Title: Individual Differences in Perception of the Speech-to-Song Illusion are Linked to Musical Aptitude but not Musical Training

Authors: Adam Tierney[1†], Aniruddh D. Patel[2,3], Kyle Jasmin[1,] 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

Word count: 10346

Abstract

In the speech-to-song illusion certain spoken phrases are perceived as sung after repetition. One possible explanation for this increase in musicality is that, as phrases are repeated, lexical activation dies off, enabling listeners to focus on the melodic and rhythmic characteristics of stimuli and assess them for the presence of musical structure. Here we tested the idea that perception of the illusion requires implicit assessment of melodic and rhythmic structure by presenting individuals with phrases that tend to be perceived as song when repeated, as well as phrases that tend to continue to be perceived as speech when repeated, measuring the strength of the illusion as the rating difference between these two stimulus categories after repetition. Illusion strength varied widely and stably between listeners, with large individual differences and high split-half reliability, suggesting that not all listeners are equally able to detect musical structure in speech. Although variability in illusion strength was unrelated to degree of musical training, participants who perceived the illusion more strongly were proficient in several musical skills, including beat perception, tonality perception, and selective attention to pitch. These findings support models of the speech-to-song illusion in which experience of the illusion is based on detection of musical characteristics latent in spoken phrases.

Public Significance Statement

In the speech-to-song illusion certain spoken phrases begin, after repeated presentation, to sound as if they were sung. People vary widely in how vividly they experience this illusion, but the sources of this variability are still poorly understood. We find that people show stable individual differences in the perception of the illusion and that this variability is largely unrelated to factors such as degree of musical training and language background. Instead, we find that individuals who vividly perceive the illusion have high proficiency in music perception skills related to timing and pitch perception. These findings contribute to understanding the relations between language and music processing in the human mind.

1. Introduction

1.1. Domain-level classification

One of the most common perceptual tasks accomplished by individuals is categorization, in which complex input varying continuously in many dimensions is binned into one of several discrete categories. In speech perception, for example, word recognition involves accurate speech sound categorization, while in music perception the recognition of meter involves perceiving strong vs. weak beats (Patel, 2008). For decades researchers have investigated the cues that listeners use to distinguish between categories in speech (Toscano & McMurray, 2010) and music (Prince, 2014). However, comparatively little is known about the processes by which listeners decide to which domain a stimulus belongs—to speech or to music, for example. Prior research on domain-level categorization (i.e. speech, music, or environmental sounds) has largely focused on classification of sounds that differ in the physical source used to produce the sound (for example, tones produced by musical instruments, vowels produced by the human vocal tract, and environmental sounds produced by a variety of sources). This research has found, for example, that listeners can use a variety of acoustic cues to distinguish between sounds from these three categories within tens of milliseconds, as reflected in both behavioral responses and magnetoencephalographic patterns (Bigand, Delbé, Gérard, & Tillmann, 2011; Ogg, Slevc, & Idsardi, 2017; Ogg, Carlson, & Slevc, 2020). However, sounds from different domains do not always differ in their physical source, as in the case of speech and song, both of which are produced by the human vocal tract. Another way to investigate domain-level classification is to make use of ambiguous spoken stimuli which can be heard as spoken or sung, depending on context (Deutsch, Honthorn, & Lapidis, 2011). For these ambiguous cross-domain stimuli, domain-level categorization can be a much more protracted process lasting tens of seconds rather than tens of milliseconds, suggesting that listeners' categorization may be influenced by acoustic patterns on a longer time scale.

1.2. Sound-to-music illusions

Music often has acoustic patterns which distinguish it from speech or other sounds (Ding et al., 2017), and also elicits distinct neural responses compared to other sounds (Norman-Haignere et al. 2015; Zuk et al., 2020). Nevertheless, studies over the past decade have demonstrated that listeners sometimes report perceiving non-musical sounds (i.e., sounds not originally intended to be heard as music) as sounding like music. An important phenomenon in this regard is the speech-to-song illusion. In this illusion, first published in recorded form in 2003 (Deutsch 2003) and in the academic literature ten years ago (Deutsch et al., 2011), a phrase which is perceived as spoken when presented once tends to be heard as sung after repetition. (The key phrase occurs in an even earlier recording from 1995 (Deutsch 1995), but it was not isolated and presented as an illusion until 2003.) Building on this work, Simchy-Gross & Margulis (2018) and Rowland, Kasdan, & Poeppel (2019) demonstrated that repeated environmental sounds are rated as more musical after repetition, and Tierney, Patel, & Breen (2018a) found that pitch contours extracted from speech and resynthesized as complex tones increased in perceived musicality with repetition. However, not all non-musical stimuli sound musical after repetition, with some undergoing dramatic perceptual changes and others continuing to be heard as non-musical (Tierney, Patel, & Breen, 2018b). This raises the key question of why listeners perceive certain non-musical stimuli as music when they are played repeatedly.

1.3. Proposed mechanisms of the speech-to-song illusion

To explain the speech-to-song illusion (henceforth, "song illusion"), Deutsch et al. (2011) suggested that pitch salience is inhibited by default during speech perception, given that pitch plays a secondary role in communicating lexical information relative to other acoustic dimensions (in non-tonal languages). Based on this hypothesis, they predicted that an area anterolateral to primary auditory cortex which had previously been linked to pitch salience (Patterson, Uppenkamp, Johnsrude, & Griffiths, 2002; Penagos, Melcher, & Oxenham, 2004; Schneider et al., 2005) would show greater activation for stimuli which perceptually transform into song when repeated. This prediction was confirmed in an fMRI study by Tierney, Dick, Deutsch, & Sereno (2013), suggesting

that perception of the illusion is indeed linked to increased pitch salience. Deutsch et al. (2011) also hypothesized that perception of speech as song involves perceptual distortion of pitch, such that the perceived pitches of syllables are warped to fit a musical scale template. This hypothesis was supported by Vanden Bosch der Nederlanden, Hannon, & Snyder (2015a), who found that when participants perceived the song illusion in repeated spoken phrases, pitch changes which departed from the perceived melodic template were easier to detect than pitch changes which moved towards it.

The explanation for the song illusion advanced by Deutsch et al. (2011), however, does not specify why pitch perception should be disinhibited by stimulus repetition. A potential explanation was advanced by Castro, Mendoza, Tampke, & Vitevich (2018), who proposed that Node Structure Theory (MacKay, Wulf, Yin, & Abrams, 1993) could account for the song illusion. According to this theory, repeated activation of lexical nodes temporarily reduces their ability to become activated. Castro et al. (2018) suggest that while deactivation of lexical nodes diminishes the perception of speech, continued activation of syllable nodes leads to the emergence of a song percept. Supporting a role for lexical deactivation in the illusion, they find that words from denser phonological neighborhoods and words from an unfamiliar language are rated as more song-like after repetition. Similarly, Vitevitch, Ng, Hatley, & Castro (2020) report that words with a high phonological clustering coefficient (which measures the tendency for phonological neighbors of a word to also be neighbors of one another) are perceived as more song-like after repetition.

The Node Structure Theory explanation for the song illusion, however, struggles to explain why deactivation of lexical nodes along with continued activation of syllable nodes should lead to perception of song, rather than speech leeched of semantic content (as in semantic satiation; Smith, 1984). The explanation advanced in Castro et al. (2018) is that since syllables are the unit of rhythmic structure in speech, and rhythm is also an important aspect of music, then syllable node activation without lexical node activation should lead to a song-like percept. According to this explanation, all spoken stimuli should transform into song when repeated, as lexical nodes are satiated and syllable

nodes continue to be active. Moreover, according to this explanation, the extent to which stimuli do or do not give rise to the illusion should be primarily tied to phonological rather than acoustic or musical characteristics. Finally, this theory cannot explain increases in musicality with repetition of non-verbal stimuli.

These predictions, however, are not borne out by the literature. First, there is ample evidence that not all spoken phrases transform into song when repeated. Tierney et al. (2013), for example, assembled a corpus of "illusion" phrases, which listeners report transform into song when repeated, and "control" phrases, which continue to be perceived as speech when repeated. (These two types of phrases were matched for speakers, syllable rate, and duration.) The relative lack of transformation in the control phrases was replicated in Graber, Simchy-Gross, & Margulis (2017), Tierney et al. (2018a), and Tierney et al. (2018b). Second, several studies have demonstrated that a range of acoustic and musical stimulus characteristics modulate the strength of the illusion. Tierney et al. (2013) showed that the illusion stimuli featured flatter pitch contours than the control stimuli. Falk, Rathcke, & Dalla Bella (2014) manipulated the stability of pitch contours within syllables and found that song percepts were more common for the more stable contours. Tierney et al. (2018b) found that higher song ratings for song illusion stimuli were linked to more isochronous rhythmic structure and a greater tendency for syllable pitches to follow musical melodic statistics. Rathcke, Falk, & Dalla Bella (in press) found that recordings with greater sonority led to earlier, stronger, and more frequent transformations to song. Third, Tierney et al. (2018a) found that eliminating linguistic content from the illusion/control stimulus sets (Tierney et al., 2013) preserves the relative differences across these stimuli in the extent to which they give rise to the sound-to-music illusion, suggesting that the primary factors driving cross-stimulus differences in illusion strength cannot be phonological. Fourth, several studies have demonstrated that environmental sounds and non-verbal complex tone sequences increase in musicality after repetition (Simchy-Gross & Margulis, 2018; Tierney et al., 2018a; Rowland et al., 2019), a finding that cannot be explained by lexical processes.

Thus Node Structure Theory alone is not a sufficient explanation of the song illusion. Given this observation, what theoretical framework could account for both the influence of phonological neighborhood and the influence of acoustic and musical characteristics on the perception of song in speech? One possibility is a hybrid of the accounts advanced in Deutsch et al. (2011) and Castro et al. (2018) in which deactivation of lexical nodes is *necessary but not sufficient* for perception of the illusion. According to this account, deactivation of lexical nodes frees up attention which can then be focused on acoustic characteristics of the stimuli, including pitch and timing patterns. The degree to which a stimulus transforms into song after lexical node deactivation, therefore, will depend on the melodic and rhythmic characteristics of the stimulus. One implication of this hybrid model is that there may exist stable individual differences across participants in the extent to which the vividness of the illusion varies across stimuli, and these individual differences may relate to musical training and musical aptitude. In two experiments we tested this implication by presenting participants with repeated phrases drawn from one of two stimulus sets, a stimulus set which listeners consistently report strongly transforms into song after repetition and a stimulus set which transforms to a much lesser degree ("Illusion" and “Control” stimuli, as documented in Tierney et al., 2013; Graber et al., 2017; Tierney et al., 2018b). We predicted that participants would reliably vary in the difference in perceived musicality between Illusion and Control stimuli after repetition, and that this variability would relate to musical training (Experiment 1) and musical aptitude (Experiment 2).

2. Experiment 1

2.1. Introduction

Musical training is one potential factor which could underly individual differences in the song illusion: if perception of the illusion requires detection of musical characteristics latent in non-musical stimuli, then musicians may be better able to detect these characteristics. Yet prior research has reported that musicians are somewhat less susceptible to auditory illusions in which stimulus details are lost in a top-down percept or gestalt (Craig, 1979; Davidson, Power, & Michie, 1987; Brennan & Stevens, 2002; Pressnitzer, Graves, Chambers, de Gardelle, & Egré, 2018), perhaps

because musicians are better able to attend to low-level acoustic characteristics of stimuli. The existing evidence regarding musical training and the speech-to-song illusion is inconsistent. Falk, Rathcke, & Dalla Bella (2014) found that participants with more years of musical training were no more likely to perceive the illusion, and in fact that their perception of the illusion was delayed relative to musically untrained participants. However, Vanden Bosch der Nederlanden et al. (2015b) found that musicians reported hearing all stimulus presentations as more musical than non-musicians, regardless of stimulus repetition or transposition. Given that Vanden Bosch der Nederlanden et al. (2015b) used only the original example from Deutsch et al. (2011) as a stimulus, this finding could indicate either that musicians are better able to detect the latent musical characteristics of this stimulus or that musicians are biased to perceive musicality regardless of stimulus characteristics.

