The Emergence of Statistical Music Universals through Iterated Pitch Learning: A Pilot Study

Jack Goffinet
jackgoffinet@gmail.com

Abstract

Music exhibits statistical universals, cross-cultural similarities that exist well above chance levels, yet little is known about why such regularities exist. Recently, researchers have addressed analogous issues in language using the natural repeated memorization and transmission of cultural information between individuals (*iterated learning*) to model the process of cultural transmission. Errors accumulate as artificial languages are repeatedly memorized and transmitted, resulting in highly structured cultural material that more closely resembles language. These results raise the possibility that many statistical universals in music are the result of learning and memory biases compounded by the process of cultural transmission. Here, iterated learning methodology is employed to model the cultural transmission of isochronous pitch sequences. Musically experienced participants are asked to transcribe aurally presented pitch sequences from memory, their transcription then becoming the sequence the next participant is asked to transcribe. Participants' cognitive biases shape initially random pitch sequences into dense, structured, and learnable structures that exhibit regularities resembling those found in music around the world. These results are consistent with the hypothesis that universal cognitive biases underlie many statistical music universals.

1 Introduction

Language has been studied fruitfully from an evolutionary standpoint for over two decades (Steels 1997). From such a perspective, words, phonemes, and grammatical structures are viewed as competing and cooperating with one another to be reproduced. Large-scale language change, in which language becomes better suited for production, processing, and memorization, results (Christiansen et al. 2016). The study of music has begun to follow suit, but has largely focused on the local transmission of musical memes (Jan 2017; Albrecht and Huron 2012; Mauch et al. 2015; Broze and Shanahan 2013; Huron 1994) rather than the global adaptation of musical systems to cognitive constraints (Lumaca, Ravignani, et al. 2018; Ravignani, Delgado, et al. 2016). The present paper investigates this global structuring of musical systems, borrowing a tool from the study of language evolution: *iterated learning*.

Iterated learning is a simple model of cultural transmission which is used to study cultural evolution (see Kirby, T. Griffiths, et al. 2014, for a review). The basic experimental structure involves a single *transmission chain* made up of several *transmission generations*,

each with a single *agent*, either a simulated or human participant. The first generation agent memorizes and transmits a randomly generated collection of cultural material. Each subsequent agent in the chain then attempts to memorize and transmit the previous agent's material, typically introducing errors. Figure 1 depicts the first generations of a transmission chain presented here. Agents' common biases are compounded generation after generation so that individually small selective pressures are amplified (Brighton 2002; Kirby, Dowman, et al. 2007). Laboratory experiments consistently find that iterated learning increases the structure and aids the learnability of complex adaptive systems: language (Kirby 2001; T. L. Griffiths and Kalish 2007; Kirby, Cornish, et al. 2008), language/music hybrid (Lumaca and Baggio 2017), birdsong (Feher et al. 2009), and abstract symbolic systems (Cornish et al. 2013; Tamariz and Kirby 2015). Although these dynamics are studied for their own right, the primary worth of the iterated learning model, like any model, stems from its ability to account for and predict empirical data. In a particularly striking example, a transmission chain of culturally isolated songbirds evolved its own characteristically wild-type songs within several transmission generations, starting from the raspy, uncharacteristic song of a songbird raised in isolation (Feher et al. 2009).

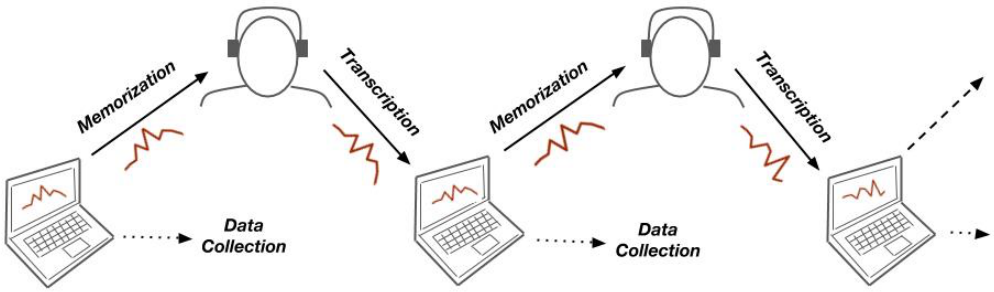


Figure 1: The first generations of the transmission chain.

Statistical musical universals (hereafter simply *universals*) are those features of music that occur with exception but significantly above chance in diverse musical cultures. Recently, Savage, Brown, et al. (2015) have conducted a large cross-cultural survey identifying 18 musical features that can be considered universals. The five universals directly relevant to the use of pitch are:

- Music uses discrete pitches.
- Pitches form nonequidistant scales.
- Scales contain seven or fewer scale degrees per octave.
- Descending or arched melodic contours are prevalent.
- Melody is composed of small intervals less than 750 cents[1].

[1] 100 cents = 1 semitone ≈ a frequency ratio of 1.06

This paper is organized as follows. First, hypothesized explanations of pitch-related music universals and related iterated learning investigations are reviewed. Second, an iterated learning experiment designed to isolate cognitive biases from other biases of cultural transmission is described. Then, the results of the experiment are described. The emergence of four of the five pitch-related universals identified by Savage, Brown, et al. (2015) is investigated and evidence is found that they can be accounted for by memory biases, providing a compelling alternative to motor explanations. Lastly, experimental limitations and the state of motor explanations of melodic contour universals are discussed.

2 Related Work

2.1 Explaining Music Universals

There is no shortage of theories concerning the origins of putative music universals. Here we describe two sets of relevant explanations, neither exclusive nor exhaustive. A first set of explanations considers the physical act of making music. Many gestural theories of music hold that music is perceived by way of physical metaphors (see Schneider 2010, for a review), yet this view by itself does not yield falsifiable predictions. In contrast, the *motor constraint hypothesis* of Tierney et al. (2011) claims arched and descending melodic contours are prevalent because they are more energetically efficient to sing than V-shaped and ascending contours. The use of small intervals in melody can also be explained in this way, although the hypothesis runs counter to the use of discrete pitches, as the octave leap in any lackluster rendition of *Happy Birthday* will remind you. Interestingly, birdsong displays a slightly weaker inclination towards small intervals, which is predicted by the motor constraint hypothesis given differences in human and avian anatomy (Tierney et al. 2011).

