scientific reports

Check for updates

OPEN

Music is scaled, while speech is not: A cross-cultural analysis

Elizabeth Phillips✉ & Steven Brown

Music is well-known to be based on sets of discrete pitches that are combined to form musical melodies. In contrast, there is no evidence that speech is organized into stable tonal structures analogous to musical scales. In the current study, we developed a new computational method for measuring what we call the "scaledness" of an acoustic sample and applied it to three cross-cultural ethnographic corpora of speech, song, and/or instrumental music (n = 1696 samples). The results confirmed the established notion that music is significantly more scaled than speech, but they also revealed some novel findings. First, highly prosodic speech—such as a mother talking to a baby—was no more scaled than regular speech, which contradicts intuitive notions that prosodic speech is more "tonal" than regular speech. Second, instrumental music was far more scaled than vocal music, in keeping with the observation that the voice is highly imprecise at pitch production. Finally, singing style had a significant impact on the scaledness of song, creating a spectrum from chanted styles to more melodious styles. Overall, the results reveal that speech shows minimal scaledness no matter how it is uttered, and that music's scaledness varies widely depending on its manner of production.

One of the hallmark features of music is that it is a combinatorial system that uses sets of discrete pitches as its building blocks[1–6]. Such pitch-sets are referred to as musical scales and are typically comprised of 5–7 pitches per octave, which get combined to form musical melodies. The composition of scales varies both between and within cultures, leading to a remarkable degree of musical diversity globally[7–10], although the use of a scale itself is a statistical universal[11]. Since scales are ubiquitous in vocal music, one might ask whether this same organizing principle applies to speech. While individual syllables in speech are often associated with particular pitches—sometimes fairly discrete pitches[6,12]—there is no evidence that the collection of pitches across the syllables of an utterance comprise a fixed set of recurring tones akin to a musical scale. While speech is indeed melodious—rather than being purely monotonic—its melodies follow general pitch contours or basic intonation patterns, rather than scales with fixed musical intervals[13–15]. For this reason, we introduce the concept of pitch "scaledness" in this article in order to distinguish it from the common concept of pitch "discreteness," since the latter may be applied to not just melodies as a whole, but to individual pitches (i.e., their degree of flatness). Our focus is on the recurrent use of scaled pitches across an audio sample.

Speech and music, despite their different principles for organizing pitch, find a universal coupling in the form of songs with words, which are vocalizations in which people *sing their linguistic utterances*, rather than speak them in the typical manner of conversational speech. Hence, the melodies of these utterances employ scales and sound like musical melodies, just like instrumental music. Sung speech is distinct from another form of singing in which the syllabic carriers for pitch are not meaningful words but are instead "vocables" (i.e., nonsense syllables)[7,16]. Such vocable singing occurs, for example, when people sing a melody using la-la-la. Figure 1 shows an overall picture of the relationship between speech and music, where sung speech comprises a *joint* function in which words and music are combined with one another, creating composite utterances in which the melody of the speech is musical[17–20]. At the far left side of the figure are purely linguistic forms of speech that do not involve musical scales, and at the far right side are purely musical forms that do not involve words, namely vocable singing and instrumental music. The arrow at the bottom of the figure depicts a predicted continuum of scaledness, i.e., tonal structure arising from the use of a recurrent set of pitches. This continuum spans from standard speech to sung speech to instrumental music.

Figure 1 presents further distinctions for both speech and sung speech with regard to the scaledness continuum. First, prosodic forms of speech are typically considered to be more "musical" than standard speech. This applies to the way that a parent talks to an infant[21–25] or that a professional actor or orator declaims their utterances during a public performance[26]. This accentuated manner of speaking is not only louder and slower than standard conversational speech[27,28], but is also higher-pitched and wider in range[23], involving the use of large pitch contours. For example, a mother saying "You're such a *gooood* girl" to her baby would not only utter the phrase in a higher-than-normal register, but would likely put an exaggerated pitch-contour on the focus-