In addition to our primary research question regarding musical training, we also investigated several other possible predictors of individual differences in perception of the illusion. For example, another possible factor underlying variability in the song illusion is age. As described above, Castro et al. (2018) speculated that one important precondition for perception of the illusion is deactivation of lexical nodes due to satiation. A similar explanation has also been advanced for the verbal transformation effect, in which repetition of a word eventually leads to an unstable percept that swaps between different words with different meanings (Warren & Gregory, 1958): MacKay et al. (1993) suggested that this illusion occurs when lexical nodes are satiated but phonological nodes remain activated. The verbal transformation effect is weaker in older participants (Warren & Warren, 1966; Pilotti & Khurshid, 2004; Pilotti, Simcox, Baldy, & Schauss, 2011), suggesting that satiation of lexical nodes due to repetition may be less rapid or less robust in older age. If so, then older participants may demonstrate less robust perception of the speech-to-song illusion as well. However, Mullin, Norkey, Kodwani, Vitevitch, & Castro (2021) reported no relationship between age and the strength of the speech-to-song effect, as perceived in the stimulus from Deutsch et al. (2011).

A third possible factor underlying individual differences in the song illusion is gender[1]. There is some limited prior evidence for sex/gender differences in the perception of auditory illusions. For example, Irwin, Whalen, & Fowler (2006) reported that female participants perceived the McGurk illusion more strongly (i.e. greater influence of visual stimuli on auditory perception), but only when the visual stimuli were relatively brief. Women were also found to be more likely to report the illusory percept (i.e. "yanny") after hearing the Yanny/Laurel stimuli (Pressnitzer et al. 2018), while Gwilliams & Wallisch (2020) similarly found more reports of "yanny" among female participants. Given this prior evidence for sex/gender differences in perception of auditory illusions, as well as evidence for sex differences in subcortical auditory processing (Krizman, Bonacina, & Kraus, 2019), we compared perception of the illusion between participants identifying as male versus as female.

Finally, there is some prior evidence that language background can affect perception of the speech-to-song illusion. Jaisin, Suphanchaimat, Candia, & Warren (2016) reported that native speakers of tonal languages did not perceive the speech-to-song illusion, although given the small sample sizes used (10 participants in each group) this finding needs replication. Castro et al. (2018) reported that repeated speech in an unfamiliar language was perceived as more musical than repeated speech in a familiar language. Rathcke et al. (2021) reported that bilingual participants reported more speech-to-song transformations in their second language. Finally, Margulis, Simchy-Gross, & Black (2015) reported that repeated speech in an unfamiliar language was perceived as particularly musical for languages that would be difficult for a participant to pronounce. Given these somewhat mixed prior findings, we investigated effects of language experience by determining both whether participants spoke English as their dominant language as well as whether participants were monolingual (only able to speak one language) or bilingual (able to speak at least two languages).

2.2. Methods

[1] Sex and gender are highly correlated but dissociable. In the current study we focused on gender, and will use that term throughout, but we describe prior results using the language that was used by the study authors.

227 2.2.1. Participants

228 One hundred seventy-one (171) participants were initially recruited from a mixture of
229 undergraduate psychology students and Mechanical Turk. Twenty-three (23) participants were
230 removed from the dataset because they failed catch trials, not reporting an increase in song
231 perception after stimuli switched from actual speech to actual song (see below for details), or
232 because they failed to produce at least one response after the first through fourth repetitions and
233 after the fifth through eighth repetitions during the speech-to-song illusion rating experiment.
234 Responses from 148 remaining participants were analyzed. A series of power analyses were
235 conducted in GPower to determine the smallest effect sizes which could be detected at a power of
236 0.8, given this sample size and an alpha of 0.05. This sample size resulted in a power of 0.8 to detect
237 a correlation of r = 0.23. For the analysis of the speech-to-song ratings, given a 2 (stimulus set) X 8
238 (repetition) RMANOVA design, the sample size resulted in a power of 0.80 to detect main effects and
239 interactions of F = 0.078. Ninety-three (93) participants identified as male, while 55 identified as
240 female. (Participants were also given the additional options "Other" and "Prefer not to say", but no
241 participants selected these options.) Participants reported a mean age of 29.89 (SD 10.14) years old
242 (range 19-61). 120 participants reported that their dominant language was English, while 28
243 participants reported an alternate dominant language. These other dominant languages were Arabic
244 (n = 1), Bulgarian (n = 1), French (n = 1), German (n = 1), Greek (n = 2), Italian (n = 6), Polish (n = 3),
245 Portuguese (n = 3), Romanian (n = 2), Russian (n = 2), Slovak (n = 2), Spanish (n = 3), and Swedish (n =
246 1). Sixty-seven (67) participants reported being bilingual (i.e. able to speak at least two languages),
247 while 81 participants reported being monolingual. Participants reported a mean of 3.19 (SD 4.24)
248 years of musical training (range 0-17). Study procedures were approved by the ethics board of the
249 Department of Psychological Sciences at Birkbeck College.

250 2.2.2. Speech-to-song illusion experiment

251 Participants were presented with 48 spoken phrases drawn from the stimulus set first
252 reported in Tierney et al. (2013). Phrases were produced by a mix of American and British English

speakers, but illusion and control datasets were perfectly matched for speakers. Prior studies (Tierney et al., 2013; 2018a; 2018b; Graber et al., 2017) have indicated that 24 of these stimuli tended to be heard as song after repetition (henceforth "illusion" stimuli), while 24 tended to be heard as speech after repetition (henceforth "control" stimuli). The mean length of stimuli was 6.6 (SD 1.5) syllables and 1.38 (SD 0.41) seconds. Stimuli were spoken by three different male talkers, represented in equal portions among illusion and control stimuli. These two stimulus categories did not differ significantly in syllable rate (illusion stimuli 5.05 syllables/second, control stimuli 4.99 syllables/second, $t(46) = 0.16$, $p = 0.877$) or duration (illusion stimuli 1.32 seconds, control stimuli 1.44 seconds, $t(46) = -1.05$, $p = 0.301$).

In each trial, eight repetitions of each phrase were presented. After each repetition, participants pressed buttons labelled 1 through 10 to indicate whether the phrase sounded like speech or song, where 1 indicated completely speech-like, while 10 indicated completely song-like. Participants were given 2 seconds to respond to each phrase, after which time the program automatically went on to the next repetition. Any missing rating (due to a non-response) was replaced by the mean of the previous and following ratings. Missing first/last repetitions were replaced by the second/seventh repetitions.

In addition, four catch trials were presented. In each catch trial, a spoken phrase was presented during the first four repetitions, while a matched sung phrase (with the same words, roughly same rate, and similar pitch contour) was presented during the second four repetitions. Catch trials were produced by the first author. Analysis was limited to data from those participants whose average rating of the last four repetitions exceeded that of the first four (by any amount) on these catch trials. As noted above, the exclusion criteria resulted in the removal of 23 out of 171 participants from the dataset, leaving 148 for analysis. The experiment lasted around 20 minutes.

To establish whether an overall speech-to-song illusion effect was present, a repeated measures ANOVA was conducted with two within-subjects factors, repetition (eight levels) and stimulus set (illusion versus control). Next, metrics summarizing illusion perception were compared

to demographic characteristics. With eight ratings for each illusion and control stimulus, this is a high-dimensional dataset, potentially exacerbating the multiple comparisons problem. To reduce the dimensionality of the data, therefore, we summarized each participant's responses in two ways for the purpose of investigating individual differences in perception of the illusion. First, the difference between the ratings of illusion and control stimuli after the eighth repetition was calculated and will be referred to as *illusion strength*. (In other words, for each participant we averaged their final rating for all illusion stimuli and all control stimuli, and then took the difference between these two mean values to compute illusion strength.) Second, the average rating across all repetitions and all stimuli was calculated and will be referred to as *musical prior*.

Any variables not normally distributed (Jarque & Bera, 1980) were transformed prior to statistical analysis. A square root transformation was used for musical prior, and age and years of musical training were converted to ranks. Pearson's correlations were used to assess the strength of the relationship between musical prior and illusion strength versus age and years of musical training. Musical prior and illusion strength were also compared between dominant English speakers and non-dominant English speakers, as well as between bilinguals and monolinguals and between male and female participants, using unpaired t tests. Analyses predicting illusion strength and musical prior were corrected separately for multiple comparisons using false discovery rate (Benjamini & Hochberg, 1995).

2.2.3. Available materials

Data from both Study 1 and Study 2 and test materials from Study 2 are available at https://osf.io/hegxs/. All stimuli from the speech-to-song illusion corpus are available at https://osf.io/t4pjq/.

2.3. Results

Overall, across participants the illusion was reliably experienced, as seen in the strong interaction between repetition and stimulus set ($F(7,1029) = 216.3$, $p < 0.001$). There was, in addition, a main effect of repetition ($F(1,1029) = 146.8$, $p < 0.001$) and of stimulus set ($F(1,137) =$

426.4, $p < 0.001$). Overall, participants rated that the illusion stimuli sounded much more musical than the control stimuli after repetition; in fact, after averaging across participants, there was no overlap in the ratings between the two stimulus sets (in other words, no control stimulus was rated higher than any illusion stimulus). However, there were very large individual differences across participants in the strength of the illusion, ranging from participants who reported no difference between illusion and control stimuli whatsoever, to participants who reported that the illusion stimuli sounded exactly like song when repeated while the control stimuli continued to sound exactly like speech. These individual differences are displayed in Figure 1, which plots average ratings of illusion and control stimuli across all eight repetitions from all 148 participants, sorted so that participants with larger ratings of illusion stimuli after the eighth repetition are closer to the top. (A color plot version of the same figure is presented in the supplementary information, see Figure S1. Additional plots showing individual differences for individual illusion and control stimuli are also available in supplementary information, see Figures S2 and S3. Also see the supplementary information for sound examples of the illusion stimuli and control stimuli with the least and most variability in rating across participants, see Sound Examples S1 through S4.) In addition, Figure 2 displays the difference between the average rating of the last repetition and the average ratings of the first repetition in illusion and control stimuli, as well as the difference in average rating between illusion and control stimuli after the first repetition and the last repetition. The difference between the average last and first repetition rating for illusion stimuli ranged from -0.17 to 8 (mean 1.91, std 1.45), while this difference for control stimuli ranged from -1.17 to 4.92 (mean 0.38, std 0.90). The difference in average rating between illusion and control stimuli after the first repetition ranged from -0.83 to 4.75 (mean 1.33, std 1.04), while this difference after the last repetition ranged from -0.25 to 7.75 (mean 2.85, std 1.61).

To assess the reliability of this speech-to-song testing battery as a measure of individual differences in musicality perception, a Monte Carlo procedure was used. Across 100 iterations, the illusion and control stimuli were randomly divided into two halves. The average rating after the

eighth repetition was then calculated separately for the illusion and control stimuli. The difference between ratings of the illusion and control stimuli was then calculated separately for each half of the trials (the *illusion strength* measure described above). The estimates from the two halves were correlated using a Pearson's correlation. The resulting correlation was corrected using the Spearman-Brown correction, to account for the fact that reliability would be expected to be slightly higher for a test with twice the number of trials. Finally, the median split half reliability across the 100 iterations was calculated. This procedure revealed a reliability of 0.89, suggesting that this testing battery provides a highly reliable measure of individual differences in the perception of musicality in spoken stimuli. A similar procedure revealed that reliability for the *musical prior* measure was also high, reaching 0.96. Interestingly, however, illusion strength and musical prior only weakly correlated (r(146) = 0.19, p = 0.023), suggesting that these two measures, although highly reliable, index dissociable aspects of the speech-to-song illusion.

