Weighting of cues to categorization of song versus speech in tone-language and non-tone-language speakers

Cognition 246 (2024) 105757

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Weighting of cues to categorization of song versus speech in tone-language and non-tone-language speakers

Magdalena Kachlicka [a], Aniruddh D. Patel [b,c], Fang Liu [d], Adam Tierney [a,*]

[a] Department of Psychological Sciences, Birkbeck, University of London, Malet Street, London, United Kingdom
[b] Department of Psychology, Tufts University, 419 Boston Ave, Medford, USA
[c] Program in Brain, Mind, and Consciousness, Canadian Institute for Advanced Research, 661 University Avenue, Toronto, Canada
[d] School of Psychology and Clinical Language Sciences, University of Reading, Whiteknights, Reading, United Kingdom

ARTICLE INFO

Keywords:
Speech
Song
Illusion
Categorization
Cue-weighting

ABSTRACT

One of the most important auditory categorization tasks a listener faces is determining a sound's domain, a process which is a prerequisite for successful within-domain categorization tasks such as recognizing different speech sounds or musical tones. Speech and song are universal in human cultures: how do listeners categorize a sequence of words as belonging to one or the other of these domains? There is growing interest in the acoustic cues that distinguish speech and song, but it remains unclear whether there are cross-cultural differences in the evidence upon which listeners rely when making this fundamental perceptual categorization. Here we use the speech-to-song illusion, in which some spoken phrases perceptually transform into song when repeated, to investigate cues to this domain-level categorization in native speakers of tone languages (Mandarin and Cantonese speakers residing in the United Kingdom and China) and in native speakers of a non-tone language (English). We find that native tone-language and non-tone-language listeners largely agree on which spoken phrases sound like song after repetition, and we also find that the strength of this transformation is not significantly different across language backgrounds or countries of residence. Furthermore, we find a striking similarity in the cues upon which listeners rely when perceiving word sequences as singing versus speech, including small pitch intervals, flat within-syllable pitch contours, and steady beats. These findings support the view that there are certain widespread cross-cultural similarities in the mechanisms by which listeners judge if a word sequence is spoken or sung.

1. Introduction

To use sound to learn about the world listeners must map continuously varying acoustic input onto discrete categories. One reason why this is a daunting task is that the nature of the categories and the relevant acoustic cues vary depending on the domain to which a sound belongs. If a sound is English speech, for example, listeners need to distinguish the high-frequency content which characterizes phonemic contrasts such as [r] vs. [l] (Eimas, 1975; Iverson et al., 2003) and which helps distinguish stressed from unstressed syllables (Chrabaszcz, Winn, Lin, & Idsardi, 2014). If a sound is song, on the other hand, listeners need to track the lower-frequency content which conveys musical features such as the pitch intervals which define the tonal structure of music (Krumhansl, 2001; McPherson & McDermott, 2018; Sankaran, Carlson, & Thompson, 2020; Schellenberg & Trehub, 1996). The categorization of sound domain, therefore, is vital for listeners, because it guides them to the most important information to focus on and extract, and this process of implicit categorization may trigger specialized neural processing tuned to specific domains, e.g., to speech vs. song (Harris, Niven, Griffin, & Scott, 2023; Norman-Haignere et al., 2022).

Prior research on sound categorization has almost exclusively focused on categorization within a domain (for example, speech or music). This work has shown that speech and music often employ highly redundant signals in which multiple cues signal the presence of any given sound category. For example, in English speech the distinction between voiced and unvoiced stop consonants is conveyed by a combination of several cues including voice onset time (the time between release of the articulators and the onset of voicing) and the fundamental frequency (F0) of the following vowel (Hanson, 2009). Similarly, in music, strong and weak beats are distinguished by multiple acoustic

* Corresponding author.
E-mail address: a.tierney@bbk.ac.uk (A. Tierney).

https://doi.org/10.1016/j.cognition.2024.105757
Received 21 September 2023; Received in revised form 9 February 2024; Accepted 20 February 2024

cues, including duration, amplitude, and pitch (Ellis & Jones, 2009; Hannon, Snyder, Eerola, & Krumhansl, 2004; Prince, 2014). This redundancy is important because categories can overlap along any single acoustic dimension, particularly in real-world listening situations where sound can be degraded by the influence of noise, reverberation, and other sound sources (Jiang, Chen, & Alwan, 2006). Learning to perceive sound categories, therefore, requires listeners to learn the relative usefulness of different cues as evidence for the existence of a particular category (cue weighting), while also being able to flexibly adjust cue weighting to fit varied listening situations (Idemaru & Holt, 2014).

Categorization of a sound pattern as speech or music might seem a trivial task compared to distinguishing between voiced and unvoiced consonants or strong and weak beats. Indeed, *instrumental* tones and spoken vowels can be distinguished in <100 ms and are not generally acoustically overlapping perceptual categories (Bigand, Delbé, Gerard, & Tillmann, 2011; Ogg, Carlson, & Slevc, 2020; Ogg, Slevc, & Idsardi, 2017). However, cross-cultural comparison of speech and *song* suggests that these two domains are overlapping acoustic categories. While song tends to have greater within-syllable pitch stability, larger pitch intervals (between the median pitches of syllables), greater average pitch height, and more regular rhythms compared to speech, the two domains overlap on each of these characteristics, and thus none alone is sufficient to be diagnostic of speech versus song (Ozaki et al., 2023). As a result, listeners must weight cues to guide their categorization of sound as speech versus song, just as in within-domain categorization.

One way to study weighting of cues to a particular set of categories is to make use of stimuli that are on the boundary between categories, because participants may settle on contrasting interpretations of these stimuli depending on their individual weighting of cues. There are naturally-occurring examples of stimuli on the border between speech and song, as demonstrated by the speech-to-song illusion (Deutsch, Henthorn, & Lapidis, 2011). In this illusion, certain spoken phrases, which sound like speech in their original context, begin to be perceived as song when removed from context and repeated. Although the original study of this illusion focused on a single striking example in British English, a follow-up study introduced a corpus of twenty-four illusion phrases in British and American English which sound significantly more song-like after repetition, matched to twenty-four control phrases produced by the same talkers which continue to be heard as speech after repetition (Tierney, Dick, Deutsch, & Sereno, 2013). These matched illusion and control stimuli can be used to study the cues which drive perception of speech versus song. Across native English speakers, for example, perception of song is linked to greater pitch stability within syllables, lower beat interval variability, and smaller pitch interval size between syllables (Falk, Rathcke, & Dalla Bella, 2014; Tierney, Patel, & Breen, 2018; Tierney, Patel, Jasmin, & Breen, 2021).

Prior research, therefore, has identified a set of cues linked to perception of song versus speech in native English speakers. However, it remains an open question whether the weighting of these cues is similar across individuals, or whether it instead varies depending on an individual's linguistic and cultural background. On the one hand, certain acoustic cues are linked to the presence of song versus speech cross-culturally, including the predominance of spectral versus temporal modulation (Albouy, Mehr, Hoyer, Ginzburg, & Zatorre, 2023), within-syllable pitch stability, and a beat-based temporal structure (Ozaki et al., 2023). On the other hand, prior research suggests that language background can lead to salient changes in cue weighting which can generalize to music perception. For example, when perceiving the location of a phrase boundary in English speech, and when distinguishing between stressed and unstressed syllables, native Mandarin speakers place greater emphasis on pitch (and lesser emphasis on other cues such as duration) relative to native English speakers and to speakers of non-tonal languages learning English as a second language (Jasmin, Sun, & Tierney, 2021; Yu & Andruski, 2010; Zhang & Francis, 2010; Zhang, Nissen, & Francis, 2008). This pitch-biased perceptual strategy is somewhat domain-general, as native Mandarin speakers also up-weight pitch and down-weight duration when categorizing musical beats as strong versus weak (Jasmin et al., 2021; Petrova, Jasmin, Saito, & Tierney, 2024).

Here we asked whether cues to the perception of song vary with linguistic and cultural background. We investigated this issue by presenting illusion and control stimuli from the speech-to-song corpus (consisting of phrases of spoken English) to native English speakers and native speakers of two tone-languages, Mandarin and Cantonese. In tone-languages, unlike in non-tone languages, spoken pitch patterns help determine lexical meaning (i.e., a word can have entirely different meanings, such as "bag" or "wing" depending on its pitch pattern). Over half of the world's languages are tone languages, and such languages can differ in their number of tones, ranging from two to seven or more. (See Patel, 2008 Ch. 2 for a detailed comparison of pitch structure in tone languages and music.) For example, Mandarin has four tones while Cantonese has six (Khouw & Ciocca, 2007; Yip, 2002), although linguists have reported that some Cantonese tones are merging due to the influence of Mandarin, leading Cantonese to have fewer than six tones (Fung & Lee, 2019; Mok, Zuo, & Wong, 2013; Ong, Wong, & Liu, 2020). We recruited two groups of Mandarin and Cantonese speakers, with one group living in the United Kingdom and the other in China; by recruiting participants with and without experience living in an English-speaking country, we could also investigate whether adult language experience can modulate the cues listeners use when evaluating if a phrase is song.

