tonic. It was for this reason that we gave both the fourth and fifth elevated weights in the structured-random target discrimination function. Although all three

[4] Listener 6's curve was not included in the average. None of the subjects in the experiment of Krumhansl & Kessler (1982) had perfect pitch whereas this listener did. Moreover, her impact function was quite different from those achieved by the other listeners raising the possibility that listeners with perfect pitch may yield systematically different results in the major-minor task than listeners without perfect pitch.

of the listeners who performed the structured-random task achieved decision statistics with heightened sensitivity to the fifth, none of these listeners was able to give positive weight to the fourth. Why not?

3. For Listeners 1, 2, 3, 4 and 5 (but not Listener 6) in the major-minor task, a high tonic (G_6) that occurred on the last pip of the stimulus scramble exerted a strong "major" influence on the listener's judgment. Why is this? No similar effect is seen for other high notes (e.g., F's or $F\sharp$'s) occurring on the last pip, suggesting that this effect is specific for the tonic rather than a general last-pip effect of pitch height. If pitch height is not the crucial factor, however, then one might expect the low tonic to exert a similar effect. However, only one listener (Listener 1) showed the same effect for the low tonic.

4. Four out of six listeners who participated in the major-minor task had F_{last}'s that differed dramatically from F_{tonic} and F_{other}. Although we have speculated that these differences may be features of the optimal strategy available to human listeners, we have no theory to offer of what this strategy might be and why it might impose these apparently counterproductive constraints.

5.3 Final remarks

We have introduced a feedback-driven, experimental paradigm requiring listeners to strive to classify rapid, randomly scrambled sequences of pure tones in accordance with various different feedback rules. Here we tested listeners using three different feedback rules (the "structured-random," "major-minor" and "squeezed-spread" rules), but many other feedback rules remain to be investigated. If scramble perception is analogous to color perception, then (ignoring the complexities introduced by the last-pip-special and the resolution-interaction models)

1. the number of computationally distinct decision statistics listeners can achieve across classification tasks using different feedback rules is equal to the number of distinct tonality mechanisms resident in the auditory system of the listener.

2. the space spanned by the non-terminal impact functions achieved in different conditions will be equal to the space spanned by the sensitivity functions characterizing the different tonality mechanisms.

A great deal of work lies ahead to complete the task of surveying this space.

Acknowledgements

This work was supported by NSF BCS-0843897.

REFERENCES

Blechner MJ (1977) Musical skill and the categorical perception of harmonic mode. *Haskins Laboratories Status Report on Speech Perception* (SR-51/52), 139-174.

Crowder RG (1984) Perception of the Major/Minor Distinction: I. Historical and Theoretical Foundations. *Psychomusicology*, *4*:1/2, 3-12.

Crowder RG (1985a) Perception of the Major/Minor Distinction: II. Experimental Investigations. *Psychomusicology*, *5*:1/2, 3-24.

Crowder RG (1985b) Perception of the Major/Minor Distinction: III. Hedonic, musical, and affective discriminations. *Bulletin of the Psychonomic Society*, *23*:4, 314-316.

Dalla Bella S, Peretz I, Rousseau L, Gosselin N (2001) A developmental study of the affective value of tempo and mode in music. *Cognition*, *80*, B1-B10.

Gagnon L, Peretz I (2003) Mode and tempo relative contributions to "happy-sad" judgements in equitone melodies. *Cognition and Emotion*, *17*:1, 25-40.

Gerardi GM, Gerken L (1995) The Development of Affective Responses to Modality and Melodic Contour. *Music Perception*, *12*:3, 279-290.

Halpern AR (1984) Perception of structure in novel music. *Memory and Cognition 12*:163-170.

Halpern AR, Bartlett JC, Dowling WJ (1998) Perception of mode, rhythm, and contour in unfamiliar melodies: Effects of age and experience. *Music Perception15*:335-356.

Halpern, A., Martin, J., and Reed, T. (2008). An ERP study of major-minor classification in melodies. *Music Perception* 25(3), 181-191.

Heinlein CP (1928) The Affective Character of the Major And Minor Modes In Music. *Comparative Psychology*, *VIII*:2, 101-142.

Hevner K (1935) The affective character of the major and minor mode in music. *American Journal of Psychology 47*:103-118.

Kastner MP, Crowder RG (1990). Perception of the major/minor distinction: IV. Emotional connotations in young children. *Music Perception 8*(2):189-202.

Krumhansl CL, & Kessler EJ (1982). Tracing the dynamic changes in perceived tonal organization in a spatial map of musical keys. *Psychological Review*, 89, 334368.

Krumhansl CL, & Shepard RN (1979). Quantification of the hierarchy of tonal functions within a diatonic context. *Journal of Experimental Psychology: Human Perception and Performance*, 5, 579-594.

Leaver AM, Halpern AR (2004) Effects of training and melodic features on mode perception. *Music Perception 22*:117-143.

Howard D, Major/Minor Triad Identification and Discrimination by Musically Trained and Untrained Listeners , *Music Perception*, 10:2 (1992:Winter) p.205

Leaver AM, Halpern AR (2004) Effects of training and melodic features on mode perception. *Music Perception 22*:117-143.