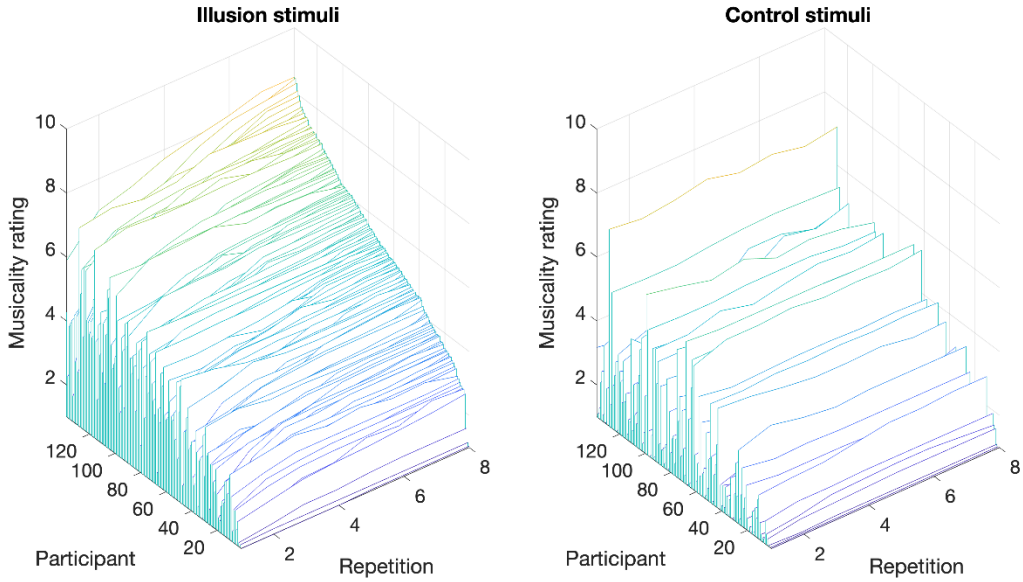

Figure 1: Waterfall plot displaying ratings of illusion and control stimuli across all 148 participants, sorted by the final rating of illusion stimuli. Each row shows the averaged ratings of a participant across all 24 stimuli in that category. The illusion stimuli matrix is sorted from highest to lowest final

rating, and the control stimuli matrix shows each participant's corresponding average rating profile for control stimuli.

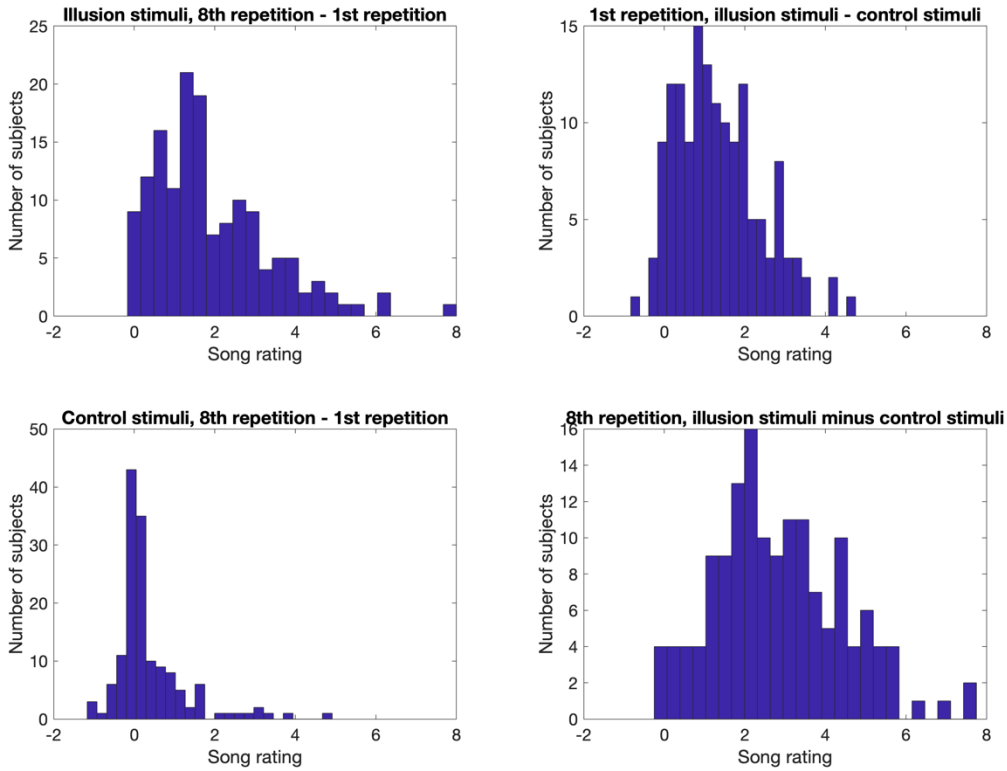

Figure 2: Top-left: histogram displaying the difference between the average rating after the last repetition minus the average rating after the first repetition for illusion stimuli. Bottom-left: histogram displaying similar data for control stimuli. Top-right: histogram displaying the difference between the average initial rating for illusion stimuli minus the average initial rating for control stimuli. Bottom-right: histogram displaying similar data for final ratings of illusion and control stimuli.

Years of musical training did not correlate with illusion strength ($r(146) = 0.10$, p(corrected) = 0.531). However, years of musical training did correlate with musical prior ($r(146) = 0.25$, p(corrected) = 0.013), such that participants with a greater degree of musical training rated *all* stimuli (both illusion and control) as more song-like across all repetitions. Figure 3 displays

scatterplots relating degree of musical training to both illusion strength and musical prior, as well as average responses to each repetition of illusion and control stimuli in non-musicians and musicians. (For display purposes musicians were defined as having at least six years of training, following Zhang et al. (2020), while non-musicians were defined as having no years of training.)

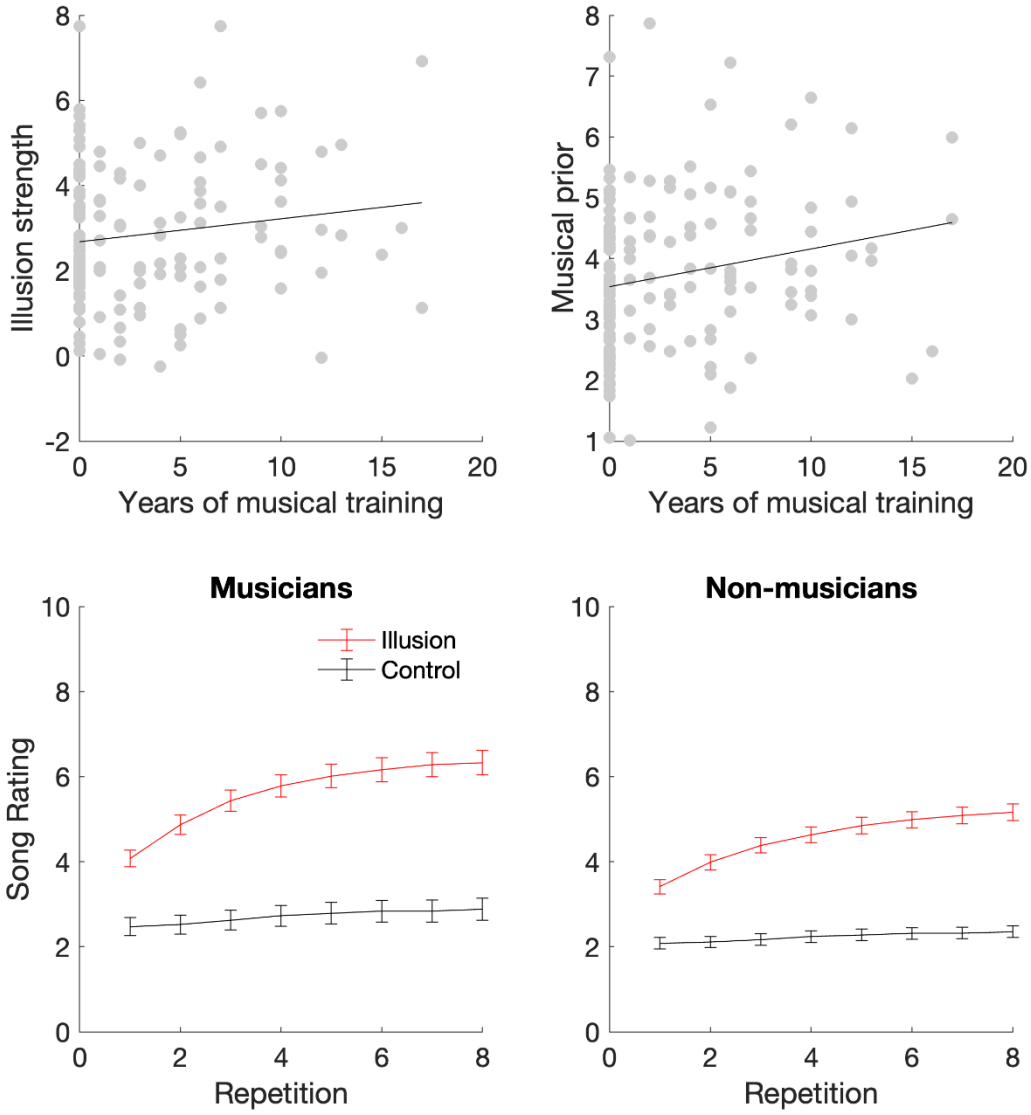

Figure 3: Top-left: scatterplot displaying years of musical training versus illusion strength. Illusion strength was calculated as the rating difference between the illusion and control stimulus sets after the eighth repetition. Top-right: scatterplot displaying years of musical training versus musical prior.

Musical prior was calculated as the grand average of ratings across all experimental trials over both stimulus sets. Bottom-left: average responses to illusion and control stimuli in musicians (n = 35). Bottom-right: average responses to illusion and control stimuli in non-musicians (n = 64). Error bars indicate standard error of the mean.

There was no correlation between age and illusion strength (r(146) = -0.08, p(corrected) = 0.589). There was similarly no correlation between age and musical prior (r(146) = -0.02, p(corrected) = 0.823. There was no difference in illusion strength (t(146) = 2.19, p(corrected) = 0.152) between participants who identified as female (mean = 3.07, std = 1.54) versus participants who identified as male (mean = 2.48, std = 1.68). Similarly, there was no difference in musical prior (t(146) = 1.25, p(corrected) = 0.533) between participants who identified as female (mean = 3.83, std = 1.25) compared to participants who identified as male (mean = 3.59, std = 1.28). There was no difference in illusion strength (t(146) = 0.09, p(corrected) = 0.927) between bilinguals (mean = 2.87, SD = 1.69) versus monolinguals (mean = 2.84, SD = 1.56). Similarly, there was no difference in musical prior (t(146) = 0.235, p(corrected) = 0.823) between bilinguals (mean = 3.77, SD = 1.30) versus monolinguals (mean = 3.71, SD = 1.24). There was no difference in illusion strength (t(146) = -0.57, p(corrected) = 0.710) between participants whose dominant language was English (mean = 2.82, std = 1.68) compared to participants whose dominant language was not English (mean = 3.01, std = 1.31). Similarly, there was no difference in musical prior (t(146) = -0.22, p(corrected) = 0.823) between participants whose dominant language was English (mean = 3.72, std = 1.25) compared to participants whose dominant language was not English (mean = 3.78, std = 1.32).

2.4. Discussion

We find large individual differences in the magnitude of the song illusion across a diverse sample of participants drawn from the general population. There was widespread agreement across participants that the musicality of the control stimuli did not change with repetition. However, there

was a high degree of variability across participants in the extent to which the illusion and control stimuli differed in musicality, and this variability increased as the stimuli were repeated: some participants reported large increases in musicality in the illusion stimuli with repetition, whereas others reported more modest increases.

