frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 26 May 2023
DOI 10.3389/fpsyg.2023.1167003

Check for updates

OPEN ACCESS

EDITED BY
Dan Zhang,
Tsinghua University, China

REVIEWED BY
Yue Ding,
Shanghai Mental Health Center, China
Juan Huang,
Johns Hopkins University, United States

*CORRESPONDENCE
Christina Vanden Bosch der Nederlanden
✉ c.dernederlanden@utoronto.ca

RECEIVED 15 February 2023
ACCEPTED 09 May 2023
PUBLISHED 26 May 2023

CITATION
Yu CY, Cabildo A, Grahn JA and Vanden Bosch der Nederlanden CM (2023) Perceived rhythmic regularity is greater for song than speech: examining acoustic correlates of rhythmic regularity in speech and song. *Front. Psychol.* 14:1167003.
doi: 10.3389/fpsyg.2023.1167003

Perceived rhythmic regularity is greater for song than speech: examining acoustic correlates of rhythmic regularity in speech and song

Chu Yi Yu[1,2], Anne Cabildo[3], Jessica A. Grahn[1,2] and Christina M. Vanden Bosch der Nederlanden[1,2,3]*

[1]The Brain and Mind Institute, Western University, London, ON, Canada, [2]Department of Psychology, Western University, London, ON, Canada, [3]Department of Psychology, University of Toronto, Mississauga, ON, Canada

Rhythm is a key feature of music and language, but the way rhythm unfolds within each domain differs. Music induces perception of a beat, a regular repeating pulse spaced by roughly equal durations, whereas speech does not have the same isochronous framework. Although rhythmic regularity is a defining feature of music and language, it is difficult to derive acoustic indices of the differences in rhythmic regularity between domains. The current study examined whether participants could provide subjective ratings of rhythmic regularity for acoustically matched (syllable-, tempo-, and contour-matched) and acoustically unmatched (varying in tempo, syllable number, semantics, and contour) exemplars of speech and song. We used subjective ratings to index the presence or absence of an underlying beat and correlated ratings with stimulus features to identify acoustic metrics of regularity. Experiment 1 highlighted that ratings based on the term "rhythmic regularity" did not result in consistent definitions of regularity across participants, with opposite ratings for participants who adopted a beat-based definition (song greater than speech), a normal-prosody definition (speech greater than song), or an unclear definition (no difference). Experiment 2 defined rhythmic regularity as how easy it would be to tap or clap to the utterances. Participants rated song as easier to clap or tap to than speech for both acoustically matched and unmatched datasets. Subjective regularity ratings from Experiment 2 illustrated that stimuli with longer syllable durations and with less spectral flux were rated as more rhythmically regular across domains. Our findings demonstrate that rhythmic regularity distinguishes speech from song and several key acoustic features can be used to predict listeners' perception of rhythmic regularity within and across domains as well.

KEYWORDS
rhythmic regularity, beat, speech, song, music information retrieval, periodicity, rhythm

Introduction

Rhythm is crucial for the perception and production of vocal communication in both music and language. In language, syllable rhythms aid in the segmentation of speech (Cutler and Butterfield, 1992; Dilley and McAuley, 2008), convey the meaning of the speaker through prosodic stress (e.g., sarcasm, Cheang and Pell, 2008), illustrate the presence of a foreign speakers' accent (Polyanskaya et al., 2017), and support simultaneous acquisition of multiple languages in infancy (Werker and Byers-Heinlein, 2008). In music, rhythm contributes to melodic identity (Jones et al., 1987; Hébert and Peretz, 1997), enables beat perception (Povel and Essens, 1985; Parncutt, 1994), impacts perceived groove in music (Matthews et al., 2019), and provides the structure that allows synchronization with music or other people (Fitch, 2016). Rhythm is clearly an important feature for both language and music, but the way that rhythm is realized in each domain—that is, how rhythm unfolds in time—is different.

Rhythm, in both music and language, can be defined as the pattern of 'events' in time (McAuley, 2010; Ravignani and Madison, 2017). Events in language typically occur at the syllable level, and events in music occur at the note level. Music and language differ in how the time intervals between events are structured. In musical rhythms, events are usually structured around a beat, or an underlying pulse (Drake, 1998; McAuley, 2010). Even though individual events are not equally spaced, the intervals between events relate to the beat, which means that durations are most commonly related by small integer ratios like 1:2 (e.g., quarter note:half note). The beat in music leads to the perception that the intervals between beats are roughly the same duration (i.e., isochronous; Ravignani and Madison, 2017; Ravignani and Norton, 2017) and gives listeners the sense of periodicity, or the perception of a pattern repeating regularly at a fixed period or interval in time (Patel, 2003; Patel et al., 2005; Kotz et al., 2018). Periodicity is present in music even despite natural tempo fluctuations or expressive timing that make a strictly isochronous beat improbable in human produced music (Fraisse, 1982; Epstein, 1985; Bharucha and Pryhor, 1986). In contrast, speech rhythms do not have a beat. It is this presence of a beat that we call rhythmic regularity.

Despite a long history of searching for strictly periodic intervals at the syllable or stress level in speech, no one has found regularly repeating patterns of equal duration in speech (Grabe and Low, 2002; Patel, 2003; Patel et al., 2005; Cummins, 2012; Goswami and Leong, 2013; Brown et al., 2017). Although speech sounds are generally considered rhythmic, those rhythms are constrained to the length of the word, linguistic stress pattern, syntactic rules, or prosodic emphasis in a sentence (Cutler and Foss, 1977; Hay and Diehl, 2007; Turk and Shattuck-Hufnagel, 2013), which does not lend well to rhythmic regularity. These temporal regularities are crucial for speech intelligibility (Shannon et al., 1995) and more crucial than spectral characteristics of speech (Albouy et al., 2020). Speakers learn the typical rhythmic patterns of their language and this knowledge gives rise to temporal predictability in speech (Rosen, 1992; Hawkins, 2014; Jadoul et al., 2016; Rathcke et al., 2021), rather than any rhythmic regularities in the speech signal (Beier and Ferreira, 2018). The differences in regularity between music and language are especially salient when comparing sensorimotor synchronization to speech and song, where speech has much greater variability in the alignment of taps to syllable events in speech (30%) compared to note events song (4%, Lidji et al., 2011; Cummins, 2012; Dalla Bella et al., 2013).

In each domain, there is considerable research characterizing the degree or type of rhythmic information in the signal. These studies ask, for instance, whether language is rhythmic at all (e.g., Nolan and Jeon, 2014) or what acoustic factors contribute to the strength of perceived regularity in music (e.g., Bouwer et al., 2018). A range metrics have been used to characterize rhythm and/or regularity within each domain and, in a few cases, across domains. These metrics include the calculation of inter-onset-intervals between successive notes or syllables (e.g., stressed and unstressed IOIs; Vanden Bosch der Nederlanden et al., 2022a,b), durational contrastiveness between pairs of successive notes or syllables (Pairwise Variability Index; Grabe and Low, 2002; Patel and Daniele, 2003; Hannon, 2009; Hannon et al., 2016), the proportion of vocalic intervals in an utterance (vowel reduction; Grabe and Low, 2002; Wiget et al., 2010; Arvaniti, 2012), acoustic feature extraction using music information retrieval techniques (e.g., Lartillot and Toiviainen, 2007; Lartillot et al., 2008; Alluri and Toiviainen, 2010; Burger et al., 2013, 2014), autocorrelations to detect self-similarity in the envelope of a signal (Leong, 2012; Suppanen et al., 2019), clock timing evidence and counter-evidence (Povel and Essens, 1985), and integer multiple relatedness (Roeske et al., 2020; De Gregorio et al., 2021). These metrics have been useful within their own contexts of identifying, for example, whether a composer's language background influenced the musical rhythms they employed (Patel and Daniele, 2003; Van Handel, 2006) or determining the strength of a beat in one musical rhythm compared to another (Henry et al., 2017; Matthews et al., 2019). However, not all speech-rhythm metrics have proven to be reliable or strong predictors of perceived speech rhythms (White and Mattys, 2007; Arvaniti, 2012; Jadoul et al., 2016). In music, the task of beat extraction is difficult (McKinney et al., 2007; Grosche et al., 2010), even if humans do it spontaneously (Grahn and Brett, 2007; Honing, 2012). The goal of the current paper is to examine whether some of the above metrics used to characterize rhythmic regularity in music or language separately can characterize the differences in rhythmic regularity *between* language and music.