Maxwell, J. C. (1855). Experiments on color as perceived by the eye, with remarks on color blindness. Transaction of the Royal Society of Edinburgh. v. XXI., part II, 275 - 298.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). Introduction to the theory of Statistics. McGraw-Hill.

Oram N & Cuddy LL (1995). Responsiveness of western adults to pitch-distributional information in melodic sequences. *Psychological Research*, 57, 103118.

Shepard, R. N. Geometrical approximations to the structure of pitch. *Psychological Review*, 1982, 89, 305-333.

Smith NA & Schmuckler MA (2004). The perception of tonal structure through the differentiation and organization of pitches. *Journal of Experimental Psychology: Human Perception and Performance* 30, 268286.

Tymoczko, D. (2011) *A Geometry of Music*. Oxford University Press.

Whissell R, Whissell C (2000). The emotional importance of key: Do Beatles songs written in different keys convey different emotional tones? *Perceptual and Motor Skills 91*:973980.

Table 1: Base Note-Count Vectors and Target Discrimination Functions in Different Conditions

Structured-Random

	G	G♯	A	A♯	B	C	C♯	D	D♯	E	F	F♯	G
B_1 (Structured)	4	1	1	1	1	4	1	5	1	1	1	1	4
B_2 (Random)	2	2	2	2	2	2	2	2	2	2	2	2	2
D (Target Function)	2	-1	-1	-1	-1	2	-1	3	-1	-1	-1	-1	2

Major-Minor

	G_5	G♯	A	A♯	B	C	C♯	D	D♯	E	F	F♯	G_6
B_1 (Major)	5	1	1	1	3	1	1	5	1	3	1	1	2
B_2 (Minor)	5	1	1	3	1	1	1	5	3	1	1	1	2
D (Target Function)	0	0	0	-2	2	0	0	0	-2	2	0	0	0

Squeezed-Spread

	G_5	G♯	A	A♯	B	C	C♯	D	D♯	E	F	F♯	G_6
B_1 (Squeezed)	5	1	1	1	3	1	1	5	3	1	1	1	2
B_2 (Spread)	5	1	1	3	1	1	1	5	1	3	1	1	2
D (Target Function)	0	0	0	-2	2	0	0	0	2	-2	0	0	0

Table 2: Histogram Perturbations

G_5	G♯	A	A♯	B	C	C♯	D	D♯	E	F	F♯	G_6
1	-1	1	-1	1	1	1	-1	-1	-1	1	-1	0
1	-1	-1	1	-1	1	1	1	-1	-1	-1	1	0
1	1	-1	-1	1	-1	1	1	1	-1	-1	-1	0
1	-1	1	-1	-1	1	-1	1	1	1	-1	-1	0
1	-1	-1	1	-1	-1	1	-1	1	1	1	-1	0
1	-1	-1	-1	1	-1	-1	1	-1	1	1	1	0
1	1	-1	-1	-1	1	-1	-1	1	-1	1	1	0
1	1	1	-1	-1	-1	1	-1	-1	1	-1	1	0
1	1	1	1	-1	-1	-1	1	-1	-1	1	-1	0
1	-1	1	1	1	-1	-1	-1	1	-1	-1	1	0
1	1	-1	1	1	1	-1	-1	-1	1	-1	-1	0

Tests of null hypothesis that $F_{tonic} = F_{other}$

Task	Listener	$\chi^2_{(11)}$	p
structured-random	1	4.963	0.933
structured-random	2	19.616	0.051
structured-random	3	28.913	0.002*
major-minor	1	21.874	0.025
major-minor	2	13.976	0.234
major-minor	3	22.597	0.020
major-minor	4	9.154	0.608
major-minor	5	14.740	0.195
major-minor	6	13.047	0.290
squeezed-spread	1	4.711	0.944
squeezed-spread	2	14.094	0.228

Table 3: *Results of likelihood ratio tests of the null hypothesis that* $F_{other} = F_{tonic}$ *in the resolution-interaction model.*

Tests of null hypothesis that $\text{Avg}(W_{tonic}) = \text{Avg}(W_{other})$

Task	Listener	$\chi^2_{(1)}$	p
structured-random	1	1.522	0.217
structured-random	2	0.940	0.332
structured-random	3	0.820	0.365
major-minor	1	10.135	0.001*
major-minor	2	4.624	0.032
major-minor	3	8.074	0.004*
major-minor	4	4.679	0.031
major-minor	5	4.573	0.032
major-minor	6	10.945	0.001*
squeezed-spread	1	0.002	0.964
squeezed-spread	2	0.002	0.966

Table 4: *Results of likelihood ratio tests of the null hypothesis that* $Avg(W_{tonic}) = Avg(W_{other})$ *in the resolution-interaction model.*

Correlations between the pip-note impact functions for the structured-random and major-minor tasks.

Listener	Correlation	Lower CI	Upper CI
1	0.2217	0.1283	0.3119
2	0.2786	0.1532	0.4148
3	0.7056	0.6435	0.7613

Table 5: The correlations are derived from the fits of the last-pip-special model of Eq. 6. Specifically, the correlations in the table are the inner products of the non-terminal pip-note impact functions (F in Eq. 6) derived from the fits to the data for the major-minor and structured-random tasks. Note that these two pip-note impact functions have low correlations for Listeners 1 and 2 but not for Listener 3.

