Title: Acoustic foundations of the speech-to-song illusion

Authors: Adam Tierney[1†], Aniruddh D. Patel[2,3], Mara Breen[4]

Affiliations

[1]Department of Psychological Sciences, Birkbeck, University of London, London, UK

[2]Department of Psychology, Tufts University, Medford, MA

[3]Azrieli Program in Brain, Mind, & Consciousness, Canadian Institute for Advanced Research (CIFAR), Toronto

[4]Department of Psychology, Mount Holyoke College, South Hadley, MA

†Corresponding author

Adam Tierney, Ph.D.

Birkbeck, University of London

Malet Street, London, WC1E 7HX

United Kingdom

Email: a.tierney@bbk.ac.uk

Phone: +44-020-7631-6368

Running title: Cues to song illusion

Word count: 7,495 words

Abstract

In the “speech-to-song illusion”, certain spoken phrases are heard as highly song-like when isolated from context and repeated. This phenomenon occurs to a greater degree for some stimuli than for others, suggesting that particular cues prompt listeners to perceive a spoken phrase as song. Here we investigated the nature of these cues across four experiments. In Experiment 1, participants were asked to rate how song-like spoken phrases were after each of eight repetitions. Initial ratings were correlated with the consistency of an underlying beat and within-syllable pitch slope, while rating change was linked to beat consistency, within-syllable pitch slope, and melodic structure. In Experiment 2, the within-syllable pitch slope of the stimuli was manipulated, and this manipulation changed the extent to which participants heard certain stimuli as more musical than others. In Experiment 3, the extent to which the pitch sequences of a phrase fit a computational model of melodic structure was altered, but this manipulation did not have a significant effect on musicality ratings. In Experiment 4, the consistency of inter-syllable timing was manipulated, but this manipulation did not have an effect on the change in perceived musicality after repetition. Our methods provide a new way of studying the causal role of specific acoustic features in the speech-to-song illusion via subtle acoustic manipulations of speech, and show that listeners can rapidly (and implicitly) assess the degree to which non-musical stimuli contain musical structure.

Keywords: speech, song, rhythm, melody, pitch

Introduction

Speech and song are usually regarded as being *acoustically* distinct categories of human vocalization (Saitou et al., 2007). However, it has been recently shown that there exist recordings of spoken phrases which are reliably heard as song under certain circumstances. In this “speech-to-song illusion”, spoken phrases which were originally intended to be heard as speech (and which are perceived as speech in their original context) sound like song when removed from context and repeated. The initial report of this phenomenon (Deutsch, Henthorn, & Lapidis, 2011) presented a single spoken phrase to participants ten times and asked them to rate whether it sounded more like song or like speech in three different conditions: unmanipulated, transposed in pitch slightly between repetitions, and with the ordering of syllables jumbled. Song ratings were initially low but increased dramatically with repetition in the unmanipulated condition, with a smaller increase in the transposed condition and no increase in the jumbled condition. These same effects were later replicated in a group of participants with no musical experience (Vanden Bosch der Nederlanden, Hannon, & Snyder, 2015a), indicating that the phenomenon is widely replicable in the general population, rather than being the result of specialized training.

This illusion demonstrates that speech and song are distinct *perceptual* categories that can be derived from physically identical stimuli (i.e., the same verbal utterances). Thus, listeners must overcome perceptual ambiguity not only when making fine judgments such as in speech sound categorization (Ganong, 1980; Connine & Clifton, 1987), but also when assigning a sound to a much broader class of stimuli such as speech versus song. However, not all speech stimuli transform into song in this manner when repeated; Tierney, Dick, Deutsch, & Sereno (2013) developed a corpus of 24 spoken “Illusion” phrases which transform into song with repetition, along with 24 “Control” phrases matched for talker, speech rate, and number of syllables which persist in sounding like speech when repeated. There was, moreover, a high degree of agreement across listeners as to which stimuli transformed and which did not. This raises the question of how the characteristics of verbal utterances interact with cognitive processing to yield the percept of speech or song, thereby resolving perceptual ambiguity. The current paper explores this issue by examining characteristics of

verbal phrases that are or are not subject to the song illusion, and by manipulating specific characteristics to see which are causally related to the perception of a verbal phrase as sung.

What cues might prompt listeners to perceive a spoken phrase as song? One possibility is that within-syllable pitch contours must be sufficiently flat for listeners to assign the syllable a single static pitch. Supporting this idea, pitches within syllables tend to be flatter in song compared to speech (Lindblom & Sundberg, 2007), and an unsupervised learning model showed that the presence of flat pitches is one of the most useful cues for discriminating between speech and music (Schluter & Sonnleitner, 2012). Discrete and gliding pitches may also be processed in partially dissociable neural networks, which could explain the fact that the intraparietal sulcus has been found to be activated during the perception of musical melodies but not speech prosody (Merrill et al., 2012). Moreover, prior work has established that spoken phrases subject to the illusion have less within-syllable pitch variability than phrases not subject to the illusion (Tierney et al., 2013). The presence of flat pitches, therefore, may be one of the cues which primes the speech-to-song illusion (henceforth, "song illusion"). Recent work by Falk, Rathcke, & Dalla Bella (2014) has presented evidence that pitch contour variability has a causal effect on the magnitude of the song illusion: flattening the pitch contour between tonal targets--the pitch targets hypothesized by autosegmental approaches to intonation (Ladd, 2008)--led to earlier and more frequent reports of song transformations for two German sentences. However, the resulting artificially flat pitch contours are uncharacteristic of natural speech, which leaves open the question of whether within-syllable pitch variability can explain variance in the song illusion among naturalistic stimuli, as well as whether a more subtle manipulation could modulate the song illusion. Here we investigate this possibility by comparing the magnitude of the song illusion when manipulating the average pitch slope within syllables, a factor which has a strong influence on the perception of within-syllable pitch contours as static tones versus pitch glides (d'Alessandro & Mertens, 1995).

A second possible cue underlying the song illusion is musical scale structure, as naïve listeners prefer musical tone sequences which conform to scale structure to tone sequences that do not (Cross et al., 1983). Moreover, listeners are better at detecting pitch changes within song illusion stimuli when they are perceived as sung than when they are perceived as spoken, but only when those changes would violate Western musical scale structure (Vanden Bosch der Nederlanden, Hannon, & Snyder, 2015b). This finding suggests that the song illusion causes listeners to hear syllable pitches in terms of musical scale tones, a process which may be facilitated if a stimulus' pitches are an easier fit to a musical key template. Yet Falk et al. (2014) found that changing two pitch intervals into perfect fifths (a pitch interval which forms a backbone of scale structure in many cultures) had only a trending effect on the frequency of reports of the song illusion, and thus the role of melodic structure in driving the song illusion remains in question. Here we investigate this possibility by comparing the magnitude of the song illusion with the extent to which each phrase's sequence of pitches fits a computational model of melodic structure (Temperley, 2007).

Another possible cue underlying the song illusion is rhythmic structure. In particular, the existence of a steady beat may be an important cue supporting the song illusion, as the presence of a steady pulse is one of the most robust features useful for computational discrimination of speech and music (Scheirer & Slaney, 1997). Falk et al. (2014) found that sentences with both a regular distribution of accents and isochronous inter-vowel and inter-accent intervals led to earlier and more frequent song transformations. This suggests that pulse regularity may be an important cue driving the song illusion in naturalistic stimuli as well. Here we investigate this possibility by comparing the magnitude of the song illusion with the variability in inter-beat intervals in a spoken phrase as calculated by a computational model of musical rhythm (Ellis, 2007).

Although the song illusion shows that listeners can evaluate the musicality of a non-musical stimulus, the fact that repetition is necessary for speech to be perceived as song suggests that the musical qualities of these stimuli are not immediately apparent. This is somewhat surprising, as intuition suggests that music is not normally mistaken for speech even after a single repetition. One possible explanation for why repetition is necessary to elicit the speech/song transformation is that when hearing speech listeners are by default operating in a 'speech perception mode' which increases the salience of information more relevant to speech than music (such as spectral envelope) and decreases the salience of information more relevant to music than speech (such as pitch). According to this account, repetition satiates speech perception resources, freeing listeners to focus on acoustic cues such as pitch and rhythm which generally play a secondary role in speech perception. As a result, repetition may cause listeners to switch from a 'speech perception mode' to a 'music perception mode' (Margulis, 2013a). This perceptual mode account is also supported by the finding that variability in inter-stimulus intervals is easier to detect for stimuli which transform into song (Graber, Simchy-Gross, & Margulis, 2017), suggesting that music perception is linked to an increase in the awareness of temporal patterns.

This account is rendered plausible by several experimental paradigms which have demonstrated effects of verbal satiation on speech perception. Massed repetition of a single word, for example, impairs subsequent judgment of whether words fit the same semantic category (Smith, 1984; Smith & Klein, 1990; Pilotti, Antrobus, & Duff, 1997; Pilotti & Khurshid, 2004), decreases N400 effects (Kounios, Kotz, & Holcomb, 2000), and inhibits subsequent recall from memory (Kuhl & Anderson, 2011). Moreover, massed repetition of a single word or phrase can also lead to verbal transformations, as the word eventually turns into a variety of semantically and phonologically related words and nonwords (Warren & Gregory, 1958; Warren, 1961; Warren, 1968; Goldstein & Lackner, 1973; Kaminska & Mayer, 2002; Bashford, Warren, & Lenz, 2006; Bashford, Warren, & Lenz, 2008). Node Structure Theory explains this phenomenon by positing that lexical notes are satiated by repetition but phonological nodes are not; as a consequence the phonological nodes eventually activate neighbouring lexical nodes (MacKay, Wulf, Yin, & Abrams, 1993). A slight modification of this theory could explain the song illusion: repeated activation of a lexical node could lead to satiation of phonological nodes but not of representations of speech prosody, since (in English) there are not links between most lexical nodes and pitch/rhythmic patterns (lexical stress being an exception). Thus repetition may leave prosodic representations unmoored from connections with lexical and phonological processing, causing them to be re-analyzed and their musical qualities assessed.