Past work has examined where in the acoustic signal the beat is located in speech and song, finding consistent tapping in speech and song at p-centers (but see conflicting takes on p-centers Morton et al., 1976; *cf.* Marcus, 1981; Vos and Rasch, 1981; Pompino-Marschall, 1989; Scott, 1998; Villing et al., 2007), vowel onsets (Rathcke et al., 2021), or at peaks in the acoustic envelope (Kochanski and Orphanidou, 2008; Lartillot and Grandjean, 2019). Still others have used cochlear models of acoustic salience to find the beat location in vocally-produced songs (Ellis, 2007; Coath et al., 2009). While these approaches are germane to the current question, our goal is to determine whether acoustic features of speech and song can eventually provide evidence of rhythmic regularity—in the form of an equally-spaced, repeating pulse—in a range of communicative and non-communicative domains. For instance, there is increasing evidence that regularity is a salient feature in the sensory landscape (Aman et al., 2021), with listeners detecting regularity within a single cycle of it emerging from a random background (Southwell and Chait, 2018) or preferentially attending to a visual stream with statistical regularities despite having no conscious perception of that regularity (Zhao et al., 2013). Stimuli in studies like these are created with careful control over what features should give rise to regularity, but a wide range of natural stimuli, including non-human animal vocalizations (Kotz et al., 2018; Roeske et al., 2020; De Gregorio et al., 2021) and environmental sounds (e.g., Gygi et al., 2004; Rothenberg, 2013) also give rise to regularity in a variety of different acoustic characteristics. Our goal is to find a metric that indexes the differences in regularity between speech and song with the future goal of using this metric to detect the degree of regularity in a range of naturally occurring sounds.

Acoustic features that differentiate temporal regularity in speech and song will also feed into perceptual and cognitive questions related to how humans differentiate speech and song in development (Vanden Bosch der Nederlanden et al., 2022a,b). Rhythmic regularity is an important feature for speech-to-song or environmental sound-to-song transformations (Simchy-Gross and Margulis, 2018; Tierney et al., 2018; Rowland et al., 2019), but spectral features seem to be better predictors of a listeners' perception of an utterance as speech or song (Hilton et al., 2022; Vanden Bosch der Nederlanden et al., 2022a,b; Albouy et al., 2023; Ozaki et al., 2023). Given the importance of rhythmic differences between and among languages for helping children acquire language (Ramus et al., 1999; Nazzi et al., 2000; Jusczyk, 2002), and for bringing about a transformation from speech to song, a clear acoustic metric of rhythmic regularity may prove useful for understanding the development of distinct domains of communication.

We address the goals in the current study by first obtaining subjective ratings of the differences in rhythmic regularity between spoken and sung utterances. After establishing this subjective metric, acoustic features of spoken and sung utterances were related to subjective ratings of rhythmic regularity to examine which features are most predictive of perceived rhythmic regularity.

Experiment 1

Participants

Thirty-three 18- to 24-year-old participants (16 males) participated in the study. An additional 7 people participated in the study but were excluded because they did not complete the study ($N = 5$ did not provide a rating for at least 90% of the rating trials, $N = 2$ did not pass attention checks within the survey; see Procedure). A third of participants reported taking music lessons and a third of participants self-reported being bilingual, but most participants were English monolinguals who learned English from birth (see Supplementary Table S1). About half of participants identified as white. Participants were recruited from the University of Western Ontario undergraduate psychology participant pool and were required to speak English fluently and have no known hearing deficits. All participants were compensated with course credit and provided informed consent to participate. All materials were approved by Western University's Research Ethics Board (REB).

Stimuli

One set of sung and spoken utterances was used for Experiment 1. We used a stimulus set generated for a different study (see Vanden Bosch der Nederlanden et al., 2022a,b). For purposes related to the previous studies' need for acoustic control, the spoken and sung utterances were acoustically matched on several features, including the sentence texts (see Appendix A), speaker identity, total duration (utterance length), tempo (syllable rate), pitch contour, RMS amplitude, and number of syllables. In total, this stimulus set included 96 stimuli (48 unique texts), 48 spoken, 48 sung, with 3 male speakers (American and British English accents). The stimuli ranged from 1.62 to 3.86 s in length with an average of approximately 2.46 s. For details on stimulus creation please see Vanden Bosch der Nederlanden et al. (2022a).

Procedure

Participants accessed the online study using Qualtrics (2021) and completed a regularity rating task and a background demographics questionnaire. In the rating task, participants heard each spoken or sung sentence presented in random order in a single block. The presentation order of spoken and sung utterances was not constrained, so participants could hear multiple spoken or sung utterances in a row. On each trial, participants rated each audio clip according to how rhythmically regular it sounded (see Appendix B1), using a rating scale of 1 (not very regular) to 9 (very regular). Two catch trials were randomly presented to ensure participants were paying attention. The audio in these catch trials gave explicit instructions for ratings. For example, if the catch trial audio said "This is a test trial. Please select number 3 on the slider below," the participant should have moved the slider to 3 before proceeding to the next trial. Immediately after the rating task, participants were asked to write out their own definition of rhythmic regularity in an open text box. Participants completed a demographic background questionnaire at the end. On average, participants completed the study in 33.61 min.

Results

Rhythmic regularity ratings were averaged separately for spoken and sung utterances. Ratings were normally distributed, with skewness and kurtosis ratings between +/−3. Average ratings were submitted to a one-way repeated-measures Analysis of Variance (ANOVA) with Utterance (Speech, Song) as the main factor. As illustrated in Figure 1A, regularity ratings did not differ between speech and song, $F(1, 32) = 1.044$, $p = 0.314$, $\eta^2 = 0.032$. However, we provided no training or guidance on what rhythmic regularity was. To capture whether participants' definition of rhythmic regularity influenced their ratings, we thematically coded each listener's self-reported definition of "rhythmic regularity" and identified 3 groups: beat-based, normal-prosody, and unclear definitions. Participants were grouped into beat-based definitions if they mentioned the words "beat" or "meter" and/or discussed the importance of rhythmic consistency (e.g., even spacing). Participants were grouped into normal-prosody definitions if they discussed linguistic stress, prosodic pitch, rhyme, and that regularity depended on sounding normal for conversation (e.g., normal speed/tempo/flow for speech). Finally, participants were placed in the unclear definition group if their definition was not based on acoustic factors (e.g., annoyance, familiarity), was not a definition (e.g., about what the goal of the study was), or had a definition that could be either beat or prosody based (see Supplementary Table S2). In the end, 12 listeners had beat-based definitions, 11 listeners had normal-prosody definitions, and 10 listeners had unclear definitions of rhythmic regularity. A follow-up 2 (Utterance: speech, song) by 3 (Definition: beat, prosody, unclear) ANOVA again showed no main effect of utterance type (speech vs. song), $F(1,30) = 1.934$, $p = 0.175$, $\eta_p^2 = 0.061$, but there was a significant interaction with definition, $F(2, 30) = 6.606$, $p = 0.004$, $\eta_p^2 = 0.306$. As illustrated in Figure 1B, the normal-prosody group rated speech as more rhythmically regular than song, $F(1, 10) = 7.085$, $p = 0.024$, $\eta^2 = 0.415$, while the beat-based group rated song as more rhythmically regular than speech, $F(1, 11) = 4.963$, $p = 0.048$, $\eta^2 = 0.311$, and the unclear group did not reliably differentiate regularity in speech and

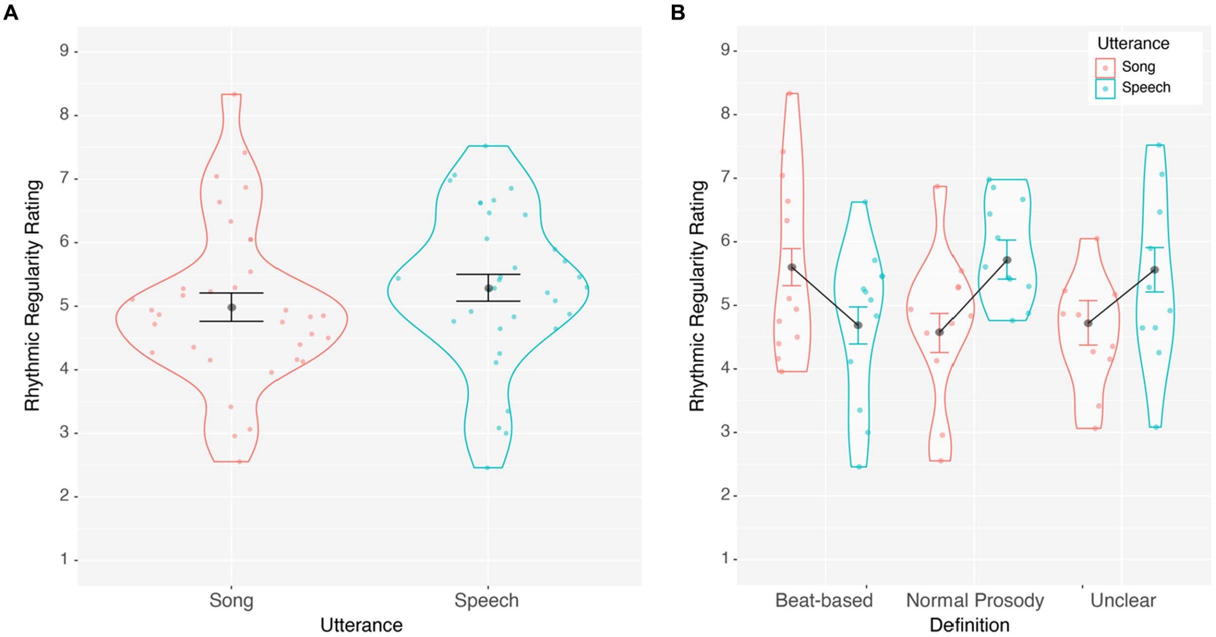

FIGURE 1
(A) Average rhythmic regularity rating of song and speech illustrating no difference in regularity ratings and **(B)** a significant interaction illustrating that speech and song regularity ratings were dependent on participants' definition of rhythmic regularity. Standard Error is within-subjects error (Morey, 2008).

song, $F(1, 9) = 2.846$, $p = 0.126$, $\eta^2 = 0.240$. These results suggest that the perceived rhythmic regularity of speech and song differed based on participants', sometimes inaccurate, definition of rhythmic regularity.