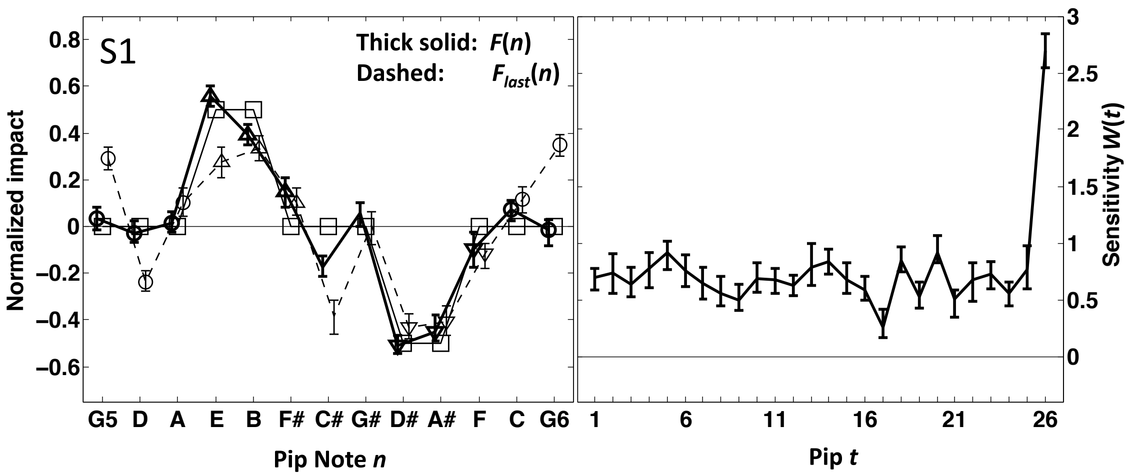

Figure 1: *The fit of the "last-pip-special" model to the results of Listener 1 in the major-minor task.* Listener 1 was the best performer of this task out of 6 listeners. Left panel gives the estimated pip-note impact functions F and F_{last} (of Eq. 6). Notes of the scale follow the circle of fifths along the horizontal axis because pip-note impact functions tend to take simpler, smoother forms when plotted on the circle of fifths than when plotted as a function of notes ordered by pitch height. The thin line marked by square tokens gives the major-minor target function. The thick black line gives the pip-note impact function F for all pips other than the last one, and the dashed line gives the pip-note impact function F_{last} for the last pip. All three functions are constrained to sum to zero, and all three are normalized. Each of F and F_{last} has several different markers: circles mark notes contained in both the major and minor diatonic scales with tonic G; upward-pointing triangles mark points contained in only the major scale; downward-pointing triangles mark points contained in only the minor scale; and unmarked points are contained in neither the major nor minor scale. The figure on the right shows the estimated pip-location sensitivities $W(t)$ for pips $t = 1, 2, \cdots, N_{pips}$, where we have set $W(N_{pips}) = W_{last}$. Error bars are 95% Bayesian credible intervals.

