RESEARCH REPORT

Learning Novel Musical Pitch via Distributional Learning

Jia Hoong Ong, Denis Burnham, and Catherine J. Stevens
Western Sydney University

Because different musical scales use different sets of intervals and, hence, different musical pitches, how do music listeners learn those that are in their native musical system? One possibility is that musical pitches are acquired in the same way as phonemes, that is, via distributional learning, in which learners infer knowledge from the distributional structure of their input. In this study, we investigate whether novel musical pitch can be acquired distributionally. Nonmusician adults were trained on a continuum spanning a novel musical chord minimal pair (i.e., a novel chord and a mistuned version of that chord) in which the continuum was presented either in a bimodal distribution, with a modal peak at each end of the continuum, or in a unimodal distribution, with a single central modal peak. Discrimination of target minimal pairs was assessed before and after exposure to the distribution. Distributional learning would be said to occur if learners in the bimodal condition, but not those in the unimodal condition, showed evidence of learning, as indexed by improvement in discriminating the minimal pair from pretest to posttest. This indeed was the outcome, suggesting that the building blocks of musical melody—musical pitch—can be acquired using distributional learning.

Keywords: learning, distributional learning, statistical learning, musical pitch, language and music

Musical structure is constrained by many regularities, and music listeners can access these regularities to learn about the underlying organization of music (Tillmann, 2008). This form of learning is termed statistical learning, that is, acquisition of knowledge by tracking the statistics of the perceptual input, ranging from simple event frequency counts to computing complex transitional probabilities (i.e., the probability of Event B following Event A). Statistical learning is implicated in the acquisition in many aspects of high-level musical knowledge, such as musical grammar and syntax, tonality, and expectations (e.g., Jonaitis & Saffran, 2009; Krumhansl, 1990; Loui, Wessel, & Hudson Kam, 2010; Pearce, Ruiz, Kapasi, Wiggins, & Bhattacharya, 2010; Tillmann & Poulin-Charronnat, 2010). What is not known, however, is whether the building blocks of musical melody—musical pitch—can be acquired using statistical learning, which is the focus of this study.

The size and number of musical intervals, and thus the number of musical pitches, vary considerably across different musical scales (Burns, 1999). For instance, the tones in Western major scale are spaced differently and more closely than tones in Japanese slendo scale or in Ugandan scale. Such differences in the division of scales result in perceptual differences (e.g., Cooke, 1992; Lynch, Eilers, Kimbrough Oller, & Urbano, 1990; Renninger & Wilson, 2006; Trainor & Trehub, 1992; Wong, Roy, & Margulis, 2009). For example, Ugandan musicians are more likely to tolerate a mistuning of a note in their native pentatonic scale (i.e., a scale that has five equal intervals) relative to Western musicians, presumably because the Ugandan musical scale has larger intervals between notes than the intervals in the Western scale (Cooke, 1992).

This perceptual difference is analogous to how listeners in different languages experience difficulty in responding to a distinction between nonnative phonemes because of the way in which they have learned to organize speech sounds in their phonological inventory (Fry, Abramson, Eimas, & Liberman, 1962; Goto, 1971; Lisker & Abramson, 1964; Pisoni, Aslin, Perey, & Hennessy, 1982). For example, although native Thai listeners have no difficulty discriminating prevoiced and short-lag bilabial stops as Thai distinguishes the two sounds (i.e., the two speech sounds are phonemic in Thai), native English listeners find it more difficult because the two sounds are not phonemic in English (Pisoni et al., 1982).

Such perceptual differences in music and in language only arise as learners gain experience in their (musical and linguistic) environment. This increasing sensitivity to native speech or musical input is termed perceptual attunement. In terms of learning native language phonemes, attunement was originally proposed to be...
based on minimally contrasted words. For example, English-learning infants may learn that /b/ and /p/ are two distinct phonemes because bear and pear refer to two different objects (e.g., Lalonde & Werker, 1995). Similarly, perhaps because infants encounter different melodies with different pitches, they may learn the distinct musical pitches that exist in their musical system. However, the minimal contrast proposal does not account for the fact that, at least in speech, infants already discriminate nonnative and native phonemes in syllables and words without knowing the referents (Stager & Werker, 1997).