A second set of explanations stresses perceptual and memory-related considerations (see Patel 2008; Snyder 2000, for reviews). The universal of octave equivalence, by which pitches separated by an octave are heard as two variants of the same psychological entity, is thought to be innate (Demany and Armand 1984) and some version of the effect is even found in rats (Blackwell and Schlosberg 1943). The use of a small set of discrete pitches per octave is widely believed to reduce memory load in much the same way that a small set of phonemes aids the memorization of words (Dowling 1978). Also, in an information theoretical sense, discrete signals offer a noise robustness that continuous signals do not (Zuidema and De Boer 2009). Small melodic intervals can be explained as adaptations to the task of auditory stream segregation, which is aided by proximity of frequency (Huron 2001). There are several available explanations of nonequidistant scales, the tendency for the fundamental interval of the octave to be split up into unequal intervals to form a scale. Infants prefer consonance over dissonance (Hannon and L. J. Trainor 2007), which may help explain why scale systems are often optimized to maximize consonances between pitches (Huron 1994). This preference happens to give rise to unequal intervals when dividing the octave, assuming harmonic sound sources[2] (Krantz and Douthett 1994). Relaxing this assumption, scale systems are generally better predicted by the frequency content of musical instruments than that of a perfect harmonic source (Sethares 2005) because frequency content affects an interval's perceived consonance (Sethares 1993). Nonequidistant scales also contain asymmetries that could help a listener categorize a scale degree by providing distinct interval relationships between surrounding notes (Balzano 1980).

2.2 Iterated Learning Investigations

Three iterated learning experiments relating to the emergence of music universals are compared to the present study in table 1. Only the work of Ravignani, Delgado, et al. (2016) explicitly investigates the emergence of music universals, while the two others are concerned more broadly with the emergence of structure in pitched acoustical transmission chains. Ravignani, Delgado, et al. (2016) demonstrate that universal rhythmic properties of music emerge from initially random rhythmic patterns. Verhoef et al. (2014) study the emergence of combinatorial structure in transmission chains of slide whistle sounds and its application to evolutionary phonology. Lumaca and Baggio (2017) investigate the emergence of the Gestalt principles of proximity, good continuation, and equivalence in short isochronous melodic systems of discrete pitches. Despite varying goals and methodologies, all three report that sequences become more learnable and structured through transmission generations.

Table 1

Comparison of relevant work.

Reference	Domain	Semantic Component	Immediate Recall	Expressivity Pressure	Motor Component
Ravignani, Delgado, et al. (2016)	continuous rhythm		✓		✓
Verhoef et al. (2014)	continuous pitch			✓	✓
Lumaca and Baggio (2017)	discrete pitch	✓		✓	
This work	discrete pitch		✓		

Ravignani, Delgado, et al. (2016) report the emergence of discrete rhythms exhibiting low-integer durational ratios from initially random continuous rhythmic patterns. They suggest working memory constraints provide the necessary pressure, which is a promising account for both the emergence of discrete pitch and discrete rhythm. This account is complicated somewhat by the work of Verhoef et al. (2014), in which slide whistles are used by participants to imitate signals produced by the previous participant. Signals develop discrete gestural patterns such as "long arched contour" followed by "short repeated notes,"

[2] A perfect harmonic source sounds a frequency called the *fundamental* simultaneously with precise integer multiples of this frequency.

but the pitches themselves do not become discretized. Playing a slide whistle is an unfamiliar task for most people, so playing discrete pitches may have been prohibitively difficult. A similar experiment using the human voice would be useful to test this explanation.

Lumaca and Baggio (2017) report decreasing absolute interval sizes across generations, which is consistent with the universal of small melodic intervals. They also report contour shape data consistent with the arched and descending contour universal (Lumaca and Baggio 2017, online supplement). This agreement between experiment and universals is encouraging, but the experimental design hinders interpretation of these results as the emergence music universals. For one, the pitch sequences were made to take on concrete referential meaning. The participants' goal was to establish a communication system in which pitch sequences referred to one of five photographs of a woman's face, each with a different facial expression. Although the nature of musical meaning is debated (Patel 2008), it is likely not concretely referential in this way (Ravignani and Verhoef 2017). Additionally, signals are not simply transmitted from one generation to the next, but subject to game-like negotiation between generations. Overall, these features of the experiment obscure the conclusions that can be drawn regarding the emergence of music universals.

3 The Present Study

This study investigates the emergence of pitch-related music universals when random pitch sequences are subjected to repeated memorization and transcription. Structural features that emerge are interpreted as adaptations to cognitive biases. The aim of the study is to determine the extent to which these cognitive biases can account for several pitch-related universals: small intervals, descending and arched contours, and nonequidistant scales containing seven or fewer pitches per octave.

The experiment is designed to dissociate cognitive effects from a variety of other unwanted effects. To this end, the procedure resembles a simple memory task: participants listen to a pitch sequence twice and then attempt to recall it. Effects of the physical act of music-making are minimized by having participants transcribe each sequence via a computer keyboard, as opposed to singing or playing a musical instrument. The evaluation of whether scales of seven or fewer scale degrees per octave emerges necessitates the use of long sequences of pitches, which in turn constrains several other experimental parameters. First, musically experienced participants are used. Ideally young and musically inexperienced participants would have been used to dissociate cultural effects, but pitch recall ability among nonmusicians is extremely limited (Williamson et al. 2010). Second, investigation of the emergence of discrete pitch requires a continuous pitch domain, but this is also expected to also limit memory span (Snyder 2000). The same consideration motivates the use of the standard equal temperament chromatic system as opposed to a less culturally significant, less familiar pitch system, such as that used by Lumaca and Baggio (2017).

Another important aspect of the experimental design is its use of immediate recall. Participants recall each signal before they hear the next rather than recalling them in batches as in Verhoef et al. (2014). Immediate recall decreases pressures of systematicity that make a set of signals more self-similar in ways that improve recall of the whole set (Cornish et al.