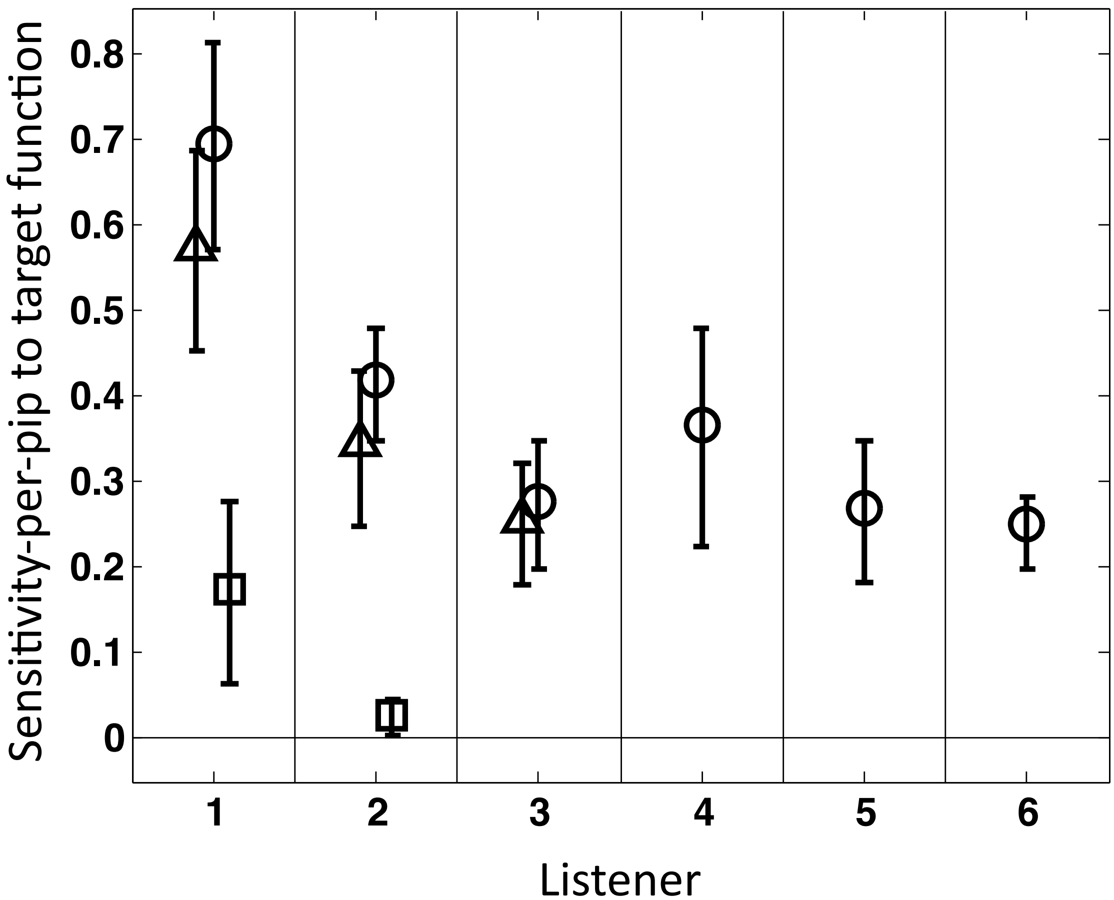

Figure 2: *Sensitivity-per-pip of listeners to the target functions in different tasks.* For a task with target function D, sensitivity-per-pip (Eq. 8) is the average value, across all N_{pips} pips t, of the correlation of D with the listener's pip-note impact function for pip t (which is F_{last} if $t = N_{pips}$ and F for all other pips) scaled by $W(t)$, the participant's sensitivity to information from pip $\triangle$'s give the results for the structured-random task; $\bigcirc$'s give results for the major-minor task; $\square$'s give the results for the squeezed-spread task. Error bars are 95% Bayesian credible intervals. Note that for listeners 1 and 2, Sensitivity-per-pip is much higher for the structured-random and major-minor tasks than it is for the squeezed-spread task.