In support of the 'speech perception mode' explanation for the increase in song perception with repetition, timbral variation inhibits accurate pitch perception (Allen & Oxenham, 2014; Caruso & Balaban, 2014; Warrier & Zatorre, 2002), suggesting an inverse relationship between the salience of spectral information (which is more important for speech perception) and pitch processing (which is more important for song perception). Furthermore, the song illusion is stronger for unfamiliar languages which are more difficult for a listener to pronounce vs. easier to pronounce (Margulis, Simchy-Gross, & Black, 2015), suggesting that the salience of speech-specific information may be decreased for languages that are less similar to a listener's native language, thereby enhancing the song illusion.

However, several characteristics of the song illusion do not fit an explanation based on satiation of speech perception resources. First, it occurs very rapidly: Falk et al. (2014), for example, reported that the song transformation commonly takes place by the third stimulus repetition. This is far more rapid than any demonstrated semantic satiation effects; indeed, most reports of semantic satiation

have contrasted the effects of three repetitions of a stimulus with thirty repetitions of a stimulus (Smith, 1984; Smith & Klein, 1990; Pilotti et al., 1997; Pilotti & Khurshid, 2004). Similarly, Kuhl & Anderson (2011) found inhibited recall from memory only after twenty seconds of repetition, whereas between five and ten seconds of repetition enhanced recall. Second, the verbal transformation effect is perceptually unstable; listeners report that words oscillate back and forth between a number of different percepts (Warren & Gregory, 1958; Warren, 1961; Natsoulas, 1965; Warren, 1968; Goldstein & Lackner, 1973; Ditzinger, Tuller, & Kelso, 1997; Kaminska & Mayer, 2002; Bashford et al., 2006; Bashford et al., 2008). No such instability has been reported for the song illusion, although its perceptual stability has not been formally investigated. Finally, the finding that the song illusion is abolished when the order of syllables is changed between repetitions (Deutsch, 2011) does not favour an explanation based on satiation of speech resources, since although changing syllable order might decrease semantic and phonological satiation (due to multi-syllabic words losing their meaning), it should not abolish it entirely (since monosyllabic words would be unaffected).

An alternative explanation for the repetition effect is that listeners need repeated exposure to stimuli in order to extract musical structure. Listeners' judgments of the exact intervals of a tone sequence, for example, are relatively poor after a single hearing but improve after a few repetitions (Deutsch, 1979). On the other hand, the melodic contour of a tonal sequence (i.e. whether each note is higher or lower than the preceding note, regardless of the size of the pitch jump) can be extracted after only a single presentation (Dowling, 1978). Thus, assigning a tonal schema to a sequence of pitches, which requires knowledge about exact intervals, may necessitate several repetitions, which may explain why repetition enhances enjoyment ratings for unfamiliar music and musicality judgments for random tone sequences (Margulis, 2013b; Margulis & Simchy-Gross, 2016). This application of a tonal schema may, then, distort the perception of the pitch sequence in the direction of the perceived key, which would explain why pitch changes within the song illusion stimuli are easier to detect if they move away from the perceived key (Vanden Bosch der Nederlanden et al., 2015b). According to this explanation, the song illusion is not analogous to the verbal transformation effect and is instead similar to the perceptual transformation that takes place when listeners are exposed to a rapid sequence of vowel sounds (Warren, Bashford, & Gardner, 1990; Warren, Healy, & Chalikia, 1996). These vowel sequences quickly transform into verbal sequences which tend to follow the phonotactic rules of English, as listeners apply their top-down knowledge of speech to make sense of a rapid sequence of acoustic cues. Similarly, the speech-to-song transformations tend to fall within diatonic musical keys (Deutsch et al., 2011; Vanden Bosch der Nederlanden et al., 2015b), even where this would conflict with the underlying stimulus acoustics. Moreover, in both cases the percept is highly stable, and there is relative agreement between participants as to the illusory content of the stimulus.

Here we evaluated these two explanations for the delay in song perception in the song illusion by investigating the acoustic cues driving both initial stimulus ratings and ratings after repetition. The 'speech perception mode' explanation for the role of repetition in the song illusion would predict that song perception judgments after a single repetition of a phrase should be minimal and unrelated to musical characteristics of the stimuli, due to the dominance of the 'speech perception mode' when perceiving novel spoken phrases. The 'musical structure' account, on the other hand, would predict that judgments of musical structure would begin immediately after a single presentation, but would initially be somewhat crude and would be refined after stimulus repetition. This account, therefore, would predict variation in perceived musicality across stimuli both after a single presentation and after repetition, and that some of the same factors which predict the increase in musicality with repetition would also underlie initial differences in musicality. Moreover,

the 'musical structure' account would predict that initial song ratings would correlate with the increase in song rating with repetition.

To address the above issues, we examined the acoustic characteristics of spoken phrases which either are or are not subject to the song illusion using the corpus of Tierney et al. (2013). This corpus consists of 24 song illusion stimuli (which listeners report sound more like song than like speech when repeated) and 24 control stimuli (which continue to sound like speech when repeated). In the current study each stimulus was repeated eight times and listeners with little musical training were asked to judge how song-like the phrase sounded after each repetition. Although prior work has established that the song illusion stimuli sound more song-like than the control stimuli after repetition (Tierney et al., 2013; Vanden Bosch der Nederlanden et al., 2015a), these previous studies investigated song perception using a binary classification in which percepts were categorized as either song or speech. It remains an open question, therefore, whether the strength of song perception varies significantly *within* the Illusion and Control stimuli, whether the Illusion stimuli sound more song-like than the Control stimuli after a single repetition, and what acoustic factors predict song perception both before and after repetition. To investigate these questions, in the present study we asked participants to rate the extent to which a phrase sounded like song after each of eight repetitions, on a scale of 1-10. We studied relations between these ratings and acoustic characteristics of the stimuli, including musical beat structure, melodic structure, and within-syllable pitch variability. In addition, we investigated whether the relationships between stimulus features and song perception differed before and after stimulus repetition.

Experiment 1

The purpose of this experiment was to investigate participants' song ratings after each of eight repetitions of phrases taken from the song illusion corpus. These ratings were then compared to stimulus characteristics including within-syllable pitch variability, melodic structure, and beat variability to determine how the contribution of these characteristics to song perception changed as stimuli were repeated.

Methods

Participants

45 participants (24 female) completed the experiment. Participants' average age was 33.1 years (standard deviation 9.2), and they reported 1.88 (4.08) years of musical training. Based on their performance, as described below, data from five participants were excluded from analysis, meaning that 40 contributed data to the analyses.

Stimuli

Stimuli consisted of 48 short phrases (mean (sd) 5.5 (1.5) syllables, 1.33 (0.41) seconds) taken from audiobooks. Given that the phrases came from spoken (not sung) passages from audiobooks, it can be assumed that they were originally intended to be heard as speech. The phrases were selected by the first author with the intention that 24 of the phrases would be heard as song when played repeatedly (the Illusion stimuli) and the other 24 would not transform in this way (the Control stimuli). The phrases were spoken by three different male talkers represented in equal portions among the Illusion and Control stimuli. The two stimulus sets did not differ in syllable rate (Illusion mean rate 5.13 syllables/second, Control mean rate 5.00 syllables/second) or duration (Illusion 1.29 seconds, Control 1.42 seconds) according to unpaired t-tests ($p > 0.05$). The median pitch of syllables in the Illusion stimuli (141.75 Hz) was about 5% higher than that of the control stimuli (134.83 Hz;

Mann-Whitney U test $p < 0.05$). Prior work has confirmed that (on average) the Illusion stimuli are heard as song after repetition, while the Control stimuli continue to be heard as speech (Tierney et al., 2013). The waveform, spectrogram, and pitch track of one Illusion and one Control stimulus is shown in Figure 1, and these stimuli are available as audio files in the supplementary information. The full set of stimuli and data from all experiments can be found at https://osf.io/t4pjq/.

Procedure

Participants were recruited using Mechanical Turk, a website for recruiting workers for internet-based tasks, and were run on the Ibex Farm system for web-based experiments (Drummond, 2013). Participants were compensated three dollars for their time. All participants completed online informed consent prior to beginning the experiment and all procedures were approved by the ethics boards at Mount Holyoke College and Birkbeck College.

Participants were presented with eight repetitions of each of the 48 phrases (24 Illusion and 24 Control phrases).The two types of phrases were intermingled in a single list, and the order of items in this list was randomized for each participant. After each repetition of a phrase, participants were asked to press a key between 1 and 10 to indicate the extent to which the phrase sounded like speech or song, with 1 indicating completely speech-like and 10 indicating completely song-like. They were given 2 seconds to respond to each repetition of a phrase, after which time (if no response had occurred) the program automatically went on to the next repetition. If participants responded after less than 2 seconds then the next repetition was immediately presented. Given this procedure, stimulus repetitions were not spaced at regular temporal intervals, unlike in previous work with this corpus (Tierney et al., 2013). However, this is unlikely to significantly impact song ratings since previous work has shown that presentation at irregular temporal intervals does not weaken the song illusion (Margulis et al., 2015). The entire procedure took between 30 and 60 minutes.

To ensure that participants were not simply responding randomly, four additional catch trials were included. For each catch trial, a spoken phrase was presented during the first four repetitions. During the last four repetitions, however, a sung phrase was presented containing the same words at the same rate and roughly the same pitches. These phrases were specially created for this study and were comparable in syllable number, mean pitch, and duration to the experimental stimuli. These spoken and sung phrases were recorded by the first author. The first and last ratings were compared for each of the four catch trials. Any participant for whom the average difference between the last rating and first rating was not greater than 0 was excluded from analysis. This procedure resulted in the exclusion of 5 participants from Experiment 1 (out of 45), 3 participants from Experiment 2a (out of 43), 3 participants from Experiment 2b (out of 43), 1 participant from Experiment 3a (out of 41), 1 participant from Experiment 3b (out of 41), 3 participants from Experiment 4a (out of 43), and 2 participants from Experiment 4b (out of 42).