2013). Immediate recall increases the difficulty, and therefore significance, of any emergence of systematicity. An extreme form of systematicity in which signals become indistinguishable from one another can be inhibited by incorporating expressivity pressures. Expressivity pressures are those pressures that require distinctive signals, for example, through the use of referential signals (e.g. Lumaca and Baggio 2017) or the retention of only dissimilar signals (e.g Verhoef et al. 2014). Fortunately, immediate recall methodology has been found to hamper the emergence of indistinguishable signals (Cornish et al. 2013; Ravignani, Delgado, et al. 2016), so expressivity pressures are not employed. An added benefit of this approach is its simplicity. If expressivity pressures were employed, it may be difficult to delineate emergent features of pitch sequences due to memory biases, expressivity pressures, and interactions of the two.

It is worth emphasizing that the iterated learning methodology presented here is not intended to be a complete model of cultural transmission. Rather, memory constraints and cognitive biases should be considered an important component of any serious model of cultural transmission. The experiment investigates which aspects of music these biases can account for, not whether the model's output is musical, which could be expected of a complete model of cultural transmission.

4 Materials and Methods

4.1 Participants

Seven participants took part in the study (mean age: 21, range: 19-26). All had significant formal musical training (mean: 12 years, range: 8-16), had previously performed melodic transcription tasks, and were either pursuing an undergraduate or graduate degree in music or had recently completed one. All participation was voluntary.

4.2 Pitch Sequences

Twenty consecutive chromatic equal temperament pitches are used, with fundamental frequencies corresponding to those of the pitches C4 through G5, inclusive, on a standard keyboard. The pitch audio is generated by superimposing each fundamental frequency's first five perfect harmonic partials with relative amplitudes given by 0.8^n for the n^{th} partial. Each tone is 500ms in duration and is subjected to a piecewise linear amplitude envelope. During the first 50ms of each tone the amplitude envelope increases linearly from zero to its maximum. During the last 100ms the amplitude envelope decreases linearly to zero. A full sequence of ten pitches takes five seconds to sound.

4.3 Transmission Chains

Each participant recalled eight sequences of ten pitches each. The first participant recalled a set of sequences with each pitch drawn uniformly at random from the 20 available pitches. Each successive participant recalled the previous participant's responses, forming eight parallel transmission chains of eight generations, counting the initial seeded generation. The

participants were unaware they were transcribing the previous participant's transcription.

4.4 Procedure

Testing was performed at a MacBook Pro laptop and audio was played through connected audio-technica ATH-A700 headphones. Participants were provided written instructions and then transcribed two practice sequences via a terminal application pictured and described in fig. 2. After the practice sequences were transcribed, participants were given the opportunity to ask clarifying questions about the testing procedure. The express goal given to the participants was to transcribe each sequence of pitches (termed *melody* in the instructions) as accurately as possible.

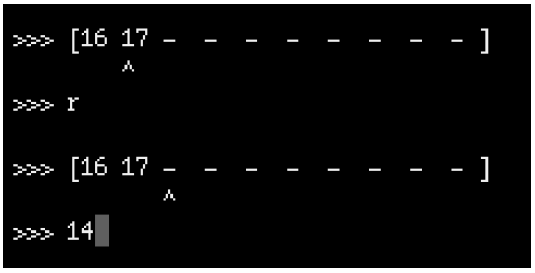


Figure 2: A screenshot of the transcription interface. Pitches are notated using the integers 1 to 20, inclusive, so that 1 denotes C4 and 20 denotes G5. The first line shows the current state of a participant's transcription: 16 (D♯/E♭5) is the first pitch and 17 (E5) is the second. The subsequent eight pitches, each represented by a dash, have not been transcribed yet. The caret, ∧, under the 17 is a pointer. On the next line, the participant enters 'r' as in 'right' and the result is immediately displayed on the next line, in which the pointer has moved to the right, under an unspecified pitch. The last line shows the participant has entered 14 as a guess of this third pitch. Once the 14 is entered by pressing the return key, the pitch is played through the participant's headphones and the 14 is displayed above the pointer on a new line. Additional commands are available to the participant: 'l' moves the pointer left, 'p' as in 'play' plays back the transcribed sequence, 'h' as in 'help' displays a list of possible commands, and 'n' as in 'next' submits the current transcription.

The following is a step-by-step description of the testing procedure: First, the participant hears a pitch sequence played twice. Both playbacks are cued by the participant and they are given no time restrictions. Immediately after the second sounding of the pitch sequence the transcription phase begins. In this phase they are given as much time as needed to transcribe the pitch sequence using the interface pictured and described in fig. 2. They then press a key to submit their transcribed sequence, after which the correct transcription is displayed above a reproduction of their response. This procedure is completed for each of the eight transmission chains. The entire task takes approximately 25 minutes to complete.

4.5 Data Analysis

Bayesian statistical methods with weakly informative, symmetric priors are employed. Bayesian methods do not rely on asymptotic assumptions and are therefore well-suited to performing

inference with small sample sizes. Three models are employed:

$$Y_i = \beta_0 + \epsilon_i \qquad \text{(Constant Model)}$$
$$Y_i = \beta_0 i + \beta_1 + \epsilon_i \qquad \text{(Linear Model)}$$
$$Y_i = \beta_0 i^2 + \beta_1 i + \beta_2 + \epsilon_i \qquad \text{(Quadratic Model)}$$

with priors given by

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, 2, 3, \ldots$$
$$\beta_j \sim \mathcal{N}(0, V), \quad j = 0, 1, 2$$
$$\sigma^2 \sim \mathcal{HN}(0, V)$$
$$V = 10$$

where $\mathcal{N}$ is a normal distribution, $\mathcal{HN}$ is a half-normal distribution, and i is an index, either the generation number or the pitch number, depending on context. All data is z-scored, including indices. Bayes factors and equal-tailed 95% credible intervals for β_0, the leading coefficient, are reported for each statistical test. A Bayes factor is simply the relative odds of the observed data given alternative hypotheses. Here, the alternative hypotheses are $\beta_0 < 0$ and $\beta_0 > 0$. Note that the priors are unbiased with respect to the hypotheses. Credible intervals are reported in original units, not normalized units. Monte-Carlo sampling with Adaptive Metropolis steps is used to estimate posterior distributions using the open source python module PyMC (Patil et al. 2010). Parameter traces are inspected by hand to monitor convergence. Burn-in periods of 20,000 out of a total 220,000 samples were found to be sufficient for all statistics. A robustness check of the model priors in which V is varied is presented in the appendix. Code reproducing all statistical results is available online[3].