Interim discussion

Experiment 1 illustrated that participants had varying definitions of rhythmic regularity when we left it undefined and did not provide training examples. Initially it appeared that our acoustically matched stimuli did not differ in perceived rhythmic regularity, but after taking participants' definitions into account (whether their definition was beat-based, normal-prosody, or unclear), regularity was greater for song than speech for beat-based definitions and greater for speech than song for normal-prosody definitions. Note that the normal-prosody definition group did not describe prosodic rhythmic regularity or a beat in speech, but rather participants in this group largely based their definitions only on the regular part of the term rhythmic regularity. Instead, these participants focused on how normal the speech sounded for everyday conversations. Although definition groupings explained a significant amount of variability in regularity ratings, it is also possible that the acoustic constraints placed on the stimuli reduced the differences in rhythmic regularity between spoken and sung exemplars. In this case, different profiles of regularity for speech and song in Experiment 1 may mean stimuli did not differ or only weakly differed in rhythmic regularity. We designed Experiment 2 to determine whether providing a clear definition of rhythmic regularity would shift participants' ratings to align with the beat-based definition of rhythmic regularity we set out to examine in addition to determining whether regularity ratings were consistent across different stimulus sets.

We improved on Experiment 1 in three ways: (1) We provided a concrete rhythmic regularity rating scale "How easy would it be to tap or clap along to that clip?" (2) We provided training examples before participants began the rating task consisting of spoken and sung clips that would be easy and not easy to tap or clap to using familiar stimuli, and (3) We added a second unmatched stimulus set of spoken and sung stimuli that were not acoustically matched to examine regularity differences between unconstrained spoken and sung exemplars.

A second goal of Experiment 2 was to relate participants' regularity ratings to acoustic features of spoken and sung exemplars. To achieve this goal, speech- and music-based acoustic features were extracted from all stimuli using Praat, MIR Toolbox, and custom music-inspired scripts (see OSF). We used standard acoustic features that are known to differ between speech and song (Vanden Bosch der Nederlanden et al., 2022a,b), as well as several features described in the introduction related to temporal regularity (see Appendix D for full feature list).

Experiment 2

Participants

Fifty-one participants (13 males) between the ages of 17–24 years of age participated. An additional 6 individuals participated but were excluded because they did not pass all attention checks (see Procedure). Note that one included participant passed attention checks but did not respond to 2 trials in the acoustically matched stimulus set. About a quarter of the participants reported musical training (see Supplementary Table S3). Almost a third of participants self-reported being bilingual, but most participants were English monolinguals and learned English from birth (see Supplementary Table S3). About half of participants identified as white (see Supplementary Table S3). Participants were recruited from the University of Western Ontario undergraduate psychology participant pool and were required to be English speakers and have no known hearing deficits. All participants were compensated with course credit and provided informed consent to participate. All materials were approved by Western University's Research Ethics Board (REB).

Stimuli

Experiment 2 included the acoustically matched stimulus set from Experiment 1 and an unmatched stimulus set created for this study. This additional stimulus set addressed the possibility that matched spoken and sung utterances did not differ on rhythmic regularity because of the constraints placed on tempo, duration, contour in their recording process. The unmatched stimulus set consisted of short clips pulled from several free sources on the internet including audiobooks.org ($N = 15$), looperman.com ($N = 7$), ccmixter.org ($N = 12$), Soundcloud.com ($N = 2$), the SiSEC database ($N = 8$; Liutkus et al., 2017), and a previous paper examining music and language comparisons ($N = 1$; Albouy et al., 2020). Podcast recordings ($N = 15$) were sampled from spotify.com under the fair dealing and educational exceptions to copyright (Copyright Act, R.S.C., 1985). The unmatched stimuli ranged from 1.84 to 3.71 s in length, with an average of 2.38 s in duration, on average. A total of 60 sentences (see Appendix C) were retrieved from the above sources, with half spoken and half sung recordings of solo voices (no instruments in the sung versions). Sentence text and speaker were not matched in this unmatched set, so no sentences were repeated. Although these stimuli were not matched for overall duration, pitch, etc., they were equated for total RMS amplitude. The acoustic features and derived rhythm metrics are reported for each stimulus set separately in Table 1, and the description and method for extracting each feature is reported in Appendix B.

Procedure

The procedure was similar to Experiment 1, except that the stimuli from the unmatched and matched datasets were blocked and rated separately from one another. Participants were asked to wear headphones and complete the surveys in a distraction-free environment. The same order–matched stimulus set, followed by the unmatched stimulus set–was used for all participants so as not to increase variability in ratings across stimulus sets and for maximal comparison to Experiment 1. Prior to each rating task, participants heard a training section with 4 training stimuli that provided examples of spoken and sung utterances that were easy and hard to clap to. Training utterances were spoken and sung by a single male speaker using the text and melody of the familiar children's song "Twinkle, Twinkle, Little Star" (Taylor and Taylor, 1806), and were labeled as "Song" or "Speech" and "Easy to tap or clap along to" or "Not easy to clap or tap along to." Easy to tap/clap utterances were sung with a strict metrical pulse or spoken like a poem with a clear prosodic metrical foot alternation. The other stimuli were performed with temporal irregularities including saying words quickly and with irregular pauses between words to disrupt any perception of a beat. Participants could listen to these examples as many times as they wanted and had to listen to all 4 to move forward in the survey. For each stimulus in the rating task, participants rated "How easy would it be to clap or tap to that clip?" with a rating scale of "1 = Not Very Easy" through to "9 = Very Easy." As before, participants could listen to the clips as many times as they wanted but had to listen at least once to move forward. Participants completed an unrelated task [the SSS test reported in Assaneo et al. (2019)] between the matched and unmatched ratings, but those data are beyond the scope of the current paper and are not reported here. The same two catch ("attention check") trials were used from Experiment 1 and were randomly incorporated in each block (4 in total). Finally, participants filled out a demographic background questionnaire.

Results

Rhythmic regularity ratings were averaged separately for spoken and sung utterances in both the matched and unmatched stimulus sets and submitted to a 2 (Utterance: speech, song) by 2 (Stimulus set: matched, unmatched) repeated-measures ANOVA. Song was rated as more rhythmically regular than speech, $F(1, 50) = 39.490$, $p < 0.001$, $\eta_p^2 = 0.441$, and matched stimuli had higher regularity ratings than unmatched stimuli, $F(1, 50) = 21.089$, $p < 0.001$, $\eta_p^2 = 0.297$. However, a significant interaction between stimulus set and utterance, $F(1, 50) = 13.899$, $p < 0.001$, $\eta_p^2 = 0.218$, suggested that the effect of utterance type was larger in the unmatched than the matched set, as illustrated in Figure 2. Simple effects revealed that for matched stimuli, song ratings were higher than speech ratings by 0.874 units on the rating scale, $F(1, 50) = 20.863$, $p < 0.001$, $\eta^2 = 0.294$. For the unmatched stimuli, song ratings were higher than speech by 1.696 units on the rating scale, $F(1, 50) = 40.338$, $p < 0.001$, $\eta^2 = 0.447$. Overall, song was consistently rated as more rhythmically regular than speech, but this difference was larger for unmatched compared to matched utterances. These findings indicate that a clear definition of rhythmic regularity allows listeners to be sensitive to rhythmic regularity as a distinguishing feature between music and language. Participants were sensitive to differences in rhythmic regularity in acoustically constrained settings as well, when features that are typically correlated with regularity, like tempo, are held constant across spoken and sung exemplars.

Correlating rhythmic measures with subjective ratings

To examine which acoustic features best predicted listeners' rhythmic regularity ratings, we included features that were correlated with regularity ratings in a linear mixed effects model. First, we performed first order correlations among all the extracted metrics (see Method and Supplementary Table S4) despite redundancy across rhythmic measures. Unmatched spoken and sung utterances differed greatly in the number of syllables (fewer for song than speech), which affected several other metrics including average syllable duration and metrics related to syllable or vocalic/consonant onsets. We performed separate first order correlations for matched and unmatched stimulus sets to ensure that features correlated in one set but not another due

TABLE 1 Acoustic features extracted for all matched and unmatched stimuli, using Praat-based linguistic metrics, Music Information retrieval metrics from MIR Toolbox, and music-inspired regularity metrics.