Department of Psychology, Neuroscience & Behaviour, McMaster University, 1280 Main St. West, Hamilton, ON L8S 4K1, Canada. ✉email: phille10@mcmaster.ca

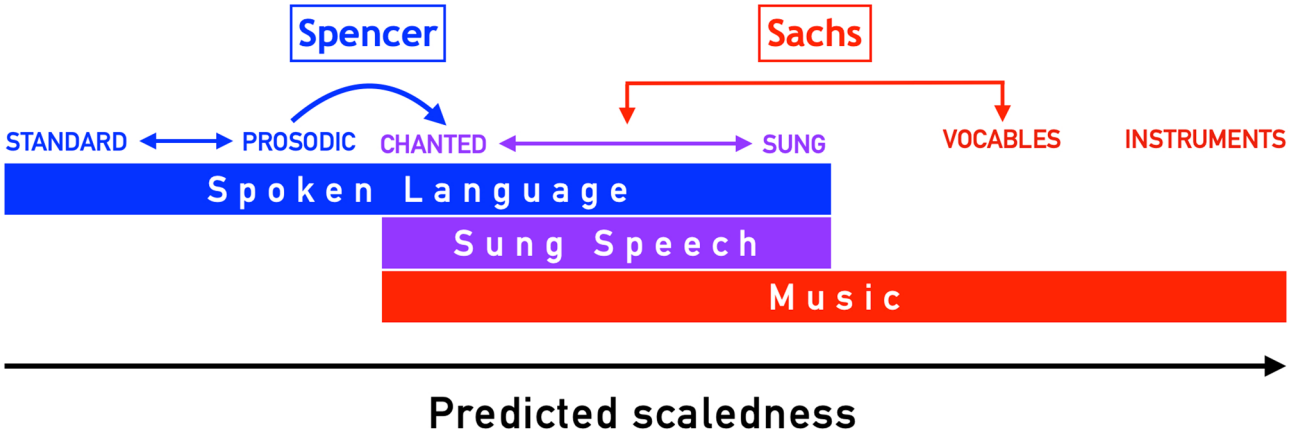

Fig. 1. The scaledness continuum from speech to sung speech to instrumental music. Speech and music can be either separate from one another or they can be combined to create composite forms of sung speech. Speech itself can be produced in either the standard manner of adult conversation or in a more prosodic manner, such as the style of a stage actor. Sung speech can be produced in either a chanted manner than is relatively speech-like or in a more musical manner that is sung with high pitch-precision. The arrow along the bottom depicts a presumed continuum of scaledness from pure speech to sung speech to pure music. Along the top of the figure are depicted Spencer's (1857) "speech theory" of the origins of song/music and Sachs' (1943) dual-origins model of the origins of song. Note that the spacing between forms is only approximate and that the figure is not meant to make quantitative predictions.

word good[24]. The evolutionary theorist Herbert Spencer, in his 1857 essay *The Origin and Function of Music*[29], proposed that, during the course of human evolution, singing (and hence music) evolved from a prosodic accentuation of speaking during times of heightened emotion. As he wrote: "While calm speech is comparatively monotonous, emotion makes use of fifths, octaves, and even wider intervals" (p. 399) and "what we regard as the distinctive traits of song are simply the traits of emotional speech intensified and systematized" (p. 402)[29]. These ideas formed the basis of Spencer's "speech theory" of the origins of music. Darwin's evolutionary theory of music[30] posited the reverse progression, claiming that speech evolved from a primitive form of hominin singing. However, Darwin's conception of singing was not well-specified in his writings, and included very non-scaled vocalizations like the territorial calls of gibbon apes. In other words, Darwin did not distinguish scaledness from prosodic accentuation.