Analysis

The main measures of interest were the ratings across all eight repetitions. However, occasionally participants failed to generate a response for a given repetition. (For example, occasionally a participant's first response was to the second repetition rather than the first presentation.) To account for this, any missing rating was replaced by the mean of the previous and following repetitions. For example, if a participant failed to produce a rating for the fourth repetition, it would be replaced by the mean of the third and fifth repetitions. Missing first and last repetitions were

replaced by the second and seventh repetitions, respectively. On average 1.2% of target responses were missing across all experiments (necessitating the use of adjacent responses).

A series of correlation and regression analyses were conducted on the ratings to determine to what extent three distinct stimulus characteristics influenced initial song ratings and rating changes. These characteristics were pitch slope within syllables, beat variability, and melodic structure, as detailed below.

The absolute *pitch slope* within syllables was measured by extracting the slope of each syllable's pitch contour using linear regression (in semitones per second). (Note that our use of semitone as units here and elsewhere in this study does not convert continuous Hz values into discrete pitch classes, it simply transforms values in Hz to values along a continuous semitone scale, i.e., one which can contain non-integer values such as 1.36 st. This was done because perceptual research suggests that the semitone scale is more relevant for speech intonation perception than the Hz scale [Nolan, 2013].) Pitch was measured in Praat (Boersma & Weenink, 2016) using the autocorrelation method with default settings, which results in one pitch measurement every 10 ms.

Beat variability was calculated using a computational model designed to detect the beat times in music (Ellis, 2007). Once the beat times were extracted, beat variability was calculated as the standard deviation of the inter-beat intervals. (As a result, this measure could not be calculated on 7 of the shortest 48 phrases, for which the algorithm extracted only two beats. These phrases were removed from analyses which included the beat variability data.)

The beat finding algorithm works as follows. First, the sound recording is divided into 40 equally-spaced Mel frequency bands. Next, within each band the onset envelope is extracted as the first derivative of amplitude with respect to time. The envelopes within each band are then averaged to form a global measure of onset strength over time. The global tempo of the recording is then estimated by multiplying the autocorrelation of the onset strength vector by a Gaussian weighting function; the peak of the weighted autocorrelation function indicates the global tempo. For the current analysis the mean of the weighting function was set at 120 beats per minute (sd 1.5 octaves). Finally, a dynamic programming algorithm chooses the beat onset times by attempting to maximize both onset strength and fit to global tempo. The relative weighting of fit to global tempo and onset strength is set by a variable called "Tightness", which for the present analysis was set at 100, allowing a moderate degree of variation from the global tempo. (A higher degree of Tightness will lead the dynamic programming algorithm to prioritize global tempo over onset strength, while a lower degree of Tightness will lead the algorithm to prioritize onset strength over global tempo.)

When this algorithm is applied to speech, the beat times chosen tend to occur on stressed syllable onsets, but they can occur at other times (including during silence) if there is sufficient evidence for beat continuation elsewhere in the phrase. This is consistent with models of beat perception of music (Large, 2008) in which beat percepts, once initiated, are somewhat resistant to contradictory information, and will continue for a time even in silence. (It is this characteristic of musical beats that makes possible the musical phenomenon of syncopation, in which, during brief passages, beats are aligned with silence rather than note onsets.) One advantage of this algorithm is that it permits a degree of non-isochronicity in beat times, which makes it suitable for extracting beat times from natural speech (Schultz, O'Brien, Phillips, & McFarland, 2016).

This algorithm has been previously validated by comparing its output to perceived beat times in a musical corpus (Ellis, 2007). However, although the algorithm has been used to estimate beat locations in speech (Schultz et al., 2016), its validity for this assessment has never been formally

evaluated. Here, to confirm that the beat locations produced by the algorithm are congruent with listeners' perception of beats in repeated speech, we asked two drummers to drum along to the 48 stimuli, repeated eight times with a 2-second interstimulus interval. Different listeners often perceive the musical beat at different metrical levels (Drake, Riess Jones, & Baruch, 2000), and so a direct one-to-one comparison of beat times and drum hits is not straightforward. Instead, participant's drum hit times were compared to the output of the model by selecting, for each stimulus, whichever of the two sets of times (drum hit times or beat times) was smaller, and then calculating the absolute value of the temporal difference between each of the times in the smaller set and the closest time in the larger set. Participant 1's drum hits were, on average, 53.9 ms from the algorithm's beats, while Participant 2's drum hits were 71.6 ms away from the algorithm's beats. Inter-subject variability in beat perception and production adds uncertainty in beat timing unrelated to the accuracy of the beat finding model. To account for this, we calculated a baseline measure of inter-subject beat consistency by using the same beat comparison method described above to compare the drum beats of participants 1 and 2. On average, the two participants' drumming was 48.6 ms apart. As a result, on average the beat tracking model performed 29% worse, relative to when two human drummers were compared, suggesting that the beats output by the model closely match the beats listeners perceive in the stimuli.

Melodic structure was calculated using a Bayesian model (Temperley, 2007) which assesses the extent to which pitch sequences fit characteristics of melodies from Western tonal music. The model makes use of four different statistical profiles generated from analysis of the Essen Folksong Collection, which contains 6217 European folk songs. The first profile is a Central Pitch Profile, which is normally distributed with a mean corresponding to Ab4 (415.3 Hz) and a variance of 13.2 semitones. This gives a higher melodic rating to a sequence whose notes fall closer to the mean of the profile. The second profile is a Range Profile, which is normally distributed with a mean equal to the first note of the sequence and a variance of 29 semitones. This gives a higher melodic rating to a sequence which includes pitches which are more tightly clustered relative to the first note of the sequence. The third profile is a Proximity Profile, which is normally distributed with a mean equal to the previous note and a variance of 8.69 semitones. This gives a higher melodic rating to a sequence which contains smaller pitch intervals. Finally there is a Key Profile, which corresponds to the probability of occurrence of the 12 scale tones given a particular key. This gives a higher melodic rating to a sequence whose distribution of pitches matches one of the 24 musical keys. The Range, Proximity, and Key profiles are combined to form the RPK profile, which calculates the probability of a particular note, given the notes that preceded it.

The model calculates a probability for each of the 24 (major and minor) diatonic keys, given a set of notes, using the following equation:

$$P(pitch\ sequence) = \sum_{k,c}\left(P(c)\prod_{n} RPK_n\right)$$

P(c) is the probability of a particular central pitch being chosen, based on the Central Pitch Profile, and the RPK profile indicates the probability of a given note according to the selected Key Profile, proximity to the previous note, and proximity to the central pitch. Melodic structure was defined as the best fit of each sequence to the key that maximized key fit, as determined by this equation. The pitch sequences used for each phrase consisted of the median pitch of each syllable.

The Temperley model is a multidimensional model of melodic structure, such that sequences can receive high melodic structure ratings if they a) feature small intervals between adjacent pitches (thereby maximizing the Proximity Profile) or b) are composed of pitches that are frequent within a

specific key (thereby maximizing the Key Profile parameter). To investigate which of these subcomponents of this model independently predicts song perception, we measured the *mean interval size* and *key fit* of each phrase. First, the median pitch of each syllable was measured, and then the mean interval size was calculated as the mean of the absolute value of the intervals in semitones between the median pitches of adjacent syllables. Key fit was calculated for each of the 24 major and minor keys, and the key fit value for the best-fitting key was returned. First, each note was given a value equal to the prevalence of that note in the key (following the key profiles given in Krumhansl, 1990). For a C major scale, for example, a C-sharp would be given a smaller value than a C. For notes with intermediate values between note classes, interpolation was used to assign the note a key fit. For example, a note halfway between a C-sharp and a C would be given a value equal to the average of the prevalence values for C-sharp and C. Our rationale for using interpolation rather than rounding to the nearest semitone was that listeners perform better on detection of mistunings in musical intervals for consonant as opposed to dissonant intervals (Hall & Hess, 1984), suggesting that perceptual assimilation does not occur for musical intervals.

Several variables were non-normally distributed, as assessed by the Jarque-Bera test, and were therefore transformed prior to correlational analyses. Beat variability, key fit, and interval size values were log-transformed, and initial song perception ratings underwent a 1/x transform.

In order to determine whether Experiment 1 had sufficient power to detect differences between the two stimulus sets across repetitions, we used the simr library (Green & MacLeod, 2016) implemented in the R statistical framework (R Core Team, 2017; RStudio Team, 2014) to implement a power analysis. We used the responses from Experiment 1 to first generate a linear mixed-effects regression model which included fixed effects of Repetition and Stimulus Set and an interaction between the two. Using a Monte Carlo method, we simulated new values for the Ratings variable, refit the model to the simulated responses, and tested whether the inclusion of the interaction term provided a better fit than a model without the interaction term using a likelihood ratio test. With a sample size of 40 participants and a significance level of 0.05, 100 simulations revealed 62% power to detect a significant interaction between Repetition and Stimulus Set, calculated as a ratings increase of 0.03 for each level of repetition.

Using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) we used linear mixed-effects regression to test whether the strength of the illusion differed significantly between the two stimulus sets. The fixed effects in the analysis were stimulus set (Illusion versus Control), repetition (One through Eight), and the interaction between these two factors. The model fitting procedure for all experiments was as follows: We centered the fixed effects and tested the model with a fully-saturated random effects structure, including random effects for subjects and items and random slopes for the two fixed effects and their interaction. If the model with maximal random effects structure failed to converge, we iteratively removed terms accounting for the least variance from the random effects structure until the model converged. We used the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017) to determine the significance of each fixed effect in the final model.