Given prior evidence that tone-language speakers up-weight pitch domain-generally when perceiving speech and music, as well as prior evidence for enhanced domain-general discrimination of and memory for pitch or pitch sequences in tone-language speakers (Creel, Weng, Fu, Heyman, & Lee, 2018; Hutka, Bidelman, & Moreno, 2015; Liu, Hilton, Bergelson, & Mehr, 2023), we predicted that when categorizing speech versus song Mandarin and Cantonese speakers would place greater importance on pitch-based cues, including pitch stability and pitch interval size. Moreover, Mandarin and Cantonese speech have lower average durational contrast between adjacent syllables compared to English (Mok, 2009). For English speakers, less durational contrast between adjacent syllables may be a more diagnostic feature of song, given that this measure (as quantified by the normalized pairwise variability index, or nPVI) tends to be slightly lower in song than in speech cross-culturally (Ozaki et al., 2023). For Mandarin and Cantonese speakers, on the other hand, low durational contrast between adjacent syllables would not be unusual for speech. Moreover, there is some evidence that Chinese folk song features particularly high durational contrast between adjacent syllables compared to English folk song (Yang & Ding, 2021). As a consequence, we predicted that native English speakers would be more likely to classify stimuli with lesser pairwise syllable duration variability as song, while native Mandarin and Cantonese speakers would not use pairwise duration variability as a cue. Finally, there is some cross-cultural overlap in scale structure: for example, Chinese traditional music uses the pentatonic scale, which is also commonly used in European folk music (Van Khe, 1977). We predicted, therefore, a similar degree of importance placed on musical key fit by tone-language and non-tone-language speakers.

As a secondary research question, we asked whether the strength of the speech-to-song illusion differs between native tone-language and non-tone-language speakers. Two lines of research lead one to expect that our native English-speaking participants should experience the song illusion more strongly than our native Mandarin and Cantonese-speaking participants. Jaisin, Suphanchaimat, Candia, and Warren (2016) hypothesized that native tone-language speakers are more likely to interpret speech pitch patterns in terms of lexical/semantic (vs. post-lexical, prosodic) contrasts and that this would attenuate song illusion strength (cf. Bidelman & Lee, 2015). Consistent with this hypothesis, Jaisin et al. (2016) found a significant reduction in song illusion strength in native tone-language vs. non-tone language speakers. Indeed, close inspection of their data (Fig. 4 of their paper) shows that none of their

native Mandarin-speakers experienced the song illusion when hearing phrases in either Mandarin or English. However, Jaisin et al. (2016) noted that their results should be considered preliminary since their comparison was based on relatively small samples and since the overall strength of the song illusion in their stimuli was modest. (Their study had ten tone-language and ten non-tone-language speakers – 5 native speakers each of Mandarin, Thai, German, and Italian – who heard five spoken phrases: one each in English, German, Mandarin, Italian, and Thai.) Apart from Jaisin et al.'s work, another line of work favoring the idea of a stronger song illusion effect in native non-tone-language vs. tone-language speakers comes from behavioral and neural research on linguistic tone perception (Wong, Chandrasekaran, & Zheng, 2012). Based on this work it has been suggested that native speakers of non-tone languages process pitch and segmental content more separately than do native tone-language speakers (Caldwell-Harris, Lancaster, Ladd, Dediu, & Christiansen, 2015). If this is the case, it should be easier for native non-tone-language speakers to focus on the pitch of utterances and thus perceive their musical characteristics, which would lead one to expect that they would experience the song illusion more strongly than native tone-language speakers.

However, two other lines of research lead one to expect that our Chinese participants should experience the song illusion in our stimuli more *strongly* than our native English-speaking participants. There is evidence that the speech-to-song illusion is stronger in languages that are more difficult for a participant to pronounce (Margulis, Simchy-Gross, & Black, 2015), possibly reflecting stronger activation of speech perception resources for easily pronounceable stimuli and, therefore, inhibition of song perception mechanisms. These findings lead one to expect that the song illusion would be stronger in the Cantonese and Mandarin participants in our study, given the phonological differences between these languages and English, which the participants did not speak natively. Additionally, there is evidence that song illusion strength is stronger for phrases in a non-native language (Rathcke, Falk, & Dalla Bella, 2021), perhaps because semantic representations are not as strongly activated as when hearing one's native language, allowing more focus on prosodic representations.

Given that prior work leads to contrasting expectations regarding the relative strength of the song illusion in our native tone-language vs. non-tone-language speakers, we had no predictions about differences in illusion strength in the two groups. Our study contributes to research on this topic by investigating the issue using a relatively large sample, with approximately ten times as many stimuli and participants as in Jaisin et al. (2016).

2. Methods

2.1. Participants

A total of 231 participants completed the experiment: 95 native speakers of English (54 female; mean (STD) age = 33.8 (11.8) years), and 136 native speakers of tonal languages. None of the English speakers were familiar with any tonal language. The tone-language speakers were native speakers of Mandarin Chinese ($N = 103$) or Cantonese ($N = 33$). Prior to analysis, data from 5 participants based in China (2 Mandarin and 3 Cantonese native speakers) were removed due to a high number of missed trials in the main task (i.e., >130 missed responses across all stimuli and their repetitions), which prevented us from accurately representing the trend in their response patterns. Thus data from 131 tone-language speakers were analyzed.

Of the Mandarin speakers, 47 were based in the UK (39 female; age = 19.5 (1.8) years) and 54 in China (34 female; age = 25.9 (6.8) years). Mandarin-speaking participants based in the UK reported a mean of 1.16 years of residence in English-speaking countries (SD = 1.71), while Mandarin-speaking participants based in China reported 0.43 years of residence in English-speaking countries (SD = 1.00). Of Cantonese speakers 16 were UK-based (13 female; age = 21.4 (3.8) years), and 14 were based in China (10 female; age = 24.9 (9.1) years). Cantonese-speaking participants based in the UK reported 3.23 years of residence in English-speaking countries (SD = 3.96), while Cantonese-speaking participants based in China reported 0.51 years of residence in English-speaking countries (SD = 1.28). None of the tone-language speakers were raised bilingual. Participants were asked to self-assess their English listening and speaking skills on a scale from 1 to 7. (Self-assessment data were not available for two Mandarin speakers based in the UK.) Self-assessed English skills were greater for Mandarin speakers based in the UK (listening, M = 5.6, SD = 1.0; speaking, M = 4.8, SD = 1.3) than for those based in China (listening, M = 3.5, SD = 1.5, $t(97) = 7.68$, $p < .001$; speaking, M = 3.5, SD = 1.4, $t(97) = 4.63$, $p < .001$). Similarly, self-assessed English skills were greater for Cantonese speakers based in the UK (listening, M = 6.3, SD = 0.6; speaking, M = 5.8, SD = 0.9) than for those based in China (listening, M = 4.0, SD = 1.5, $t(28) = 5.78$, $p < .001$; speaking, M = 3.7, SD = 1.4, $t(28) = 4.92$, $p < .001$).

Regarding years of musical training, native English speakers reported 5.56 (10.13), UK-based Mandarin speakers reported 5.98 (4.50), China-Based Mandarin speakers reported 3.04 (4.93), UK-based Cantonese speakers reported 6.72 (4.20), and China-based Cantonese speakers reported 5.97 (6.67). Together, the tone-language speakers reported 4.84 (5.09) years of musical training. The native English and tone-language speaker groups did not differ in their degree of musical training (unpaired *t*-test, $t(224) = 0.70$, $p = .48$). See Table 1 for a summary of demographics for all participants.

2.2. Stimuli

All the phrases in the speech-to-song illusion stimulus set were extracted from audiobooks in English (i.e., they were meant to be perceived as speech) and were read by 3 different male voice actors (1 British, 2 American), equally represented across control and illusion speech samples. There were 24 phrases that sound significantly more song-like after repetition ("illusion stimuli") and 24 stimuli that continue to sound like speech after repetition ("control stimuli"), as established in Tierney et al. (2013) and Tierney et al. (2018). Control and illusion stimuli had 5.5 syllables on average (SD = 1.5, range 4 to 9 syllables), and did not differ significantly in rate or duration (mean rate 5.00 and 5.13 syllables/s; mean duration 1.42 and 1.29 s, for control and illusion stimuli, respectively). Illusion stimuli were slightly but significantly higher than control stimuli in median fundamental frequency: 141.75 Hz vs. 134.83 Hz (<1 semitone; for more details about the corpus, see Tierney et al. (2013) and Tierney et al. (2018)). All stimuli from the speech-to-song illusion corpus (including audio files and transcriptions) can be found at https://osf.io/t4pjq/ and https://osf.io/kbj7u/.

2.3. Procedure

English native speakers were recruited from Prolific (www.prolific.co) and were compensated for their time. As participants in this online platform, they were likely residents of a variety of different English-speaking countries. Mandarin and Cantonese speakers living in the UK were recruited via the SONA recruitment platform and rewarded with course credits. Mandarin and Cantonese participants living in China were recruited via word of mouth and social and professional networks of the third author and volunteered their time to complete the experiment.