Importantly, our split-half reliability calculations showed a reliability of 0.88 for our measure of illusion strength, which was calculated as the difference in musicality between the illusion and control stimuli after repetition. This confirms a key prediction of the hybrid model of the speech-to-song illusion (see Section 1.2), which suggests that deactivation of lexical nodes after repetition frees up cognitive resources, enabling participants to extract the melodic and rhythmic characteristics of the phrases and thus assess the presence of musical structure. This model predicts that there will be strong differences across stimuli in the extent to which they give rise to the illusion (due to differences in degree of musical structure), but that the magnitude of the differences across stimuli will vary across participants (due to individual differences in the ability to detect musical structure). The first prediction was confirmed by previous research (Tierney et al., 2013; 2018b; Graber et al., 2017). The second prediction is confirmed here, as we find that there are highly stable individual differences in the disparity in musicality between the illusion and control stimuli. Interestingly, however, we also find that there are stable individual differences in musicality ratings across the entire stimulus set, and across all repetitions (our "musical prior" measure). Nonetheless, although we find both illusion strength and musical prior are reliable measures, they correlated only very weakly. This suggests that an individual's perception of the musicality of a phrase after repetition is determined by at least two main factors: their overall tendency to perceive musicality in sound, and their ability to assess musical structure in complex sounds.

We predicted that musical training would relate to the difference in musicality ratings between illusion and control stimuli after repetition (i.e. to illusion strength). This prediction was based on the idea that individuals with musical training would be better able to extract melodic and rhythmic characteristics from stimuli to assess the presence of musical structure. However, we

found that while degree of musical training did not relate to illusion strength, it did relate to musical prior, such that participants with more musical training were more likely to produce high musicality ratings across all stimuli after all repetitions. This difference can be clearly seen in Figure 3, in which musicians' musicality ratings are higher than those of non-musicians even for control stimuli after the first repetition. This result aligns with and clarifies Vanden Bosch der Nederlanden et al.'s (2015b) finding that when musicians and non-musicians provided ratings of the original Deutsch et al. (2011) speech-to-song stimulus, musicians provided higher ratings overall, i.e. they heard all stimulus presentations more musically regardless of repetition or transposition. One possible explanation for the relationship between musical training and musical prior is that musicians have sufficiently extensive knowledge of a wide variety of musical styles and examples that they can map any random sequence of pitches and durations onto a possible musical framework to some degree.[2] The lack of a relationship between illusion strength and musical training has two possible explanations. One possibility is that perception of the illusion does not require extraction of the melodic and rhythmic characteristics of stimuli. A second possibility is that individual differences in melody and rhythm perception (i.e. music aptitude) exist independently of and prior to engagement in formal musical training (Kragness et al., 2020), possibly driven in part by genetic differences (Niarchou et al., 2021). Experiment 2 was designed to examine the relationship between music aptitude and illusion strength (see below).

Turning to the effects of age, we did not find any differences between younger and older participants in illusion strength. Prior research has suggested that lexical node satiation decreases with age, as can be seen in decreased verbal transformation effects in older participants (Warren & Warren, 1966; Pilotti & Khurshid, 2004; Pilotti et al., 2011). This result suggests that individual differences in the readiness with which lexical nodes satiate is not a primary factor driving differences in perception of the song illusion in these stimuli. However, this conclusion may not

[2] An extreme example of this ability can be seen when jazz accompaniment brings out the musical characteristics of natural speech, as in Henry Hey's orchestration of political speeches and interviews (https://www.youtube.com/watch?v=9nlwwFZdXck).

generalize beyond the current stimulus set, and stimulus sets designed to elicit cross-stimulus differences in lexical satiation (by manipulating, for example, phonological neighborhood density; see Castro et al., 2018) could lead to larger age effects. We also found no differences in either illusion strength or musical prior between participants who identified as male versus female, suggesting that perception of the speech-to-song illusion is relatively unaffected by gender.

In terms of the effect of language background, we found no differences in illusion strength between monolinguals and bilinguals, nor between dominant and non-dominant speakers of English. This result suggests that detecting musical characteristics in speech is not enhanced by exposure to multiple languages, and that lack of proficiency in a spoken language does not necessarily modulate perception of the speech-to-song illusion in that language. Initially this may seem to contradict prior research suggesting that language history can modulate the speech-to-song illusion, with song perception diminished for stimuli drawn from unfamiliar languages (Castro et al., 2018), especially difficult-to-pronounce languages (Margulis et al., 2015), but enhanced in familiar second languages in bilinguals (Rathcke et al. 2021). Our findings, however, do not necessarily contradict these findings, as our participants were all somewhat familiar with English, even when English was not their dominant language (although we do not have detailed information about participants' proficiency, limiting somewhat the conclusions that can be drawn about language experience and perception of the illusion). Moreover, these prior studies were not designed to investigate individual differences in the variability in illusion perception between strongly versus weakly transforming stimuli. Our finding of a null effect of language background, however, should be qualified by the lack of tone language speakers in our sample, given previous evidence that tone language speakers perceive the illusion differently (Jaisin et al. 2016) and that tone language ability correlates with melody perception skills (Swaminathan, Kragness, Schellenberg, 2021). Our stimuli could be useful for studying effects of tone language experience on perception of the song illusion in future research, provided the participants are bilingual in English.

Overall, we find that although our paradigm can reliably measure individual differences in the strength of the speech-to-song illusion, none of the demographic predictors we measured (including musical training) were significantly associated with illusion strength. What factors, then, might drive individual differences in illusion strength? We hypothesized that individual differences in the ability to extract rhythmic and melodic information from speech could drive variability in illusion strength and investigated this possibility in Experiment 2.

3. Experiment 2

3.1. Introduction: investigating musical aptitude and selective attention to pitch as drivers of the song illusion

One factor driving individual differences in perception of the song illusion could be musical aptitude, i.e., musical abilities which vary between individuals independently of any effect of formal musical training (Kragness et al., 2020) and which may partly reflect genetic differences (Wesseldijk, Mosing, & Ullén, 2021; Niarchou et al., 2021). Indeed, prior work has suggested that certain perceptual and cognitive abilities are more strongly tied to musical aptitude than to musical training (Swaminathan, Schellenberg, & Khalil, 2017; Swaminathan & Schellenberg, 2019; Wesseldijk, Gordon, Mosing, & Ullén, 2021). The extent to which speech stimuli sound musical after repetition has been linked to a variety of stimulus characteristics, including the presence of a steady beat and tonal structure (Tierney et al., 2018b) as well as the flatness of pitch within syllables (Tierney et al., 2013; Falk et al., 2014). Detection of these characteristics may be more successful in participants with greater musical aptitude and more precise auditory perception, leading to more robust perception of the speech-to-song illusion. Another possibility is that individual differences in the ability to direct attention to pitch information help determine the strength of the illusion. Prior research has indicated a link between perception of the illusion and increased pitch salience, as shown via increased activation in pitch-sensitive cortical areas (Tierney et al., 2013). That song perception is linked to increased pitch salience is also supported by the finding that imitation of pitch content is enhanced for song relative

to speech stimuli (Pfordresher, Mantell, & Pruitt, 2021). Participants who can more readily direct attention to pitch information in speech when instructed to do so, therefore, may be better able to detect latent musical structure in stimulus pitch patterns and therefore more robustly perceive the illusion.

To investigate these ideas, in Experiment 2 we examined whether illusion strength was linked to performance on six tests of auditory perception: two tests each of musical aptitude, dimension-selective attention, and psychophysics (Table 1). The musical aptitude tests were the Beat Alignment Test (BAT; Iversen & Patel, 2008), which asks participants to judge whether a series of tone pips is aligned with the beat or shifted away from the beat of clips of music, and the Tonality Alignment Test (TAT), a novel test constructed for the current study which asks participants to judge whether a sung melody is aligned with a tonal grid or misaligned (by being either compressed or expanded). The dimension-selective attention tests (first introduced in Jasmin, Sun, & Tierney, 2021) presented participants with pairs of spoken words which varied orthogonally in relative pitch and relative loudness. Participants were either asked to attend to pitch, indicating which of the two words was higher, or attend to loudness, indicating which of the two words was louder. We predicted that performance on the attention-to-pitch test, but not the attention-to-loudness test, would be linked to the robustness of perception of the speech-to-song illusion. The psychoacoustic tests were pitch and amplitude rise time discrimination (using parameters taken from Kachlicka, Saito, & Tierney, 2019), as we reasoned that participants with more accurate auditory perception may be better able to detect subtle cues to musicality such as slight differences in pitch contour flatness or subtle amplitude envelope cues to beat location. (Amplitude rise time is one of the primary cues conveying the exact timing of musical beats (Danielsen et al., 2019)).

Test	Source
Attention to Amplitude	Jasmin et al., 2021
Attention to Pitch	Jasmin et al., 2021
BAT	Iversen & Patel, 2008
TAT	New
Frequency Discrimination	Kachlicka et al., 2019
Rise Time Discrimination	Kachlicka et al., 2019

Table 1. Summary of measures included as possible predictors of illusion strength.

3.2. Methods

3.2.1. Participants

Ninety-five (95) participants were initially recruited from the online recruitment service Prolific. Because catch trials were not included in the speech-to-song illusion perception test in this experiment, participant inclusion was based on performance exceeding at least 55% across both dimension-selective attention tests, the BAT, and the TAT, as well as having at least one response for the first and second half of the repetitions during each trial of the speech-to-song illusion paradigm. One additional participant was excluded due to receiving the lowest possible score on both the frequency and amplitude rise time discrimination tests. Seventy-six (76) participants passed these criteria. Forty-eight of these identified as female, while 28 identified as male. A series of power analyses were conducted in GPower to determine the smallest effect sizes which could be detected at a power of 0.8, given this sample size and an alpha of 0.05. This sample size resulted in a power of 0.8 to detect a correlation of r = 0.31. For the analysis of the speech-to-song ratings, given a 2 (stimulus set) X 8 (repetition) RMANOVA design, the sample size resulted in a power of 0.80 to detect main effects and interactions of F = 0.109. For the regression analysis, the sample size resulted in a power of 0.80 to detect effect sizes of F = 2.23 across six predictors. Participants reported a mean age of 30.7 years (SD 10.5, range 18-67). 74 of the participants reported that their dominant language was English, while 2 participants reported other dominant languages (1 Hindi and 1 Dutch). 62 participants reported being monolingual while 14 participants reported being bilingual (i.e. being able to speak at least two languages). Participants reported a mean of 5.22 years of musical training (std 7.72, range 0-36). Participants were tested using the Gorilla platform for

online testing (Anwyl-Irvine et al., 2020). The entire experiment lasted approximately 45 minutes. All participants completed the tasks in the same order, as follows: attention-to-amplitude, attention-to-pitch, song illusion experiment, TAT, BAT, frequency discrimination, and amplitude rise time discrimination. Participants were allowed to rest between tasks, but these breaks were optional rather than mandated. Study procedures were approved by the ethics board of the Department of Psychological Sciences at Birkbeck College.

3.2.2. Dimension-Selective Attention Tests

The ability to attend to individual acoustic dimensions in speech was assessed with a pair of tests first presented in Jasmin et al. (2021). Stimuli in this test are drawn from a pair of recordings of sentence fragments which are identical lexically but differ in the position of word emphasis: “STUDY music” and “study MUSIC”. These recordings were morphed onto one another using STRAIGHT (Kawahara & Irino, 2005) so that the extent to which individual acoustic dimensions resembled one versus the other recording could be precisely controlled, while all other acoustic characteristics were set to be constant across stimuli. For these tests, a 4 X 4 grid of stimuli was constructed in which pitch and amplitude varied in the extent to which they resembled the pattern in the initial-emphasis recording (“STUDY music”) versus the final-emphasis recording (“study MUSIC”). (Note that phonetic content was fixed across trials, enabling participants to focus on the target dimensions.) Specifically, the pitch levels used were 0%, 33%, 67%, and 100%, where 0% indicates patterns identical to the initial-emphasis recording, 100% indicates patterns identical to the final-emphasis recording, and 33% and 67% indicate intermediate values. Similarly, the amplitude levels used were 0%, 33%, 67%, and 100%. To summarize, a set of 16 stimuli were constructed in which the extent to which pitch patterns versus amplitude patterns implied the existence of initial versus final word emphasis was independently varied (See Figure S4 for a schematic of the stimulus design).