The musicologist Curt Sachs presented a compromise position[7] by arguing that song had not one but two origins, one in music itself and the other in speech. He referred to the former style as melogenic (i.e., born from melody) and latter style as logogenic (i.e., born from words). This dual-origins model parallels an important distinction in ethnomusicology between different manners of singing[17]. This is shown in Fig. 1 as a distinction between a more "sung" style that adheres to a musical scale and a more "chanted" style that places a greater emphasis on the text. In the continuum shown in Fig. 1, the chanted (logogenic) style is presumed to have a lower scaledness than the sung (melogenic) style, since it is more speech-like. This intermediacy of chanted speech is mentioned by Spencer as well: "recitative, or musical recitation, is in all respects intermediate between speech and song" (p. 402)[29]. One implication of Sachs' theory is that the scaledness of song should vary with the manner of production (e.g., chanted vs. sung). Additionally, our previous work has suggested that the scaledness of vocal music should differ from that of instrumental music. We have demonstrated that the voice is an *imprecise* pitch-generating device, especially compared to musical instruments that can be engineered to produce highly precise pitches and thus musical scales[31]. Hence, the scaledness continuum shown in Figure 1 indicates that vocal music should be less scaled than instrumental music, even when producing the same melodic material.

It is important to note that scaledness is not the same thing as "pitch discreteness"[32–34], which measures the degree to which melodic notes are comprised of a stable pitch curve with clear onsets and offsets. Scaledness is instead an indicator of pitch-class discreteness, since it ignores how a melody unfolds over time, and instead measures how well the overall pitch of the audio clusters into stable scale degrees. Although scaledness is a measure of the overall tonal structure of a sample, it is not the same thing as the tonality measures (such as key strength) used in most automatic music information retrieval packages, such as MIRtoolbox. These measures are typically calculated by matching the pitch chromagram—which is a representation of pitch relying on octave equivalence—to an existing Western music-theoretic template, such as the Krumhansl-Schmuckler pitch profile[35,36]. Many speech/music classifiers operate based on the assumption that music-theoretic features do not apply to speech[37]. However, we cannot assume that such features obligatorily apply to the music of non-Western cultures. Scaledness is thus a culturally unbiased and more fundamental measure of a sample's underlying pitch structure.

In the present study, we present a new computational method for measuring the overall scaledness of an acoustic sample. Briefly, the method involves automatically clustering the pitches of the sample and taking the best solution's silhouette score, which captures both the average flatness of the clusters (i.e., the extent to which pitches within a class are similar) and the average steppiness between clusters (i.e., the extent to which pitch-

classes are spaced apart). We tested this measurement by applying it to a set of 1696 acoustic samples from across three cross-cultural ethnographic corpora containing samples of speech, song, and/or instrumental music. We formulated a number of predictions for the study. The most general prediction was that music would be shown to be more scaled than speech. Likewise, based on our previous work on the imprecision of the voice as a pitch-generating instrument, we predicted that instrumental music would be more scaled than vocal music (i.e., song). Next, we predicted that, within the realm of speech itself, prosodic speech (e.g., infant-directed speech) would reveal a greater level of scaledness than standard speech (e.g., speech among adults). In tandem, we predicted that speech produced using a tonal language would show greater scaledness than speech uttered using a nontonal language. Finally, with regards to singing, we explored Sachs' distinction between a more melodious (or melogenic) style of singing and a more chanted (or logogenic) style, and predicted that melodious singing would show a greater degree of scaledness than chanting. Along similar lines, we predicted that, within the realm of vocal music, songs produced with greater pitch-class imprecision by the singer[31] would show lower degrees of scaledness. We first tested these predictions using descriptive statistics, and then used predictive classification to investigate if scaledness was sufficiently informative to correctly classify samples along the speech-music continuum.