Participants listened to all 48 stimuli (24 control and 24 illusion) in random order. On each trial, they heard each stimulus 8 times in a row. Prior to starting the experiment, participants were told "Your job will be to rate how much the phrase sounds like speech or song on scale from 1 to 10." After each repetition, they were presented with a rating scale of 10 numbered boxes with the leftmost box labeled "nonmusical" and the rightmost box labeled "musical." Participants pressed a button from the

Table 1
Demographics across all five participant groups. Parentheses indicate standard deviation. For the speaking and listening self-assessments, the low end of the scale (1) was labeled "very bad", while the high end of the scale (7) was labeled "very good".

	Native English	Native Mandarin, UK resident	Native Mandarin, China resident	Native Cantonese, UK resident	Native Cantonese, China resident
N	95	47	54	16	14
Age	33.8(11.8)	19.5(1.8)	25.9(6.8)	21.4(3.8)	24.9(9.1)
Years musical training	5.56(10.1)	5.98(4.50)	3.04(4.93)	6.72(4.20)	5.97(6.67)
Self-assessed English listening (1–7)		5.6(1.0)	3.5(1.5)	6.3(0.6)	4.0(1.5)
Self-assessed English speaking (1–7)		4.8(1.3)	3.5(1.4)	5.8(0.9)	3.7(1.4)
Years residence in English-speaking countries		1.16(1.71)	0.43(1.00)	3.23(3.96)	0.51(1.28)

scale after each repetition. During repetitions of a stimulus the maximum interstimulus interval (ISI) was 2 s, after which the participants heard another repetition of the phrase even if they failed to provide a rating. If a rating was provided, the next repetition started as soon as participants responded so ISIs were often shorter than 2 s. After finishing the main task, participants were asked to complete a short demographic and language experience questionnaire. Completion of the experiment took approximately 30 min. The experiment was designed and hosted on the Gorilla platform (www.gorilla.sc, Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020) and was distributed to participants for completion online.

The procedures of this experiment were approved by the Ethics Committee for the Department of Psychological Sciences at Birkbeck, University of London.

2.4. Data processing

2.4.1. Speech-to-song illusion measures

The main measures of interest were the musicality ratings after each stimulus repetition. There were few missed responses (3.46% of total responses with roughly 1/3 of these missing trials occurring after the first repetition of the stimulus). If the first or the last rating was missing we replaced these ratings with the rating from the second or penultimate trial, respectively. If a single rating was missing after any other repetition, we replaced it with the average of its neighbors. In case of two consecutive missing answers (919 responses in total, 1% of total responses), we interpolated these values with the na_interpolation function from the imputTS R package (Moritz & Bartz-Beielstein, 2017). To accurately represent the trend in participants' response patterns, we only allowed up to 2 missing trials in either the first half or the second half of the trials (i.e., if four or more responses were missing, we excluded these trials from analysis; 130 trials were removed). Similar to the procedures reported by Tierney et al. (2021), we reduced the dimensionality of the data by calculating the following metrics for each participant: *illusion strength* (the difference between the average ratings of illusion and control stimuli after the last repetition) and *musical prior* (averaged ratings across all repetitions and all stimuli). The illusion strength measure was designed to capture the extent to which a phrase transformed into song, controlling for more general repetition effects. The musical prior measure was designed to capture a participant's overall tendency to rate all stimuli as musical, regardless of stimulus class (illusion vs. control) or repetition.

2.4.1.1. Pitch contour measures. For all stimuli, pitch was measured in Praat (Boersma & Weenink, 2022) using the autocorrelation method with default parameters, which detects the periodicity in the windowed input signal and returns one fundamental frequency (F0) measurement every 10 ms (Boersma, 1993). Two aspects of pitch contour were quantified for analysis. ***Pitch slope*** was measured by extracting the absolute value of the slope of each syllable's F0 contour with linear regression, then averaging across syllables in a phrase. ***Pitch interval size*** is the mean of the absolute value of the intervals in semitones between the median pitches of adjacent syllables in a phrase.

2.4.1.2. Melodic structure. We computed ***musical key fit*** for all stimuli using Krumhansl (2001). The purpose of this measure is to quantify the extent to which the pitches in each stimulus fit a musical key (rather than to pinpoint the key of best fit). First, the median pitch of each syllable was measured. Next, the difference between the mean pitch averaged across all notes and the nearest concert pitch was subtracted from all notes; this aligned the note sequence as much as possible with a single musical key. Next, each note was given a value equal to the prevalence of that note in the key (following the key profiles given in Krumhansl, 2001). For a C-major scale, for example, a C-sharp would be given a smaller value than a C. For notes with intermediate values between note classes, interpolation was used to assign the note a key fit. For example, a note halfway between a C-sharp and a C would be given a value equal to the average of the prevalence values for C-sharp and C. Finally, prevalence values were averaged across notes. This procedure was conducted for all 24 major and minor keys, and the key fit value for the best-fitting key was returned.

2.4.1.3. Rhythmic variability measures. We employed two different measures of rhythmic variability. The first measure of rhythmic variability was the ***normalized Pairwise Variability Index*** (nPVI; Low, Grabe, & Nolan, 2000), which quantifies the pairwise variability of successive durations (e.g., vowels, syllables, musical notes, etc.).

$$nPVI = \frac{100}{m-1} \times \sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{\frac{d_k + d_{k+1}}{2}} \right|$$

Our measured unit was syllable duration, so in the above equation m represents the number of measured syllable durations in the sequence and d_k is the duration of the kth syllable. The formula computes the difference between the duration of each syllable and its successor and then divides this value by the average duration of the two syllables, and this procedure is repeated for each successive pair of syllables in the sequence (i.e., syllables 1 and 2, then syllables 2 and 3, then syllables 3 and 4, etc.). The average of these values is the nPVI for the sequence, where higher nPVI values indicate greater average durational contrast between successive syllables. (See the Appendix of Daniele & Patel, 2013 for an example of nPVI computation from a sequence of durations.) The nPVI measure was first developed to study linguistic rhythm (Grabe & Low, 2002; Nolan & Asu, 2009), and has been used to show that "syllable-timed" languages tend to have lower nPVI than "stress-timed" languages, but it has also been used to measure musical rhythm and its relationship to speech rhythm (McGowan & Levitt, 2011; Patel, 2008; Patel, Iversen, & Rosenberg, 2006).

As a second measure of rhythmic variability we calculated ***beat variability*** based on a computational model used to detect beat intervals in music (Ellis, 2007), following the procedures and parameters described in Tierney et al. (2018). Although the algorithm was intended to detect musical beats, it can also be used on speech since detected beats are based on pitch and intensity changes, similar to stress in speech (Sluijter, Van Heuven, & Pacilly, 1997). The model has previously been used to investigate beats in speech (Schultz et al., 2016), and its validity

for finding beats in our speech stimuli was confirmed as only 29% worse than human drummers (Tierney et al., 2018). Once the model estimated the beat timings for each spoken phrase, beat variability was calculated as the standard deviation of the inter-beat intervals. Due to the short length of six phrases, beat variability could not be calculated and these phrases were removed from analyses involving beat variability measures.

2.4.2. Cue weighting measurement

We measured each participant's weighting of pitch slope, pitch interval size, musical key fit, nPVI, and beat variability as cues to the perception of speech versus song. For each participant, a linear regression was run with these five stimulus features as predictors and the musicality rating of each stimulus after the final repetition as the outcome variable. Beta coefficients for each predictor were extracted as the individual's weighting of that cue. Bayesian ANOVAs and *t*-tests were then used to compare cue weights across groups.

According to the Jarque-Bera test estimating the goodness of fit of the data to a normal distribution, beat variability and pitch interval size cue weights were not normally distributed, so they were log transformed prior to analysis. Similarly, illusion strength was not normally distributed and so was rau transformed prior to analysis. The Jarque-Bera test was computed using DescTools R package (Signorell et al., 2019). All correlations were corrected for multiple comparisons using false discovery rate correction (FDR, Benjamini & Hochberg, 1995) from the R stats package. All the data from the current study are available at: https://osf.io/kbj7u/.