On each trial, participants were presented with a single stimulus. The two tests involved presentation of stimuli drawn from the same set of 16 possible stimuli but differed in the instructions given to participants. For the Attention to Amplitude test, they were asked to press a

button to indicate whether the first or second word was louder, ignoring any differences in pitch. For the Attention to Pitch test, they were asked to press a button to indicate whether the first or second word was higher in pitch, ignoring any differences in amplitude. Performance on each test was summarized as portion correct. For each test 40 trials were presented, with each trial pseudorandomly drawn from the set of 16 possible stimuli (a single randomization was performed for each test and used across all participants). This test has previously been shown to be sensitive to individual differences in dimensional salience: Mandarin speakers were found to display better performance on the Attention to Pitch test and worse performance on the Attention to Amplitude test, compared to native English speakers, suggesting that their experience speaking a tone language led to increased pitch salience across languages (Jasmin et al. 2021). Stimuli from these tests are available at https://osf.io/hegxs/.

3.2.3. Song Illusion Experiment

The song illusion experiment was conducted in a manner almost identical to that of Experiment 1, the only difference being that the catch trials were not included. Instead, the first four trials of the test consisted of 8 repetitions of practice items. These practice items were taken from the same speakers as the illusion/control stimuli but were different recorded phrases. (None of these four practice stimuli were drawn from the main stimulus set, and so cannot be defined as strictly being included among either the illusion or control stimuli. However, the first author, who assembled the original stimulus set (Tierney et al., 2013), perceived two of them as clearly transforming into song and two of them as continuing to be perceived as speech when repeated.) The inclusion of these items was meant to help participants get accustomed to rating the musicality of stimuli; the ratings of these stimuli were not analyzed.

3.2.4. TAT

To assess participants' ability to determine whether a stream of sung pitches was aligned with a tonal template, we used a new tonality perception test devised for this study, the TAT. (We

599 elected not to use the mistuning perception test (Larrouy-Maestri, Harrison, & Müllensiefen, 2019)
600 because it assesses whether listeners can detect a pitch shift between a vocal line and an
601 instrumental accompaniment, whereas the illusion induces melody perception in a single
602 unaccompanied vocal recording.) Stimuli were drawn from the DSD100 (Liutkus et al., 2016), a
603 dataset of 100 publicly available music recordings in which each component instrument can be
604 downloaded separately. A single short portion of the vocal track from each of 22 different recordings
605 was extracted. The duration of these musical passages ranged from 9.5 to 24.6 seconds (mean 15.7,
606 SD 3.9). These 22 stimuli were then divided into three different sets that underwent three different
607 types of pitch morphing in Praat (Boersma & Weenink, 2021). For 6 of the stimuli, the pitch contour
608 was expanded by 30% on a time-point-by-time-point basis (Zatorre & Baum, 2012). This was done by
609 first extracting the fundamental frequency (F0) of each phrase using Praat (default settings, with a
610 time step of 0.01 seconds), then comparing the ratio between the F0 value at each time point and
611 the median F0 across all time points. The F0 of each time point was then adjusted to be equal to 1.3
612 times its original distance from the median F0, in semitones. (Previous perceptual research suggests
613 that the semitone scale is more relevant to the perception of pitch in speech than the Hz scale
614 [Nolan, 2003].) For 5 of the stimuli, F0 contours were contracted by 30% in a similar manner: the F0
615 of each time point was adjusted to be equal to 0.7 times its original distance from the median F0.
616 For 11 of the stimuli, the F0 values were set to be equivalent to their original values. However, to
617 ensure that any distortions due to the F0 manipulations were present across all stimuli, the
618 "unaltered" stimuli were constructed by first expanding the F0 contours by 30% and then
619 contracting them back to their original values.
620 Participants were told they would hear 22 melodies, some of which would be in tune, and
621 some of which would not be in tune. They were asked to indicate whether the melody they heard
622 was in tune (by pressing a button marked "in tune") or out of tune (by pressing a button marked
623 "out of tune"). Task performance was summarized by calculating sensitivity (d'). Stimuli from this
624 test are available at https://osf.io/hegxs/.

3.2.5. BAT

This beat perception test used 22 stimuli drawn from Iversen & Patel (2008). These were instrumental musical excerpts from a broad variety of musical genres, including rock, classical, and jazz, onto which a sequence of 1 KHz 100 ms pure tones was superimposed. These tones were either aligned with the beat or shifted away by 25% of the inter-beat-interval. Participants were asked to indicate whether the tones were aligned with the beat or shifted away from the beat (by pressing buttons marked “on the beat” or “off the beat”). As soon as participants selected their response, the next stimulus was presented. A difference between this test and that reported in Iversen & Patel (2008) is that only a short excerpt of each stimulus was presented (long enough to contain seven beats, average duration 3.81 seconds), to keep the experiment relatively brief. Task performance was summarized by calculating sensitivity (d’). Stimuli from this test are available at https://osf.io/hegxs/.

3.2.6. Self-assessment of Music Perception Skills

Because the TAT is a novel measure, it is of interest to establish its validity by comparing performance on this test to self-assessments of tonality perception ability. In addition, we were interested in comparing the validity of the TAT to that of the BAT, an already established measure in the field. Participants were asked to indicate on a scale from 1 to 7 the extent to which they agreed with each of four statements, with 1 indicating “not at all” and 7 indicating “agree completely”. These statements were: 1) “I can tell when people sing or play out of tune”, 2) “When I sing I have no idea whether I’m in tune or not”, 3) “I can tell when people sing or play out of time with the beat”, and 4) “When I clap to music I have no idea whether I’m on the beat or not”. A composite measure of self-assessed tonality perception was calculated by averaging the response to question 1 with the inverted response to question 2. Similarly, a composite measure of self-assessed beat

perception was calculated by averaging the response to question 3 with the inverted response to question 4.

3.2.7. Psychoacoustic Discrimination

Two tests were conducted to examine the precision of participants' perception of frequency and amplitude rise time. Each test consisted of a single 2-down 1-up adaptive staircase design. In each test, stimuli were drawn from a continuum of 101 stimuli, representing 101 different levels of the target acoustic continuum (i.e. either frequency or rise time). In each trial, participants were presented with 3 sounds, with either the first or the third being different from the other two, which were identical. The two identical sounds always corresponded to level 1 of the target continuum. Participants were asked to indicate which of the three sounds was different by pressing either a button labelled "1" or a button labelled "3". Initially, the target sound level (in other words, the level of the different sound) was set at 50. After every two correct responses, the target level decreased, becoming more similar to the comparison level, while after every incorrect response, the target level increased, becoming more dissimilar to the comparison level. The size of these level increases/decreases changed across the block, becoming smaller after each subsequent "reversal" or inflection point (in other words, two correct responses following a set of incorrect responses, or one incorrect response following a set of at least two correct responses). These step sizes were 10, 5, 2, 1, 1, 1, and 1 before the first through seventh reversals, respectively. The block continued until 75 trials were presented or participants reached the seventh reversal, whichever came first.

The stimulus continua for the two tests were constructed as follows. For the frequency discrimination test, the comparison stimulus was always presented at a fundamental frequency (F0) of 330 Hz, while the target stimulus ranged from 330.3 to 360 Hz in 100 equal linear steps. Frequency discrimination stimuli were 500-ms four-harmonic complex tones with equal amplitude across harmonics with a 0.015 linear amplitude ramp at the beginning and end to avoid perception of transient clicks. For the amplitude rise time test, the comparison stimulus was always presented

676 with a linear amplitude ramp at the beginning of the stimulus of duration 15 ms, while the target
677 stimulus ranged from a rise time of 17.8 to 300 ms in 100 equal linear steps. Rise time stimuli were
678 500-ms four-harmonic complex tones with equal amplitude across harmonics and an F0 of 330 Hz.
679 Thresholds for each test were calculated as the mean of the target stimulus levels at the second
680 through final reversals.

681 Any variables that were shown to be non-normally distributed according to a Jacque-Bera
682 test were transformed prior to analysis. A rau transformation was used for the attention-to-pitch
683 test and beat and tonality self-assessments, a log transformation was used for the frequency and
684 rise time discrimination thresholds, and years of musical training were converted to ranks.

685 *3.3. Results*

686 As in Experiment 1, across participants the song illusion was very reliably experienced, as
687 seen in the strong interaction between repetition and stimulus set ($F(7,525) = 85.3$, $p < 0.001$). There
688 was once again a main effect of repetition ($F(1,525) = 50.9$, $p < 0.001$) and of stimulus set ($F(1,75) =$
689 182.0, $p < 0.001$). Illusion strength and musical prior did not correlate ($r(74) = -0.12$, $p = 0.318$),
690 suggesting that these two measures index dissociable aspects of the speech-to-song illusion.

691 To test the validity of the TAT and BAT, performance on these measures was correlated with
692 self-ratings of tonality perception (the ability to detect whether one's own singing and others'
693 singing is in tune) as well as self-ratings of beat perception (the ability to detect whether one's own
694 singing/playing and others' singing/playing is on the beat). BAT performance was correlated with
695 beat perception self-rating ($r(74) = 0.42$, $p < 0.001$) but not tonality perception self-rating ($r(74) =$
696 0.21, $p = 0.072$). TAT performance was correlated with both tonality perception self-rating ($r(74) =$
697 0.42, $p < 0.001$) and beat perception self-rating ($r(74) = 0.27$, $p = 0.018$). Figure 4 displays the
698 relationship between TAT and BAT performance and tonality and beat perception self-rating. The
699 finding that performance on the TAT relates to self-assessment of both beat and tonality processing
700 suggests that while this measure may be a valid measure of musical aptitude, its specificity in
701 assessing tonality per se may not be optimal.