Instead, evidence suggests that learners acquire native phonemes via distributional learning, a specific form of statistical learning in which the statistics tracked by the learner refer to the item frequency count of the input (e.g., Escudero, Benders, & Wanrooij, 2011; Escudero & Williams, 2014; Maye & Gerken, 2000; Maye, Weiss, & Aslin, 2008; Maye, Werker, & Gerken, 2002; Ong, Burnham, & Escudero, 2015). For example, given a continuum that spans a stop consonant minimal pair, such as [d] (as in day) and [t] (as in stay), learners can correctly infer the consonants based on the distributional structural input. If distributional learning occurs, then learners exposed to a bimodal distribution with modal peaks at values corresponding to [d] and [t] toward each end of the continuum would learn that there are two different speech sounds, whereas those exposed to a unimodal distribution with the modal peak in the center of the continuum would not learn the difference between [d] and [t] from the input. Consequently, learners exposed to bimodal input, but not those exposed to unimodal input, should show improved performance on discriminating the minimal pair after training. This twofold outcome is taken as evidence for distributional learning (Ong et al., 2015). Such learning is not restricted to stop consonants but also operates with vowels and lexical tones, which suggests that speech sounds that are perceived in a more continuous manner (as opposed to those that are believed to be perceived in a categorical manner) could also be acquired via distributional learning.

Given the many similarities between speech and music (Lerdahl & Jackendoff, 1985; Patel, 2003, 2008; Perrachione, Vedrenco, Vinke, Gibson, & Dilley, 2013; Slevc, Rosenberg, & Patell, 2009; Tillmann, 2014), could the same learning mechanism be involved in the acquisition of building blocks in both domains? This is assumed in the shared sound category learning mechanism (SSCLM) hypothesis, in which the learning of auditory building blocks is claimed to rely upon a common mechanism (Patel, 2008). The hypothesis seems even more plausible if one accepts the premise that spoken language evolved from music (Darwin, 1871), or the fact that it would be unfamiliar to the participants, who will not have any prior knowledge of that scale. For each chord, a minimal pair is formed between the chord and a 2.5% mistuned version of the middle note of that chord (i.e., Chord X minimal pair, X1X2X3-X1X2 X3; and Chord Y minimal pair, Y1Y2Y3-Y1Y2 Y3), similar to that used previously by Koelsch, Schröger, and Tervaniemi (1999). The frequencies of each note for the chord are displayed in Table 1.

To synthesize the minimal pairs, a program was written in MaxMSP 5 to specify the frequencies of the notes using MIDI, which was then sent to LogicPro 7. The MIDI was played using a female choir preset (Astral Choir) and a male choir preset (Choir Male Chant) on the Alchemy plugin. The duration of each stimulus was approximately 700 ms, with duration being exactly equivalent within each minimal pair. All the stimuli were normalized for intensity at 70dB. Four minimal pairs, which constitute the test stimuli for the present experiment, form a 2 × 2 factorial complex: Chord (Chord X and Chord Y) × Choir Gender (female and male), analogous to the 2 × 2 factorial test stimuli in a related study with lexical tones (2 syllables × 2 speaker’s gender; Ong et al., 2015).
Table 1
Prime Number Notes (Dean, 2009) Used in This Study, Forming Two Chords: Chord X (X1, X2, X3) and Chord Y (Y1, Y2, Y3)

<table>
<thead>
<tr>
<th>Prime number used</th>
<th>X1</th>
<th>X2</th>
<th>X2'</th>
<th>X3/Y1</th>
<th>Y2</th>
<th>Y2'</th>
<th>Y3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prime Number Note</td>
<td>17</td>
<td>23</td>
<td>23</td>
<td>31</td>
<td>41</td>
<td>41</td>
<td>47</td>
</tr>
<tr>
<td>Note. X2 and Y2 are mistuned by 2.5% (i.e., X2' and Y2', respectively) to form a minimal pair for each chord (see Appendix for more details on how the scale is formed).</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Based on a pilot study, we determined that the Female Chord X minimal pair was the most difficult to discriminate, and so this minimal pair was used as the training stimulus pair. To form an eight-step training continuum, we first calculated the difference between the middle notes (X2 and X2') of the chord, and then divided the difference by seven to determine the step sizes for the frequency of the middle note for the intermediate training tokens in the continuum (see Table 2). These frequencies, along with the frequencies of the other two notes in the chord (notes X1 and X3), were sent to LogicPro 7 via MaxMSP 5 and synthesized. Thus, an eight-equal-step training continuum was formed, with Token 1 being Female X1X2X3 and Token 8 being Female X1X2'X3.