Results

Participants' ratings confirmed that the song illusion was stronger for the Illusion stimuli than the Control stimuli. After the final repetition the mean Illusion rating was 5.0 while the mean Control rating was 2.5. However, there was substantial variation in ratings within these stimulus sets, which overlapped slightly. Taken together, this set of stimuli span a large portion of the range of possible final song ratings, from almost entirely speech-like to almost entirely song-like. Figure 2 (top panel)

displays mean ratings across all eight repetitions for a single example Illusion and Control stimulus (the same stimuli displayed in Figure 1) and (bottom panel) mean ratings across repetitions for all illusion and control stimuli. Figure 3 (left panel) displays mean initial and final ratings for Illusion and Control stimuli. Figure 3 (right panel) displays values for the change in song rating with repetition for each stimulus sorted by the size of this difference, with Illusion stimuli plotted in grey and Control stimuli plotted in black. Figure 4 displays a histogram of final ratings across all 48 stimuli, which are sorted by mean final song rating.

The final model of Experiment 1 contained random effects of stimulus set and repetition. The fixed effects of this model appear in Table 1.There was a main effect of stimulus set, B = 1.97, $t(80.9) = 5.24$, $p < 0.001$, indicating that the Illusion stimuli were perceived as more song-like than the Control stimuli across both initial and final ratings. There was also a main effect of repetition, B = 0.11, $t(43.9) = 7.36$, $p < 0.001$, indicating that song ratings increased with repetition. Finally, there was an interaction between stimulus set and repetition, B = 0.20, $t(46) = 12.86$, $p < 0.001$, indicating that the increase in song ratings with repetition was larger for the Illusion than for the Control stimuli. A follow-up analysis revealed that, across the entire dataset, initial song ratings correlated with the change in song rating after repetitions ($r(46) = 0.53$, $p < 0.001$).

To determine the extent to which song ratings were stable between participants, we conducted a Monte Carlo analysis in which the participants were randomly divided into two groups of 20 participants and song ratings were compared across groups using Spearman correlations. This procedure was repeated 100 times, resulting in an average correlation between song ratings of $r(46) = 0.94$. This result suggests that a relatively small number of participants suffices for the elicitation of reliable song ratings.

To examine the factors driving differences between stimuli in the strength of song perception, we conducted Pearson's correlations between stimulus characteristics and both initial song ratings and change in song ratings with repetition. P-values reported here were False Discovery Rate corrected. Higher initial song ratings were linked to lower beat variability ($r(39) = -0.42$, $p < 0.05$) and flatter within-syllable pitch slope ($r(46) = -0.40$, $p < 0.05$), but were only marginally linked to melodic structure ($r(46) = 0.28$, $p < 0.1$). Greater change in song ratings with repetition was linked to lower beat variability ($r(39) = -0.39$, $p < 0.05$), flatter within-syllable pitch slopes ($r(46) = -0.64$, $p < 0.001$), and greater melodic structure ($r(46) = 0.50$, $p < 0.01$).

The model we used to assess the degree of melodic structure in the stimuli is multidimensional, reflecting the influence of both pitch interval size and conformity to the distribution of pitches in Western musical scales. To investigate which components of the model are contributing to the relationship with song perception, we ran additional correlations between stimulus ratings and both pitch interval size and fit to Western key structure. Higher initial song ratings were marginally linked to key fit ($r(46) = 0.25$, $p < 0.1$) and were not correlated with pitch interval size ($r(46) = 0.01$, $p > 0.1$). Greater rating change was marginally linked to greater key fit ($r(46) = 0.28$, $p < 0.1$) and significantly correlated with smaller pitch interval size ($r(46) = -0.51$, $p < 0.001$). Pitch interval size was significantly more correlated with rating change than with initial ratings ($z(45) = 2.28$, $p < 0.05$), but this was not the case for key fit ($z(45) = 0.15$, $p > 0.05$). Correlations between stimulus characteristics and song ratings (including 95% confidence intervals) can be found in Table 2.

Hierarchical regression was performed to examine the extent to which within-syllable pitch slope, beat variability, and melodic structure explained variance in rating change. The model predicting rating change solely from beat variability explained 0.15 of the variance in song ratings ($F(1,40) = 6.9$,

$p < 0.05$). Adding pitch slope significantly improved model fit ($F(1,40) = 14.2$, $p < 0.001$), explaining an additional 0.23 of the variance. Adding melodic structure once again significantly improved model fit ($F(1,40) = 9.2$, $p < 0.01$), explaining an additional 0.12 of the variance, for a total r-squared value of 0.50. The relationship between predicted rating change based on the full model and actual rating change is displayed in Figure 5.

Discussion

Our results confirm the existence of the song illusion and show that specific aspects of stimulus structure are associated with the illusion. Participants rated certain phrases as more song-like after repetition, and this transformation was greater for the examples that were pre-selected as being likely to be subject to the song illusion (Tierney et al., 2013). Moreover, a number of the Illusion examples were rated as sounding more like song than speech after repetition. Thus, not only does repetition increase the musicality of certain spoken phrases, but in certain cases this effect can be so strong as to cause speech to be re-categorized as song.

Song ratings were highly stable between randomly selected groups of participants, suggesting that musicality judgments were driven by a reliable set of cues. One of these cues appears to be within-syllable pitch slope, which correlated with the increase in song perception with repetition. Both beat variability and melodic structure also correlated with the extent of the speech/song transformation. These findings, therefore, suggest that a variety of melodic and rhythmic cues contribute to the song illusion, indicating that participants can rapidly (and implicitly) assess the degree to which non-musical stimuli contain musical structure, and use this information to make judgments of musicality.

Initial song ratings for some stimuli were relatively high, with a few stimuli being rated as more like song than speech after only a single repetition. Moreover, initial song ratings and increase in song rating with repetition were correlated, and beat variability and pitch slope were correlated with initial song ratings. Melodic structure, however, only significantly predicted change in song ratings with repetition. Overall, these findings suggest that speech perception does not entirely inhibit the ability to make musical judgments about spoken phrases, even after a single stimulus presentation.

These findings suggest that flat within-syllable pitch slopes, conformity to the melodic characteristics of western music, and stable beats can cause verbal stimuli to sound song-like after repetition. However, the correlational design of Experiment 1 cannot demonstrate that these factors play a causal role. To assess whether these factors can directly affect song perception in speech stimuli we ran three follow-up experiments in which stimuli from the song illusion corpus were manipulated and participants were asked to rate how song-like the stimuli sounded before and after repetition. In Experiment 2, we collected ratings of the stimuli for which we manipulated within-syllable pitch contours by increasing or decreasing the pitch slope of each syllable. In Experiment 3, we collected ratings of all stimuli for which we manipulated between-syllable pitch contours to be more or less melodic. In Experiment 4, we collected ratings of the stimuli for which we manipulated rhythmic structure by increasing or decreasing the variability of inter-beat intervals. With this method, we can investigate whether within-syllable pitch slope, between-syllable melodic structure, and rhythmic variability play a causal role in determining the strength of the song illusion.

Experiment 2

The purpose of this experiment was to investigate the effects of manipulating within-syllable pitch slope on song ratings. Within-syllable pitch contours were manipulated to have flatter or steeper slopes. Participants' song ratings after each of eight repetitions were then compared between the two sets of manipulated stimuli.

Methods

Participants

For the Flat condition, 40 participants were tested (19 female). Participants' average age was 35.2 years (standard deviation 11.6), and they reported 0.6 (1.5) years of musical training. For the Sloped condition, 40 participants were tested (12 female). Participants' average age was 32.6 years (standard deviation 8.6), and they reported 1.8 (2.8) years of musical training.

Stimuli

To investigate the effect of within-syllable pitch slope on judgments of the musicality of speech, we used Praat to manipulate the pitch of the stimuli, creating Flat and Sloped versions. To create the Flat stimuli, within-syllable pitch contours were extracted and detrended to remove any linear trend from the contour (i.e., any overall rising or falling "tilt"), and the stimuli were resynthesized. To create the Sloped stimuli, within-syllable pitch contours were modified such that all stimuli had pitch slopes with an absolute value equal to the average pitch slopes of the Control stimuli. Pitch change was measured in st/s^2, based on psychoacoustic research suggesting that the threshold for perception of a pitch glissando versus a steady tone in speech scales with the square of the duration (Mertens, 2004). The average pitch slope of control stimuli was 0.77, well over the glissando threshold for speech of 0.32 reported in previous work (Rossi, 1971; t'Hart, Collier, & Cohen, 1990; Mertens, 2004), whereas the average pitch slope of illusion stimuli was 0.29, under the glissando threshold. Thus, for the Sloped condition, all syllables in all stimuli were given a slope with an absolute value of 0.77 st/s^s, either ascending or descending depending on the original direction of the syllable's slope. The effects of this process on the pitch contours of the stimuli are illustrated in Figure 6, which plots the pitch contours for the Flat and Sloped versions of a single illusion and a single control stimulus (specifically, the stimuli in Figure 1). As evident from this figure, the "Flat" versions of a phrase did not have truly flat (monotonic) pitch contours within syllables (which would have made them sound highly unnatural): rather, they had pitch contours within syllables which did not have an *overall* upward or downward "tilt" in their pitch pattern.

Procedures

Procedures were identical to Experiment 1.

Analysis

In order to determine whether Experiments 2-4 had sufficient power to detect differences between the acoustic manipulations across repetitions, we used the simr library (Green & MacLeod, 2016) to implement a power analysis. We used the responses from Experiment 2 to first generate a linear mixed-effects regression model which included fixed effects of Repetition and Manipulation and an interaction between the two. Using a Monte Carlo method, we simulated new values for the Ratings variable, refit the model to the simulated responses, and tested whether the inclusion of the interaction term provided a better fit than a model without the interaction term using a likelihood ratio test. With a sample size of 80 participants (40 in each Manipulation group) and a significance level of 0.05, 100 simulations revealed 92% power to detect a significant interaction between Repetition and Manipulation, calculated as a ratings increase of 0.03 for each level of repetition.