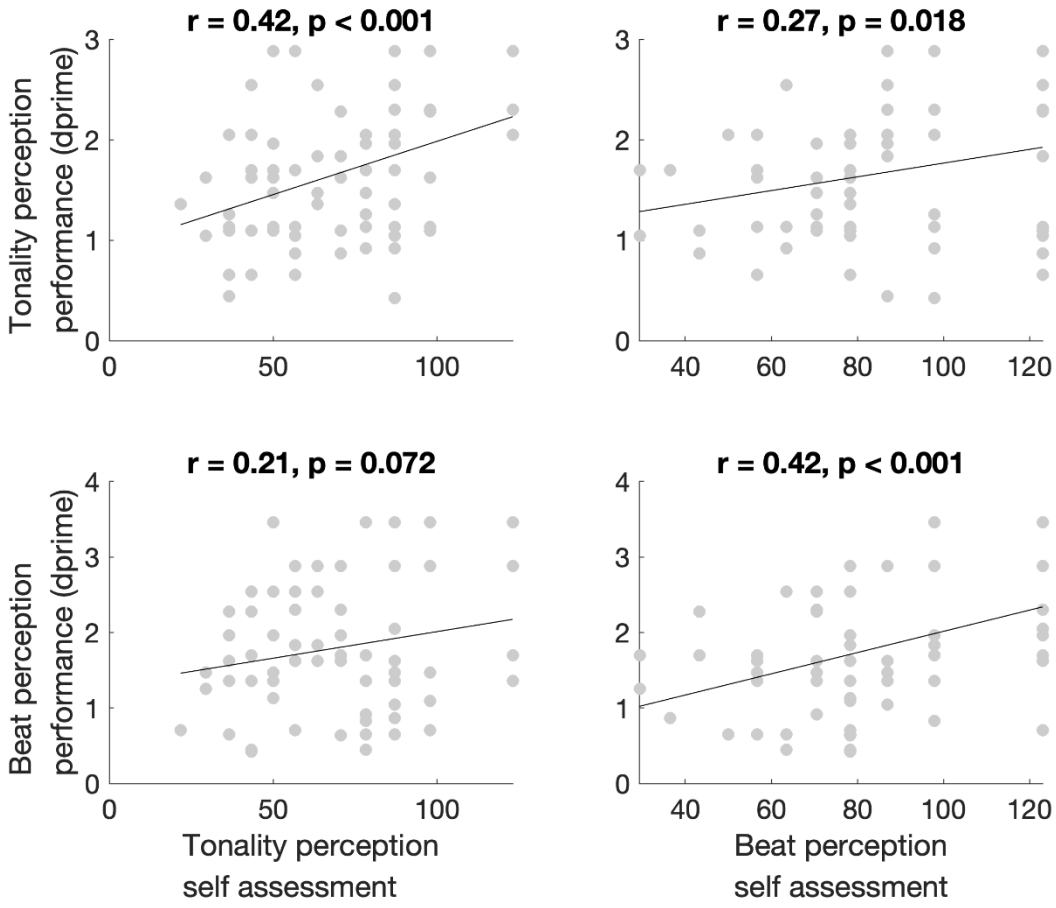

Figure 4. Upper left panel, relationship between tonality perception self-assessment and performance on the TAT. Upper right panel, relationship between beat alignment perception self-assessment and performance on the TAT. Lower left panel, relationship between tonality perception self-assessment and performance on the BAT. Lower right panel, relationship between beat perception self-assessment and performance on the BAT.

To investigate whether performance on the musical aptitude and auditory perception tests was associated with musical training, Pearson correlations were used to investigate the relationship between musical training and these measures. These correlations were corrected for multiple comparisons using false discovery rate, Benjamini & Hochberg 1995. Years of musical training were not significantly correlated with performance on any of the musical and auditory tests, including

attention to amplitude (r(74) = -0.05. p(corrected) = 0.674), attention to pitch (r(74) = 0.25, p(corrected) = 0.092), TAT (r(74) = 0.30, p(corrected) = 0.053), BAT (r(74) = 0.22, p(corrected) = 0.116), frequency discrimination (r(74) = -0.06, p(corrected) = 0.674), and amplitude rise time discrimination (r(74) = -0.13, p(corrected) = 0.413). These nonsignificant correlations underscore the distinction between musical aptitude and musical training as measures of musical ability (Kragness et al., 2020).

Pearson correlations were used to investigate the relationship between musical aptitude and the strength of perception of the speech-to-song illusion, which was operationalised as the difference in the average ratings of illusion and control stimuli after the eighth repetition (i.e., illusion strength in Experiment 1). These correlations were corrected for multiple comparisons using false discovery rate (Benjamini & Hochberg 1995). Illusion strength was correlated with the ability to direct attention to pitch within spoken phrases (r(74) = 0.26, p(corrected) = 0.044) but not with the ability to direct attention to amplitude within spoken phrases (r(74) = 0.09, p(corrected) = 0.454). Illusion strength was correlated with performance on the TAT (r(74) = 0.30, p(corrected) = 0.027) and performance on the BAT (r(74) = 0.38, p(corrected) = 0.005). Psychophysical discrimination thresholds were not correlated with illusion strength, including frequency discrimination (r(74) = 0.19, p(corrected) = 0.147) and amplitude rise time discrimination (r(74) = -0.14, p(corrected) = 0.256). Figure 6 displays the relationship between illusion strength and musical aptitude.

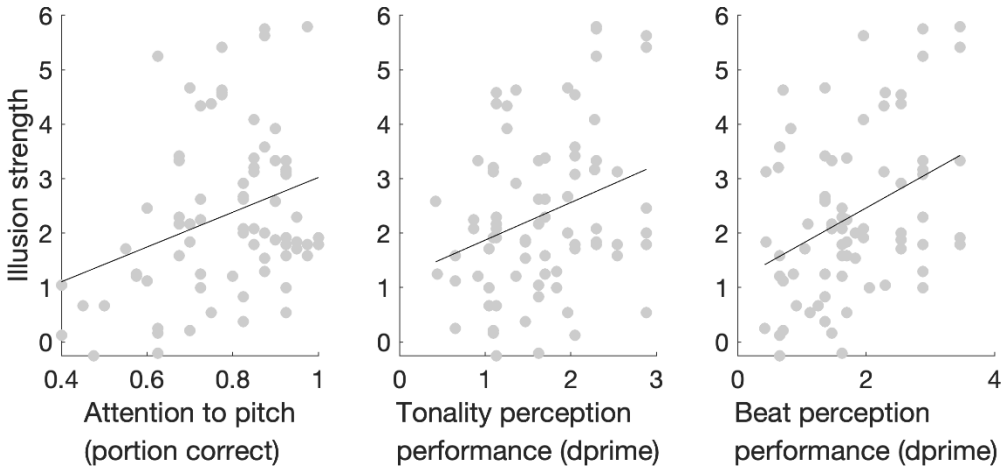

Figure 5. Left panel, relationship between attention to pitch and final illusion-control rating difference. Middle panel, relationship between TAT performance and final illusion-control rating difference. Right panel, relationship between BAT performance and final illusion-control rating difference.

Pearson correlations were used to investigate the relationship between musical aptitude and musical prior, which was operationalised as mean rating across all stimuli over all eight repetitions. These correlations were corrected for multiple comparisons using false discovery rate (Benjamini & Hochberg 1995). Musical prior was not correlated with performance on any of the auditory or musical tests, including attention to amplitude ($r(74) = -0.27$, p(corrected) = 0.114), attention to pitch ($r(74) = 0.20$, p(corrected) = 0.273), TAT ($r(74) = 0.02$, p(corrected) = 0.894), BAT ($r(74) = 0.11$, p(corrected) = 0.495), frequency discrimination ($r(74) = 0.02$, p(corrected) = 0.894), and amplitude rise time discrimination ($r(74) = -0.16$, p(corrected) = 0.332).

To examine whether the predictors explain independent variance in illusion strength, backwards linear regression was used, starting with a full model containing all (centered, scaled) predictors and removing predictors based on AIC. (The Variance Inflation Factor for all predictors was less than 1.6 in the full and reduced models, suggesting that multicollinearity was not a major problem in this dataset.) The resulting model (see Table 2) explained 22% of variance in illusion strength ($F(3,72) = 6.80$, $p < 0.001$). Predictors contained in the model included frequency discrimination threshold (beta = 0.306, t = 1.964, p = 0.053), BAT performance (beta = 0.449, t = 2.764, p = 0.007), and TAT performance (beta = 0.340, t = 2.083, p = 0.041). See Figure 7 for a depiction of predicted versus actual illusion strength values across participants.

	Coefficient	Std. error	t value	p value
(Intercept)	2.327	0.154	15.149	< 0.001
Frequency discrimination	0.306	0.156	1.964	0.053
BAT	0.449	0.162	2.764	0.007
TAT	0.340	0.163	2.083	0.041

Table 2. Regression predicting cross-participant differences in illusion strength.

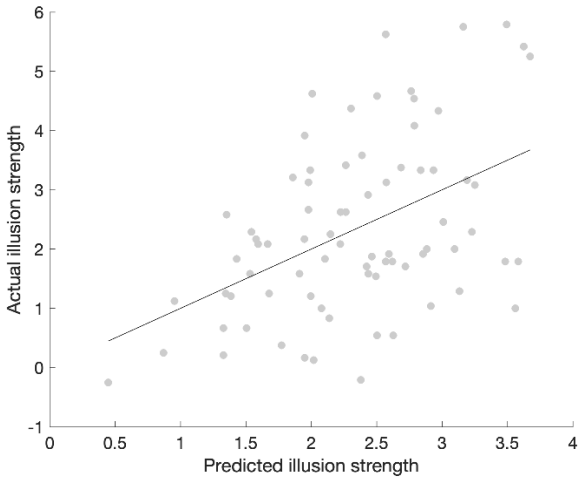

Figure 6. Scatterplot displaying predicted (x-axis) versus actual (y-axis) illusion strength values across participants.

3.4. Discussion

We found that musical aptitude—including performance on the Tonality Alignment Test (TAT) and Beat Alignment Test (BAT)—as well as the ability to direct attention to the pitch of speech were linked to the strength of the speech-to-song illusion (the difference in musicality ratings between strongly transforming versus weakly transforming stimuli) but not to musical prior (the overall tendency to rate stimuli as musical). This result suggests that a key factor driving the song illusion is perceptual sensitivity to the degree of musical structure present in speech. Our finding of a link between individual differences in musical aptitude and individual differences in perception of illusion strength confirms a key prediction of the hybrid account of the speech-to-song illusion, in

which satiation of lexical nodes frees up cognitive resources so that listeners can focus on extraction of rhythmic and melodic information from stimuli. Our results suggest that not all listeners are equally able to assess the musical structure of acoustic stimuli, and that this assessment skill may be driven by sensitivity to the details of musical pitch and rhythmic patterns. Importantly, we find that musical aptitude is not linked to individual differences in musical prior, suggesting that our results are not driven by an overall tendency for certain listeners to perceive stimuli as musical, but truly reflect differences in the ability to assess musical structure in non-musical stimuli.

The strongest predictor of the strength of the speech-to-song illusion was performance on the BAT. The presence of an isochronous beat is one of the most reliable characteristics distinguishing speech from music cross-culturally (Savage, Brown, Sakai, & Currie, 2015). Moreover, the presence of a steady pulse has also been shown to be a useful feature when constructing automatic classifiers of speech versus music (Scheirer & Slaney, 1997). Listeners may, therefore, be broadly aware that song tends to contain an isochronous beat while speech does not and use the presence or absence of a musical beat as relevant evidence when categorizing a stimulus as speech versus song. Indeed, we have previously shown that within the stimulus set used in the current study, stimuli with more isochronous beats (as extracted via a computational model of beat perception) are perceived as more musical after repetition (Tierney et al., 2018b). Beat perception is not limited to musicians, but rather is a broadly present skill in the general population: although neural entrainment to musical beats is enhanced by musical training, it is clearly present in non-musicians (Doelling & Poeppel, 2015). There is some evidence that beat perception emerges very early in development: infants coordinate their movements with the musical tempo based on musical pulse clarity (Zentner & Eerola, 2010) and can integrate auditory and somatosensory/proprioceptive information when deciding which notes carry musical beats (Phillips-Silver & Trainor, 2005). Nevertheless, although basic competence in beat perception is widespread, there are large individual differences in this ability (Müllensiefen, Gingras, Musil, & Stewart, 2014; Tranchant, Vuvan, & Peretz, 2016; Tranchant, Lagrois, Bellemare, Schultz, & Peretz, 2021), and impaired beat

perception may limit individuals' ability to detect musical structure in speech and other non-musical stimuli.

Performance on the TAT was another significant predictor of the strength of the speech-to-song illusion. The use of discrete scales is another characteristic that distinguishes speech from music cross-culturally (Savage, Brown, Sakai, & Currie, 2015). Listeners may, as a result, take the presence of tonality as evidence that a stimulus should be categorized as song rather than speech. Tonality perception, like beat perception, is widespread in the general population: Western European musicians and non-musicians show a preference for scale structure in melodic sequences (Cross, Howell, & West, 1983). Our findings, however, suggest that not everyone is equally able to detect the presence of scale structure, and that individual differences in this ability help determine variability in the robustness of musicality perception in non-musical stimuli. The correlation between tonality perception and speech-to-song perception (as well as with self-assessment of tonality perception skills) demonstrates the validity of this novel measure, which may be useful in a variety of future studies of tonality perception, including investigations of the relationship between musical and language skills, studies of the effects of musical training, and research on the development of musical abilities.