		Matched			Unmatched		
		Speech	Song	*P*	Speech	Song	*P*
Praat-based metrics	F0	138.45 (20.09)	138.15 (11.41)	0.930	158.88 (55.86)	277.53 (75.59)	<0.001
	F0 instability	1.40 (0.50)	0.68 (0.14)	<0.001	1.23 (0.38)	0.97 (0.34)	0.006
	Total duration	2.43 (0.33)	2.49 (0.37)	0.381	2.29 (0.23)	2.48 (0.42)	0.030
	Syllable duration	0.26 (0.04)	0.27 (0.04)	0.196	0.21 (0.04)	0.39 (0.11)	<0.001
	Stressed duration	0.37 (0.08)	0.37 (0.09)	0.771	0.31 (0.12)	0.43 (0.24)	0.020
	Vocalic nPVI	53.61 (14.49)	54.44 (16.35)	0.792	59.66 (16.73)	72.02 (26.37)	0.035
	Consonantal PVI	117.87 (39.83)	108.93 (32.83)	0.233	95.20 (51.39)	184.16 (71.25)	<0.001
	Stress syllable nPVI	51.07 (15.24)	51.88 (13.74)	0.784	51.95 (21.11)	67.59 (32.94)	0.033
	Syllable nPVI	61.96 (15.32)	57.00 (15.16)	0.114	55.39 (14.83)	65.06 (23.23)	0.060
	%V	0.49 (0.07)	0.55 (0.08)	<0.001	0.48 (0.08)	0.66 (0.09)	<0.001
	ΔC	0.08 (0.02)	0.07 (0.02)	0.154	0.07 (0.03)	0.08 (0.04)	0.288
	ΔV	0.07 (0.02)	0.07 (0.02)	0.002	0.06 (0.02)	0.21 (0.10)	<0.001
Music information retrieval	Spectral flux	45.17 (4.42)	38.81 (4.12)	<0.001	107.44 (40.93)	90.44 (20.45)	0.048
	Sub-band flux 1	1.36 (1.37)	1.01 (0.56)	0.107	1.36 (0.67)	1.397 (1.01)	0.880
	Sub-band flux 2	1.36 (1.37)	1.01 (0.56)	0.107	4.85 (3.89)	1.11 (0.37)	<0.001
	Sub-band flux 3	7.13 (2.03)	6.33 (2.42)	0.080	40.58 (32.16)	13.74 (17.82)	<0.001
	Sub-band flux 4	13.27 (3.01)	10.71 (2.47)	<0.001	46.03 (21.90)	33.14 (16.98)	0.014
	Sub-band flux 5	20.56 (4.44)	17.75 (4.21)	0.002	28.92 (8.87)	26.65 (12.85)	0.429
	Sub-band flux 6	12.95 (3.63)	11.28 (3.58)	0.026	19.36 (7.57)	24.93 (15.61)	0.085
	Sub-band flux 7	10.84 (3.59)	10.45 (3.92)	0.614	13.99 (7.15)	20.47 (9.31)	0.004
	Sub-band flux 8	6.56 (2.37)	5.76 (2.25)	0.091	8.21 (4.24)	10.53 (6.28)	0.099
	Sub-band flux 9	2.82 (1.42)	2.10 (0.90)	0.004	7.92 (6.73)	12.33 (9.52)	0.022
	Pulse clarity (Max)	0.22 (0.10)	0.23 (0.10)	0.757	0.23 (0.08)	0.23 (0.09)	0.948
	Pulse clarity (Min)	0.16 (0.06)	0.16 (0.06)	0.571	0.20 (0.06)	0.19 (0.05)	0.603
	Tempo (autocorr)	127.87 (36.19)	132.70 (36.93)	0.524	116.90 (26.70)	110.47 (30.63)	0.390
	Tempo (spectrum)	146.14 (29.10)	144.77 (26.68)	0.812	141.08 (28.74)	126.55 (27.83)	0.054
Music-inspired metrics	Integer multiple	0.35 (0.20)	0.36 (0.20)	0.893	0.37 (0.16)	0.36 (0.27)	0.925
	Asynchrony	0.12 (0.12)	0.11 (0.13)	0.703	0.16 (0.12)	0.16 (0.13)	0.366
	Asynchrony SD	0.12 (0.11)	0.12 (0.12)	0.743	0.14 (0.12)	0.12 (0.12)	0.489
	Signed asynchrony	0.04 (0.14)	0.02 (0.15)	0.444	0.11 (0.19)	0.06 (0.16)	0.324
	Signed SD	0.14 (0.12)	0.13 (0.13)	0.791	0.16 (0.12)	0.14 (0.13)	0.458

Average (st dev) value for spoken and sung exemplars, the *p*-value (uncorrected paired samples *t*-test) characterizes whether the metric differed for speech and song.

to syllable number had the opportunity to be entered into the model (see Supplementary Table S4). Several first order correlation features were highly correlated with other predictors, such that F0, syllable duration, stressed interval, %V, consonantal PVI, and ΔV were all correlated with one another (all *rs* > 0.3, see Supplementary Table S5). To reduce multicollinearity, the feature that was most highly correlated with rhythmic regularity was entered for model testing (i.e., average syllable duration, see Supplementary Table S4). Spectral flux was correlated with each sub-band flux metric. Total spectral flux was chosen for model testing over any sub-band measure because overall flux correlated consistently with rhythmic regularity in each stimulus set, while sub-band flux correlations were present or absent depending on the stimulus set. The final features entered into the model were F0 instability, total duration, average syllable duration, and spectral flux (but see Supplementary Table S6 for additional analyses using consonantal PVI and %V instead of syllable duration). All measures were mean-centered and any measures with kurtosis or skewness (+/−3) were log-transformed and mean-centered before being entered into the model.

Participant ID and Stimulus ID were entered as random effects, with 1 spectral and 3 temporal features added as fixed effects. These fixed effects significantly improved the fit of the basic model (see Table 2, Model 1), but duration did not uniquely contribute to the model. After removing duration, Model 2 accounted for a significant amount of variance compared to the random effects model and Model 1 did not account for more variance than the Model 2 ($p = 0.743$).

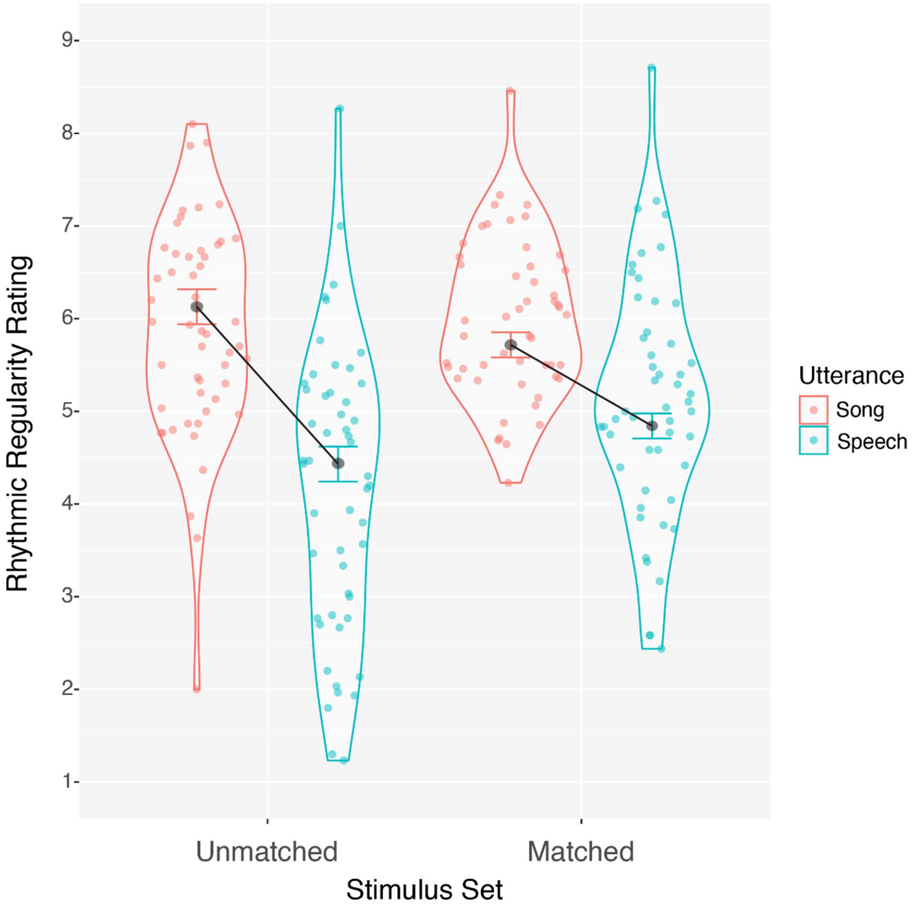

FIGURE 2
Average rhythmic regularity ratings for song and speech grouped by matched and unmatched stimulus sets, within-subjects standard error (Morrey, 2008).

Model 3 included syllable count to ensure that predictors were robust to the small number of syllables present in sung utterances from the unmatched condition. Syllable count did not significantly improve fit compared to Model 2 (see Table 2, Model 3), and did not change the significance of average syllable duration. Finally, Model 4 examined whether the acoustic features from Model 2 would remain significant even after adding speech and song labels into the model (utterance type). F0 Instability was no longer significant in this final model, presumably because F0 stability was more predictive of speech-song differences than regularity within stimulus classes. Thus, in addition to songs having greater rhythmic regularity than speech, stimuli with longer syllable durations and less spectral flux were rated as more rhythmically regular (Figure 3).