In addition, as practice stimuli to familiarize the participants with the discrimination task, we synthesized a 440-Hz sinewave tone and a 440-Hz sawtooth tone, both 800 ms in duration, using Praat (Boersma & Weenink, 2013). The sinewave tone was also used as beep tones to be identified by the participants during the training phase (see Procedure) in a secondary task (Ong et al., 2015).

Procedure
Participants completed three tasks (distributional learning task, language and music background questionnaire, and another task that is part of a larger project, which will not be considered further here), the order of which was randomized. All the tasks were presented using MATLAB 2012b on an Acer TravelMate P653 laptop. The auditory stimuli were presented using a pair of Sennheiser HD650 headphones connected to an Edirol USB Audio Capture UA-25EX audio interface. The general procedure for each task is described next.

Distributional Learning Task
There were three phases in this task: pretest, training, and posttest. At pretest and at posttest, using an ABX discrimination task (i.e., a discrimination task in which participants have to decide whether the third sound “X” is similar to either the first sound “A” or the second sound “B”), participants were asked to discriminate all four test minimal pairs. For instance, in a typical test trial, a participant might encounter the following: Chord Y—Chord X—Chord Y. In this case, the participant should respond “A” as the third sound (X, or in the example, Chord Y) matches with the first sound (A). Each minimal pair was tested eight times, resulting in a total of 32 trials in each session, the order of which were randomized and not blocked. Participants were required to respond within 1 s, with no replacement trials. Prior to performing the pretest, participants were given four practice trials with feedback using sinewave and sawtooth tones to familiarize them with the format of ABX discrimination task.

During training, participants were randomly assigned to the bimodal or the unimodal condition. For the bimodal condition, Tokens 2 and 7 of the training continuum were heard most frequently, whereas for the unimodal condition, Tokens 4 and 5 were heard most often (see Figure 1). Crucially, the number of times both conditions heard Tokens 1 and 8 (i.e., the test stimuli for the Female Chord X minimal pair) were the same. To ensure that the participants were perceptually attentive during training, they were required to complete a cover task; they were instructed that they would hear a total of 288 sounds and that some of those sounds, occurring randomly, would be beeps (Ong et al., 2015). They were required to follow along the sound number being played and circle the sound number on a response sheet with the numbers from 1 to 288 when, and only when, they heard a beep. The distributional learning task took approximately 30 min to complete.

Language and Music Background Questionnaire
Participants were required to indicate the languages known to them and rate them on a 5-point scale how well they (a) speak, (b) understand, (c) write, and (d) read in each of those languages. They were also asked to indicate whether they had any musical training (self-taught, private, or formal), and if so, the type of training that they had, the age of commencement, and the duration of training.

Results
All participants performed the cover task accurately during training (i.e., all identified the beeps with 100% accuracy); therefore, none were excluded from analysis. Prior to analyzing the main results, an independent t test revealed that the two distribution conditions did not differ on their pretest accuracy score

Table 2
Frequency (in Hz) of the Prime Number Note X2 for the Eight-Step Training Continuum

<table>
<thead>
<tr>
<th>Token 1 (X2)</th>
<th>Token 2</th>
<th>Token 3</th>
<th>Token 4</th>
<th>Token 5</th>
<th>Token 6</th>
<th>Token 7</th>
<th>Token 8 (X2')</th>
</tr>
</thead>
<tbody>
<tr>
<td>345.000</td>
<td>343.768</td>
<td>342.536</td>
<td>341.304</td>
<td>340.071</td>
<td>338.839</td>
<td>337.607</td>
<td>336.375</td>
</tr>
</tbody>
</table>

Note. Note that the frequencies of the other two prime number notes in the chord (X1 and X3) are kept constant from Token 1 to Token 8.
Figure 1. Distribution of the training continuum heard by the unimodal condition and the bimodal condition.