Results

Descriptive analysis

Figure 2 shows the mean silhouette values for each of the three corpora and their component categories. In keeping with the principal hypothesis of the study, song showed a significantly higher scaledness value than speech. This result was observed by comparing the Song and Speech categories within Hilton-Mehr, and by comparing Song to both Described and Recited speech within Ozaki-Savage. Another predicted result was that instrumental music showed a significantly higher scaledness value than song, as seen within the Ozaki-Savage dataset. In fact, this was the single largest category difference in the entire dataset. An unexpected finding was that prosodic speech showed barely any difference from standard speech, despite intuitions that prosodic speech is more "musical" than regular speech. This was seen in the comparison between Infant- and Adult-directed speech within Hilton-Mehr (which was non-significant), and between Described and Recited speech within

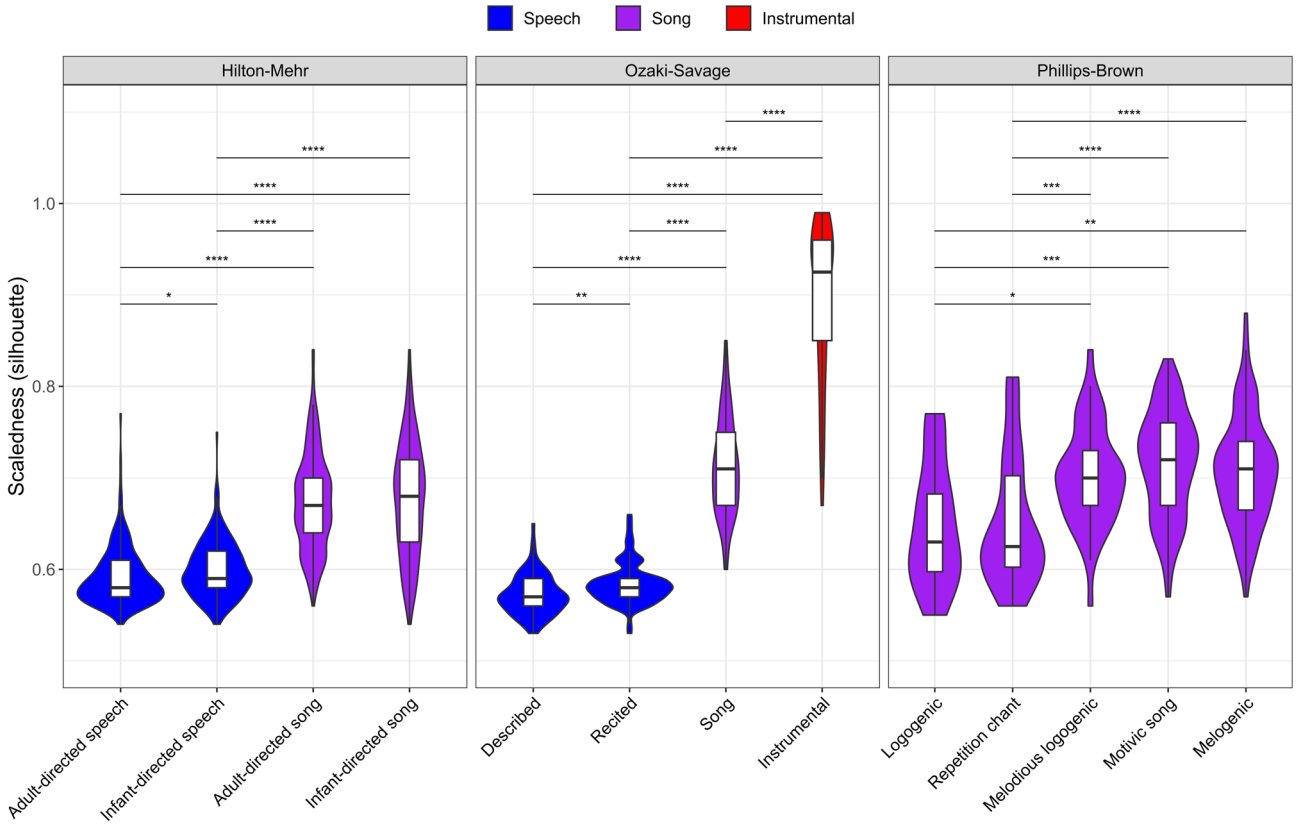

Fig. 2. Scaledness values across the three corpora, shown as violin plots. The violin plots are color-coded for overall category as per Fig. 1. The outer colored plot presents a mirrored kernel density estimation (i.e., a smoothed histogram of the data distribution). The inner boxplot presents the median as a horizontal line, the interquartile range (IQR; 25th and 75th percentiles) as a box, and whiskers that extend to the maximum and minimum values (excluding outliers further than 1.5× the IQR from the 25th or 75th percentiles). Statistical comparisons within each corpus are *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Ozaki-Savage (which was significant, although extremely small; see Extended Data Fig. 3 for visualizations of the pairwise differences between means).