5 Results

Figure 3 displays the results of the experiment. The analysis described below show that iterated learning of pitch sequences increases sequence learnability. The emergence of four relevant statistical music universals (small interval sizes, arched and falling contours, nonequidistant scales, and scales of seven or fewer scale degrees per octave) are investigated. Evidence is found that all four can be accounted for by the iterated learning process.

5.1 Transcription Accuracy

Figure 4a shows two measures of transcription accuracy by pitch index. Pitch accuracy is a note-by-note comparison of target sequences against transcriptions. Contour accuracy is a comparison of general motion (rising, falling, or stationary) in neighboring pitches. Consistent with previous studies, contour is recalled more accurately than exact pitch (Dowling and Fujitani 1971; Edworthy 1985). Both measures exhibit the primacy effect, the improved recall for items near the beginning of presented lists, which is well-replicated in immediate

[3] https://github.com/jackgoffinet/iterated_pitch_learning

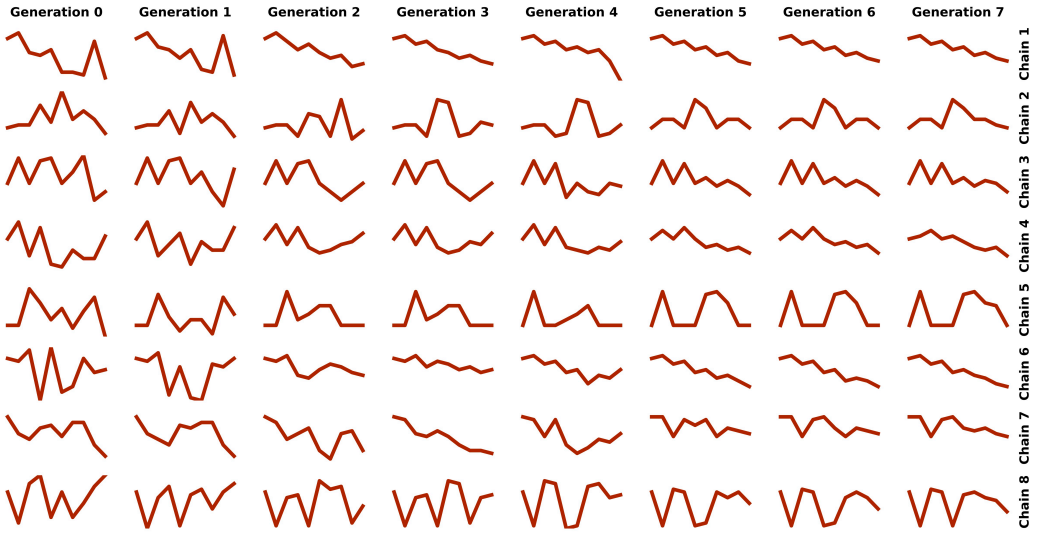


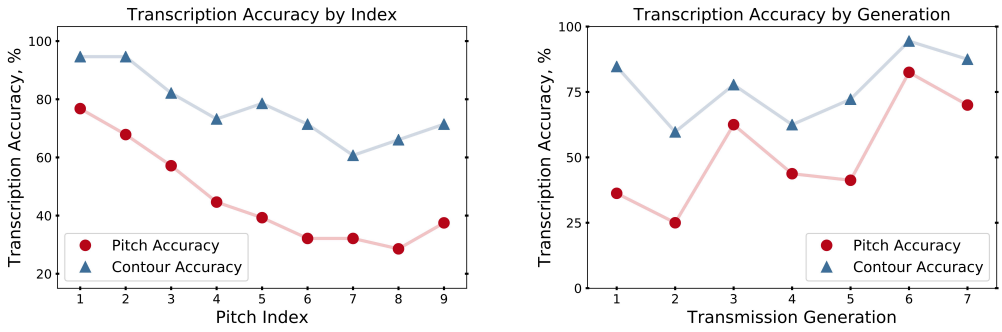
Figure 3: Experiment overview. Red lines represent pitch sequences. Each row shows how a sequence is transformed by the process of repeated memorization and transcription. Each column shows the sequences transcribed by a single participant, except the leftmost column, which was randomly generated.

serial recall experiments (e.g. Henson 1998), and is relevant to concerns here because it predicts increased regularization pressure for later pitches compared to earlier pitches. This effect, which is verified in several ways below, allows windows into more and less regular regimes simply by investigating the final and initial halves of the sequences, respectively. This is used to predict how sequences would evolve given stronger regularization pressures.

Sequences become more learnable as they are repeatedly memorized and transcribed, indicated by the improved transcription accuracy over transmission generations in fig. 4b. Linear regression confirms a strong positive correlation between logit transformed pitch accuracy and generation (BF=22.6, CI=[-0.06, 0.67]) as well as a significant positive correlation for logit transformed contour accuracy (BF=5.07, CI=[-0.24, 0.62]). Participant pitch accuracy never exceeds 83%, which is far enough from perfect recall to dismiss the possibility that the final sequences were influenced only by the biases of the first several participants. Although increasing accuracy over generations implies that later generations made fewer modifications to the sequences, it does not necessarily indicate that the biases of later generations are less influential. Rather, it more likely indicates that the biases of earlier generations are shared by later generations to some degree. The final generation sequences are therefore taken to be the products of all the participants' biases (cf. Navarro et al. 2017).