(unimodal: $M = .511$, $SE = .019$ vs. bimodal: $M = .469$, $SE = .019$), $t(48) = 1.564$, $p = .124$.

We conducted a 2 (distribution condition: unimodal vs. bimodal) × 2 (session: pretest vs. posttest) × 2 (test chord: trained vs. novel) × 2 (test gender: trained vs. novel) mixed ANOVA on accuracy scores of the distributional learning task to determine whether the two conditions differed in their discrimination performance as a product of exposure to the training stimuli (see Figure 2a–d). There was a main effect of session, $F(1, 48) = 18.763$, $p < .001$, $\eta^2_p = .281$, and a main effect of test gender, $F(1, 48) = 7.075$, $p = .011$, $\eta^2_p = .128$. In general, posttest scores were higher than pretest scores ($M = .554$, $SE = .016$ vs. $M = .490$, $SE = .014$) and performance on the male choir test stimuli was higher than on the female choir test stimuli ($M = .546$, $SE = .015$ vs. $M = .499$, $SE = .015$). Crucially, we found a significant two-way interaction between session and distribution condition, $F(1, 48) = 4.246$, $p = .045$, $\eta^2_p = .081$. Simple main effect analyses revealed that for the unimodal condition, posttest scores were not significantly different from pretest scores ($M = .554$, $SE = .022$ vs. $M = .511$, $SE = .019$), $F(1, 24) = 1.989$, $p = .171$, whereas for the bimodal condition, posttest scores were significantly higher than the pretest scores ($M = .564$, $SE = .022$ vs. $M = .469$, $SE = .019$), $F(1, 24) = 29.037$, $p < .001$.

As in previous studies (Escudero et al., 2011; Ong et al., 2015), we also conducted a series of one-sample $t$ tests on difference scores (posttest–pretest) with the Holm-Bonferroni correction method to determine whether the distribution conditions improved on the four test dimensions (trained chord, novel chord, trained gender, and novel gender; see Figure 2e). For the unimodal condition, none of the difference scores were significantly above zero (trained chord: $t[24] = .911$, $p = .371$; novel chord: $t[24] = .792$, $p = .436$; trained gender: $t[24] = .898$, $p = .378$; novel gender: $t[24] = 1.317$, $p = .200$), whereas for the bimodal condition, all the scores were significantly above zero (trained chord: $t[24] = 2.986$, $p = .006$; novel chord: $t[24] = 3.149$, $p = .004$; trained gender: $t[24] = 3.036$, $p = .006$; novel gender: $t[24] = 3.798$, $p = .001$). This indicates that there was significant distributional learning: Only the bimodal condition, and not the unimodal condition, improved as a result of training.

Discussion

This study examined whether distributional learning can be used to acquire musical pitch. The results suggest that nonmusicians are able to learn novel musical pitch simply by tracking the distributional structure of the input: Those exposed to input with two peripheral modal peaks (bimodal distribution), but not those exposed to input with a single central modal peak (unimodal distribution), improved in the discrimination of musical pitch that span the distribution. Results of participants in the unimodal condition are especially noteworthy, as those participants did not show any significant improvement despite the fact that the current pretest–training–posttest experimental design will most likely introduce practice effects among all participants. Furthermore, participants in the bimodal condition were able to generalize learning to a different chord structure (Chord Y) that has different relational properties to that on which they were trained (Chord X). This suggests that distributional learning not only facilitates learning of the encountered input—the mechanism also allows learners to extract the relevant critical feature (i.e., the meaningfulness of pitch) and generalize it to novel stimuli. Previous distributional learning studies have demonstrated that the product of distributional learning is abstract, allowing learners to generalize their knowledge to different contexts (Maye et al., 2008) and to different speakers (Escudero & Williams, 2014). Furthermore, converging evidence from research in psycholinguistics reveals that learners are able to apply meaningful features that they acquire from their native phonology to a separate class of nonnative speech sounds. For example, Korean listeners, whose language uses duration to distinguish short and long Korean vowels, are able to transfer such knowledge to perceive short and long nonnative consonants more accurately than listeners whose native language does not use duration meaningfully (Pajak & Levy, 2014). Thus, we propose that distributional learning enables learners to acquire not only the elements in the input (e.g., speech sounds, musical pitch) but also meaningful features that define those elements (e.g., pitch, duration), which enables them to generalize their knowledge beyond what they encountered.