3. Results

3.1. Speech-to-song illusion effect

First, we used a linear mixed-effects regression for each of the five language groups (native English, native Mandarin UK residents, native Mandarin China residents, native Cantonese UK residents, native Cantonese China residents) to determine whether musicality ratings changed with stimulus repetition and whether ratings differed significantly between the two stimulus sets. The fixed effects in the model

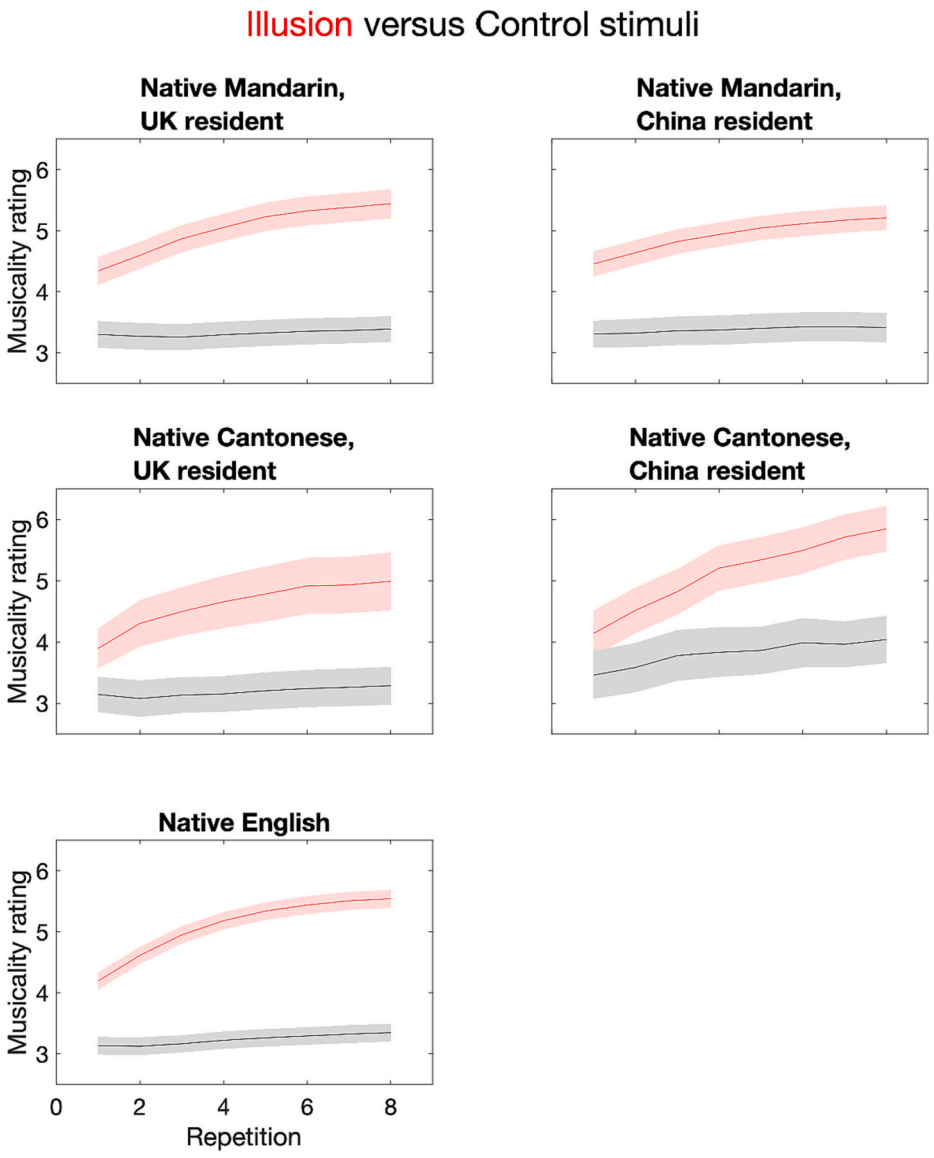

Fig. 1. Musicality rating as a function of stimulus repetition for song illusion (red) and control (black) stimuli for all five groups. (Native Mandarin, UK resident, $n = 47$; Native Mandarin, China resident, $n = 54$; Native Cantonese, UK resident, $n = 16$, Native Cantonese, China resident, $n = 14$; Native English, $n = 95$.) The shaded region indicates one standard error of the mean. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

included stimulus set (illusion vs control), repetition (1 to 8), and their interaction, whereas subject and item were defined as random effects. Repetition was entered as a continuous variable and standardized by centering and dividing by 2 standard deviations using the rescale function from the arm R package (Gelman & Hill, 2007). (Standardization by 2 standard deviations facilitates the interpretation of coefficients and enables coefficients to be directly comparable across a variety of different types of predictors, as explained in Gelman, 2008.) Stimulus set was entered as a categorical variable and sum coded so that contrast was centered at zero (i.e., −0.5, 0.5). Analysis was conducted using the lme4 R package (Bates, Mächler, Bolker, & Walker, 2015). See Fig. 1 for plots of changes in musicality ratings with repetition across all five groups.

For all five groups, the song illusion stimuli were rated as more musical than the control stimuli (native English speakers, $\beta = 1.84$, $t = 6.69$, $p < .001$; native Mandarin UK residents, $\beta = 1.66$, $t = 6.59$, $p < .001$; native Mandarin China residents, $\beta = 1.54$, $t = 5.48$, $p < .001$; native Cantonese UK residents, $\beta = 1.72$, $t = 5.47$, $p < .001$; native Cantonese China residents, $\beta = 1.28$, $t = 4.75$, $p < .001$). Furthermore, the musicality rating increased with stimulus repetition across all five groups (native English speakers, $\beta = 0.50$, $t = 25.5$, $p < .001$; native Mandarin UK residents, $\beta = 0.39$, $t = 14.7$, $p < .001$; native Mandarin China residents, $\beta = 0.49$, $t = 10.3$, $p < .001$; native Cantonese UK residents, $\beta = 0.47$, $t = 9.78$, $p < .001$; native Cantonese China residents, $\beta = 0.66$, $t = 13.6$, $p < .001$). Importantly, this repetition effect was larger for the illusion stimuli than for the control stimuli across all five groups (native English speakers, $\beta = 0.69$, $t = 17.6$, $p < .001$; native Mandarin UK residents, $\beta = 0.63$, $t = 11.8$, $p < .001$; native Mandarin China residents, $\beta = 0.42$, $t = 7.37$, $p < .001$; native Cantonese UK residents, $\beta = 0.65$, $t = 6.878$, $p < .001$; native Cantonese China residents, $\beta = 0.64$, $t = 6.63$, $p < .001$). Overall, therefore, the speech-to-song illusion effect was heard robustly across all five language groups. Finally, we ran separate regressions on illusion and control stimuli to determine whether there was a significant change in musicality rating with repetition, collapsing across all five groups. The musicality ratings increased with stimulus repetition for both the illusion stimuli ($\beta = 0.74$, $t = 38.2$, $p < .001$; mean increase of 1.16 (standard deviation 2.07)) and (to a much lesser extent) for the control stimuli ($\beta = 0.13$, $t = 8.65$, $p < .001$; mean increase of 0.18 (standard deviation 1.35)). Moreover, we ran an additional regression on the 1st repetition of illusion and control stimuli, collapsing across all five groups. There was a main effect of stimulus class ($\beta = 1.03$, $t = 4.47$, $p < .001$), indicating that the illusion and control stimuli differed in musicality after only a single presentation.

3.2. Differences in musical prior and illusion strength

To determine whether there was evidence for a null hypothesis of no differences between the tone-language groups, we compared musical prior (averaged ratings across all repetitions and all stimuli) and illusion strength (the difference between the average ratings of illusion and control stimuli after the last repetition) across the four tone-language groups with two one-way Bayesian ANOVA models using JASP software (2023) and interpretations of Bayes Factor proposed by Lee and Wagenmakers (2014). For musical prior, the resulting Bayes Factor (BF01) was 12.24 supporting strong evidence in favour of the null hypothesis. For illusion strength, the Bayes Factor was 15.51 providing strong evidence in support of the null hypothesis (Fig. 2). In addition, to test the hypothesis that the strength of the speech-to-song illusion is modulated by second language proficiency (Rathcke et al., 2021), we compared the illusion strength of the UK-resident and China-resident groups with a Bayesian independent samples *t*-test using JASP, collapsing across tone language (Mandarin and Cantonese). (As reported in the Participants section above, UK-resident participants had significantly higher self-reported English speaking and listening proficiency than China-resident participants.) The resulting Bayes Factor was 4.15. This provides moderate evidence in favour of the null hypothesis, rather than supporting the idea that second language proficiency influences the strength of the song illusion when listening to phrases in a non-native language.

Because there were no differences between the tone-language groups in illusion strength or musical prior, analyses comparing tone-language and native English groups on these measures collapsed across all four tone-language groups. We compared musical prior and illusion strength between tone-language and non-tone-language speakers with two Bayesian independent samples *t*-tests using JASP. For musical prior, the

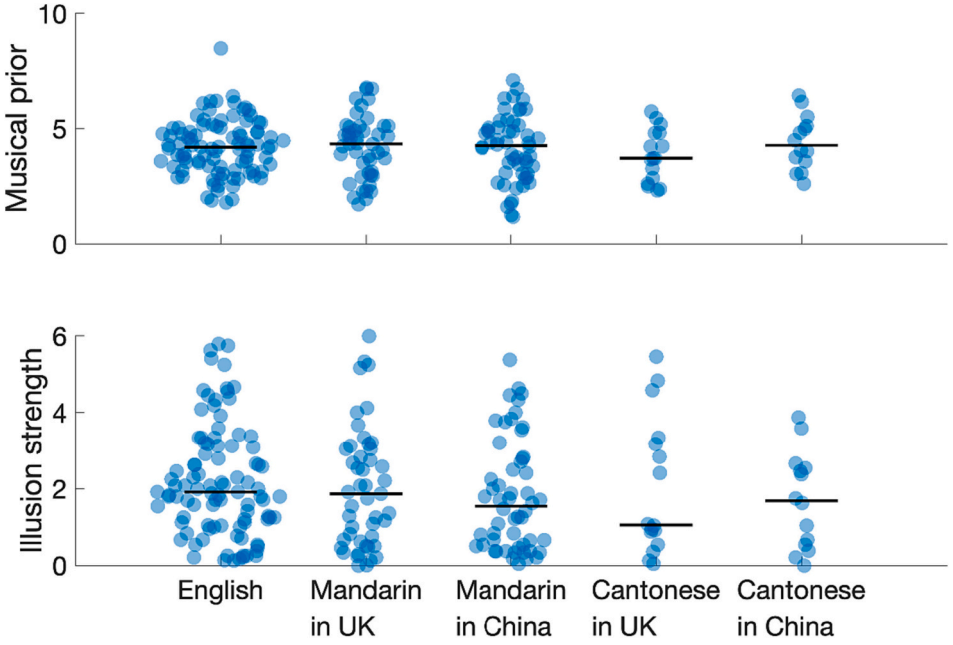

Fig. 2. Musical prior (top) and illusion strength (bottom) values across all five groups. Black horizontal lines indicate the median for each group, and dots show data for individual participants. Musical prior was calculated by averaging across all eight repetitions for all forty-eight stimuli. Illusion strength was calculated as the difference between average ratings for the illusion versus control stimuli after the eighth repetition.