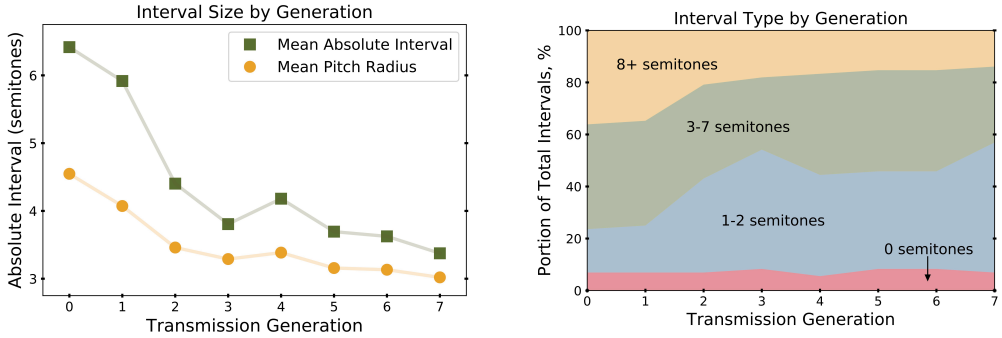
5.2 Small Intervals

Figure 5a presents two measures of sequence density. An absolute interval is the distance in semitones from one pitch to the next and the radius of a pitch is the distance in semitones from that pitch to the average pitch of the sequence, not necessarily an integer. Regres-



(a) *(b)*

Figure 4: Measures of transcription accuracy by (a) pitch index and (b) transmission generation.



(a) *(b)*

Figure 5: Evolution of (a) interval size and (b) interval types by generation.

sion confirms individual participants are biased towards decreasing mean absolute interval (BF=15.1, CI=[-0.98, 0.16]) and mean pitch radius (BF=24.2, CI=[-0.45, 0.04]). Mean intervals in the latter halves of sequences decrease 66% across generations compared to a decrease of only 32% for the initial halves (57% vs. 33% for pitch radii), evidence of heightened regularization pressures for later pitches as predicted by the primacy effect.

In Savage, Brown, et al. (2015), melodies composed of intervals less than 750 cents (here, less than 8 semitones) are reported to be a statistical music universal. Figure 5b shows the portion of intervals larger than this threshold falls from 36% to 14% from the first to last generation and appears to be decreasing slowly. By the last generation, three of eight sequences contain large intervals, compared with eight of eight for the first generation. As predicted by the primacy effect, large intervals are disproportionately found towards the beginning of pitch sequences (by the last generation, all occur within the first five intervals). Consistent with the findings of Lumaca and Baggio (2017), participants are not significantly biased towards increasing or decreasing the portion of 0-semitone intervals (BF=1.49 in favor of increasing CI=[-0.23,0.29]).

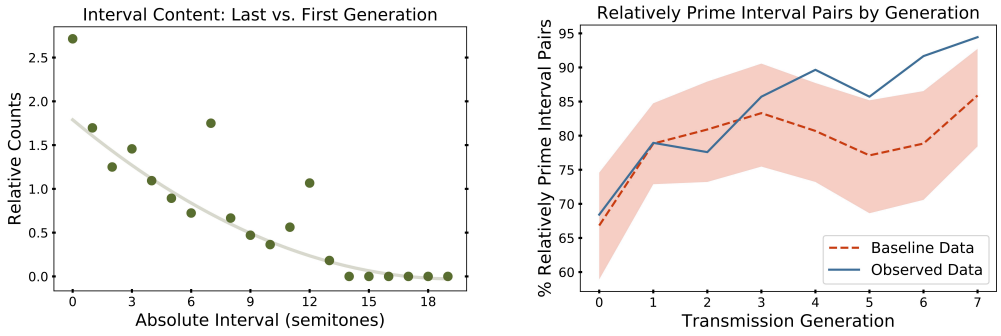
5.3 Arched and Descending Pitch Contour

In Savage, Brown, et al. (2015), melodic phrases with arched or descending contours, as opposed to V-shaped or ascending contours, are reported as statistical music universals. In this context a phrase is defined as "a self-contained series of notes in one or multiple vocal parts. Phrases are usually separated by breaths or long pauses, but can also be separated by more complex grouping principles" (Savage, Merritt, et al. 2012). For simplicity, each pitch sequence is interpreted as a single musical phrase. Qualitative inspection of fig. 3, chains one, three, four, six, and seven, reveals a pervasive trend towards descending contours, especially in the sequence's latter halves, while transmission chains two and five display more arched contours. Following Savage, Tierney, et al. (2017), these trends are assessed quantitatively by performing linear and quadratic regressions on an artificial melody designed to summarize the contour information of the available phrases. First, each phrase is normalized in time and then displaced in pitch (log frequency) to achieve the same mean frequency. Then the median pitch for each timestep is taken to represent the collection of phrases. The sign of the leading coefficient of a linear regression determines whether the collection of phrases exhibits generally ascending or descending contours, while the sign of the leading coefficient of a quadratic regression determines whether the phrases exhibit arched or V-shaped contours. The only deviation from this procedure taken here is to sample the mean pitch, rather than the median, at each timestep to compensate for the comparatively small dataset. The regressions show a clear trend of descending contours over ascending contours (BF=148, CI=[-0.85, -0.13]) and arched contours over V-shaped contours (BF=148, CI=[-0.29, -0.05]) in the last generation. By chance, the randomly generated initial generation is significantly descending (BF=77.2, CI=[-0.79, -0.07]) and arched (BF=67.9, CI=[-0.27, -0.02]), but not as significantly or as strongly as in the final generation. Note that the primacy effect is not immediately relevant to sequence contours because contour is a property of the entire sequence. Inspecting both halves of the sequence separately would obscure this fact.

5.4 Scales of Seven or Fewer Nonequidistant Pitches per Octave

The last relevant statistical universal described by Savage, Brown, et al. (2015) is the use of nonequidistant scales containing seven or fewer scale degrees per octave. This universal is difficult to address using sequences of only ten pitches and is not directly observed (the last generation of sequences contain between 5 and 10 unique pitches and between 4 and 10 unique pitch classes), so its emergence is necessarily more speculative than others. There is nevertheless evidence that nonequidistant scales of seven or fewer scale degrees per octave would emerge given more extreme selection pressures such as longer sequences or musically naive participants.