Interim discussion

A major goal of Experiment 2 was to standardize participants' interpretation of rhythmic regularity by providing a concrete definition centered on ease of clapping or tapping along with the stimulus. With this definition, rhythmic regularity ratings were significantly higher for sung than spoken utterances. Experiment 2 also expanded on the acoustically matched stimulus set from Experiment 1 by including an additional unmatched stimulus set more representative of speech and song in everyday settings. Participants rated song as more rhythmically regular than speech for both sets, but the difference was larger for the acoustically unmatched than the matched set. Naturally recorded utterances may emphasize the differences in regularity between song and speech compared to recordings that equate tempo, pitch contour, and average pitch between speech and song. However, regularity differences are apparent even in carefully acoustically matched stimulus sets, suggesting that regularity helps differentiate speech and song. Finally, we estimated which acoustic features across both stimulus sets were most predictive of regularity ratings. Although the type of stimulus (speech or song) was a significant predictor of regularity, longer syllable durations and less spectral flux also predicted higher rhythmic regularity ratings.

General discussion

The goal of this work was to obtain a subjective metric of rhythmic regularity—an equally-spaced, repeating pulse—and examine acoustic features that predict participants' ratings of regularity. Experiment 1 illustrated that the term rhythmic regularity was interpreted differently across participants, leading to different patterns of regularity across

TABLE 2 LME models predicting rhythmic regularity.

Model	Variable	Estimate	t-value	P
Model 1:	Duration	0.074	0.323	0.7469
	Syllable duration	**3.270**	**4.668**	**<0.0001**
	spectral flux	**−0.008**	**−3.521**	**0.0006**
	F0 instability	**−0.467**	**−2.980**	**0.0034**
X^2(8, *N* = 7,954) = 61.254, *p* < 0.001, AIC = 32,857 (*compared to random intercept model*)				
Model 2:	Syllable duration	**3.372**	**5.412**	**<0.0001**
	Spectral flux	**−0.008**	**−3.625**	**0.0004**
	F0 instability	**−0.466**	**−2.986**	**0.0033**
X^2(7, *N* = 7,954) = 61.1468, *p* < 0.001, AIC = 32,855 (*compared to random intercept model*)				
Model 3:	Syllable count	−0.018	−0.299	0.7718
	Syllable duration	**3.121**	**2.985**	**0.0033**
	Spectral flux	**−0.008**	**−3.546**	**0.0005**
	F0 instability	**−0.467**	**−2.985**	**0.0033**
X^2(8, *N* = 7,954) = 0.0921, *p* = 0.7615, AIC = 32,857 (*compared to model 2*)				
Model 4:	Utterance type (speech)	**−0.980**	**−5.297**	**<0.0001**
	Syllable duration	**1.483**	**2.194**	**0.0298**
	Spectral flux	**−0.008**	**−4.363**	**<0.0001**
	F0 instability	−0.095	−0.531	0.5965
X^2(8, *N* = 7,954) = 26.464, *p* < 0.0001, AIC = 32,830 (*compared to model 2*)				

Model 4 is the best fitting model, with syllable duration and spectral flux predicting rhythmic regularity even after accounting for stimulus type (speech vs. song). Model 3 and Model 4 illustrate syllable duration and spectral flux are robust predictors of rhythmic regularity even after accounting for the number of syllables in an utterance and stimulus type. Bold variables indicate significant predictors to the model.

spoken and sung exemplars. Experiment 2 operationalized the definition of rhythmic regularity by asking how easy it would be to tap or clap to the stimulus. With this definition, participants rated song as more regular–or easier to clap or tap to–than speech in both acoustically matched and acoustically unmatched stimulus sets. Subjective regularity ratings were significantly affected by acoustic features of syllable duration and spectral flux, with longer durations and less flux related to higher regularity ratings. These results add to the literature by (1) highlighting the salience of rhythmic regularity as a differentiator of speech and song (Patel and Daniele, 2003; Patel et al., 2005; Vanden Bosch der Nederlanden et al., 2022b) and (2) adding to the growing literature on spectral flux as a salient acoustic feature in listeners' perceptual processing of sound (Weineck et al., 2022).

Spectral flux is a metric of the distance between successive frames, or moments in time, in the frequency spectrum, with larger values indicating large changes in the spectrum from moment to moment (Alluri and Toiviainen, 2010). It logically follows that song should have less spectral flux since notes are held longer (i.e., greater proportion of the utterance is vocalic) than in speech, creating fewer changes in the spectrum on a moment-to-moment basis. The metrical framework of sung utterances may also make for fewer sudden and more evenly spaced changes in the spectrum compared to speech. Spectral flux has been described as an acoustic correlate of the beat in music, but with greater spectral flux indicating greater beat salience (Burger et al., 2013). These authors extracted spectral flux from low and high frequency bands in the spectrum corresponding to the kick drum, hi-hat, and cymbal. For this reason, large amounts of spectral flux in these bands acted as a proxy for rhythmic information from these instruments. These stimulus-specific differences help to explain the seeming paradox of greater spectral flux predicting more beat salience in music, while greater spectral flux predicts less rhythmic regularity when comparing speech to song.

Our results elucidate what features participants use to provide regularity ratings when comparing speech and song, but these features alone are unlikely to capture the presence of a beat or the integer multiple relatedness of sounds snapping to the metrical grid across a wide range of environmental stimuli. We attempted to account for listeners' subjective regularity ratings using several music- and language-inspired metrics of regularity. In particular, the proportion of intervals per sentence that were related by integer multiples (Roeske et al., 2020) was not correlated with regularity ratings. It may be that our sentence-level approach is too coarse a metric and behavioral responses like tapping or continuous regularity ratings could shed light on which features participants relied on at particular moments in time to feel a beat (similar to Rathcke et al., 2021). The consistency with which those moments align with inter-onset-interval or stimulus features could provide a path forward for creating novel metrics to characterize regularity differences in speech and song. Another set of metrics used for this study (Asynchrony, Signed Asynchrony and their variability) was inspired by the clock timing work from Povel and Essens (1985) (similar to Norton and Scharff, 2016 for birdsong). However, this metric also failed to provide any relationship to subjective regularity and may also require input from the p-center-related literature (e.g., Rathcke et al., 2021) to determine the correct beat locations and onset times used to develop the underlying "clock" for speech and song. Onset intervals related to vocalic or other salient features of the stimulus may be more fruitful than the reliance on linguistic onsets used here. Finally, music information retrieval metrics like pulse clarity and stimulus-extracted tempo had no relationship to rhythmic regularity in speech and song, suggesting that these feature extraction methods are perhaps better suited for use with multi-instrument (e.g., vocals and instrumentation) excerpts of musical pieces rather than vocal sung and spoken utterances.

Linguistic measures, including measures that have previously been used to relate speech and music to one another, such as nPVI, also did not explain additional variance in rhythmic regularity beyond average syllable duration (see Supplementary Table S6). Vocalic nPVI was originally developed to capture the vowel reduction (i.e., change in vowel quality to a "schwa" and shortened duration of vowel length) that happens in many of the so-called "stress-timed" languages (Grabe and Low, 2002; Patel et al., 2005; Cummins, 2012). This measure is not best at capturing rhythmic variability, but rather contrastiveness between pairs of syllables. Indeed, our calculations indicated that music often had more contrastiveness than speech (see Table 1, Unmatched stimuli), which is likely due to large integer-related duration differences like quarter notes to half or whole notes that speech does not employ. Comparisons of previous work from separate studies suggested that nPVIs were much higher for speech (in the 50–70 range) than instrumental music (in the 30–40 range; Patel and Daniele, 2003; Hannon et al., 2016), but these studies used musical notation to estimate nPVI durations instead of actual recordings. Studies that have used acoustic segmentation of speech and song have illustrated more

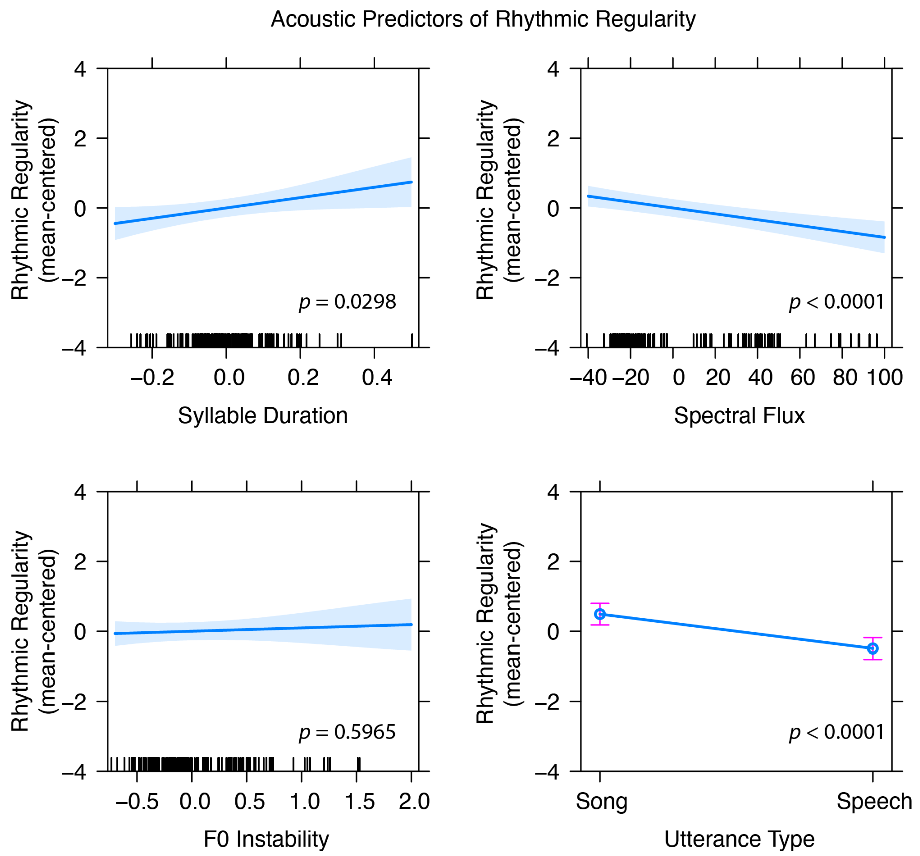

FIGURE 3
Model 4 indicates that syllable duration and spectral flux are significant predictors of perceived rhythmic regularity even after accounting for utterance type (speech vs. song), but F0 instability, which was significant in models without utterance type as a factor, is no longer a significant predictor of perceived regularity. Error bars for utterance type and shaded error regions represent standard error calculated using the Kenward-Roger coefficient covariance matrix (Effect package, R).