Bayes Factor was 6.77 providing moderate evidence in support of the null hypothesis. For illusion strength, the Bayes Factor was 2.59, and as a result there was anecdotal support for the null hypothesis. See Fig. 2 for plots of raw illusion strength and musical prior data across all five groups.

3.3. Consistency in ratings across groups

To determine whether the different groups ranked the stimuli similarly in their relative musicality after repetition, for each group we averaged ratings across all participants for the 8th repetition of each stimulus, then compared the musicality rankings across groups for all 48 stimuli using Spearman correlations. Fig. 3 displays scatterplots showing the relationship between stimulus ratings across each pair of groups and lists the associated correlations. Correlations were high, ranging from rho = 0.85 (relationship between ratings of English and Mandarin-speaking China residents) to rho = 0.94 (relationship between ratings of Mandarin-speaking UK residents and Cantonese-speaking UK residents). All correlations were significant at $p < .001$. Fig. 4 displays mean musicality ratings after the eighth repetition for each stimulus for the native English-speaking and native Chinese-language-speaking groups. Ratings differences across groups survived FDR correction for multiple comparisons for three stimuli: the native English-speaking group rated stimuli 12 and 27 as more songlike, and stimulus 30 as less songlike, compared to the native Chinese-language-speaking group.

3.4. Musical cue weighting

Fig. 5 shows a plot of cue weights across all five predictors for all five groups. To determine whether there was evidence for a null hypothesis of no differences in musical cue weighting between the tone-language groups, we compared weighting of musical key fit, pitch slope, interval size, nPVI, and beat variability across the groups with five one-way Bayesian ANOVA models using JASP software (2023). The Bayes Factors (BF01) for key fit, pitch slope, interval size, nPVI, and beat variability were 3.69, 10.2, 1.89, 6.71, and 19.15, respectively, indicating no group differences in cue weighting.

Because there were no differences between the tone-language groups in weighting of any cue to musicality, analyses comparing cue weighting in tone-language and native English groups collapsed across all four tone-language groups. To determine whether there was evidence for a null hypothesis of no difference in cue weighting between the tone-language and non-tone-language speakers, we compared weighting of cues to musicality across these two groups with five Bayesian independent samples *t*-tests using JASP. The Bayes Factors (BF01) for key fit, pitch slope, interval size, nPVI, and beat variability were 1.84, 2.33, 3.05, 0.125, and 5.62, respectively. Thus for interval size and beat variability there was moderate evidence in favour of the null hypothesis of similar use of cues across tone-language and non-tone-language speakers, and for key fit and pitch slope there was anecdotal evidence in favour of this hypothesis. Only for nPVI was there moderate evidence for a different use of cues between the native Chinese and English speakers. Specifically, contrary to our predictions, Chinese speakers rated stimuli with higher nPVIs as sounding more musical (higher musicality ratings after the eighth repetition), while native English speakers did not use this cue.

Finally, we determined which cues significantly predicted musicality ratings for the English and Chinese participants. First, for each of the 42

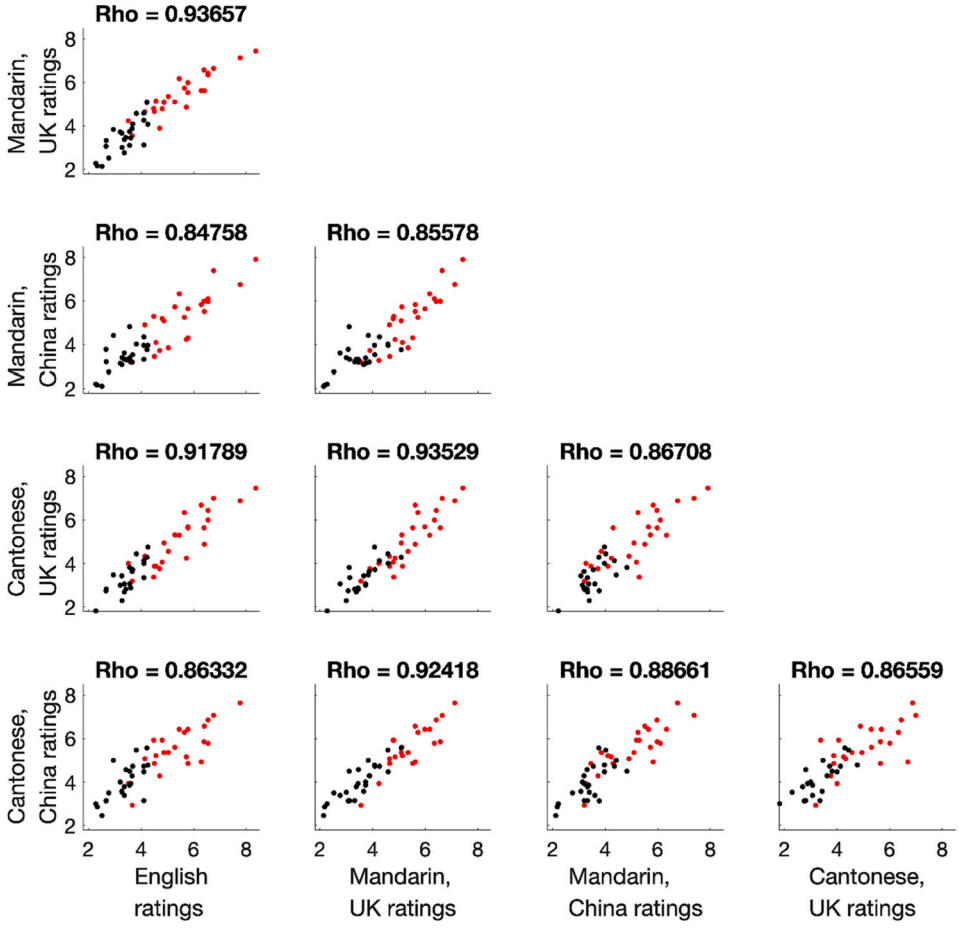

Fig. 3. Average musicality ratings across participants for each of 24 illusion (red) and 24 control (black) stimuli after 8 repetitions, compared across all five groups. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

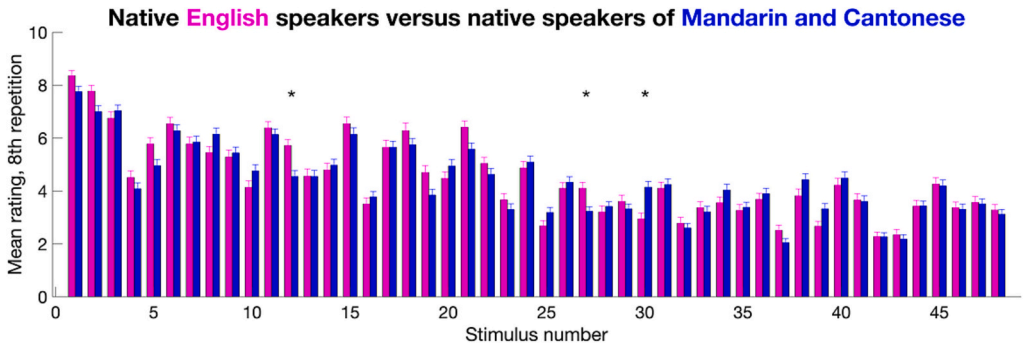

Fig. 4. Average musicality ratings across participants for each of 48 stimuli after 8 repetitions, collapsed across the native Chinese-language speakers (blue, $n = 131$) and compared to the native English speakers (magenta, $n = 95$). Error bars indicate one standard error of the mean. Stimulus numbers 1 through 24 correspond to the pre-defined illusion stimuli, while 25 through 48 correspond to the control stimuli. Asterisks over stimuli 12, 27, and 30 indicate stimuli for which there was a significant group difference in musicality rating after FDR correction for multiple comparisons. Stimulus transcriptions and audio files of stimuli are available in supplementary data (see links in Appendix A), along with a version of this figure with separate bars for native English, Mandarin and Cantonese speakers (Fig. S1). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