The first important feature of this universal is the phenomenon of octave salience. An octave (an interval of 12 semitones) corresponds to an exact doubling of frequency and is believed to be perceptually salient for innate, non-cultural reasons (Demany and Armand 1984). Pitches separated by an octave interval are, in a sense, heard as the same pitch, so it is not surprising that almost all scale systems divide the octave and not a different interval. Figure 6a shows the relative interval counts of the last generation compared to the first generation, considering all pairs of pitches in each sequence, not just adjacent pairs. The unison, perfect fifth, and octave (0, 7, and 12 semitones, approximate frequency ratios of 1, $3/2$, and 2, respectively) are disproportionately represented. The relative prominence of



(a) *(b)*

Figure 6: Evidence for the emergence of nonequidistant scales of seven or fewer scale degrees per octave. Plot (a) shows the disproportionate presence of unisons, perfect fifths, and octaves in the final generation, which is taken as indirect evidence of octave equivalence. The trendline is a quadratic ordinary least squares regression discounting the three outlying points. Plot (b) shows the emergence of relatively prime adjacent intervals, evidence of pressures toward the formation of nonequidistant scales. Baseline data is shown with a shaded equal-tailed 95% credible region, obtained by Monte-Carlo analysis of shuffled interval content (n=10^4, see section 5.4).

the octave and other low-integer frequency ratios is consistent with the principle of octave equivalence.

The second important feature of this universal is the use of seven or fewer scale degrees per octave. While this feature emerges in only six of eight transmission chains, there is evidence that more extreme selection pressures would increase that fraction. First, the average number of unique pitches per chain is negatively correlated with generation (BF=295, CI=[-2.90, -0.78]). Perhaps counterintuitively, the correlation is weaker for pitch classes (BF=10.1, CI=[-1.87, 0.46]), yet this should be expected due to the decrease in pitch radii. Also of interest, the number of pitches and pitch classes in the sequences of the final generation is negatively correlated with average interval sizes (pitches: BF=6.19, CI=[-1.48,0.52]; pitch classes: BF=14.1, CI=[-1.74,0.30]). This may indicate participants strike a tradeoff between the memory of pitches and the memory of large intervals. Thus, larger selection pressures are expected to decrease the number of scale degrees.

The last important feature of this universal is the use of nonequidistant scales. If transmission chains tended to settle into equidistant scales, then we would be able to detect this on a local level: adjacent intervals, measured in semitones, would rarely be relatively prime[4]. The standard chromatic scale used in the experiment contains 12 pitches per octave, which is readily broken down into scales of equidistant pitches separated by 2, 3, 4, or 6 semitones. Divided in this way, all intervals would be multiples of 2, 3, 4, or 6 semitones, respectively, and would therefore not be relatively prime[5]. Instead we find the opposite effect: transmission chains exhibit a strong tendency toward relatively prime adjacent intervals. Decreasing interval sizes could contribute to this effect, so a Monte-Carlo analysis was conducted in which the intervals of each sequence are shuffled and only sequences with a pitch range expressible by the 20 available pitches are included ($n = 10^4$). Therefore only the order information of the intervals is manipulated. Repeated pitches are treated like

single pitches and only contiguous intervals between three distinct pitches are counted. The results, shown in fig. 6b with the 95% credible region shaded, show that sequences exceed the upper limits of the credible region from the fourth generation onward. This strongly suggests that pitch memory is enhanced by relatively prime adjacent intervals, perhaps by allowing for asymmetries to help "orient" the listener (Balzano 1980). This may relate to the finding that infants show improved processing of unfamiliar scales with two interval sizes instead of one (Trehub et al. 1999). We hypothesize that this bias towards relatively prime interval pairs contributes to the formation of nonequidistant scales.

6 Conclusions & Discussion

This paper has presented an experiment designed to model the cultural transmission of isochronous pitch sequences via repeated memorization and transcription. Care is taken to isolate memory biases in transmission from motor, social, semantic, affective, gestural, expressive, and aesthetic biases. Repeated memorization and transcription is found to alter the structure of pitch sequences in several ways. First, interval size and pitch radius decrease, consistent with the results of Lumaca and Baggio (2017), and offering an interesting parallel to the observation of shrinking drawings in transmission chains (Tamariz and Kirby 2015). Second, arched and descending contours emerge, resembling world music (Savage, Brown, et al. 2015) and birdsong (Tierney et al. 2011). Third, unisons, perfect fifths, and octaves emerge as important structural intervals, consistent with dissonance-related theories of scale formation (e.g. Sethares 2005). Fourth, participants show a bias toward decreasing the number of unique pitches, possibly aiding the emergence of scale structures. Finally, pitch sequences show a strong bias towards relatively prime consecutive intervals, which we hypothesize contributes to the formation of nonequidistant scales. All of the above biases help account for different pitch-related music universals compiled by Savage, Brown, et al. (2015), providing preliminary evidence that the investigated music universals are adaptations to cognitive biases. Previously, it was found that iterated learning can account for the six compiled rhythmic music universals (Ravignani, Delgado, et al. 2016). A number of non-trivial aspects of world music can be reduced to cognitive considerations if these conclusions are generally correct.

6.1 Limitations

Participant memory capacity is a significant limitation of the present study. The results suggest that scale formation could possibly be accounted for by the cognitive biases relevant to pitch memorization, but sequences longer than 10 pitches are likely needed for a clear

[4]Two integers are relatively prime if no integer greater than one divides both. For example, 3 and 8 are relatively prime, but not 6 and 9.

[5]This is intentionally discounting the chromatic scale, which contains all twelve pitches within the octave. Note that each pitch sequence presented here is trivially contained within the chromatic scale. Apart from containing many more than 7 scale degrees, a chromatic scale system is unlikely to emerge as a result of a pressure towards equidistant scales because there are more reasonable alternatives, such as those containing 2, 3, or 4 scale degrees.

demonstration of emerging scale structures. Participant pitch memory is not expected to extend far beyond 10 pitches without unwanted floor effects. Nevertheless, it may be feasible to investigate the formation of longer sequences with modified transmission chains in which any single participant is only asked to recall part of a longer maintained sequence or by some form of extended rehearsal.