The consistently low scaledness values across speech types was further supported by the analysis of language tonality, as shown in Fig. 3. When the vocal samples in Hilton-Mehr and Ozaki-Savage were sorted according to language categories, there was no significant difference in scaledness between the tonal and nontonal language types. In addition, there was no significant interaction between language tonality and utterance type in either dataset that could not be fully accounted for by the main effect of type, namely song vs. speech. However, this analysis may have been limited by the small *n* of some groups.

In contrast to these null differences for the manners of speaking, we found significant differences among many of the various manners of singing in the Phillips-Brown corpus, the categorization of which is detailed in Extended Data Table 1. The right-most panel in Fig. 2 shows an ascending gradient of silhouette values, spanning from the chanted (logogenic) styles to the sung (melogenic) styles. The melodious logogenic style was an intermediate category, combining a chanted style of vocalizing words with extensive movement along a musical scale.

For the Phillips-Brown dataset, we also investigated correlations between scaledness and other vocal properties at the per-song level (Fig. 4). The correlation between scaledness and the number of scale tones was not significant ($r = 0.08$, $t = 1.58$, $df = 414$, $p > 0.1$), indicating that our measure of scaledness is not confounded by scale size. However, there were small but significant correlations between scaledness and the mean step-size between scale tones ($r = 0.20$, $t = 4.19$, $df = 413$, $p < 0.001$), as well as with the number of different of step-sizes in scales ($r = 0.19$, $t = 3.95$, $df = 414$, $p < 0.001$), suggesting that some aspects of scale structure do indirectly impact scale discreteness, as one might expect. As per our prediction, we observed a significant negative correlation between scaledness and vocal imprecision ($r = -0.29$, $t = -6.15$, $df = 406$, $p < 0.001$), suggesting that the more imprecise the production, the less discrete the scale.

Finally, a combined analysis across the three corpora revealed a mean scaledness value of 0.66 across all samples ($n = 1696$, $sd = 0.08$). The mean values were 0.59 for speech, 0.69 for song, and 0.89 for instrumental music. The differences among these categories were all significant ($df = 2$, F-value = 1174, $p < 0.001$). A further combined analysis that distinguished standard speech (Adult-directed and Descriptive) from prosodic speech (Infant-directed and Recited) yielded a small but significant difference between these two categories, although they both differed significantly from both song and instrumental music. Nearly the entire scaledness continuum