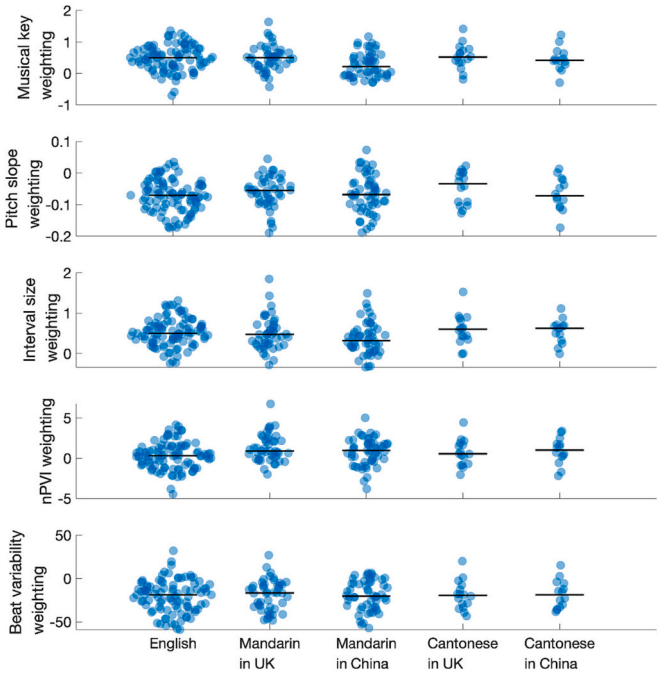

Fig. 5. Cue weights for judgment of stimulus musicality for five different features (musical key fit, pitch slope, interval size, nPVI, and beat variability) across all five groups. Black horizontal lines indicate the median for each group, and dots show data for individual participants. Y-axis values represent beta coefficients for each predictor in a linear regression predicting musicality ratings after the eighth repetition for each participant.

stimuli we averaged the musicality rating after the eighth repetition across all participants within each group. (Note that we could not extract beat variability ratings for six of the stimuli due to their brevity, and so they were not included in this analysis.) Next, for each group we ran a linear regression with musicality rating as the outcome measure and five predictors: musical key fit, pitch slope, interval size, nPVI, and beat variability. Regression coefficients and *p*-values can be found in Table 2 for the native English-speaking participants and in Table 3 for the native Chinese-language-speaking participants. For native English speakers, the model predicted 48.4% of the variance in musicality ratings ($F(1,36) = 6.75$, $p < .001$), and all predictors except for nPVI contributed significantly to explaining variance in musicality ratings. For native Chinese speakers, the model predicted 42.4% of the variance in musicality ratings ($F(1,36) = 5.29$, $p < .001$), and three predictors

Table 2
Regression coefficients and significance values for cues predicting cross-stimulus differences in musicality ratings after the eighth repetition, averaged across the native English-speaking participants.

	Coefficient	Std. error	t value	*p* value
(Intercept)	8.73	0.97	8.98	<0.001
Musical key fit	0.48	0.22	2.18	0.036
Pitch slope	−0.07	0.02	−4.19	<0.001
Pitch interval size	0.53	0.22	2.37	0.023
nPVI	0.27	1.06	0.26	0.797
Beat variability	−20.81	6.73	−3.09	0.004

Table 3
Regression coefficients and significance values for cues predicting cross-stimulus differences in musicality ratings after the eighth repetition, averaged across the native Mandarin- and Cantonese-speaking participants.

	Coefficient	Std. error	t value	p value
(Intercept)	7.64	0.95	8.07	<0.001
Musical key fit	0.41	0.22	1.88	0.068
Pitch slope	−0.06	0.02	−3.64	<0.001
Pitch interval size	0.47	0.22	2.15	0.038
nPVI	0.97	1.03	0.94	0.35
Beat variability	−18.98	6.55	−2.90	0.006

contributed significantly to explaining variance in musicality ratings (pitch slope, pitch interval size, and beat variability). One predictor (musical key fit) was marginally significant in explaining variance, and nPVI did not predict significant variance in musicality ratings. Note that the overall pattern of beta coefficients is highly similar between the two groups, even though musical key fit crosses the significance threshold in one group but not the other.

4. Discussion

We find that the speech-to-song illusion is perceived robustly by native Mandarin, Cantonese, and English speakers: a set of illusion phrases taken from audiobooks in English sounded more song-like than a set of control examples, and this difference in musicality increased with repetition in all of our participant groups (Fig. 1). Thus, speaking a tone language is not an impediment to perceiving the illusion. When the strength of the illusion was compared between native tone-language and non-tone-language speakers, there was anecdotal evidence in favour of the null hypothesis of no difference between groups. Therefore, we cannot conclusively say that the tone-language speakers heard the illusion to the same degree as non-tone-language speakers. There is a trend

for illusion strength to be slightly weaker in the former group (Fig. 2). However, any potential difference in illusion strength between the groups is at best very small. As shown in Fig. 4, across the 48 illusion and control phrases in this study there is a striking similarity in musicality ratings made by native tone-language and non-tone language speakers after hearing eight repetitions of each phrase. Indeed, only one illusion and two control phrases show significant differences in ratings by our native tone-language vs. non-tone-language speakers, after correcting for multiple comparisons.

These findings contrast with those of Jaisin et al. (2016), who found that tone-language speakers (of Mandarin and Thai) experienced the song illusion significantly more weakly than non-tone-language speakers. Several differences between our study and this previous study could help explain the discrepancy. First, Jaisin et al.'s stimuli came from five languages (one phrase each in English, German, Mandarin, Italian, and Thai), while our study used only English phrases as stimuli. Second, as noted by the authors, the magnitude of the illusion in their stimuli was relatively weak. Third, their study had relatively few stimuli and participants, so it is possible that their findings reflect these relatively small sample sizes. Fourth, our study included both transforming illusion examples and comparatively less-transforming control examples, enabling us to distinguish between rating bias (our "musical prior") and the strength of the illusion itself.

Our inclusion of native English and native Mandarin and Cantonese speakers, including participants living in China, enabled us to ask whether cues to musicality are similar across three linguistic groups. To do so, we collapsed across participants within each group to see whether the different groups ranked the musicality of the spoken phrases after repetition similarly. Across groups there was very strong agreement in which examples sounded like song after repetition, with rho values averaging around 0.9 (Fig. 3). This suggests that the cues which listeners use when deciding if a stimulus is song or speech may not only be similar across Western listeners, as has been shown in previous work (Falk et al., 2014; Tierney et al., 2018), but may also generalize to two groups who speak different Chinese tone languages. This cross-cultural agreement in the relative musicality of speech after repetition is broadly consistent with prior research indicating several consistent cross-cultural characteristics of music (Savage, Brown, Sakai, & Currie, 2015) and song (Mehr et al., 2019), and with recent cross-cultural research on the acoustic factors differentiating speech and song (Albouy et al., 2023; Hilton et al., 2022; Ozaki et al., 2023).

Having determined that the language groups broadly rank our stimuli similarly in terms of musicality, we next ran a follow-up analysis to examine the weighting of several potential cues to perception of speech versus song. For two of these factors, there was moderate evidence in favour of the null hypothesis of no difference between groups: both tone-language and non-tone-language speakers were more likely to rate a looped speech phrase as sounding like song after the final repetition if it featured small pitch intervals and steady beats. Moreover, for two other factors, there was anecdotal evidence in favour of the null hypothesis: both language groups were more likely to rate an example as sounding like song if the set of pitches across syllables fit a musical key, and if there were relatively flat pitch contours within a syllable, but we cannot say conclusively that they weighted these factors to the exact same degree. The finding that these cues are used similarly across both western and non-western cultures matches the results of a recent study of cross-cultural acoustic differences between speech and song, which found that song, compared to speech, tends to have more stable within-syllable pitches and more regular rhythms (Ozaki et al., 2023). Interestingly, however, this same study found that song does not tend to have smaller pitch intervals than speech cross-culturally. Given this, it remains to be determined why pitch interval size was consistently used as a cue to song perception across native English, Mandarin, and Cantonese speakers (as well as in the separate group of native English speakers tested in Tierney et al., 2018). One possibility is that it is not interval size per se that is important as a cue to musicality perception in song illusion phrases, but that this feature is confounded with another, more useful cue. For example, it is possible that, due to the brevity of the phrases (mean duration of ~1.4 s), having several syllables at roughly the same pitch assists with extraction of a tonal center; this may be less important for real-world song perception, as listeners would usually be exposed to much longer examples of song or speech.

Originally we predicted that Mandarin and Cantonese speakers would up-weight pitch-based cues to perception of song versus speech, including pitch interval size and pitch slope. This prediction was based on previous findings that tone-language speakers upweight pitch cues during auditory categorization, both within the speech domain (Jasmin et al., 2021; Yu & Andruski, 2010; Zhang et al., 2008; Zhang & Francis, 2010) as well as when categorizing musical beats as strong versus weak (Jasmin et al., 2021; Petrova et al., 2024). One explanation for this finding, based on attention-to-dimension theories of speech categorization (Francis & Nusbaum, 2002; Gordon, Eberhardt, & Rueckl, 1993; Holt, Tierney, Guerra, Laffere, & Dick, 2018), is that for Mandarin speakers pitch is more salient, i.e. tends to exogenously capture attention. If this is so, we would predict pitch-based cues to be broadly upweighted across many different categorization tasks within multiple domains, including speech, music, and environmental sounds. However, here we find that in a cross-domain categorization task, there is no difference between tone-language and non-tone-language speakers in pitch weighting. This argues against tone-language experience being linked to a global increase in pitch salience. What, then, could explain the finding that tone-language speakers up-weight pitch as a cue during musical beat perception? One possible explanation is that this reflects a transfer of perceptual strategies from the perception of lexical stress (for which tone-language speakers tend to up-weight pitch) to the perception of musical beats, given the similar role which musical beat strength and lexical stress play in hierarchically marking prominence. If so, one might expect tone-language speakers to up-weight pitch only for categorization tasks for which there is a clear analogue in speech, such as perception of beat strength and musical phrase boundaries, but not for categorization tasks without a speech analogue, such as cross-domain categorization or perception of the material of objects giving rise to impact sounds (Lutfi & Liu, 2007).