To determine whether the pitch variability manipulation affected the perception of the song illusion, we analysed musicality ratings using the model-fitting procedure described in Experiment 1 with Stimulus Set (Illusion versus Control) and Repetition (One through Eight) as within-subjects factors and Pitch Slope (Flat versus Sloped) as a between-subjects factor.

Results

Ratings of the Flat and Sloped manipulations were compared to determine the effect of the within-syllable pitch slope manipulation on participants' song ratings (Figure 7, Table 3). For the Illusion stimuli, participants in the Flat condition produced initial ratings of 3.36 and final ratings of 4.98, while participants in the Sloped condition produced initial ratings of 3.16 and final ratings of 4.15. For the Control stimuli, participants in the Flat condition produced initial ratings of 2.40 and final ratings of 2.68, while participants in the Sloped condition produced initial ratings of 2.90 and final ratings of 3.42 (all ratings represent average values).

In the final model, there was a main effect of stimulus set, B = 1.26, $t(72) = 4.58$, < 0.001, reflecting the tendency for Illusion stimuli to be given higher song ratings. There was also a main effect of repetition, B = 0.12, $t(78) = 9.02$, $p < 0.001$ and an interaction between stimulus set and repetition, B = 0.11, $t(30390) = 11.93$, $p < 0.001$, indicating that song ratings increased with repetition and that this increase was larger for Illusion stimuli. Although there was no main effect of pitch slope on song ratings, B = 0.03, $t(110) = 0.10$, $p > 0.05$, there was a three-way interaction between repetition, stimulus set, and pitch slope, B = 0.11, $t(30390) = 5.73$, $p < 0.001$, due to the fact that the pitch slope manipulation differentially affected the Illusion and Control stimuli. Follow-up analyses on the Illusion stimuli alone demonstrated a marginal effect of pitch slope, B = 0.71, $t(91.9) = 1.82$, $p = 0.07$, indicating higher song ratings for the Flat versions of the Illusion stimuli than for the Sloped versions. In addition, there was a main effect of repetition, B = 0.17, $t(78) = 10.42$, $p < 0.001$, and an interaction between repetition and pitch slope, B = 0.08, $t(78) = 2.35$, $p < 0.05$, such that song ratings increased with repetition, and that this effect was greater for the Flat Illusion stimuli than the Sloped Illusion stimuli. Conversely, analysis of the Control stimuli alone demonstrated a main effect of pitch slope, B = -0.65, $t(97.16) = -2.36$, $p < 0.05$, with higher song ratings for the *Sloped* versions of the Control stimuli than for the Flat versions. Moreover, there was a main effect of repetition, B = 0.06, $t(78) = 5.10$, $p < 0.001$, but no interaction between repetition and pitch slope, $p > 0.05$.

Discussion

We found that manipulating the within-syllable pitch slope of the Illusion and Control stimuli had no effect on initial song ratings, but did change song ratings after repetition. Specifically, for the Illusion stimuli, musicality increases were greater for the Flat stimuli than for the Sloped stimuli. Moreover, differences in the size of the repetition effect between the Illusion and Control stimuli were greater for the Flat stimuli than for the Sloped stimuli. This is in line with the findings of Falk et al. (2014), who found that flattening the pitch contour between tonal targets can enhance song perception in speech. However, for the Control stimuli we found higher song ratings for Sloped than Flat stimuli. This seemingly contradictory result may reflect a difference in the range of musicality perceived across stimuli in the Flat versus Sloped conditions (which were heard by different participants). If we assume that participants use a strategy in which the average musicality across all stimuli is assigned a value near the middle of the rating scale, then an increase in perceived musicality in the Illusion stimuli would be paired, all else being equal, with a decrease in ratings for the Control stimuli, due to participants assigning a greater degree of musicality to the middle of the scale. Similarly, a decrease in perceived musicality in the Illusion stimuli would be a paired with an increase in ratings for the Control stimuli, due to participants assigning a lesser degree of musicality to the middle of the scale.

The larger repetition effect for the Sloped Control stimuli, therefore, may simply reflect a tendency for participant ratings to drift towards the middle of the rating scale when musicality varies to a lesser extent across a stimulus set.

Overall, then, these results indicate that manipulating within-syllable pitch slope can modulate the size of the increase in song perception with repetition. This lends support to our finding in Experiment 1 that within-syllable pitch slope is correlated with the size of the repetition effect. Steep within-syllable pitch slopes may exceed the threshold for perception of a glissando within a spoken syllable, rather than a static pitch (Rossi, 1971; t'Hart et al., 1990; Mertens, 2004). This could affect musicality judgments in several ways. First, given that music is characterized by relatively flat pitch contours within notes (Schluter & Sonnleitner, 2012), listeners may simply be using the relative frequency of flat versus sloped pitches when making a judgment about whether a given stimulus should be classified as song versus as speech. Second, flat pitches may be a precondition for the detection of other musical characteristics in stimuli. For example, perception of pitches and pitch intervals in glissandos may not be accurate enough to make judgments about pitch interval size and key fit.

Experiment 3

The purpose of this experiment was to investigate the effects of manipulating melodic structure on song ratings. Between-syllable pitch contours were manipulated to provide better or worse fits to musical structure. Participants' song ratings after each of eight repetitions were then compared between the two sets of manipulated stimuli.

Methods

Participants

For the Strong Musical Structure condition, 40 participants were tested (17 female). Participants' average age was 34.8 years (standard deviation 10.4), and they reported 1.2 (2.2) years of musical training. For the Weak Musical Structure condition, 40 participants were tested (19 female). Participants' average age was 36.9 years (standard deviation 11.2), and they reported 1.6 (3.5) years of musical training.

Stimuli

First, the median pitch of each syllable was calculated, and the syllable's entire pitch contour was slightly shifted so that the median pitch was aligned with the nearest semitone (relative to 440 Hz). Next, the pitch of each syllable was randomly shifted up or down with a magnitude of -3, -2, -1, 0, +1, +2, or +3 semitones. This procedure was carried out 250 times, and the Temperley algorithm was then used to select the iteration which minimized melodic structure (for the Weak Musical Structure stimuli) or maximized melodic structure (for the Strong Musical Structure stimuli). The manipulation did not alter the pitch contour within syllables; see Figure 8 for an illustration of the results of this process on the pitch contour of an Illusion and Control stimulus.

To confirm that this manipulation was successful in controlling melodic structure, a set of tonal melodies were constructed based on the output of the minimization versus maximization of the Temperley algorithm, as described above. In other words, the pitch of each tone of the melodies was equal to the median pitch of the pitch contours of each syllable in the Strong Musical Structure and Weak Musical Structure stimuli. The tones were five-harmonic complex tones, with cosine ramping at onset and offset to avoid transients. Two participants (both male, one 35 years old, the other 23 years old) were presented with matched pairs (i.e. a Strong Musical Structure and Weak Musical Structure version of the same stimulus) and were asked to rate which of the two melodies sounded more melodic. For one of the two participants, the Strong Musical Structure version was rated as more melodic for all 48 stimuli. For the second participant, the Strong Musical Structure version was

rated as more melodic for 45 out of the 48 stimuli. Thus, this procedure was clearly successful in manipulating the musicality of melodies.

Procedures

Procedures were identical to Experiment 1.

Analysis

To determine whether the melodic structure manipulation affected the perception of the song illusion, we analysed musicality ratings using the model-fitting procedure described in Experiment 1 with Stimulus Set (Illusion versus Control) and Repetition (One through Eight) as within-subjects factors and Musical Structure (Strong versus Weak) as a between-subjects factor.

Results

Ratings of the Strong Musical Structure and Weak Musical Structure manipulations were compared to determine the effect of the melodic structure manipulation on participants' song ratings (Figure 9, Table 4). Song ratings were lower overall in the Weak Musical Structure condition, especially for Illusion stimuli (see Figure 9). For the Illusion stimuli, participants in the Weak Musical Structure condition produced average initial ratings of 3.36 and final ratings of 4.61, while participants in the Strong Musical Structure condition produced initial ratings of 3.84 and final ratings of 5.06. For the Control stimuli, participants in the Weak Musical Structure condition produced normalized initial ratings of 2.34 and final ratings of 2.61, while participants in the Strong Musical Structure condition produced normalized initial ratings of 2.56 and final ratings of 2.72.

In the final model, there was a main effect of stimulus set, $B = 1.88$, $t(65) = 5.51$, $p < 0.001$, reflecting the tendency for Illusion stimuli to be given higher song ratings. There was also a main effect of repetition, $B = 0.10$, $t(78) = 7.98$, $p < 0.001$, and an interaction between stimulus set and repetition, $B = 0.13$, $t(30390) = 14.50$, $p < 0.001$, indicating that song ratings increased with repetition and that this increase was larger for Illusion stimuli. However, there was no main effect of the melodic structure manipulation on song ratings, and this manipulation did not interact with the other factors (all p's > 0.05).

Discussion

We found that manipulating the degree to which each phrase fit the characteristics of Western music (as measured using the model of Temperley, 2007) did not significantly change overall song ratings, the increase in song ratings with repetition, or the difference in song ratings between Illusion and Control stimuli. Given that there was a non-significant trend towards greater song ratings for the Strong Musical Structure stimuli and that our statistical power was sufficient only to detect a medium-sized effect, we cannot draw strong conclusions from these results.