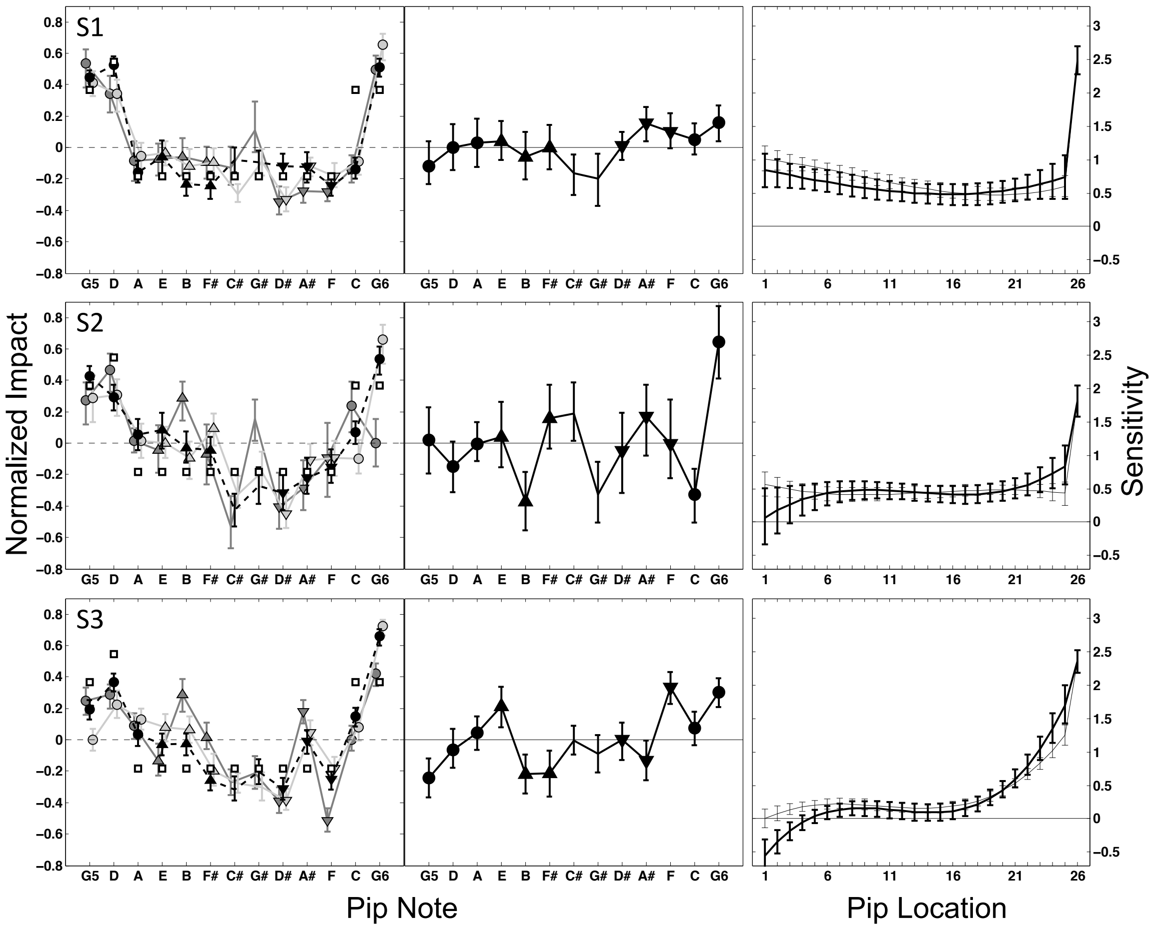

Figure 3: *Fits of the "resolution-interaction" model to the data of Listeners 1, 2 and 3 in the structured-random task.* The panels on the left show the estimated pip-note impact functions F_{tonic} (dark gray markers), F_{other} (light gray markers) and F_{last} (dashed line, black markers) of Eq. 7. All three functions are constrained to sum to zero, and all three are normalized. The small, open squares mark the target function for the structured-random task. Each of the three pip-note impact functions has several different markers: circles mark notes contained in both the major and minor diatonic scales with tonic G; upward-pointing triangles mark points contained in only the major scale; downward-pointing triangles mark points contained in only the minor scale; and unmarked points are contained in neither the major nor minor scale. Panels in the center plot $F_{other} - F_{tonic}$, and panels on the right plot the estimated pip-location sensitivity functions W_{tonic} (thick line) and W_{other} (thin line) for $t = 1, 2, \cdots, N_{pips}$, where $W_{other}(N_{pips}) = W_{tonic}(N_{pips}) = W_{last}$. Error bars are 95% Bayesian credible intervals.

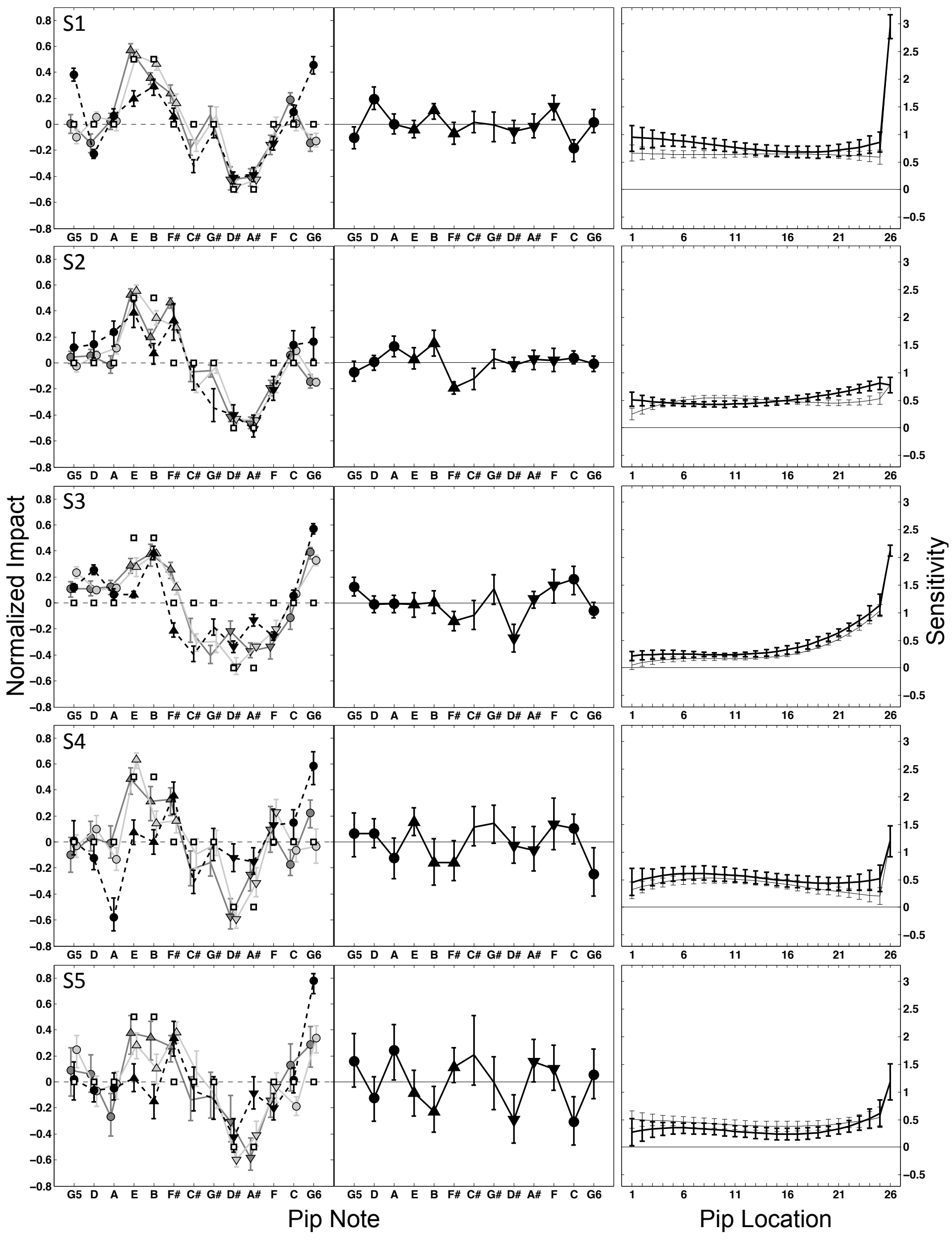

Figure 4: *Fits of the "resolution-interaction" model to the data of Listeners 1, 2, 3, 4 and 5 in the major-minor task.* Plotting conventions are the same as those used in Fig. 3.