Our results indicate that distributional learning may be the mechanism for the formation of musical pitch and, thus, musical interval representations. If so, then it may be the case that music listeners become sensitive to the intervals of their musical system by learning from their native musical input the number of pitches that exist in their musical scales. This would imply that when listeners who are familiar with a particular musical system encounter an unfamiliar musical system, which may have a different set of intervals, perceptual differences arise (Cooke, 1992; Lynch et al., 1990; Renninger & Wilson, 2006; Trainor & Trehub, 1992; Wong et al., 2009). The participants in this study were adults, who already have an established representation of their musical systems; so in order to investigate the developmental progression and locus of any distributional learning for native versus nonnative scales, studies with infants are required to determine whether distributional learning plays a role in enculturation to the surrounding musical system(s).

Nonetheless, our results are in line with literature that suggests that certain aspects of musical knowledge are acquired statistically. Although the majority of the evidence thus far is limited to higher order knowledge such as musical grammar and syntax (Jonaitis & Saffran, 2009; Loui et al., 2010; Pearce et al., 2010; Tillmann & Poulin-Charron, 2010), we have now demonstrated that even low-level knowledge such as musical pitch can also be acquired statistically. Specifically, we demonstrated empirically that distributional learning may cause attunement to scales in learners’
musical systems analogous to perceptual attunement to learners’ native language(s). Native musical scale attunement would act as a template to which learners acquire tonality (i.e., the relations between tones), which is also believed to be based on statistical distribution of tones in melodies (Krumhansl, 1990). Furthermore, when our results and the results of studies on distributional learning of speech sounds (Escudero & Williams, 2014; Maye & Gerken, 2000; Maye et al., 2008; Ong et al., 2015) are taken together, this provides support for the SSCLM hypothesis, which states that auditory building blocks, be they linguistic or musical, are acquired using a common learning mechanism (Patel, 2008).

Not only does support for the SSCLM hypothesis contribute to our understanding of the origin of music and language—it also has implications for second language learning and music learning.

One may question our inclusion of attentive listening via the cover task during training, as previous distributional learning studies on consonants and vowels tend to use a passive listening training task (e.g., Escudero & Williams, 2014). However, we have found that distributional learning for lexical tones is more readily observed when an attentive listening training task is used (Ong et al., 2015). An attentive task during training is more naturalistic—for example, infants learn about language from

Figure 2. Pretest and posttest accuracy by distribution condition for (a) trained chord, (b) novel chord, (c) trained gender, and (d) novel gender. Figure 2(e) illustrates difference scores (posttest minus pretest) on the four test dimensions by the unimodal condition and the bimodal condition. Error bars represent 95% confidence intervals. *p < .01.
infant-directed speech (IDS), which has attention-getting characteristics such as heightened pitch and exaggerated prosody (Burnham, Kitamura, & Vollmer-Conna, 2002; Kuhl et al., 1997; Liu, Kuhl, & Tsao, 2003; Song, Demuth, & Morgan, 2010; Xu Rattanasone, Burnham, & Reilly, 2013). Infants’ earliest musical exposure is infant-directed singing by caregivers, which shares characteristics with IDS in attracting infants’ attention (Trainor, 1996; Trainor, Clark, Huntley, & Adams, 1997). Furthermore, music is ubiquitous and engaging—although we may not be consciously aware of it, we are often attending to background music, which affects our concurrent or subsequent behavior (e.g., Milliman, 1986; North, Hargreaves, & McKendrick, 1999). Having a cover task to potentiate attention during training, although appearing artificial, may be argued to be more ecologically valid than having learners passively listen to the input.