We find that the ability to direct attention to pitch in speech is linked to the robustness of perception of the speech-to-song illusion. This relationship, however, does not reflect the influence of general attentional skills, as the ability to direct attention to amplitude in speech was not linked to perception of the illusion. This finding aligns with the proposal by Deutsch et al. (2011) that perception of musical characteristics in speech is inhibited by default because pitch salience is down-regulated during speech perception, but that repetition of speech disinhibits attention to pitch, making possible the detection of musical characteristics in stimuli. Several findings support the possibility that variation in other acoustic dimensions can interfere with pitch perception, possibly by diverting attention away from pitch. For example, Warrier & Zatorre (2002), Allen & Oxenham (2014), and Caruso & Balaban (2014) showed that the presence of variation in timbre can interfere

with pitch perception, while Russo, Vuvan, & Thompson (2019) found that vowel content can interfere with relative pitch perception in speech stimuli. The idea that perceiving the speech-to-song illusion can lead to increased pitch salience is supported by the finding that pitch-sensitive cortical areas increase in activation when illusion stimuli are repeated (Tierney et al., 2013). This increase in salience may be greater in individuals who can more readily direct attention to pitch, leading to more robust perception of the illusion.

4. General discussion

Writing in 55 BCE, Cicero remarked in *De Oratore* that "even in speaking there may be a concealed kind of music." Over 2,000 years later, the discovery of the song illusion provided compelling evidence for this claim (Deutsch, 2003; Deutsch et al., 2011). The current work on individual differences in the experience of the illusion suggests that Cicero's phrase should be updated: "even in speaking there may be a concealed kind of music, *though only some can hear it*." Specifically, our results suggest that the ability to focus on and extract several different kinds of auditory and musical cues is linked to the strength of the song illusion. These results align with prior work on other auditory illusions, which has generally shown that the salience of (and ability to perceive) different stimulus characteristics is linked to the influence these characteristics have on the final percept. In the "Yanny/Laurel" illusion, for example, whether individuals perceive "Yanny" versus "Laurel" is linked to their degree of prior exposure to lower versus higher auditory frequencies in their environment (Gwilliams & Wallisch, 2020). Similarly, the magnitude of the McGurk effect correlates with lipreading skill (Strand, Cooperman, Rowe, & Simenstad, 2014; Brown et al., 2018), as well as the amount of time individuals spend fixating their gaze on a talker's mouth during audiovisual speech perception (Gurler, Doyle, Walker, Magnotti, & Beauchamp, 2015). The song illusion, therefore, like other illusions, may involve the weighting of multiple potential cues to which individuals are differentially sensitive.

Our findings support theoretical models of the song illusion which suggest that perception of the illusion requires extraction of musical characteristics from the stimuli (Deutsch et al. 2011,

Tierney et al. 2018b). The need to extract musical characteristics from the stimuli could partly explain the increase in musicality perception with repetition. Prior work has shown that perception of pitch interval size is surprisingly poor after a single presentation of a melody, and that melody repetition is necessary before participants achieve consistent interval perception (Deutsch, 1979). The idea that the increase in song illusion strength with repetition is driven by the time it takes to extract musical characteristics from the stimuli is supported by the finding that the increase in musicality with repetition is similarly sized when speech stimuli and synthesized complex tone stimuli with identical pitch contours are presented (Tierney et al., 2018a).

Our findings constrain theories of the song illusion in significant ways. The finding that individual differences in perception of the illusion are linked to musical aptitude is not predicted by models which suggest that the illusion is driven by deactivation of lexical nodes and continued activation of syllable nodes (Castro et al., 2018). While our results indicate that Node Structure Theory does not provide a *sufficient* explanation of the song illusion, our results do not imply that deactivation of lexical nodes is *not* relevant to perception of the illusion. Our experiment was not designed to test the theory that individual differences in the speed or robustness of lexical satiation drive individual differences in perception of the illusion, an idea worth exploring in future research.

Here we find that musical aptitude predicts the strength of the song illusion, but musical training does not. This may seem somewhat contradictory since the main goal of musical training is to boost musical skills. Indeed, prior work has shown that musicians demonstrate more precise mistuning perception (Hutchins, Roquet, & Peretz, 2012; Larrouy-Maestri, 2018) and stronger neural entrainment to musical beats (Doelling & Poeppel, 2015; but see Hickey, Merseal, Patel, & Race, 2020). However, the lack of a relationship between musical training and illusion perception aligns with other recent work finding that cognitive/perceptual abilities are more strongly linked to musical aptitude than to degree of musical training (Swaminathan, Schellenberg, & Khalil, 2017; Swaminathan & Schellenberg, 2019). One possibility is that strong individual differences in musical aptitude exist prior to music training, and in fact help determine whether individuals begin (and stick

with) musical training (Kragness et al., 2020). Overall, our results suggest that individual differences in the perception of musicality in speech are tied to musical skills that do not depend strongly on formal musical training, possibly reflecting stable traits influenced by genetic differences between individuals (Wesseldijk et al., 2021; Niarchou et al., 2021). It should be noted that we measured musical training, rather than musical experience, and therefore it remains an open question whether aspects of perception of the speech-to-song illusion are related to musical experience (which future research could investigate using the Gold-MSI: Müllensiefen et al., 2014).

Although our finding of a relationship between musical aptitude and illusion strength represents an initial step towards understanding the factors driving individual differences in perception of the speech-to-song illusion, the results of our linear regression explained only 22% of the variance in illusion strength. It is likely, therefore, that other major factors driving individual differences in perception of the illusion remain to be uncovered. One possibility is that individuals who tend to focus on the F0 rather than spectral information when assessing pitch contour may more robustly perceive the illusion. Research on perception of the missing fundamental illusion has shown that there exist large individual differences in the extent to which listeners rely on F0 versus spectral information when judging the interval between two notes (Schneider et al., 2005; Ladd et al., 2013). These divergent listening strategies may have consequences for the ability to extract musical information from sound sequences. Although pitch contour can be as easily extracted from inharmonic compared to harmonic sounds, the ability to extract F0 information facilitates judgment of exact pitch intervals and tonality (McPherson & McDermott, 2018) as well as memory for pitch (McPherson & McDermott, 2020). As a result, spectrally-biased listeners may have difficulty detecting musical regularities latent in the speech-to-song illusion stimuli. Another possible factor driving individual differences in perception of the illusion is the extent to which individuals find music rewarding or absorbing: listeners with a greater hedonic or absorptive response to music may be more keen to seek out musical characteristics in non-musical stimuli. This hypothesis could be tested in future work using the Barcelona Musical Reward Questionnaire (Mas-Herrero, Marco-

Pallares, Lorenzo-Seva, Zatorre, & Rodriguez-Fornells, 2013) or the Absorption in Music Scale (Sandstrom and Russo, 2013). Answers to specific questions within these scales may prove especially useful for helping predict susceptibility to the song illusion, e.g., "I like to find patterns in everyday sounds" from the Absorption in Music Scale. Individual differences in the vividness and control of auditory imagery could help predict illusion strength as well (Halpern, 2015).

In conclusion, we find that there are strong, reliable individual differences in the tendency to perceive certain repeated spoken phrases as sung. These differences seem to be largely independent of demographic characteristics, including language and musical training, but are linked to individual differences in specific perceptual and musical skills. This finding suggests that individual differences in auditory abilities may strongly affect perceptual categorization, influencing not only categorization judgments within a domain (such as speech perception; Jasmin, Dick, Holt, & Tierney, 2020) but even perceptual categorization between domains (here, speech versus music). A deeper understanding of these individual differences would help researchers predict whether a given listener will experience a particular spoken phrase as sung when repeated. This in turn would provide a powerful tool for exploring the cognitive and brain mechanisms underlying selective neural responses to speech and music (Ogg, Moraczewski, Kuchinsky, & Slevc, 2019; Zuk, Teoh, & Lalor, 2020; Boebinger, Norman-Haignere, McDermott, & Kanwisher, 2021), by using the same physical stimuli to elicit categorically different perceptual experiences.

5. Acknowledgements

We would like to thank Peter Pfordresher, Jonathan De Souza, and three anonymous reviewers for their helpful comments on previous versions of this manuscript.

6. References

Allen, E., & Oxenham, A. (2014). Symmetric interactions and interference between pitch and timbre. *JASA, 135*, 1371-1379.

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. (2020). Gorilla in our midst: an online behavioral experiment builder. *Behavior Research Methods, 52,* 388-407.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*, 289-300.

Bigand, E., Delbé, C., Gerard, Y., & Tillmann, B. (2011). Categorization of extremely brief auditory stimuli: domain-specific or domain-general processes? *PLoS ONE 6,* e27024.

Boebinger, D., Norman-Haignere, S. V., McDermott, J. H., & Kanwisher, N. (2021). Music-selective neural populations arise without musical training. *Journal of Neurophysiology*, *125*(6), 2237-2263.

Boersma, P., & Weenink, D. (2021). Praat: doing phonetics by computer [Computer program]. Version 6.1.42, retrieved 15 April 2021 from http://www.praat.org/

Brennan, D., & Stevens, C. (2002). Specialist musical training and the octave illusion: analytical listening and veridical perception by pipe organists. *Acta Psychologica, 109,* 301-314.

951

952 Brown, V., Hedayati, M., Zanger, A., Mayn, S., Ray, L., Dillman-Hasso, N., & Strand, J. (2018). What
953 accounts for individual differences in susceptibility to the McGurk effect? *PLoS ONE, 13,* e0207160.
954
955 Caruso, V. C., & Balaban, E. (2014). Pitch and timbre interfere when both are parametrically varied.
956 *PLoS ONE, 9(1)*, e87065.
957
958 Castro, N., Mendoza, J., Tampke, E., & Vitevitch, M. (2018). An account of the speech-to-song illusion
959 using Node Structure Theory. *PLoS ONE, 13*, e0198656.
960
961 Cicero, M. *On Oratory and Orators. With Notes Historical and Explanatory. A New Edition, Carefully*
962 *Revised and Corrected. In Two Volumes 1808.* Translation by William Guthrie.
963
964 Craig, J. (1979). The effect of musical training and cerebral asymmetries on perception of an auditory
965 illusion. *Cortex, 15,* 671-677.
966
967 Cross, I., Howell, P., & West, R. (1983). Preferences for scale structure in melodic sequences. *Journal*
968 *of Experimental Psychology: Human Perception and Performance, 9*, 444–460.
969
970 Danielsen, A., Nymoen, K., Anderson, E., Câmara, G., Langerød, M., Thompson, M., & London, J.
971 (2019). Where is the beat in that note? Effects of attack, duration, and frequency on the perceived
972 timing of musical and quasi-musical sounds. *Journal of Experimental Psychology: Human Perception*
973 *and Performance*, *45*, 402-418.
974
975 Davidson, B., Power, R., & Michie, P. (1987). The effects of familiarity and previous training on
976 perception of an ambiguous musical figure. *Perception and Psychophysics, 41,* 601-608.

977

978 Deutsch, D. (1975). Musical illusions. *Scientific American, 233,* 92-105.

979

980 Deutsch, D. (1995). *Musical Illusions and Paradoxes*. Philomel Records.

981

982 Deutsch, D. (2003). *Phantom Words and Other Curiosities*. Philomel Records.

983

984 Deutsch, D. (1979). Octave generalization and the consolidation of melodic information. *Canadian*

985 *Journal of Psychology, 33,* 201-205.

986

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

988 *129,* 2245-2252.

989

990 Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in

991 speech and music. *Neuroscience & Biobehavioral Reviews*, *81*, 181-187.