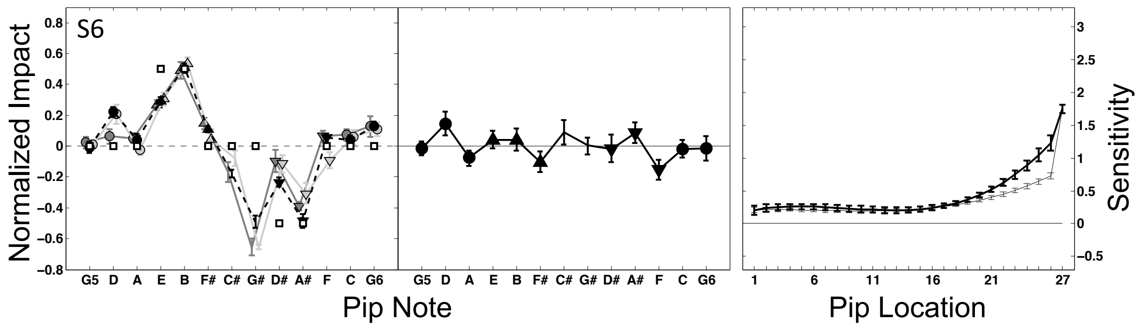

Figure 5: *Fits of the "resolution-interaction" model to the data of Listeners 1, 2, 3, 4 and 5 in the major-minor task.* Plotting conventions are the same as those used in Fig. 3.

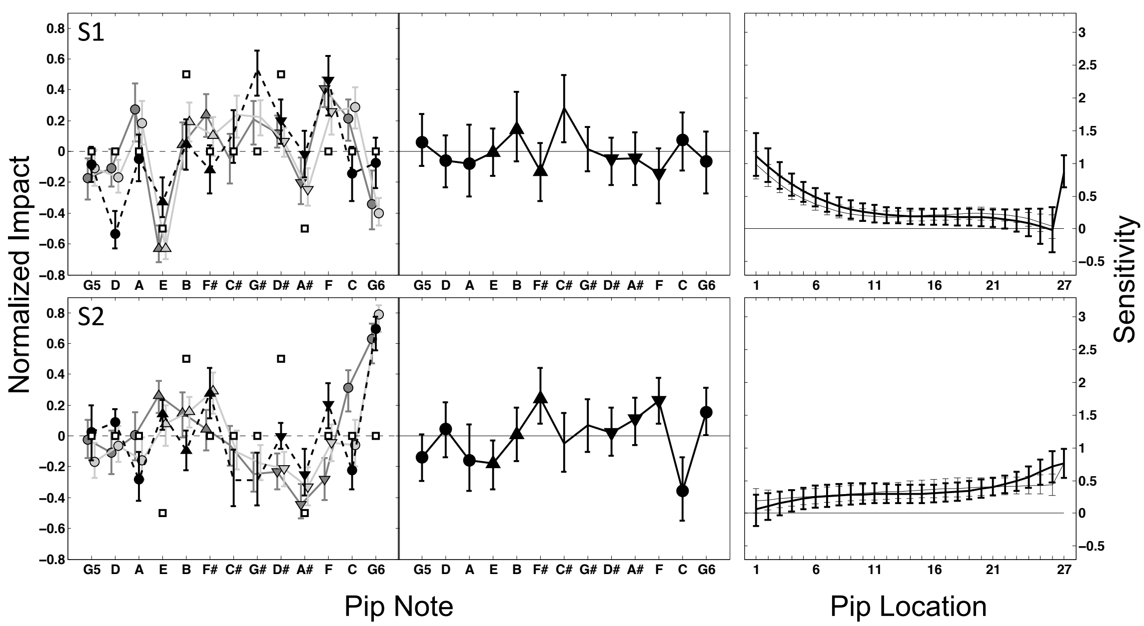

Figure 6: *Fits of the "resolution-interaction" model to the data of Listeners 1 and 2 in the squeezed-spread task.* Plotting conventions are the same as those used in Fig. 3.