We note that this study was conducted with participants with no tone language experience and no musical training. Would experience with lexical tones or musical pitch facilitate distribution learning of novel musical pitch? This seems plausible given that tone-language listeners and nontone language musicians tend to outperform nontone language nonmusicians in perceiving musical pitch (e.g., Bidelman, Hutka, & Moreno, 2013). Future studies could investigate this by comparing whether different populations (e.g., tone-language listeners and nontone language musicians) would show differential distributional learning.

In line with previous studies of statistical learning of musical knowledge, this study demonstrated that adult learners can acquire novel musical pitch based on the distributional structure of the musical input. Previous studies indicated that this learning mechanism is also implicated in the acquisition of speech sounds. Taken together, this evidence supports the hypothesis that perhaps all learners passively listen to the input.

We are often attending to background music, which is ubiquitous and engaging—although we may not be consciously aware of it, we are often attending to background music, which affects our concurrent or subsequent behavior (e.g., Milliman, 1986; North, Hargreaves, & McKendrick, 1999). Having a cover task to potentiate attention during training, although appearing artificial, may be argued to be more ecologically valid than having learners passively listen to the input.

We note that this study was conducted with participants with no tone language experience and no musical training. Would experience with lexical tones or musical pitch facilitate distribution learning of novel musical pitch? This seems plausible given that tone-language listeners and nontone language musicians tend to outperform nontone language nonmusicians in perceiving musical pitch (e.g., Bidelman, Hutka, & Moreno, 2013). Future studies could investigate this by comparing whether different populations (e.g., tone-language listeners and nontone language musicians) would show differential distributional learning.

In line with previous studies of statistical learning of musical knowledge, this study demonstrated that adult learners can acquire novel musical pitch based on the distributional structure of the musical input. Previous studies indicated that this learning mechanism is also implicated in the acquisition of speech sounds. Taken together, this evidence supports the hypothesis that perhaps all auditory primitives are acquired using the same mechanism (Patel, 2008). Further studies are required to investigate whether this mechanism is involved in the early acquisition of native musical pitch and whether the effect is modulated by extensive pitch experience in either linguistic or musical contexts.

References


Description on the Formation of Prime Number Scale Used in This Study

The Prime Number Scale (Dean, 2009) is a novel microtonal scale. The scale is formed by first choosing a base frequency (Dean suggests a base frequency from 10 Hz to 20 Hz), and then multiplying it with a series of prime numbers. This allows the composer to form a prime number scale of their choosing, that is, the scale is not “fixed” like the chromatic scale, but rather the scale is defined by the composer. In our experiment, we chose the base frequency of 15 Hz and multiplied it with the following prime numbers: 17, 23, 31, 41, and 47. We chose to start with 17 as our first prime number note (X1) because the product of 17 × 15 Hz (i.e., 255 Hz) is roughly equivalent to the frequency of middle C (C4) in the chromatic scale (261.6 Hz). From there, the notes of our prime number scale correspond to the same scale position as the Western C major scale: if 17 × 15 Hz refers to C4, then the multiplication of the base frequency with the next prime number after 17 (i.e., 19) would refer to subsequent note after C4 in the C major scale (i.e., D4), and so on. Because of this correspondence, to form the prime number version of C major and G major chords (CEG and GBD, respectively), we omitted the prime numbers 19, 29, 37, and 43 (which would refer to the chromatic notes D4, F4, A4, and C5). We could have chosen any three prime numbers as the multiplier with the base frequency for our chords, but we chose to build our prime number scale to correspond to the Western C major scale, despite the fact that the ratio between the tones in the chords are different between the two scales (see Figure A1).

Received September 11, 2015
Revision received March 4, 2016
Accepted March 12, 2016

Figure A1. Comparison of note frequency in C major (C4-E4-G4) and G major (G4-B4-D5) chords in Western C major scale (in gray), and in Chord X (X1-X2-X3) and Chord Y (Y1-Y2-Y3) in Prime Number Scale (in black).