992

993 Doelling, K., & Poeppel, D. (2015). Cortical entrainment to music and its modulation by expertise.

994 *PNAS, 112,* E6233-E6242.

995

996 Falk, S., Rathcke, T., & Dalla Bella, S. (2014). When speech sounds like music. *Journal of Experimental*

997 *Psychology: Human Perception and Performance, 40,* 1491-1506.

998

999 Graber, E., Simchy-Gross, R., & Margulis, E. (2017). Musical and linguistic listening modes in the

1000 speech-to-song illusion bias timing perception and absolute pitch memory. *Journal of the Acoustical*

1001 *Society of America, 142,* 3593-3602.

1002

Gurler, D., Doyle, N., Walker, E., Magnotti, J., & Beauchamp, M. (2015). A link between individual differences in multisensory speech perception and eye movements. *Attention, Perception, and Psychophysics, 77*, 1333-1341.

Gwilliams, L., & Wallisch, P. (2020). Immediate ambiguity resolution in speech perception based on prior acoustic experience. *PsyArXiv*. doi:10.31234/osf.io/yptfr

Halpern, A. (2015). Differences in auditory imagery self-report predict neural and behavioral outcomes. *Psychomusicology: Music, Mind, and Brain, 25,* 37-47.

Hickey, P., Merseal, H., Patel, A. D., & Race, E. (2020). Memory in time: Neural tracking of low-frequency rhythm dynamically modulates memory formation. *Neuroimage, 213*, 116693

Hutchins, S., Roquet, C., & Peretz, I. (2012). The vocal generosity effect: how bad can your singing be? *Music Perception, 30,* 147-159.

Irwin, J., Whalen, D., & Fowler, C. (2006). A sex difference in visual influence on heard speech. *Perception & Psychophysics, 68,* 582-592.

Iversen, J. R, & Patel, A. D. (2008). The Beat Alignment Test (BAT): Surveying beat processing abilities in the general population. In K. Miyazaki, M. Adachi, Y Hiraga, Y Nakajima, & M. Tsuzaki (Eds.), *Proceedings of the 10th International Conference on Music Perception & Cognition (ICMPC10)* (CD-ROM; pp. 465–468). Adelaide, Australia: Causal Productions.

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.

Jarque, C., & Bera, A. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters, 6,* 255-259.

Jasmin, K., Dick, F., Holt, L., & Tierney, A. (2020). Tailored perception: individuals' speech and music perception strategies fit their perceptual abilities." *Journal of Experimental Psychology: General, 149,* 914-934.

Jasmin, K., Sun, H., & Tierney, A. (2021). Effects of language experience on domain-general perceptual strategies. *Cognition, 206*, 104481.

Kachlicka, M., Saito, K., & Tierney, A. (2019). Successful second language learning is tied to robust domain-general auditory processing and stable neural representation of sound. *Brain and Language, 192,* 15-24.

Kawahara, H., & Irino, T. (2005). Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation. In P. Divenyi (Ed.), Speech separation by humans and machines (pp. 167–180). Boston, MA: Kluwer Academic Publishers.

Kragness, H., Swaminathan, S., Cirelli, L., & Schellenberg, G. (2020). Individual differences in musical ability are stable over time in childhood. *Developmental Science.* doi:10.1111/desc.13081

Krizman, J., Bonacina, S., & Kraus, N. (2019). Sex differences in subcortical auditory processing emerge across development. *Hearing Research, 380,* 166-174.

Ladd, R., Turnbull, R., Browne, C., Caldwell-Harris, C., Ganushchak, L., Swoboda, K., Woodfield, V., & Dediu, D. (2013). Patterns of individual differences in the perception of missing-fundamental tones. *Journal of Experimental Psychology: Human Perception and Performance, 39,* 1386-1397.

Larrouy-Maestri, P. (2018). I know it when I hear it: on listeners' perception of mistuning. *Music & Science, 1,* 2059204318784582.

Larrouy-Maestri, P., Harrison, P., & Müllensiefen, D. (2019). The mistuning perception test: a new measurement instrument. *Behavior Research Methods, 51,* 663-675.

Liutkus, A., Stöter, F., Rafii, Z., Kitamura, D., Rivet, B., Ito, N., Ono, N., & Fontecave, J. (2016). The 2016 signal separation evaluation campaign. In *International Conference on Latent Variable Analysis and Signal Separation*, pp. 323-332. Springer, Cham.

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

Mas-Herrero, E., Marco-Pallares, J., Lorenzo-Seva, U., Zatorre, R., & Rodriguez-Fornells, A. (2013). Individual differences in music reward experiences. *Music Perception, 31,* 118-138.

McPherson, M., & McDermott, J. (2018). Diversity in pitch perception revealed by task dependence. *Nature Human Behaviour, 2,* 52-66.

McPherson, M., & McDermott, J. (2020). Time-dependent discrimination advantages for harmonic sounds suggest efficient coding for memory. *PNAS, 117,* 32169-32180.

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PLoS one*, *9*(2), e89642.

Mullin, H., Norkey, E., Kodwani, A., Vitevitch, M., & Castro, N. (2021). Does age affect perception of the Speech-to-Song illusion? *PLoS one, 16,* e0250042.

Niarchou, M., Gustavson, D., Sathirapongsasuti, J., Anglada-Tort, M., Eising, E., Bell, E., McArthur, E., Straub, P., The 23andMe Research Team, McAuley, J., Capra, J., Ullén, F., Creanza, N., Mosing, M., Hinds, D., Davis, L., Jacoby, N., & Gordon, R. (2021). Unravelling the genetic architecture of musical rhythm: a large-scale genome-wide association study of beat synchronization. *bioRxiv, 836197.* doi:10.1101/8361

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

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, *88*(6), 1281-1296.

Ogg, M., Slevc, R., & Idsardi, W. (2017). The time course of sound category identification: insights from acoustic features. *Journal of the Acoustical Society of America, 142,* 3459-3473.

Ogg, M., Moraczewski, D., Kuchinsky, S., & Slevc, R. (2019) Separable neural representations of sound sources: speaker identity and musical timbre. *NeuroImage, 191,* 116-126.

Ogg, M., Carlson, T., & Slevc, R. (2020) The rapid emergence of auditory object representations in cortex reflect central acoustic attributes. *Journal of Cognitive Neuroscience, 32,* 111-123.

Patel, A. (2008). *Music, language, and the brain.* Oxford University Press.

Patterson, R., Uppenkamp, S., Johnsrude, I., & Griffiths, T. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron, 36,* 767-776.

Penagos, H., Melcher, J., & Oxenham, A. (2004). A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. *Journal of Neuroscience, 24,* 6810-6815.

Pfordresher, P., Mantell, J., & Pruitt, T. (2021). Effects of intention in the imitation of sung and spoken pitch. *Psychological Research.* doi:10.1007/s00426-021-01527-0

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

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

Pilotti, M., Simcox, T., Baldy, J., & Schauss, F. (2011). Are verbal transformation sensitive to age differences and stimulus properties? *The American Journal of Psychology, 124*, 89-97.

Pressnitzer, D., Graves, J., Chambers, C., de Gardelle, V., & Egré, P. (2018). Auditory perception: Laurel and Yanny together at last. *Current Biology, 28*, R737-R759.

Prince, J. (2014). Pitch structure, but not selective attention, affects accent weightings in metrical grouping. *Journal of Experimental Psychology: Human Perception and Performance, 40,* 2073-2090.

Rathcke, T., Falk, S., & Dalla Bella, S. (in press). Music to your ears: sentence sonority and listener background modulate the "speech-to-song illusion". *Music Perception*.

Rowland, J., Kasdan, A., & Poeppel, D. (2019). There is music in repetition: looped segments of speech and nonspeech induce the perception of music in a time-dependent manner. *Psychonomic Bulletin & Review, 26*, 583-590.

Russo, F., Vuvan, D., & Thompson, W. (2019). Vowel content influences relative pitch perception in vocal melodies. *Music Perception, 37,* 57-65.

Sandstrom, G. M., & Russo, F. A. (2013). Absorption in music: Development of a scale to identify individuals with strong emotional responses to music. *Psychology of Music, 41(2),* 216-228.

Savage, P., Brown, S., Sakai, E., & Currie, T. (2015). Statistical universals reveal the structures and functions of human music. *PNAS, 112,* 8987-8992.

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).

Schneider, P., Sluming, V., Roberts, N., Scherg, M., Goebel, R., Specht, H., Dosch, H., Bleeck, S., Stippich, C., & Rupp, A. (2005). Structural and functional asymmetry of lateral Heschl's gyrus reflects pitch perception preference. *Nature Neuroscience, 8*, 1241-1247.

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

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

Strand, J., Cooperman, A., Rowe, J., & Simenstad, A. (2014). Individual differences in susceptibility to the McGurk Effect: links with lipreading and detecting audiovisual incongruity. *JSLHR, 57*, 2322-2331.

Swaminathan, S., Schellenberg, G., & Khalil, S. (2017). Revisiting the association between music lessons and intelligence: training effects or music aptitude? *Intelligence, 62,* 119-124.

Swaminathan, S., & Schellenberg, G. (2019). Musical ability, music training, and language ability in childhood. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46,* 2340-2348.

Swaminathan, S., Kragness, H., & Schellenberg, G. (2021). The Musical Ear Test: Norms and correlates from a large sample of Canadian undergraduates. *Behavior Research Methods.* doi:10.3758/s13428-020-01528-8

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,* 240-254.

Tierney, A., Patel, A., & Breen, M. (2018a). Repetition enhances the musicality of speech and tone stimuli to similar degrees. *Music Perception, 35*, 573-578.

Tierney, A., Patel, A., & Breen, M. (2018b). Acoustic foundations of the speech-to-song illusion. *Journal of Experimental Psychology: General, 6,* 888-904.

Toscano, J., & McMurray, B. (2010). Cue integration with categories: weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science, 34,* 434-464.

Tranchant, P., Vuvan, D. T., & Peretz, I. (2016). Keeping the beat: A large sample study of bouncing and clapping to music. *PloS one, 11(7)*, e0160178.

Tranchant, P., Lagrois, M. É., Bellemare, A., Schultz, B. G., & Peretz, I. (2021). Co-occurrence of Deficits in Beat Perception and Synchronization Supports Implication of Motor System in Beat Perception. *Music & Science, 4*, 2059204321991713.

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

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

Vitevitch, M., Ng, J., Hatley, E., & Castro, N. (2020). Phonological but not semantic influences on the speech-to-song illusion. *Quarterly Journal of Experimental Psychology*. doi: 10.1177/1747021820969

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

Warren, R., & Warren, R. (1966). A comparison of speech perception in childhood, maturity, and old age by means of the verbal transformation effect. *Journal of Verbal Learning and Verbal Behavior, 5*, 142-146.

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

Wesseldijk, L., Gordon, R., Mosing, M., & Ullén, F. (2021). Music and verbal ability—a twin study of genetic and environmental associations. *Psychology of Aesthetics, Creativity, and the Arts*. doi:10.1037/aca0000401

Wesseldijk, L., Mosing, M., & Ullén, F. (2021). Why is an early start of training related to musical skills in adulthood? A genetically informative study. *Psychological Science, 32,* 3-13.

Zatorre, R., & Baum, S. (2012). Musical melody and speech intonation: singing a different tune? *PLoS Biology, 10,* e1001372.

Zentner, M., & Eerola, T. (2010). Rhythmic engagement with music in infancy. *PNAS, 107*, 5768-5773.

Zhang, J., Susino, M., McPherson, G., & Schubert, E. (2020). The definition of a musician in music psychology: a literature review and the six-year rule. *Psychology of Music, 48,* 389-409.

Zuk, N. J., Teoh, E. S., & Lalor, E. C. (2020). EEG-based classification of natural sounds reveals specialized responses to speech and music. *NeuroImage*, *210*, 116558.