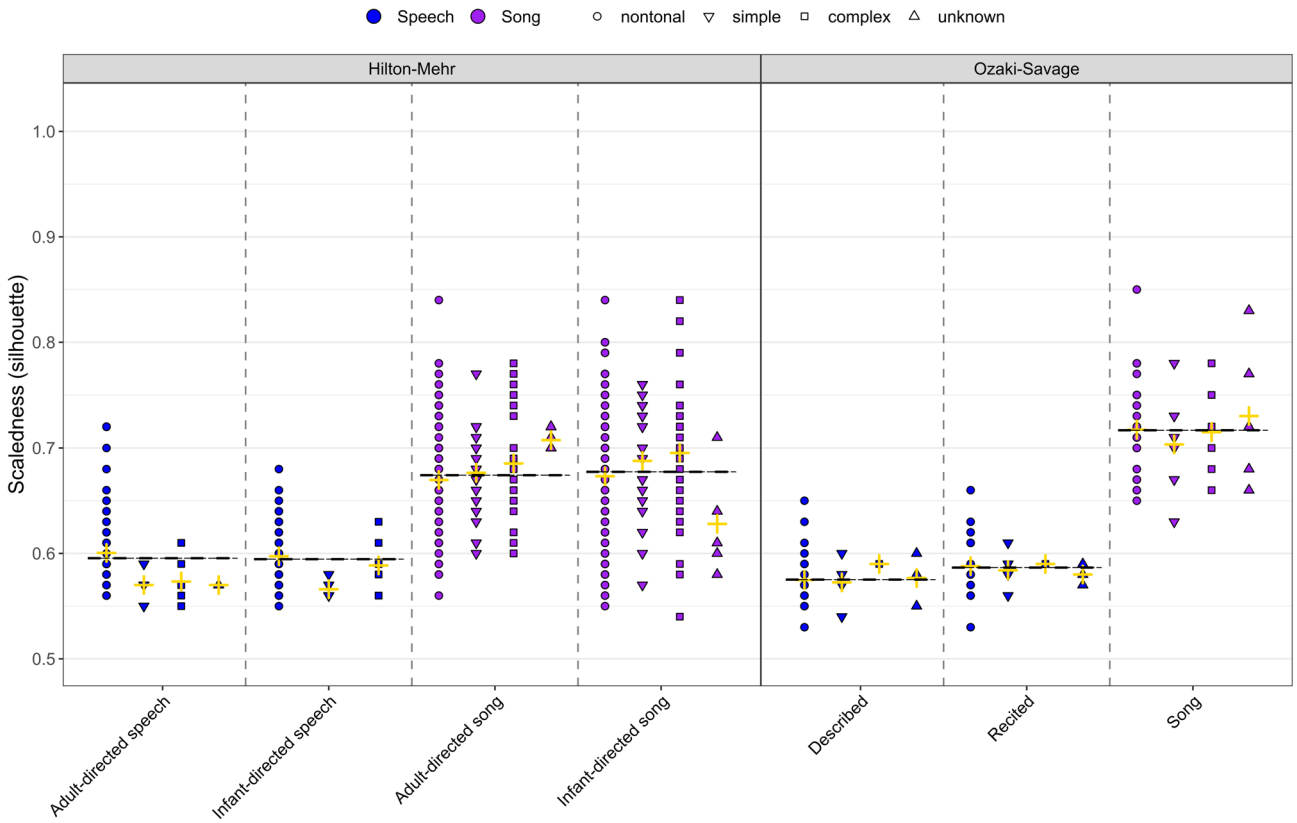

Fig. 3. Scaledness values across language categories with regard to their use of lexical tone. Silhouette values are shown for the Hilton-Mehr corpus (left) and the vocal samples of the Ozaki-Savage corpus (right). The languages are classified as either nontonal (●), simple tonal (▼), complex tonal (■), or unknown (▲). The mean for each language type is shown with a gold + and written above. The mean for each utterance type is shown with a dashed line. Speech samples are shown in blue, and song samples in purple. There were no significant differences among language classes that were not driven by utterance type (i.e., speech vs. song).