We find a moderate group difference in the extent to which pairwise syllabic durational variability is used as a cue to musicality. However, this difference contrasted with our predictions: we had originally predicted that native English speakers would rate low-nPVI examples as more musical, while Mandarin and Cantonese speakers would not use this cue. Instead, we found that native English speakers did not use this cue at all, while for tone-language speakers, higher-nPVI syllabic rhythm was linked to greater song perception. The lack of use of nPVI as a cue by native English speakers is in alignment with recent cross-cultural research on the acoustics of song versus speech, which showed only a very slight tendency for song to have lower nPVI than speech (Ozaki et al., 2023). It remains unclear, however, why the native Mandarin and Cantonese speakers have a greater tendency to rate stimuli with higher nPVI as more songlike compared to native English speakers. One possibility is that Chinese vocal music may feature higher pairwise variability than speech, resisting the overall cross-cultural trend. If so, then our Chinese participants, who grew up exposed to Chinese vocal music, may have learned to associate high nPVI with song. There is currently no direct evidence for this in the literature, but there is some preliminary evidence for greater pairwise variability of Chinese versus English folk song (Yang & Ding, 2021). Given that Mandarin speech tends to have lesser pairwise variability than English speech (Mok, 2009), this suggests that nPVI may indeed be higher for Mandarin and Cantonese song compared to speech, a prediction that could be investigated by future research.

However, we advise caution in interpreting the group difference in nPVI weighting between native English speakers and native Mandarin/Cantonese speakers for two reasons. First, it is not yet known if the nPVI reliably differentiates the temporal patterning of syllables in English vs.

Mandarin/Cantonese speech and song in large samples of data, in part because nPVI is influenced by other variables including elicitation method (read versus spontaneous speech) and inter-speaker variability (Arvaniti, 2012). Second, when collapsing across participants, we found that nPVI was not a significant cue to song perception in either the native English or the tone-language speaker groups. This suggests that the group difference in nPVI weighting, albeit significant, may not be large enough to have any practical consequences.

A limitation of our study is that our stimuli solely consisted of English phrases, which was a more familiar language to the native English vs. tone-language speakers. This difference in familiarity could potentially drive any group differences in either overall speech-to-song perception or in musicality cue weighting. Indeed, prior work has shown that the speech-to-song illusion is enhanced in less familiar languages, especially languages that are difficult to pronounce (Margulis et al., 2015), and that the illusion is enhanced when hearing phrases in a non-native language (Rathcke et al., 2021). However, our main finding is that native English speakers and tone-language speakers are strikingly similar in their perception of the illusion: they report the same degree of transformation into song, rank stimuli similarly in musicality after repetition, and largely use the same cues when deciding whether a stimulus sounds more like song or speech. This finding suggests that the use of native versus non-native (but familiar) speech is unlikely to have a strong effect on perception of the illusion. One way to test this idea would be to run a follow-up study in which native speakers of English and Mandarin (with exposure to both languages) rate the musicality of song illusion and control phrases in the two languages. Interestingly, it remains to be demonstrated that native speakers of Mandarin can experience the illusion when listening to looped phrases in their own language. Indeed, Jaisin et al. (2016) found that neither native Mandarin nor Thai speakers experienced the song illusion when hearing a looped phrase in their native language, even though native German and Italian speakers experienced the illusion when hearing these same phrases. Thus there is clearly room for more work on perception of the song illusion in tone languages by native tone-language speakers. Such work should attend to possible differences in spoken pitch patterns between tone-languages and non-tone languages that could be relevant to song illusion perception. For example, Eady (1982) compared fundamental frequency (F0) trajectories in Mandarin vs. American English speech and found that Mandarin had a greater mean rate of F0 change and more F0 fluctuations as a function of time and number of syllables (cf. Ding, Hoffmann, & Hirst, 2016; Keating & Kuo, 2012; Yuan & Liberman, 2014). This could make it harder to find song illusion phrases in Mandarin than in English, given that within-syllable pitch slope plays a role in influencing which spoken phrases transform to sounding like song when looped.

Stepping back to the larger picture, by using an acoustic illusion in which certain spoken phrases perceptually transform into song when looped, we find striking similarities in the way native tone-language and non-tone-language speakers weight acoustic cues when deciding if English phrases are speech or song. This is consistent with the idea that human brains may have evolved specialized neural mechanisms to recognize and separately process these two ubiquitous forms of human communication.

CRediT authorship contribution statement

Magdalena Kachlicka: Data curation, Formal analysis, Methodology, Writing – original draft. **Aniruddh D. Patel:** Conceptualization, Project administration, Supervision, Writing – review & editing. **Fang Liu:** Project administration, Investigation, Data curation. **Adam Tierney:** Conceptualization, Formal analysis, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

None.

Data availability

Data is shared in the manuscript via a link to OSF

Acknowledgments

The authors would like to thank Wanqi Wang (Shanghai Normal University), Cunmei Jiang (Shanghai Normal University), Xianjun Huang (Capital Normal University), Linshu Zhou (Shanghai Normal University), Li (Alice) Wang (Chinese University of Hong Kong), Emily Haoyan Ge (Hong Kong Metropolitan University), Yaoyao Ruan (University of Oxford) and Chaoqun Zheng (Concordia University) who helped us recruit participants in China and Hong Kong. Special thanks go to Yan Cai (University of Reading) and Athena Lai (University of Reading) for their help in adapting our task instructions into Mandarin and Cantonese. We also thank Pat Keating for directing us to the work of Ding, Hoffmann, and Hirst. This research was funded by an Economic and Social Research Council (ESRC) grant (ES/V007955/1) to AT and a European Research Council (ERC) Starting Grant (CAASD, 678733) to FL.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://osf.io/kbj7u/ and the audio stimuli can be found at https://osf.io/t4pjq/ https://doi.org/10.1016/j.cognition.2024.105757

References

Albouy, P., Mehr, S., Hoyer, R., Ginzburg, J., & Zatorre, R. (2023). Spectro-temporal acoustical markers differential speech from song across cultures. *bioRxiv*. https://doi.org/10.1101/2023.01.29.526133

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

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics, 40*, 351–373.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48.

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.

Bidelman, G. M., & Lee, C. C. (2015). Effects of language experience and stimulus context on the neural organization and categorical perception of speech. *Neuroimage, 120*, 191–200.

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

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam, 17*, 97–110.

Boersma, P., & Weenink, D. (2022). Praat: doing phonetics by computer [Computer program]. Version 6.3.03. retrieved 17 December 2022 from http://www.praat.org/.

Caldwell-Harris, C., Lancaster, A., Ladd, D., Dediu, D., & Christiansen, M. (2015). Factors influencing sensitivity to lexical tone in an artificial language. *Studies in Second Language Acquisition, 37*, 335–357.

Chrabaszcz, C., Winn, M., Lin, C., & Idsardi, W. (2014). Acoustic cues to perception of word stress by English, mandarin, and Russian speakers. *Journal of Speech, Language, and Hearing Research, 57*, 1468–1479.

Creel, S., Weng, M., Fu, G., Heyman, G., & Lee, K. (2018). Speaking a tone language enhances musical pitch perception in 3-5-year-olds. *Developmental Science, 21*, Article e12503.

Daniele, J. R., & Patel, A. D. (2013). An empirical study of historical patterns in musical rhythm: Analysis of German & Italian classical music using the nPVI equation. *Music Perception: An Interdisciplinary Journal, 31*(1), 10–18.

Deutsch, D., Henthorn, T., & Lapidis, R. (2011). Illusory transformation from speech to song. *The Journal of the Acoustical Society of America, 129*, 2245–2252.

Ding, H., Hoffmann, R., & Hirst, D. (2016, May). Prosodic transfer: A comparison study of Ff0 patterns in L2 English by Chinese speakers. In *Proceedings of speech prosody* (pp. 756–760). https://doi.org/10.21437/SpeechProsody.2016-155. Boston, MA.

Eady, S. J. (1982). Differences in the F0 patterns of speech: Tone language versus stress language. *Language and Speech, 25*(1), 29–42.

Eimas, P. (1975). Auditory and phonetic coding of the cues for speech: Discrimination of the [r-l] distinction by young infants. *Perception & Psychophysics, 18*, 341–347.

Ellis, D. (2007). Beat tracking by dynamic programming. *Journal of New Musical Research, 36*(1), 51–60.

Ellis, R., & Jones, M. (2009). The role of accent salience and joint accent structure in meter perception. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 264–280.