comparable nPVI values (Vanden Bosch der Nederlanden et al., 2022a,b). Thus, it is not surprising that this metric did not uniquely predict rhythmic regularity for spoken compared to sung stimuli.

Despite the ease with which humans pick up on regularity in speech, song, and environmental sounds, easily extractable acoustic features that characterize those subjective reports remain elusive. Our study confirms that participants hear more rhythmic regularity in sung compared to spoken utterances, providing concrete metrics for how best to obtain participant's subjective regularity ratings. The findings from this study also add to the literature by characterizing that regularity is easier to detect–or more likely to be present–when syllables are longer, and when there is less moment-to-moment fluctuation in the spectrum. Future work should build on these results to develop more continuous and fine-grained metrics for quantifying rhythmic regularity from the acoustic signal. There is growing evidence that rhythmic regularity is an important signal for attention, perception, development, and movement (Grahn and Brett, 2007; Gordon et al., 2014; Bedoin et al., 2016; Trainor et al., 2018; Aman et al., 2021; Lense et al., 2021) in humans, and is present in a range of human and non-human primate communicative vocalizations (Roeske et al., 2020; De Gregorio et al., 2021), as well as many environmental sounds (Gygi et al., 2004). Indeed, the perception of rhythmic regularity is key to how both human and non-human animals (e.g., cockatoos, sea lions) align their movements to a beat (Fitch, 2013). A greater understanding of what acoustic features humans rely on to perceive regularity and extract an underlying pulse in communicative signals like speech and song will contribute to theories of evolutionary origins of beat processing (e.g., are the features humans use to find a beat the same or different from animals?) and theories about perceptual biases toward regularity in everyday soundscapes.

One potential limitation of the current study is the use of lyrics in both the music and language domains. We wanted to use speech and song because they exemplify the acoustic and structural differences between domains (Vanden Bosch der Nederlanden et al., 2022a,b), while maintaining the ability to control for timbral, semantic, and other temporal or spectral acoustic features. It will be important to characterize the role linguistic content plays in the perception of rhythmic regularity in song. For instance, is song without words perceived as more strictly regular that song with words given that note durations are less dictated by word length or stress? If so, then are instrumental melodies perceived as more rhythmically regular than songs without words? Or does linguistic or semantic content help to bolster temporal prediction for what type of note and/or word will come next? Similarly, would speech without semantic content (e.g., low-pass filtered) be perceived as more or less regular than semantic speech? This and future work will help shed light on the temporal features that distinguish speech and song and, more broadly, the domains of music and language.

The current findings add to the literature on rhythm in music and language by providing a concrete subjective metric of rhythmic regularity that reliably differs between speech and song across stimulus sets. The metric is simple to understand and can be used to characterize the perception of rhythmic regularity across developmental populations, in individuals with little or no musical training, and in a range of stimulus sets beyond music and language (e.g., bird song). Our findings are important for characterizing the inherent differences in music and language that (1) may be important for learning to differentiate musical and linguistic communication early in development (Vanden Bosch der Nederlanden et al., 2022a,b) and (2) underlie many of the perceptual advantages ascribed to music over language. For instance, cross-culturally humans prefer simple integer ratios in music (Jacoby and McDermott, 2017) and remember these musical rhythms better than syncopated rhythms that disrupt the occurrence of events on a beat (Fitch and Rosenfeld, 2007). Future work comparing the prominence of features in speech compared to song could address the divergence of musical and linguistic communication in humans. For instance, does the preservation of rhythmic regularity in music come at a cost to the transmission of quick messages meant to transact information? Is strict isochrony better for promoting verbatim memory of information occurring on, but not off the beat (Jones et al., 1981; Large, 2008; Helfrich et al., 2018) while vague periodicity without strict isochrony (as in speech) is better for encoding the gist of a message? Answering seemingly simple questions like how humans perceive differences in rhythmic regularity in speech and song, has the potential to address several important areas of psychology related to human communicative development, origins of music and language, cross-species comparisons, and perceptual biases toward regularity in everyday scenes.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://osf.io/hnw5t/.

Ethics statement

The studies involving human participants were reviewed and approved by the University of Western Ontario Ethics Board. The patients/participants provided their written informed consent to participate in this study.

Author contributions

CY, JG, and CV designed the experiments. CY and CV recruited the participants and performed the data analysis. CV and AC extracted acoustic features and manually segmented stimuli. CY wrote the first draft. CV provided subsequent drafts. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by NSERC RGPIN-2016-05834 awarded to JG and NSERC RGPIN-2022-04413 and DGECR-2022-00294 awarded to CV.

Acknowledgments

We would like to thank Patrick Grzela and Olivia Jiaming Xu for their help setting up the online survey and training trials for Experiment 1. Thanks to Teia Tremblett for helping with stimulus organization and identifying stressed syllables. We also wish to thank Renee Ragguet for her assistance programming the music-inspired rhythmic regularity metrics used in this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1167003/full#supplementary-material

References

Albouy, P., Benjamin, L., Morillon, B., and Zatorre, R. J. (2020). Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science* 367, 1043–1047. doi: 10.1126/science.aaz3468

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

Alluri, V., and Toiviainen, P. (2010). Exploring perceptual and acoustical correlates of polyphonic timbre. *Music. Percept.* 27, 223–242. doi: 10.1525/mp.2010.27.3.223

Aman, L., Picken, S., Andreou, L. V., and Chait, M. (2021). Sensitivity to temporal structure facilitates perceptual analysis of complex auditory scenes. *Hear. Res.* 400:108111. doi: 10.1016/j.heares.2020.108111

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *J. Phon.* 40, 351–373. doi: 10.1016/j.wocn.2012.02.003

Assaneo, M. F., Ripollés, P., Orpella, J., Lin, W. M., de Diego-Balaguer, R., and Poeppel, D. (2019). Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning *Nat. Neurosci.* 22, 627–632. doi: 10.1038/s41593-019-0353-z

Bedoin, N., Brisseau, L., Molinier, P., Roch, D., and Tillman, B. (2016). Temporally regular musical primes facilitate subsequent syntax processing children with specific language impairment. *Front. Neurosci.* 10:e00245. doi: 10.3389/fnins.2016.00245

Beier, E. J., and Ferreira, F. (2018). The temporal prediction of stress in speech and its relation to musical beat perception. *Front. Psychol.* 9:431. doi: 10.3389/fpsyg.2018.00431

Bharucha, J. J., and Pryhor, J. H. (1986). Disrupting the isochrony underlying rhythm: an asymmetry in discrimination. *Percept. Psychophys.* 40, 137–141. doi: 10.3758/bf03203008

Bouwer, F. L., Burgoyne, J. A., Odijk, D., Honing, H., and Grahn, J. A. (2018). What makes a rhythm complex? The influence of musical training and accent type on beat perception. *PLoS One* 13:e0190322. doi: 10.1371/journal.pone.0190322

Brown, S., Pfordresher, P. Q., and Chow, I. (2017). A musical model of speech rhythm. *Psychomusicology* 27, 95–112. doi: 10.1037/pmu0000175

Burger, B., Ahokas, J. R., Keipi, A., and Toiviainen, P. (2013). Relationships between spectral flux, perceived rhythmic strength, and the propensity to move. In: *10th Sound And Music Computing Conference*. Stockholm, Sweden.