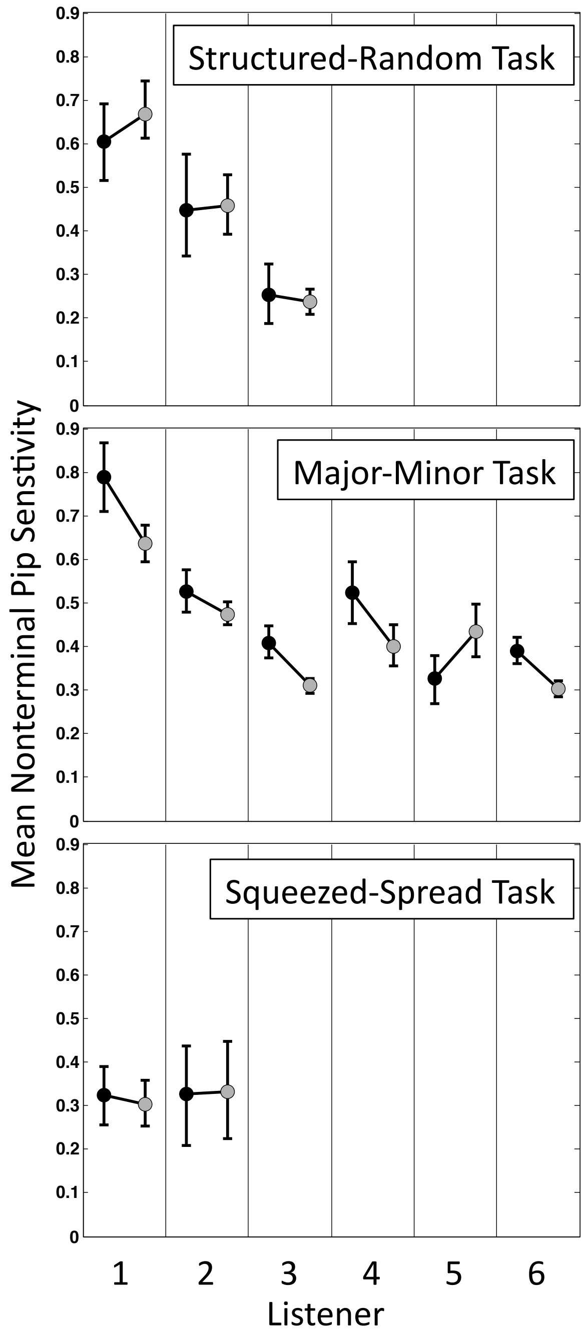

Figure 7: *The means across all pips t of $W_{tonic}(t)$ (black markers) and $W_{other}(t)$ (gray markers) for all listeners is all tasks.* Note that in the major-minor task, for five out of six listeners, the mean of $W_{tonic}(t)$ is greater than the mean of $W_{other}(t)$. This difference is significant for Listeners 1, 3 and 6.