The Temperley model is multi-dimensional, incorporating absolute pitch, pitch range, pitch interval size, and the extent to which pitches conform to Western musical scales. One possibility is that some of these dimensions are more fundamental to the speech-song illusion than others, and that a more targeted manipulation could have had a larger effect on musicality perception. In particular, our results from Experiment 1 indicated that pitch interval size was more strongly correlated with increase in musicality rating with repetition than was fit to a musical key, suggesting that interval size may be the strongest cue to musicality. Future work could investigate this possibility by independently manipulating stimuli along each of the four dimensions of the Temperley model to see which has the greatest effect on the size of the speech-song illusion.

Experiment 4

The purpose of this experiment was to investigate the effects of manipulating rhythmic structure on song ratings. Inter-beat intervals were manipulated to increase or decrease the variability of inter-beat interval timing. Participants' song ratings after each of eight repetitions were then compared between the two sets of manipulated stimuli.

Methods

Participants

For the Isochronous condition, 40 participants were tested (18 female). Participants' average age was 35.8 years (standard deviation 9.1), and they reported 0.9 (2.5) years of musical training. For the Variable condition, 40 participants were tested (17 female). Participants' average age was 33.2 years (standard deviation 9.4), and they reported 2.3 (3.0) years of musical training.

Stimuli

Praat was used to manipulate the timing of the Illusion and Control stimuli. Timing manipulations were based on the syllable rime (i.e. from the onset of the syllable's first vowel to the end of the syllable), given previous evidence that the point at which the onset of a syllable is perceived falls closer to the vowel onset than to syllable onset (Morton, Marcus, & Frankish, 1976). In the isochronous condition, stimuli were manipulated such that the duration of each rime was made identical, and equal to the mean rime duration of the phrase. In the variable condition, a Monte Carlo method was used to construct stimuli which, according to the computational model of beat times, had highly variable inter-beat intervals. The duration of each inter-rime-onset interval within a phrase was multipled by 2^n, with n for each interval drawn randomly from a continuous uniform distribution between -1 and 1. This process was repeated 250 times. For each of the resulting stimuli, beat variability was calculated using the computational model of beat timing. The exemplar with maximal beat variability was then selected. See Figure 10 for an illustration of the results of this process on the waveform and spectrogram of an Illusion and Control stimulus.

Procedures

Procedures were identical to Experiment 1.

Analysis

To determine whether the beat variability manipulation affected the perception of the song illusion, we analysed musicality ratings using the model-fitting procedure described in Experiment 1 with Stimulus Set (Illusion versus Control) and Repetition (One through Eight) as within-subjects factors and Beat Variability (Isochronous versus Variable) as a between-subjects factor.

Results

Ratings of the Isochronous and Variable timing manipulations were compared to determine the effect of the timing manipulation on participants' song ratings (Figure 11, Table 5). For the Illusion stimuli, participants in the Isochronous condition produced average initial ratings of 3.95 and final ratings of 5.63, while participants in the Variable condition produced initial ratings of 4.23 and final ratings of 5.56. For the Control stimuli, participants in the Isochronous condition produced average initial ratings of 2.52 and final ratings of 2.87, while participants in the Variable condition produced initial ratings of 2.82 and final ratings of 3.09.

Using the model-fitting procedure described in Experiment 1, we analysed song ratings on a trial-by-trial basis using mixed-effects regression with stimulus set (Illusion versus Control) and repetition (One through Eight) as within-subjects factors and manipulation (Isochronous versus Variable timing) as a between-subjects factor. In the final model, there was a main effect of stimulus set, B = 2.31, $t(66) = 6.99$, $p < 0.001$, reflecting the tendency for Illusion stimuli to be given higher song ratings. There was also a main effect of repetition, B = 0.12, $t(86) = 8.87$, $p < 0.001$, and an interaction between stimulus set and repetition, B = 0.15, $t(46) = 12.21$, $p < 0.001$, indicating that song ratings increased with repetition and that this increase was larger for Illusion stimuli. However, there was no main effect of the timing manipulation on song ratings, and this manipulation did not interact with the other factors (all p's > 0.05).

Discussion

We found no difference in the increase in song ratings with repetition between the stimuli with Isochronous versus Variable inter-rime timing. This result contrasts with our finding in Experiment 1 that stimuli with variable inter-beat-intervals increased in musicality to a lesser extent with repetition and with the finding of Falk (2014) that stimuli with isochronous inter-accent intervals were more likely to transform into song and did so more rapidly. One way to explain this seeming discrepancy is that it is rhythmic regularity at higher hierarchical levels, not at the syllable level, which is crucial for the perception of musical beats in speech and, therefore, the perception of the song illusion. If so, forcing isochrony at the level of the individual syllable could actually have increased timing variability at higher levels crucial for beat perception, given that stressed syllables (and accented words) were separated by variable numbers of syllables. Future work could investigate this issue by comparing and contrasting the effects of imposing isochrony on inter-syllable and inter-stress intervals.

General discussion

Overall, we show that naïve listeners produced highly reliable ratings when asked to assess the musicality of short spoken phrases. After a single repetition, these ratings were relatively low. However, initial ratings correlated with the increase in ratings with repetition, and were also linked to stimulus characteristics, including beat variability and within-syllable pitch slope, suggesting that listeners can begin to pick up on musical characteristics of stimuli even after a single repetition. With repetition, song ratings for certain phrases increased, and some phrases began to sound so musical that they seemed to transform into song. This increase was linked to several stimulus characteristics: within-syllable pitch slope, melodic structure, and beat variability. Three follow-up experiments investigated the extent to which these characteristics play a causal role in the perceptual transformation of speech into song, finding that manipulating within-syllable pitch slope changed the magnitude of the increase in musicality ratings.

Our findings provide an opportunity to test two competing accounts of the increase in musicality with repetition in the song illusion stimuli. The "speech perception mode" account of the repetition effect suggests that by default speech is perceived in a perceptual mode that emphasizes higher-frequency timbral components of sound and de-emphasizes pitch and rhythm patterns, but that repetition satiates speech representations, leading to a switch to a music perception mode in which pitch and rhythm are more salient. This account would predict that initial musicality ratings should be uncorrelated with ratings after repetition (since the two ratings reflect different perceptual modes) and not linked to pitch and rhythm characteristics (since these characteristics are by default

de-emphasized). The “musical structure” mode account, on the other hand, suggests that listeners are always capable of evaluating the musical characteristics of stimuli, but that this process takes time, necessitating repeated exposure to sequences as mental representations of tonal structure are fine-tuned. This account would predict that initial musical musicality ratings should correlate with ratings after repetition (given that the same basic perceptual mechanisms are at play across repetitions) and should be linked to pitch and rhythm characteristics (since music perception is not initially inhibited). We found that initial and final musicality ratings were correlated, and that initial musicality ratings correlated with both pitch and rhythm-based characteristics of the stimuli, evidence which clearly supports the musical structure account of the repetition effect. (For other studies of non-linguistic stimuli which also support this account, cf. Tierney, Patel, & Breen, in press, and Simchy-Gross and Margulis, 2018.)

The fact that initial ratings were correlated with both the change in musicality with repetition and with acoustic characteristics of the stimuli suggests that the increase in song perception with repetition is due to increasing precision of and confidence in analysis of musical characteristics of the phrase, rather than satiation of speech perception resources. However, manipulation of within-syllable pitch slope affected the increase in musicality with repetition but not musicality judgments after a single stimulus presentation. Further work, therefore, is needed to determine the relative contributions of speech satiation versus gradual extraction of musical structure to the speech/song transformation. Of course, these are not mutually exclusive explanations, and so one possibility is that both contribute to the illusion to some degree.

We find at best weak evidence that the speech/song illusion is driven by the extent to which the sequence of pitches underlying a phrase’s syllables fits a musical key. Correlations between key fit and musicality perception were only marginal, and manipulating the musical structure of the phrases using a model featuring key fit as a prominent component did not have a significant effect on musicality ratings. These results are in line with the finding of Falk et al. (2014) that the imposition of a perfect fifth interval had only a trending effect on the speech/song illusion. Given prior evidence that the speech/song illusion involves perceiving pitch sequences as conforming to scale structure (Deutsch et al., 2011, Vanden Bosch der Nederlanden et al., 2015b), these findings suggest that listeners are willing to perceptually classify pitch values within scale templates even if these pitch values do not strongly fit an existing musical key, as long as there is sufficient evidence that a sequence has other musical characteristics.

On the other hand, we found that the size of pitch intervals was linked to the change in musicality ratings after repetition. This is somewhat surprising, as small intervals predominate in both music (Von Hippel & Huron, 2000) and speech [when intervals are defined as pitch distances between the mean fundamental frequency of successive syllables] (Tierney, Russo, & Patel, 2008). Given that a predominance of small intervals is not unique to music, it is unclear why listeners would rely on interval size when making the decision whether a phrase is more characteristic of song than speech. One possibility is that large intervals may interfere with the long-distance pitch comparisons necessary for the construction of a tonal schema. Supporting this theory, Deutsch (1978) showed that larger pitch intervals interfere more with delayed pitch comparisons. Another possibility is that a phrase containing larger pitch intervals sounds less musical because large pitch intervals are more difficult to produce with the sub-semitone accuracy necessary for music production, especially for listeners who are not trained singers. Studies of the verbal transformation effect have shown that illusory percepts tend to be drawn towards sequences that are easier to produce (Sato, Schwartz, Abry, Cathiard, & Loevenbruck, 2006; Sato, Valleé, Schwartz, & Rousset, 2007), and production constraints may have a similar effect on illusory musical percepts in the speech/song illusion.

Our finding that subtly manipulating within-syllable pitch slope can modulate the song illusion may enable the construction of verbal stimuli that are closely matched acoustically but differ strikingly in the extent to which they elicit a song percept. This could be a useful tool for the investigation of the neural and cognitive mechanisms involved in music perception. Comparing music perception to other perceptual modes has been challenging due to the acoustic differences between naturalistic music and other auditory stimuli as well as the cultural preconceptions which affect listeners when perceiving real music. Manipulating the strength of the song illusion while leaving most acoustic characteristics unaltered could enable a highly controlled comparison of speech and song perception.