Burger, B., Thompson, M. R., Luck, G., Saarikallio, S. H., and Toiviainen, P. (2014). Hunting for the beat in the body: on period and phase locking in music-induced movement. *Front. Hum. Neurosci.* 8:e00903:903. doi: 10.3389/fnhum.2014.00903

Cheang, H. S., and Pell, M. D. (2008). The sound of sarcasm. *Speech Commun.* 50, 366–381. doi: 10.1016/j.specom.2007.11.003

Coath, M., Denham, S. L., Smith, L. M., Honing, H., Hazan, A., Holonowicz, P., et al. (2009). Model cortical responses for the detection of perceptual onsets and beat tracking in singing. *Connect. Sci.* 21, 193–205. doi: 10.1080/09540090902733905

Cummins, J. (2012). The intersection of cognitive and sociocultural factors in the development of reading comprehension among immigrant students. *Read. Writ.* 25, 1973–1990. doi: 10.1007/s11145-010-9290-7

Cutler, A., and Butterfield, S. (1992). Rhythmic cues to speech segmentation: evidence from juncture misperception. *J. Mem. Lang.* 31, 218–236. doi: 10.1016/0749-596X(92)90012-

Cutler, A., and Foss, D. J. (1977). On the role of sentence stress in sentence processing. *Lang. Speech* 20, 1–10. doi: 10.1177/002383097702000101

Dalla Bella, S., Białuńska, A., and Sowinski, J. S. (2013). Why movement is captured by music, but less by speech: role of temporal regularity. *PLoS One* 8:e71945. doi: 10.1371/journal.pone.0071945

De Gregorio, C., Valente, D., Raimondi, T., Torti, V., Miaretsoa, L., Friard, O., et al. (2021). Categorical rhythms in a singing primate. *Curr. Biol.* 31, R1379–R1380. doi: 10.1016/j.cub.2021.09.032

Dilley, L. C., and McAuley, J. D. (2008). Distal prosodic context affects word segmentation and lexical processing. *J. Mem. Lang.* 59, 294–311. doi: 10.1016/j.jml.2008.06.006

Drake, C. (1998). Psychological processes involved in the temporal Organization of Complex Auditory Sequences: universal and acquired processes. *Music. Percept.* 16, 11–26. doi: 10.2307/40285774

Ellis, D. P. W. (2007). Beat tracking by dynamic programming. *J. New Music Res.* 36, 51–60. doi: 10.1080/09298210701653344

Epstein, D. (1985). Tempo relations: a cross-cultural study. *Music Theory Spectr* 7, 34–71. doi: 10.2307/745880

Fitch, W. T. (2013). Rhythmic cognition in humans and animals: distinguishing meter and pulse perception. *Front. Syst. Neurosci.* 7:68. doi: 10.3389/fnsys.2013.00068

Fitch, W. T. (2016). Dance, music, meter and groove: a forgotten partnership. *Front. Hum. Neurosci.* 10:64. doi: 10.3389/fnhum.2016.00064

Fitch, W. T., and Rosenfeld, A. J. (2007). Perception and production of syncopated rhythms. *Music. Percept.* 25, 43–58. doi: 10.1525/mp.2007.25.1.43

Fraisse, P. (1982). “Rhythm and tempo” in *Psychology of music*. ed. D. Deutsch (New York: Academic Press), 149–180.

Gordon, R. L., Shivers, C. M., Wieland, E. A., Kotz, S. A., Yoder, P. J., and McAuley, J. D. (2014). Musical rhythm discrimination explains individual differences in grammar skills in children. *Dev. Sci.* 18, 635–644. doi: 10.1111/desc.12230

Goswami, U., and Leong, V. (2013). Speech rhythm and temporal structure: converging perspectives? *Lab. Phonol.* 4, 67–92. doi: 10.1515/lp-2013-0004

Grabe, E., and Low, E. (2002). Durational variability in speech and the rhythm class hypothesis. *Lab. Phonol.* 7, 515–546. doi: 10.1515/9783110197105.2.515

Grahn, J. A., and Brett, M. (2007). Rhythm and beat perception in motor areas of the brain. *J. Cogn. Neurosci.* 19, 893–906. doi: 10.1162/jocn.2007.19.5.893

Grosche, P., Müller, M., and Sapp, C. (2010). What makes beat tracking difficult? A case study on Chopin mazurkas. In: *Proceedings of the 11th International Society for Music Information Retrieval Conference*, ISMIR 2010, Utrecht, Netherlands. 649–654.

Gygi, B., Kidd, G. R., and Watson, C. S. (2004). Spectral-temporal factors in the identification of environmental sounds. *J. Acoust. Soc. Am.* 115, 1252–1265. doi: 10.1121/1.1635840

Hannon, E. E. (2009). Perceiving speech rhythm in music: listeners classify instrumental songs according to language of origin. *Cognition* 111, 403–409. doi: 10.1016/j.cognition.2009.03.003

Hannon, E. E., Lévêque, Y., Nave, K. M., and Trehub, S. E. (2016). Exaggeration of language-specific rhythms in English and French Children's songs. *Front. Psychol.* 7:939. doi: 10.3389/fpsyg.2016.00939

Hawkins, S. (2014). Situational influences on rhythmicity in speech, music, and their interaction. *Philos. Trans. R. Soc. B* 369:20130398. doi: 10.1098/rstb.2013.0398

Hay, J. S. F., and Diehl, R. L. (2007). Perception of rhythmic grouping: testing the iambic/trochaic law. *Percept. Psychophys.* 69, 113–122. doi: 10.3758/BF03194458

Hébert, S., and Peretz, I. (1997). Recognition of music in long-term memory: are melodic and temporal patterns equal partners? *Mem. Cognit.* 25, 518–533. doi: 10.3758/BF03201127

Helfrich, R. F., Fiebelkorn, I. C., Szczepanski, S. M., Lin, J. J., Parvizi, J., and Knight, R. T. et al. (2018). Neural Mechanisms of Sustained Attention Are Rhythmic. *Neuron* 99, 854–865.e5. doi: 10.1016/j.neuron.2018.07.032

Henry, M. J., Herrmann, B., and Grahn, J. A. (2017). What can we learn about beat perception by comparing brain signals and stimulus envelopes? *PLoS One* 12:e0172454. doi: 10.1371/journal.pone.0172454

Hilton, C. B., Moser, C. J., Bertolo, M., Lee-Rubin, H., Amir, D., Bainbridge, C. M., et al. (2022). Acoustic regularities in infant-directed speech and song across cultures. *Nat. Hum. Behav.* 6, 1545–1556. doi: 10.1038/s41562-022-01410-x

Honing, H. (2012). Without it no music: beat induction as a fundamental musical trait. *Ann. N. Y. Acad. Sci.* 1252, 85–91. doi: 10.1111/j.1749-6632.2011.06402.x

Jacoby,, N., and McDermott, J. H. (2017). Integer Ratio Priors on Musical Rhythm Revealed Cross-culturally by Iterated Reproduction. *Curr. Biol.* 27, 359–370. doi: 10.1016/j.cub.2016.12.031

Jadoul, Y., Ravignani, A., Thompson, B., Filippi, P., and de Boer, B. (2016). Seeking temporal predictability in speech: comparing statistical approaches on 18 world languages. *Front. Hum. Neurosci.* 10:e00586. doi: 10.3389/fnhum.2016.00586

Jones, M. R., Kidd, G., and Wetzel, R. (1981). Evidence for rhythmic attention. *J. Exp. Psychol. Hum. Percept. Perform.* 7, 1059–1073. doi: 10.1037//0096-1523.7.5.1059

Jones, M. R., Summerell, L., and Marshburn, E. (1987). Recognizing melodies: a dynamic interpretation. *Q. J. Exp. Psychol.* 39, 89–121. doi: 10.1080/02724988743000051

Jusczyk, P. W. (2002). How infants adapt speech-processing capacities to native-language structure. *Curr. Dir. Psychol. Sci.* 11, 15–18. doi: 10.1111/1467-8721.00159

Kochanski, G., and Orphanidou, C. (2008). What marks the beat of speech? *J. Acoust. Soc. Am.* 123, 2780–2791. doi: 10.1121/1.2890742

Kotz, S. A., Ravignani, A., and Fitch, W. T. (2018). The evolution of rhythm processing. *Trends Cogn. Sci.* 22, 896–910. doi: 10.1016/j.tics.2018.08.002

Large, E. W. (2008). Resonating to Musical Rhythm: Theory and Experiment. *Psychol. Time*. 189–232. doi: 10.1016/B978-0-08046-977-5.00006-5

Lartillot, O., Eerola, T., Toiviainen, P., and Fornari, J. (2008). Multi-feature modeling of pulse clarity: design, validation and optimization. In: *9th International Conference on Music Information Retrieval*. Philadelphia, USA. 521–526.

Lartillot, O., and Grandjean, D. (2019). Tempo and metrical analysis by tracking multiple metrical levels using autocorrelation. *Appl. Sci.* 9:5121. doi: 10.3390/app9235121

Lartillot, O., and Toiviainen, P. (2007). MIR in Matlab (II): a toolbox for musical feature extraction from audio. In: *Proceedings of the 10th International Conference on Digital Audio Effects*. Bordeaux, France. 127–130.

Lense, M. D., Ladányi, E., Rabinowitch, T.-C., Trainor, L. J., and Gordon, R. (2021). Rhythm and timing as vulnerabilities in neurodevelopmental disorders. *Phil. Trans. R. Soc.* 376:20200327. doi: 10.1098/rstb.2020.0327

Leong, V. (2012). Prosodic rhythm in the speech amplitude envelope: Amplitude modulation phase hierarchies (AMPHs) and AMPH models. Doctoral dissertation. University of Cambridge, Cambridge.