The other major limitation of this study is the influence of cultural effects. Many aspects of music processing are believed to be learned through musical stimuli at young ages (L. Trainor and Hannon 2013), so this limitation cannot be easily overcom. To counter these effects, immediate recall methodology was used, which may reduce culturally-biased pressures towards systematicity. The use of unfamiliar musical scales are also expected to suppress some cultural effects, but there are often cultural analogs. For example, some intervals within the unfamiliar Bohlen-Pierce scale used by Lumaca and Baggio (2017) closely approximate the equal temperament minor third, which is not a very significant interval acoustically, but is nevertheless significant culturally. Such significant confounding factors could even be grounds to doubt the present interpretation of results. Could it be that the pitch sequences grow to resemble world music only to the extent that Western music resembles world music? More precise characterizations of the differences between Western music and other music will be needed to answer this question. Ultimately, cross-cultural or infant-appropriate pitch recall tasks may be necessary to address cultural factors in a satisfying way.

6.2 Melodic Contours and the Motor Constraint Hypothesis

The finding that memory biases can account for arched and falling contours is surprising given previous work concluding that these music universals are the result of motor constraints of physical music production (Tierney et al. 2011; Savage, Tierney, et al. 2017). The *motor constraint hypothesis* holds that arched and descending contours are more energetically efficient to produce using the human voice and as a result are more common in speech and in music. One possible reconciliation of these results is that pitch memory is optimized for the tendencies of speech sounds, our primary source of pitched stimulus, via a process of statistical learning: the voice is influenced by motor constraints which in turn become reflected in the memory and perception of speech and music (see Terken 1991; Pierrehumbert 1979, for the case of speech perception). In talking of music evolution it is easy to forget that music is made by people who may have complex motivations and diverse, unconsciously formulated musical strategies. In this vein, it is plausible that descending and arched melodic contours serve a widespread aesthetic goal, for example the goal of ending phrases with listeners in a state of comparatively low psychological arousal induced by lower pitches. Of course not all of these explanations are mutually exclusive; melodic contour could simply be an overconstrained facet of music. For now, though, more investigation is necessary before making specific claims.

References

Albrecht, J. & Huron, D. (2012). On the emergence of the major-minor system: cluster analysis suggests the late 16th century collapse of the dorian and aeolian modes. In

Proceedings of the 12th international conference on music perception and cognition and the 8th triennial conference of the european society for the cognitive sciences of music (pp. 46–53).

Balzano, G. J. (1980). The group-theoretic description of 12-fold and microtonal pitch systems. *Computer music journal*, *4*(4), 66–84.

Blackwell, H. R. & Schlosberg, H. (1943). Octave generalization, pitch discrimination, and loudness thresholds in the white rat. *Journal of Experimental Psychology*, *33*(5), 407.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial life*, *8*(1), 25–54.

Broze, Y. & Shanahan, D. (2013). Diachronic changes in jazz harmony. *Music Perception: An Interdisciplinary Journal*, *31*(1), 32–45.

Christiansen, M. H., Chater, N., & Culicover, P. W. (2016). *Creating language: integrating evolution, acquisition, and processing*. MIT Press.

Cornish, H., Smith, K., & Kirby, S. (2013). Systems from sequences: an iterated learning account of the emergence of systematic structure in a non-linguistic task. In *Cogsci*.

Demany, L. & Armand, F. (1984). The perceptual reality of tone chroma in early infancy. *The journal of the Acoustical Society of America*, *76*(1), 57–66.

Dowling, W. J. (1978). Scale and contour: two components of a theory of memory for melodies. *Psychological review*, *85*(4), 341.

Dowling, W. J. & Fujitani, D. S. (1971). Contour, interval, and pitch recognition in memory for melodies. *The Journal of the Acoustical Society of America*, *49*(2B), 524–531.

Edworthy, J. (1985). Interval and contour in melody processing. *Music Perception: An Interdisciplinary Journal*, *2*(3), 375–388.

Feher, O., Wang, H., Saar, S., Mitra, P. P., & Tchernichovski, O. (2009). De novo establishment of wild-type song culture in the zebra finch. *Nature*, *459*(7246), 564–568.

Griffiths, T. L. & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive science*, *31*(3), 441–480.

Hannon, E. E. & Trainor, L. J. (2007). Music acquisition: effects of enculturation and formal training on development. *Trends in cognitive sciences*, *11*(11), 466–472.

Henson, R. N. (1998). Short-term memory for serial order: the start-end model. *Cognitive psychology*, *36*(2), 73–137.

Huron, D. (1994). Interval-class content in equally tempered pitch-class sets: common scales exhibit optimum tonal consonance. *Music Perception: An Interdisciplinary Journal*, *11*(3), 289–305.

Huron, D. (2001). Tone and voice: a derivation of the rules of voice-leading from perceptual principles. *Music Perception: An Interdisciplinary Journal*, *19*(1), 1–64.

Jan, S. (2017). *The memetics of music: a neo-darwinian view of musical structure and culture*. Routledge.

Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, *5*(2), 102–110.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*(12), 5241–5245.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current opinion in neurobiology*, *28*, 108–114.

Krantz, R. J. & Douthett, J. (1994). A measure of the reasonableness of equal-tempered musical scales. *The Journal of the Acoustical Society of America*, *95*(6), 3642–3650.

Lumaca, M. & Baggio, G. (2017). Cultural transmission and evolution of melodic structures in multi-generational signaling games. *Artificial Life*.

Lumaca, M., Ravignani, A., & Baggio, G. (2018). Music evolution in the laboratory: cultural transmission meets neurophysiology.

Mauch, M., MacCallum, R. M., Levy, M., & Leroi, A. M. (2015). The evolution of popular music: usa 1960–2010. *Royal Society open science*, *2*(5), 150081.

Navarro, D. J., Perfors, A., Kary, A., Brown, S., & Donkin, C. (2017). When extremists win: on the behavior of iterated learning chains when priors are heterogeneous. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 847–852).