Context of the research

In conclusion, listeners can make sophisticated musical judgments about stimuli which they know were not intended to be heard as music. Thus, music perception is a listening mode which can be applied to a wide variety of stimuli rather than being limited to a narrow range of cultural artefacts. This perceptual flexibility may contribute to the remarkable diversity of music. It is an open question, however, whether the cues to musicality revealed here reflect universal preferences as opposed to musical and linguistic influences shared by our subjects. Future work should examine whether these same cues are relied upon cross-culturally. Speakers of tone languages, for example, have been shown to rate speech as less song-like, and report less of an increase in song perception when spoken stimuli are repeated (Jaisin, Suphanchaimat, Candia, & Warren, 2016). This could indicate that speakers of tone languages are relying less on melodic cues and more on rhythmic cues when assessing the musicality of speech. Another interesting direction for future work would be to examine how cues to musicality are weighted differently by participants in different developmental stages. Infants are capable of making sophisticated judgments of the locations of musical beats (Phillips-Silver & Trainor, 2005), and pulse clarity increases coordination between musical and motor tempos in infants moving to music (Zentner & Eerola, 2010). Beat stability may, therefore, be a cue to the musicality of stimuli relatively early in life. On the other hand, infants are insensitive to scale structure (Trehub, Bull, & Thorpe, 1984; Trainor & Trehub, 1992). The extent to which a sequence of pitches contains melodic structure may, then, be less important as a cue to song perception at early developmental stages. Testing these predictions would require the development of a method of assessing song perception that does not require explicit ratings. One possibility is that infants would prefer to listen to more repetitions of stimuli which elicit the song illusion, compared to control stimuli.

Acknowledgements

We thank Bob Ladd for helpful comments on an earlier version of this manuscript.

References

Allen, E. J., & Oxenham, A. J. (2014). Symmetric interactions and interference between pitch and timbre. *The Journal of the Acoustical Society of America*, *135*(3), 1371-1379.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390-412.

Bashford, J., Warren, R., & Lenz, P. (2006). Polling the effective neighborhoods of spoken words with the verbal transformation effect. *JASA*, 119, EL55.

Bashford, J., Warren, R., & Lenz, P. (2008). Evoking biphone neighborhoods with verbal transformations: illusory changes demonstrate both lexical competition and inhibition. *JASA*, 123, EL32.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1).

Bigand, E., & Poulin-Charronnat, B. (2006). Are we 'experienced listeners'? A review of the musical capacities that do not depend on formal musical training. *Cognition, 100*, 100-130.

Boersma, P., & Weenink, D. (2016). Praat: doing phonetics by computer [Computer program]. Version 6.0.22, retrieved from http://www.praat.org/

Caruso, V. C., & Balaban, E. (2014). Pitch and timbre interfere when both are parametrically varied. *PloS one*, *9*(1), e87065.

Connine, C., & Clifton, C. (1987). Interactive use of lexical information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance,* 13, 291-299.

Cross, I., Howell, P., & West, R. (1983), Preferences for scale structure in melodic sequences. *Journal of Experimental Psychology: Human Perception and Performance, 9*, 444-460.

D'Alessandro, C., and Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language,* 9, 257-288.

Dalla Bella, S., Bialunska, A., & Sowinski, J. (2013). Why movement is captured by music, but less by speech: role of temporal regularity. *PLoS ONE*, 8, e71945.

Deutsch, D. (1978). Delayed pitch comparisons and the principle of proximity. *Perception and Psychophysics, 23*, 227-230.

Deutsch, D. (1979). Octave generalization and the consolidation of melodic information. *Canadian Journal of Psychology, 33*, 201-205.

Deutsch, D., Henthorn, T., & Lapidis, R. (2011). Illusory transformation from speech to song. *JASA, 129*, 2245-2252.

Ditzinger, T., Tuller, B., & Kelso, J. S. (1997). Temporal patterning in an auditory illusion: the verbal transformation effect. *Biological cybernetics*, *77(1)*, 23-30.

Dowling, J. (1978). Scale and contour: two components of a theory of memory for melodies. *Psychological Review, 85*, 341-354.

Drake, C., Riess Jones, M., & Baruch, C. (2000). The development of rhythmic attending in auditory sequences: attunement, referent period, focal attending. *Cognition*, 77, 251-288.

Drummond, A. (2013) Ibex Farm. Available: http://spellout.net/ibexfarm/.

Ellis, D. (2007). Beat tracking by dynamic programming. *Journal of New Music Research, 36,* 51-60.

Falk, S., Rathcke, T., & Dalla Bella, S. (2014). When Speech Sounds Like Music. *Journal of Experimental Psychology: Human Perception and Performance, 40*, 1491-1506.

Ganong, W. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance,* 6, 110-125.

Goldstein, L., & Lackner, J. (1973). Alterations of the phonetic coding of speech sounds during repetition. *Cognition*, 2, 279-297.

Grahn, J., & Brett, M. (2007). Rhythm and beat perception in motor areas of the brain. *Journal of Cognitive Neuroscience*, 19, 893-906.

Graber, E., Simchy-Gross, R., & Margulis, E. (2017). Musical and linguistic listening modes in the speech-to-song illusion bias timing perception and absolute pitch memory. Journal of the Acoustical Society of America, 142, 3593.

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493-498.

Grube, M., & Griffiths, T. (2009). Metricality-enhanced temporal encoding and the subjective perception of rhythmic sequences. *Cortex*, 45, 72-79.

Hall, D., & Hess, J. (1984). Perception of musical interval tuning. *Music Perception,* 2, 166-195.

Jaisin, K., Suphanchaimat, R., Candia, M., & Warren, J. (2016). The speech-to-song illusion is reduced in speakers of tonal (vs. non-tonal) languages. *Frontiers in Psychology, 7*, 662.

Kaminska, Z., & Mayer, P. (2002). Changing words and changing sounds: a change of tune for verbal transformation theory? *European Journal of Cognitive Psychology*, 14, 315-333.

Kounios, J., Kotz, S., & Holcomb, P. (2000). On the locus of the semantic satiation effect: evidence from event-related brain potentials. *Memory and Cognition*, 28, 1366-1377.

Krumhansl, C. (1990). *Cognitive Foundations of Musical Pitch*. New York: Oxford University Press.

Kuhl, B., & Anderson, M. (2011). More is not always better: paradoxical effects of repetition on semantic accessibility. *Psychon Bull Rev*, 18, 964-972.

Kuznetsova, A., Brockhoff, P.B. & Christensen, R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13), pp. 1–26.

Ladd, R. (2008). *Intonational Phonology.* New York, NY: Cambridge University Press.

Large, E. (2008). Resonating to musical rhythm: theory and experiment. In *The Psychology of Time*. S. Grondin, Ed.: 189-231. Emerald. United Kingdom.

Lindblom, B., & Sundberg, J. (2007). The human voice in speech and singing. In *Springer Handbook of Acoustics* (pp. 669-712). Springer New York.

MacKay, D., Wulf, G., Yin, C., & Abrams, L. (1993). Relations between word perception and production: new theory and data on the verbal transformation effect. *Journal of Memory and Language*, 32, 624-646.

Margulis, E. (2013a). *On Repeat: How Music Plays the Mind*. New York, NY: Oxford University Press.

Margulis, E. (2013b). Aesthetic responses to repetition in unfamiliar music. *Empirical Studies of the Arts, 31*, 45-57.

Margulis, E., Simchy-Gross, R., & Black, J. (2015). Pronunciation difficulty, temporal regularity, and the speech-to-song illusion. *Frontiers in Psychology, 6*, 48.

Margulis, E., & Simchy-Gross, R. (2016). Repetition enhances the musicality of randomly generated tone sequences. *Music Perception, 33*, 509-514.

Merrill, J., Sammler, D., Bangert, M., Goldhahn, D., Lohmann, G., Turner, R., & Friederici, A. (2012). Perception of words and pitch patterns in song and speech. *Frontiers in Psychology, 3,* 76.

Mertens, P. (2004). The Prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In Proceedings of Speech Prosody 2004, Nara, Japan, pp. 23–26.

Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centers (P-centers). *Psychological Review*, *83*, 405-408.

Natsoulas, T. (1965). A study of the verbal-transformation effect. *The American Journal of Psychology*, 78, 257-263.

Nolan, F. (2003, August). Intonational equivalence: an experimental evaluation of pitch scales. In *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona* (Vol. 39).

Palmer, C., & Krumhansl, C. (1990). Mental representations for musical meter. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 728-741.

Phillips-Silver, J., & Trainor, L. (2005). Feeling the beat: movement influences infant rhythm perception. *Science, 308,* 1430-1430.

Pilotti, M., Antrobus, J., & Duff, M. (1997). The effect of presemantic acoustic adaptation on semantic "satiation". *Memory and Cognition*, 25, 305-312.

Pilotti, M., & Khurshid, A. (2004). Semantic satiation effect in young and older adults. *Perceptual and Motor Skills*, 98, 999-1016.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rossi, M. (1971). Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole. *Phonetica*, 23, 1-33.

RStudio Team. (2014). *RStudio: Integrated Development for R*. Boston, MA. Retrieved from http://www.rstudio.org/

Saitou, T., Goto, M., Unoki, M., & Akagi, M. (2007). Speech-to-singing synthesis: converting speaking voices to singing voices by controlling acoustic features unique to singing voices. In *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 215-218).

Sato, M., Schwartz, J., Abry, C., Cathiard, M., & Loevenbruck, H. (2006). Multistable syllables as enacted percepts: a source of an asymmetric bias in the verbal transformation effect. *Perception and Psychophysics*, 68, 458-474.

Sato, M., Vallée, N., Schwartz, J., & Rousset, I. (2007). A perceptual correlate of the labial-coronal effect. *Journal of Speech, Language, and Hearing Research*, 50, 1466-1480.