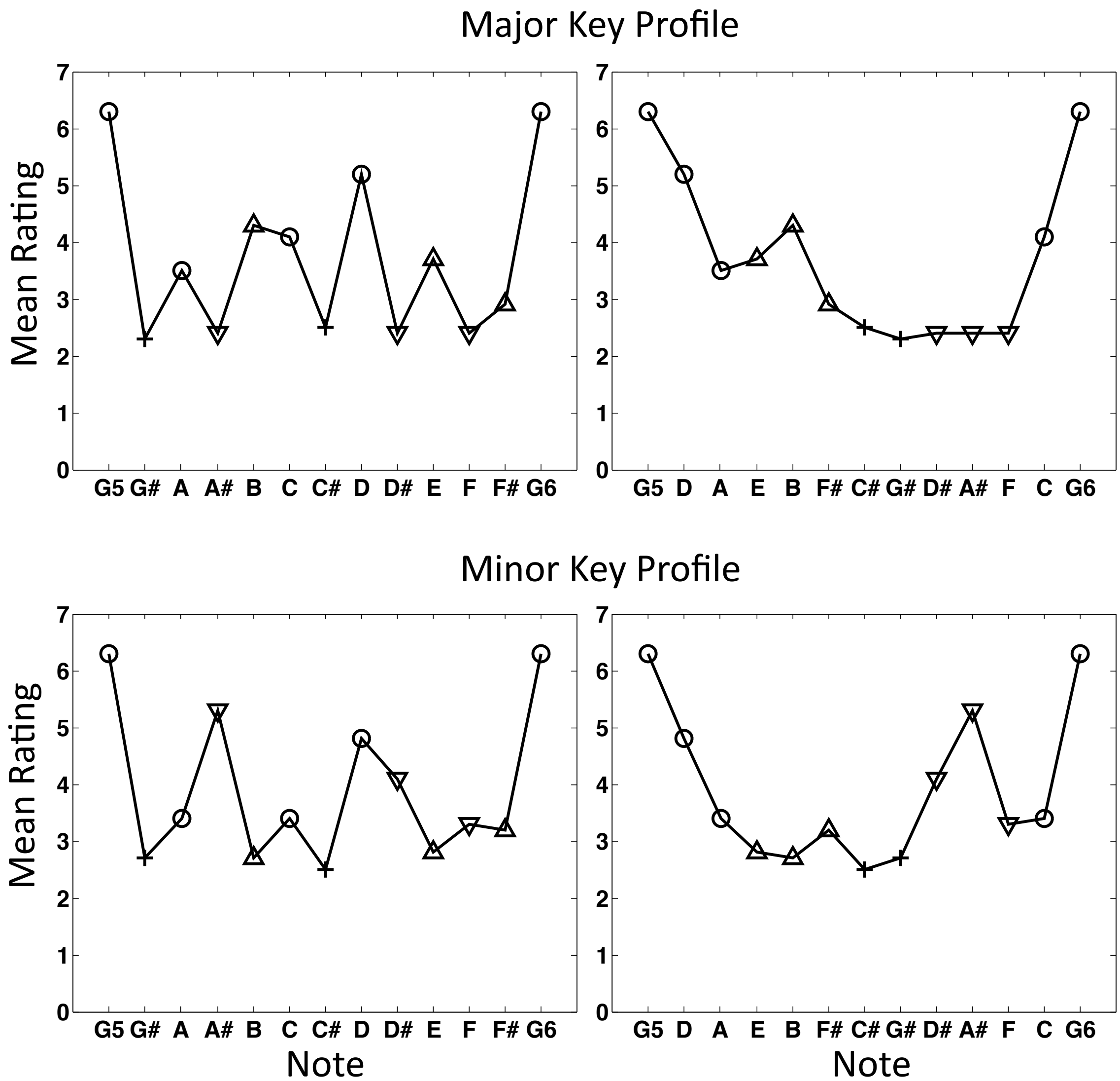

Figure 8: *The major key and minor key profiles (Krumhansl & Kessler, 1982)* The "major key profile" (top two panels) reflects the goodness with which different notes of the chromatic scale are perceived to fit into music in a major key (plotted here with tonic G to facilitate comparison with other plots) and the "minor key profile" (bottom two panels) reflects the goodness with which different notes are perceived to fit into music in a minor key. The left panels plot the profiles as in Krumhansl & Kessler (1982), with notes arranged by pitch height; the right panels plot them with notes arranged in the circle of fifths (in line with the other figures in this paper).

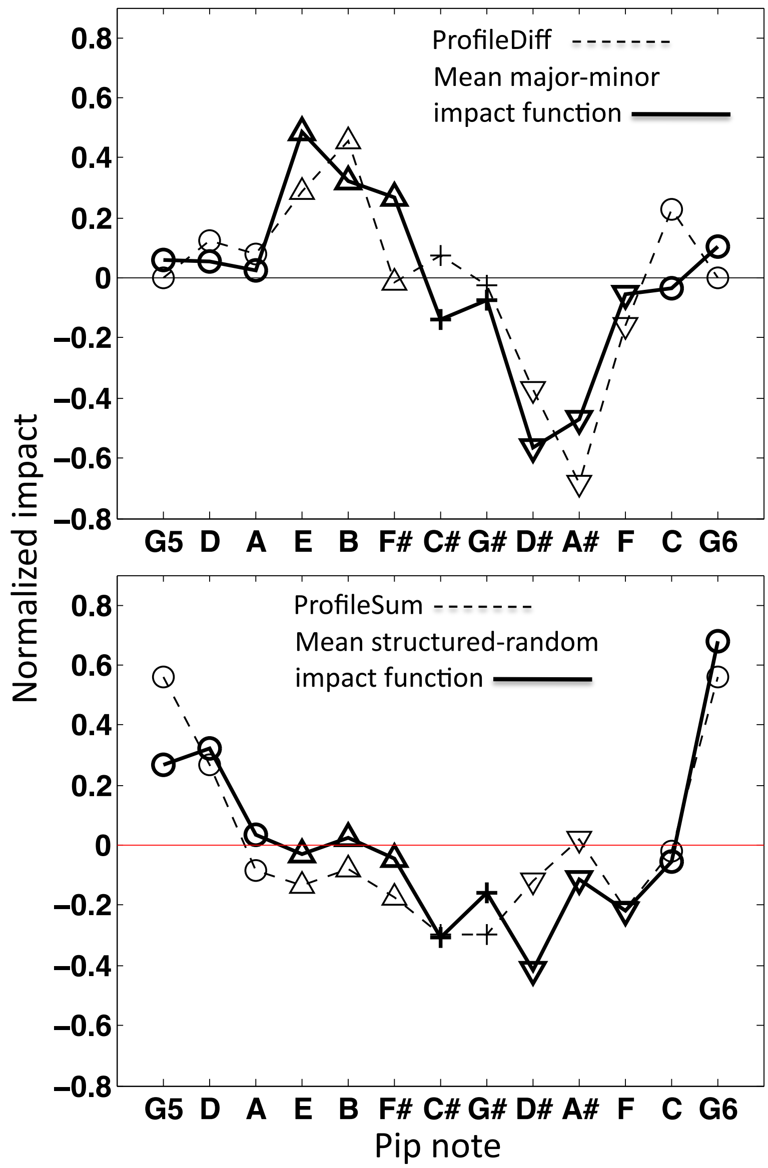

Figure 9: *The functions ProfileSum (Eq. 10) and ProfileDiff (Eq. 9) plotted with the mean non-terminal, pip-note impact functions from the major-minor and structured-random tasks.* The top panel shows the function ProfileDiff (dashed line) and the mean non-terminal, pip-note impact function from the major-minor task (solid line), where the average is taken across Listeners 1, 2, 3, 4 and 5. The bottom panel shows ProfileSum (dashed line) and the mean non-terminal, pip-note impact function from the sturctured-random task (solid line). All functions have had their means set to 0 and have been normalized to facilitate comparison with the impact functions in other plots.