Lidji, P., Palmer, C., Peretz, I., and Morningstar, M. (2011). Listeners feel the beat: entrainment to English and French speech rhythms. *Psychon. Bull. Rev.* 18, 1035–1041. doi: 10.3758/s13423-011-0163-0

Liutkus, A., Stöter, F.-R., Rafii, Z., Kitamura, D., Rivet, B., Ito, N., et al. (2017). The 2016 signal separation evaluation campaign. In: *13th International Conference on Latent Variable Analysis and Signal Separation*. Grenoble, France.

Marcus, S. M. (1981). Acoustic determinants of perceptual center (P-center) location. *Percept. Psychophys.* 30, 247–256. doi: 10.3758/bf03214280

Matthews, T. E., Witek, M., Heggli, O. A., Penhune, V. B., and Vuust, P. (2019). The sensation of groove is affected by the interaction of rhythmic and harmonic complexity. *PLoS One* 14:e0204539. doi: 10.1371/journal.pone.0204539

McAuley, J. D. (2010). “Tempo and rhythm” in *Music perception*. eds. M. R. Jones, R. R. Fay and A. N. Popper (New York, NY: Springer), 165–199.

McKinney, M., Moelants, D., Davies, M., and Klapuri, A. (2007). Evaluation of audio beat tracking and music tempo extraction algorithms. *J. New Music Res.* 36, 1–16. doi: 10.1080/09298210701653252

Morey, R. D. (2008). Confidence Intervals from Normalized Data: A correction to Cousineau (2005). *Tutor. Quant. Methods Psychol.* 4, 61–64. doi: 10.20982/tqmp.04.2.p061

Morton, J., Marcus, S., and Frankish, C. (1976). Perceptual centers (P-centers). *Psychol. Rev.* 83, 405–408. doi: 10.1037/0033-295X.83.5.405

Nazzi, T., Jusczyk, P. W., and Johnson, E. K. (2000). Language discrimination by English-learning 5-month-olds: effects of rhythm and familiarity. *J. Mem. Lang.* 43, 1–19. doi: 10.1006/jmla.2000.2698

Nolan, F., and Jeon, H. S. (2014). Speech rhythm: a metaphor? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369:20130396. doi: 10.1098/rstb.2013.0396

Norton, P., and Scharff, C. (2016). "bird Song Metronomics": isochronous Organization of Zebra Finch Song Rhythm. *Front. Neurosci.* 10:309. doi: 10.3389/fnins.2016.00309

Ozaki, Y., Kloots, M. de H., Ravignani, A., and Savage, P. E. (2023). Cultural evolution of 575 music and language *PsyArXiv.* doi: 10.31234/osf.io/s7apx

Parncutt, R. (1994). A perceptual model of pulse salience and metrical accent in musical rhythms. *Music. Percept.* 11, 409–464. doi: 10.2307/40285633

Patel, A. (2003). Rhythm in language and music. *Ann. N. Y. Acad. Sci.* 999, 140–143. doi: 10.1196/annals.1284.015

Patel, A. D., and Daniele, J. R. (2003). An empirical comparison of rhythm in language and music. *Cognition* 87, B35–B45. doi: 10.1016/S0010-0277(02)00187-7

Patel, A. D., Iversen, J. R., Chen, Y., and Repp, B. H. (2005). The influence of metricality and modality on synchronization with a beat. *Exp. Brain Res.* 163, 226–238. doi: 10.1007/s00221-004-2159-8

Polyanskaya, L., Ordin, M., and Busa, M. G. (2017). Relative salience of speech rhythm and speech rate on perceived foreign accent in a second language. *Lang. Speech* 60, 333–355. doi: 10.1177/0023830916648720

Pompino-Marschall, B. (1989). On the psychoacoustic nature of the P-center phenomenon. *J. Phon.* 17, 175–192. doi: 10.1016/S0095-4470(19)30428-0

Povel, D. J., and Essens, P. (1985). Perception of temporal patterns. *Music. Percept.* 2, 411–440. doi: 10.2307/40285311

Qualtrics (2021). Available at: https://www.qualtrics.com

Ramus, F., Nespor, M., and Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition* 73, 265–292. doi: 10.1016/s0010-0277(99)00058-x

Rathcke, T., Lin, C., Falk, S., and Dalla Bella, S. (2021). Tapping into linguistic rhythm. *Lab. Phonol.* 12:11. doi: 10.5334/labphon.248

Ravignani, A., and Madison, G. (2017). The paradox of Isochrony in the evolution of human rhythm. *Front. Psychol.* 8:1820. doi: 10.3389/fpsyg.2017.01820

Ravignani, A., and Norton, P. (2017). Measuring rhythmic complexity: a primer to quantify and compare temporal structure in speech, movement, and animal vocalizations. *J Lang. Evol.* 2, 4–19. doi: 10.1093/jole/lzx002

Roeske, T. C., Tchernichovski, O., Poeppel, D., and Jacoby, N. (2020). Categorical rhythms are shared between songbirds and humans. *Curr. Biol.* 30, 3544–3555.e6. doi: 10.1016/j.cub.2020.06.072

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philos. Trans. Biol. Sci.* 336, 367–373. doi: 10.1098/rstb.1992.0070

Rothenberg, D. (2013). *Bug music: How insects gave us rhythm and noise*. Manhattan, United States: St. Martin's Press.

Rowland, J., Kasdan, A., and Poeppel, D. (2019). There is music in repetition: looped segments of speech and nonspeech induce the perception of music in a time-dependent manner. *Psychon. Bull. Rev.* 26, 583–590. doi: 10.3758/s13423-018-1527-5

Scott, S. K. (1998). The point of P-centres. *Psychol. Res.* 61, 4–11. doi: 10.1007/PL00008162

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science* 270, 303–304. doi: 10.1126/science.270.5234.303

Simchy-Gross, R., and Margulis, E. H. (2018). The sound-to-music illusion: repetition can musicalize nonspeech sounds. *Music Sci* 1:205920431773199. doi: 10.1177/2059204317731992

Southwell, R., and Chait, M. (2018). Enhanced deviant responses in patterned relative to random sound sequences. *Cortex* 109, 92–103. doi: 10.1016/j.cortex.2018.08.032

Suppanen, E., Huotilainen, M., and Ylinen, S. (2019). Rhythmic structure facilitates learning from auditory input in newborn infants. *Infant Behav. Dev.* 57:101346. doi: 10.1016/j.infbeh.2019.101346

Taylor, J., and Taylor, A. (1806). *Rhymes for the nursery*. London: Arthur Hall, Virtue, & Co, 24.

Tierney, A., Patel, A. D., and Breen, M. (2018). Acoustic foundations of the speech-to-song illusion. *J. Exp. Psychol. Gen.* 147, 888–904. doi: 10.1037/xge0000455

Trainor, L. J., Chang, A., Cairney, J., and Li, Y.-C. (2018). Is auditory perceptual timing a core deficit of developmental coordination disorder? *Ann. N. Y. Acad. Sci.* 1423, 30–39. doi: 10.1111/nyas.13701

Turk, A., and Shattuck-Hufnagel, S. (2013). What is speech rhythm? A commentary on Arvaniti and Rodriquez, Krivokapić, and Goswami and Leong. *Lab. Phonol.* 4, 93–118. doi: 10.1515/lp-2013-0005

Van Handel, L. (2006). "Trends in/over time: rhythm in speech and musical melody in 19th-century art song" in *Sound and music computing, 2006* (Marseille: France)

Vanden Bosch Der Nederlanden, C. M., Joanisse, M. F., Grahn, J. A., Snijders, T. M., Schoffelen, J. M., and Schoffelen, J.-M. (2022a). Familiarity modulates neural tracking of sung and spoken utterances. *Neuroimage* 252:119049. doi: 10.1016/j.neuroimage.2022.119049

Vanden Bosch Der Nederlanden, C. M., Qi, X., Sequeira, S., Seth, P., Grahn, J. A., Joanisse, M. F., et al. (2022b). Developmental changes in the categorization of speech and song. *Dev. Sci.* e13346. doi: 10.1111/desc.13346

Villing, R. C., Ward, T., and Timoney, J. (2007). *A review of P-Centre models. [Conference presentation]. Rhythm production and perception workshop*, Dublin, Ireland.

Vos, J., and Rasch, R. (1981). The perceptual onset of musical tones. *Percept. Psychophys.* 29, 323–335. doi: 10.3758/bf03207341

Weineck, K., Wen, O. X., and Henry, M. J. (2022). Neural synchronization is strongest to the spectral flux of slow music and depends on familiarity and beat salience. *Elife* 11:e75515. doi: 10.7554/eLife.75515

Werker, J. F., and Byers-Heinlein, K. (2008). Bilingualism in infancy: first steps in perception and comprehension. *Trends Cogn. Sci.* 12, 144–151. doi: 10.1016/j.tics.2008.01.008

White, L., and Mattys, S. L. (2007). Calibrating rhythm: first language and second language studies. *J. Phon.* 35, 501–522. doi: 10.1016/j.wocn.2007.02.003

Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., and Mattys, S. (2010). How stable are acoustic metrics of contrastive speech rhythm? *J. Acoust. Soc. Am.* 127, 1559–1569. doi: 10.1121/1.3293004

Zhao, J., Al-Aidroos, N., and Turk-Browne, N. B. (2013). Attention is spontaneously biased toward regularities. *Psychol. Sci.* 24, 667–677. doi: 10.1177/0956797612460407