Scheirer, E., & Slaney, M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator. In *1997 IEEE International Conference in Acoustics, Speech, and Signal Processing* (pp. 1331-1334).

Schluter, J., & Sonnleitner, R. (2012). Unsupervised feature learning for speech and music detection in radio broadcasts. In *Proceedings of the 15th International Conference on Digital Audio Effects*.

Schultz, B., O'Brien, I., Phillips, N., & McFarland, D. (2016). Speech rates converge in scripted turn-taking conversations. *Applied Psycholinguistics, 37*, 1201-1220.

Simchy-Gross, R, & Margulis, L. (2018). The sound-to-music illusion: repetition can musicalize nonspeech sounds. *Music & Science*, 1, 1-6.

Smith, L. (1984). Semantic satiation affects category membership decision time but not lexical priming. *Memory and Cognition*, 12, 483-488.

Smith, L, & Klein, R. (1990). Evidence for semantic satiation: repeating a category slows subsequent semantic processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 16, 852-861.

t'Hart, J, Collier, R., & Cohen, A. (1990). *A Perceptual Study of Intonation*. Cambridge University Press, Cambridge.

Temperley, D. (2007). *Music and probability*. Cambridge, MA: MIT Press.

Tierney, A., Russo, F., & Patel, A. (2008). Empirical comparisons of pitch patterns in music, speech, and birdsong. *JASA, 123*, 3721.

Tierney, A., Dick, F., Deutsch, D., & Sereno, M. (2013). Speech versus song: multiple pitch-sensitive areas revealed by a naturally occurring musical illusion. *Cerebral Cortex, 23*, 249-254.

Tierney, A., Patel, A., & Breen, M. (in press). Repetition enhances the musicality of speech and tone stimuli to similar degrees. *Music Perception.*

Trainor, L., & Trehub, S. (1992). A comparison of infants' and adults' sensitivity to western musical structure. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 394-402.

Trehub, S., Bull, D., & Thorpe, L. (1984). Infants' perception of melodies: the role of melodic contour. *Child Development, 55*, 821-830.

Vanden Bosch der Nederlanden, C., Hannon, E., & Snyder, J. (2015a). Everyday musical experience is sufficient to perceive the speech-to-song illusion. *Journal of Experimental Psychology: General, 2*, e43-e49.

Vanden Bosch der Nederlanden, C., Hannon, E., & Snyder, J. (2015b). Finding the music of speech: musical knowledge influences pitch processing in speech. *Cognition, 143*, 135-140.

Von Hippel, P., & Huron, D. (2000). Why do skips precede reversals? The effect of tessitura on melodic structure. *Music Perception, 18*, 59-85.

Warren, R., & Gregory, R. (1958). An auditory analogue of the visual reversible figure. *The American Journal of Psychology*, 71, 612-613.

Warren, R. (1961). Illusory changes of distinct speech upon repetition—the verbal transformation effect. *Brit. J. Psychol.*, 52, 249-258.

Warren, R. (1968). Verbal transformation effect and auditory perceptual mechanisms. *Psychological Bulletin*, 70, 261-270.

Warren, R., Bashford, J., & Gardner, D. (1990). Tweaking the lexicon: organization of vowel sequences into words. *Perception & Psychophysics*, 47, 423-432.

Warren, R., Healy, E., & Chalikia, M. (1996). The vowel-sequence illusion: intrasubject stability and intersubject agreement of syllabic forms. *JASA*, 100, 2452-2461.

Warrier, C., & Zatorre, R. (2002). Influence of tonal context and timbral variation on perception of pitch. *Perception and Psychophysics, 64*, 198-207.

Zentner, M., & Eerola, T. (2010). Rhythmic engagement with music in infancy. *Proceedings of the National Academy of Sciences, 107*, 5768-5773.

Figure 1. Waveform (top), spectrogram (middle) and pitch contour (bottom) for a sample Illusion stimulus ("Somehow I can get", left) and a sample Control stimulus ("Quiet word with you, Bradstreet", right). In the bottom plots syllable onsets are marked by vertical lines.

Figure 2. (Top) Mean song ratings for eight repetitions averaged across participants for an example Illusion (solid line) and Control (dotted line) stimulus. The shaded region indicates standard error of the mean. (Bottom) Mean song ratings for eight repetitions averaged across participants for all Illusion (solid line) and Control (dotted line) stimuli.

Figure 3. (Left) Initial and final song ratings for Illusion (black) and Control (grey) stimuli. The identity line is plotted in light grey. (Right) Stem plot displaying values for the change in song rating with repetition sorted by the size of this difference, with Illusion stimuli plotted in black and Controls stimuli plotted in grey.

Figure 4. Histogram of song ratings of each phrase after the final repetition. Histogram bins are 2 points wide, with centers from 0 to 10. Phrases are arranged according to mean song rating, such that Stimulus 1 had the lowest mean song rating and stimulus 48 had the highest.

Figure 5. Relationship between the predicted and actual changes in song rating across all stimuli.

Figure 6. (Top) Pitch contour of an example Illusion stimulus after the steep pitch contour and flat pitch contour manipulations. (Bottom) Pitch contour of an example Control stimulus after the steep pitch contour and flat pitch contour manipulations. The stimuli are the same example stimuli that were displayed in Figure 1.

Figure 7. Black lines display ratings of stimuli with flattened within-syllable pitch slopes. Grey lines display ratings of stimuli with expanded within-syllable pitch slopes. Error bars indicate standard error of the mean.

Figure 8. (Top) Pitch contour of an example Illusion stimulus after the strong musical structure and weak musical structure manipulations. The pitch contour of the first syllable of the Illusion stimulus was identical for the strong musical structure and weak musical structure manipulations. (Bottom) Pitch contour of an example Control stimulus after the strong musical structure and weak musical structure manipulations. The stimuli are the same example stimuli that were displayed in Figure 1.

Figure 9. Black lines display ratings of stimuli with strong musical structure. Grey lines display ratings of stimuli with weak musical structure. Error bars indicate standard error of the mean.

Figure 10. Waveform and spectrograms of example Illusion and Control stimuli in isochronous and variable timing conditions. The stimuli are the same example stimuli that were displayed in Figure 1.

Figure 11. Black lines display ratings of stimuli with isochronous timing. Grey lines display ratings of stimuli with variable timing. Error bars indicate standard error of the mean.

	Rating				
	B	*std. Error*	*t-value*	*df*	*p-value*
Fixed Effects					
(Intercept)	3.84	0.26	14.95	72.81	<.001
Repetition	0.11	0.02	7.36	43.94	<.001
Stimulus Set	1.97	0.38	5.24	80.93	<.001
Rep:StimSet	0.20	0.02	12.86	46	<.001

Table 1. Model parameters for linear mixed effects models comparing effects of Repetition and Stimulus Set for Experiment 1.

	Initial song rating	Rating change
Within-syllable pitch slope	**-0.40 (-0.13, 0.62)**	**-0.64 (-0.44, -0.78)**
Beat variability	**-0.42 (-0.64, -0.13)**	**-0.39 (-0.62, -0.09)**
Melodic structure	0.28(-0.01, 0.52)	**0.50 (0.25, 0.69)**
Fit to musical key	0.25 (-0.04, 0.50)	**0.28 (0.00, 0.53)**
Pitch interval size	-0.05 (-0.33, 0.23)	**-0.46 (-0.66, -0.20)**

Table 2. Pearson's correlations and 95% confidence intervals, relating song ratings to stimulus characteristics. Significant correlations at $p < 0.05$ are indicated with boldface.

	Rating				
	B	*std. Error*	*t-value*	*df*	*p-value*
Fixed Effects					
(Intercept)	3.52	0.18	20.12	117	< 0.001
Repetition	0.12	0.01	9.02	78	< 0.001
Stimulus Set	1.26	0.27	4.58	72	< 0.001
Pitch Slope	0.03	0.28	0.10	110	0.92
Rep:StimSet	0.11	0.01	11.93	30390	< 0.001
StimSet:PitchSlope	1.36	0.36	3.73	114	< 0.001
Rep:PitchSlope	0.02	0.03	0.92	78	0.36
Rep:StimSet:PitchSlope	0.11	0.02	5.73	30390	< 0.001

Table 3. Model parameters for linear mixed effects models examining effects of Repetition, Stimulus Set, and within-syllable pitch slope.

					Rating
	B	*std. Error*	*t-value*	*df*	*p-value*
Fixed Effects					
(Intercept)	3.52	0.19	18.42	90	< 0.001
Repetition	0.10	0.01	7.98	78	< 0.001

Stimulus Set	1.88	0.34	5.51	65	< 0.001
Musical Structure	0.31	0.26	1.21	117	0.23
Rep:StimSet	0.13	0.01	14.50	30390	< 0.001
Rep:MelodicStruct	-0.01	0.02	-0.38	78	0.71
StimSet: MelStruct	0.36	0.39	0.93	116	0.35
Re:StimSet:MelStruct	0.00	0.02	0.28	30390	0.78

Table 4. Model parameters for linear mixed effects models examining effects of Repetition, Stimulus Set, and melodic structure.

	Rating				
	B	*std. Error*	*t-value*	*df*	*p-value*
Fixed Effects					
(Intercept)	4.00	0.19	21.31	96	< 0.001
Repetition	0.12	0.01	8.87	86	< 0.001
Stimulus Set	2.31	0.33	6.99	66	< 0.001
Beat Variability	-0.14	0.30	-0.46	120	0.65
Rep:StimSet	0.15	0.01	12.21	46	< 0.001
Rep:BeatVar	0.03	0.03	1.03	78	0.31
StimSet:BeatVar	0.25	0.49	0.51	88	0.61
Rep:StimSet:BeatVar	0.03	0.02	1.51	30340	0.13

Table 5. Model parameters for linear mixed effects models examining effects of Repetition, Stimulus Set, and beat variability.