Falk, S., Rathcke, T., & Dalla Bella, S. (2014). When speech sounds like music. *Journal of Experimental Psychology: Human Perception and Performance, 40*, 1491–1506.

Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance, 28*(2), 349.

Fung, R. S., & Lee, C. K. (2019). Tone mergers in Hong Kong Cantonese: An asymmetry of production and perception. *The Journal of the Acoustical Society of America, 146*(5). EL424-EL430.

Gelman, A. (2008). Scaling regression units by dividing by two standard deviations. *Statistics in Medicine, 27*, 2865–2873.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gordon, P. C., Eberhardt, J. L., & Rueckl, J. G. (1993). Attentional modulation of the phonetic significance of acoustic cues. *Cognitive Psychology, 25*(1), 1–42.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven, & N. Warner (Eds.), *7. Papers in laboratory phonology* (pp. 515–546).

Hannon, E., Snyder, J., Eerola, T., & Krumhansl, C. (2004). The role of melodic and temporal cues in perceiving musical meter. *Journal of Experimental Psychology: Human Perception and Performance, 30*, 956–974.

Hanson, H. M. (2009). Effects of obstruent consonants on fundamental frequency at vowel onset in English. *The Journal of the Acoustical Society of America, 125*(1), 425–441.

Harris, I., Niven, E. C., Griffin, A., & Scott, S. K. (2023). Is song processing distinct and special in the auditory cortex? *Nature Reviews Neuroscience, 24*(11), 711–722.

Hilton, C. B., Moser, C. J., Bertolo, M., Lee-Rubin, H., Amir, D., Bainbridge, C. M., … Mehr, S. A. (2022). Acoustic regularities in infant-directed speech and song across cultures. *Nature Human Behaviour, 6*(11), 1545–1556.

Holt, L. L., Tierney, A. T., Guerra, G., Laffere, A., & Dick, F. (2018). Dimension- selective attention as a possible driver of dynamic, context- dependent re-weighting in speech processing. *Hearing Research, 366*, 50–64.

Hutka, S., Bidelman, G., & Moreno, S. (2015). Pitch expertise is not created equal: Cross-domain effects of musicianship and tone language experience on neural and behavioural discrimination of speech and music. *Neuropsychologia, 71*, 52–63.

Idemaru, K., & Holt, L. (2014). Specificity of dimension-based statistical learning in word recognition. *JEP:HPP, 40*, 1009–1021.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition, 87*(1), B47–B57.

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

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

Jiang, J., Chen, M., & Alwan, A. (2006). On the perception of voicing in syllable-initial plosives in noise. *Journal of the Acoustical Society of America, 119*, 1092–1105.

Keating, P., & Kuo, G. (2012). Comparison of speaking fundamental frequency in English and mandarin. *The Journal of the Acoustical Society of America, 132*(2), 1050–1060.

Khouw, E., & Ciocca, V. (2007). Perceptual correlates of Cantonese tones. *Journal of Phonetics, 35*(1), 104–117.

Krumhansl, C. (2001). *Cognitive foundations of musical pitch*. NY: Oxford University Press.

Lee, M., & Wagenmakers, E. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Liu, J., Hilton, C., Bergelson, E., & Mehr, S. (2023). Language experience predicts music processing in a half-million speakers of fifty-four languages. *Current Biology, 33*, 1916–1925.

Low, E. L., Grabe, E., & Nolan, F. (2000). Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English. *Language and Speech, 43*(4), 377–401.

Lutfi, R. A., & Liu, C. J. (2007). Individual differences in source identification from synthesized impact sounds. *The Journal of the Acoustical Society of America, 122*(2), 1017–1028.

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

McGowan, R. W., & Levitt, A. G. (2011). A comparison of rhythm in English dialects and music. *Music Perception, 28*(3), 307–314.

McPherson, M. J., & McDermott, J. H. (2018). Diversity in pitch perception revealed by task dependence. *Nature Human Behaviour, 2*(1), 52–66.

Mehr, S., Singh, M., Knox, D., Ketter, D., Pickens-Jones, D., Atwood, S., … Glowacki, L. (2019). Universality and diversity in human song. *Science, 366*, eaax0868.

Mok, P. (2009). On the syllable-timing of Cantonese and Beijing mandarin. *Chinese Journal of Phonetics, 2*, 148–154.

Mok, P. P., Zuo, D., & Wong, P. W. (2013). Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Language Variation and Change, 25*(3), 341–370.

Moritz, S., & Bartz-Beielstein, T. (2017). inputeTS: Time series missing value imputation in R. *The R Journal, 9*, 207–218.

Nolan, F., & Asu, E. L. (2009). The pairwise variability index and coexisting rhythms in language. *Phonetica, 66*, 64–77.

Norman-Haignere, S. V., Feather, J., Boebinger, D., Brunner, P., Ritaccio, A., McDermott, J. H., … Kanwisher, N. (2022). A neural population selective for song in human auditory cortex. *Current Biology, 32*(7), 1470–1484.

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.

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.

Ong, J. H., Wong, P., & Liu, F. (2020). Musicians show enhanced perception, but not production, of native lexical tones. *The Journal of the Acoustical Society of America, 148*(6), 3443–3454.

Ozaki, Y., Tierney, A., Pfordresher, P. Q., McBride, J., Benetos, E., Proutskouva, P., … Savage, P. E. (2023). (Accepted ["Recommended"]). Globally, songs and instrumental melodies are slower, higher, and use more stable pitches than speech [Stage 2 Registered Report]. In *Peer Community In Registered Reports*. https://doi.org/10.31234/osf.io/jr9x7. Preprint.

Patel, A. D. (2008). *Music, language, and the brain*. NY: Oxford Univ. Press.

Patel, A. D., Iversen, J. R., & Rosenberg, J. C. (2006). Comparing the rhythm and melody of speech and music: The case of British English and French. *The Journal of the Acoustical Society of America, 119*(5), 3034–3047.

Petrova, K., Jasmin, K., Saito, K., & Tierney, A. (2024). *Extensive residence in a second language environment modifies perceptual strategies for suprasegmental categorization*. JEP:HPP.

Prince, J. (2014). Pitch structure, but not selective attention, affects accent weightings in metrical grouping. *JEP:HPP, 40*, 2073–2090.

Rathcke, T., Falk, S., & Dalla Bella, S. (2021). Music to your ears: Sentence sonority and listener background modulate the "speech-to-song illusion". *Music Perception: An Interdisciplinary Journal, 38*(5), 499–508.

Sankaran, N., Carlson, T. A., & Thompson, W. F. (2020). The rapid emergence of musical pitch structure in human cortex. *Journal of Neuroscience, 40*(10), 2108–2118.

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

Schellenberg, G., & Trehub, S. (1996). Natural musical intervals: Evidence from infant listeners. *Psychological Science, 7*, 272–277.

Schultz, B. G., O'Brien, I., Phillips, N., McFarland, D. H., Titone, D., & Palmer, C. (2016). Speech rates converge in scripted turn-taking conversation. *Applied PsychoLinguistics, 37*, 1201–1220.

Signorell, A., Aho, K., Alfons, A., Anderegg, N., Aragon, T., Arppe, A., Baddeley, A., Barton, K., Bolker, B., & Borchers, H. (2019). *DescTools: Tools for descriptive statistics. R package version 0.99, 28* (p. 17).

Sluijter, A. M., Van Heuven, V. J., & Pacilly, J. J. (1997). Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America, 101*, 503–513.

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*(2), 249–254.

Tierney, A., Patel, A. D., & Breen, M. (2018). Acoustic foundations of the speech-to-song illusion. *Journal of Experimental Psychology: General, 147*(6), 888–903.

Tierney, A., Patel, A. D., Jasmin, K., & Breen, M. (2021). Individual differences in perception of the speech-to-song illusion are linked to musical aptitude but not musical training. *Journal of Experimental Psychology: Human Perception and Performance, 47*(12), 1681.

Van Khe, T. (1977). Is the pentatonic universal? A few reflections on pentatonism. *The World of Music, 19*, 76–84.

Wong, P. C., Chandrasekaran, B., & Zheng, J. (2012). The derived allele of ASPM is associated with lexical tone perception. *PLoS One, 7*(4), Article e34243.

Yang, L., & Ding, H. (2021). Comparing the rhythm of instrumental music and vocal music in mandarin and English. In *12th international symposium on Chinese spoken language processing (ISCSLP)* (pp. 1–5). IEEE.

Yip, M. (2002). *Tone*. NY: Cambridge University Press.

Yu, V., & Andruski, J. (2010). A cross-language study of perception of lexical stress in English. *Journal of Psycholinguistic Research, 39*, 323–344.

Yuan, J., & Liberman, M. (2014). F0 declination in English and mandarin broadcast news speech. *Speech Communication, 65*, 67–74.

Zhang, Y., & Francis, A. (2010). The weighting of vowel quality in native and non-native listeners' perception of English lexical stress. *Journal of Phonetics, 38*, 260–271.

Zhang, Y., Nissen, S., & Francis, A. (2008). Acoustic characteristics of English lexical stress produced by native mandarin speakers. *Journal of the Acoustical Society of America, 123*, 4498–4513.