Patel, A. D. (2008). *Music, language, and the brain*. Oxford university press.

Patil, A., Huard, D., & Fonnesbeck, C. J. (2010). Pymc: bayesian stochastic modelling in python. *Journal of statistical software*, *35*(4), 1.

Pierrehumbert, J. (1979). The perception of fundamental frequency declination. *The Journal of the Acoustical Society of America*, *66*(2), 363–369.

Ravignani, A., Delgado, T., & Kirby, S. (2016). Musical evolution in the lab exhibits rhythmic universals. *Nature Human Behaviour*, *1*, 0007.

Ravignani, A. & Verhoef, T. (2017). *Which melodic universals emerge from repeated signaling games?* In press.

Savage, P. E., Brown, S., Sakai, E., & Currie, T. E. (2015). Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences*, *112*(29), 8987–8992.

Savage, P. E., Merritt, E., Rzeszutek, T., & Brown, S. (2012). Cantocore: a new cross-cultural song classification scheme. *Analytical Approaches to World Music*, *2*(1), 87–137.

Savage, P. E., Tierney, A. T., & Patel, A. D. (2017). Global music recordings support the motor constraint hypothesis for human and avian song contour. *Music Perception: An Interdisciplinary Journal*, *34*(3), 327–334.

Schneider, A. (2010). Music and gestures: a historical introduction and survey of earlier research. In *Musical gestures* (pp. 81–112). Routledge.

Sethares, W. A. (1993). Local consonance and the relationship between timbre and scale. *The Journal of the Acoustical Society of America*, *94*(3), 1218–1228.

Sethares, W. A. (2005). *Tuning, timbre, spectrum, scale*. Springer Science & Business Media.

Snyder, B. (2000). *Music and memory: an introduction*. MIT press.

Steels, L. (1997). The synthetic modeling of language origins. *Evolution of communication*, *1*(1), 1–34.

Tamariz, M. & Kirby, S. (2015). Culture: copying, compression, and conventionality. *Cognitive science*, *39*(1), 171–183.

Terken, J. (1991). Fundamental frequency and perceived prominence of accented syllables. *The Journal of the Acoustical Society of America*, *89*(4), 1768–1776.

Tierney, A. T., Russo, F. A., & Patel, A. D. (2011). The motor origins of human and avian song structure. *Proceedings of the National Academy of Sciences*, *108*(37), 15510–15515.

Trainor, L. & Hannon, E. E. (2013). Musical development. *The psychology of music*, 423–498.

Trehub, S. E., Schellenberg, E. G., & Kamenetsky, S. B. (1999). Infants' and adults' perception of scale structure. *Journal of experimental psychology: Human perception and performance*, *25*(4), 965.

Verhoef, T., Kirby, S., & de Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *Journal of Phonetics*, *43*, 57–68.

Williamson, V. J., Baddeley, A. D., & Hitch, G. J. (2010). Musicians' and nonmusicians' short-term memory for verbal and musical sequences: comparing phonological similarity and pitch proximity. *Memory & Cognition*, *38*(2), 163–175.

Zuidema, W. & De Boer, B. (2009). The evolution of combinatorial phonology. *Journal of Phonetics*, *37*(2), 125–144.

7 Appendix: Priors Robustness Check

data	model	V	2.5 quantile	97.5 quantile	BF ($\beta_0 < 0$)	BF ($\beta_0 > 0$)
mean interval differences	constant	5	-0.9434	0.1434	15.9664	0.0626
		10	-0.9810	0.1568	15.1186	0.0661
		20	-1.0280	0.1668	14.5933	0.0685
mean radius differences	constant	5	-0.4279	0.0321	24.1288	0.0414
		10	-0.4530	0.0371	24.1731	0.0414
		20	-0.4647	0.0420	22.8152	0.0438
unison interval content	constant	5	-0.2206	0.2686	0.6910	1.4471
		10	-0.2316	0.2888	0.6691	1.4944
		20	-0.2413	0.2967	0.6620	1.5106
accuracy by generation (logit transformed)	linear	5	-0.0592	0.6466	0.0438	22.8407
		10	-0.0592	0.6701	0.0443	22.5516
		20	-0.0792	0.6911	0.0483	20.7085
contour accuracy by generation (logit transformed)	linear	5	-0.2260	0.5801	0.1840	5.4342
		10	-0.2432	0.6161	0.1972	5.0722
		20	-0.2756	0.6317	0.2008	4.9809
descending contour (generation 0)	linear	5	-0.7748	-0.0698	84.5798	0.0118
		10	-0.7899	-0.0674	77.1861	0.0130
		20	-0.7863	-0.0609	67.7049	0.0148
descending contour (generation 7)	linear	5	-0.8234	-0.1255	131.0132	0.0076
		10	-0.8512	-0.1314	148.3652	0.0067
		20	-0.8467	-0.1317	142.3692	0.0070
pitch classes by average interval	linear	5	-1.6974	0.2919	14.1309	0.0708
		10	-1.7372	0.3036	14.1492	0.0707
		20	-1.7849	0.3410	13.0895	0.0764
pitch classes by generation	linear	5	-1.8205	0.4182	10.6503	0.0939
		10	-1.8741	0.4551	10.1495	0.0985
		20	-1.9211	0.5286	9.1642	0.1091
pitches by average interval	linear	5	-1.4622	0.4965	6.3225	0.1582
		10	-1.4806	0.5236	6.1855	0.1617
		20	-1.5644	0.5559	6.1930	0.1615
pitches by generation	linear	5	-2.8574	-0.7861	387.3495	0.0026
		10	-2.9036	-0.7822	295.2963	0.0034
		20	-2.9121	-0.7560	278.3296	0.0036
arched contour (generation 0)	quadratic	5	-0.2696	-0.0222	74.7289	0.0134
		10	-0.2741	-0.0197	67.9180	0.0147
		20	-0.2781	-0.0180	64.3168	0.0155
arched contour (generation 7)	quadratic	5	-0.2909	-0.0438	127.5347	0.0078
		10	-0.2964	-0.0457	148.3652	0.0067
		20	-0.2955	-0.0401	106.9331	0.0094