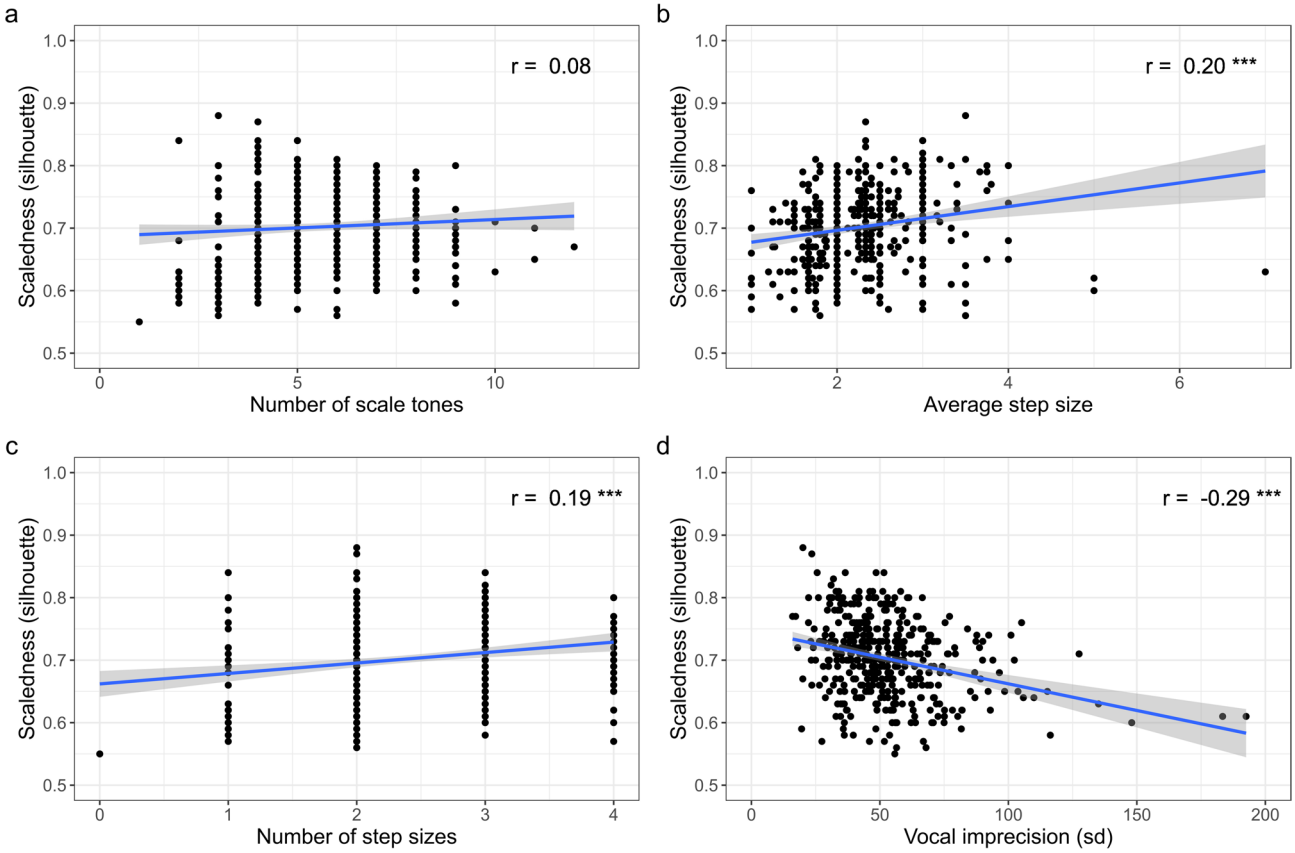

Fig. 4. The relationship between scaledness and other scale or singing features. The graphs show the relationship between scaledness and (**a**) number of tones in a scale, (**b**) the mean step-size in a scale, (**c**) the number of different step-sizes in a scale, and (**d**) the degree of vocal imprecision in production (in cents). The data are based on the Phillips-Brown corpus and the results in Phillips and Brown[31] and Brown et al. (submitted). The shaded blue line in each graph is the line of best fit with 95% CIs. ***$p < 0.001$.

(0.5–1.0; see Discussion) is spanned by these categories, with some overlap at the tail ends of their distributions (see Extended Data Fig. 2).

Predictive analysis

The model using scaledness as the sole predictor to classify the Hilton-Mehr dataset into the 4 categories specified by the authors (see Fig. 2) achieved an accuracy of 43.6% (one-sided p-value [Accuracy (Acc) > No information rate (NIR)] < 0.001) with an Akaike Information Criteria (AIC) value of 564.2 (Table 1). The most common sources of error were confusions between the speech categories (Infant- vs. Adult-directed; 29% of errors) and the song categories (Infant- vs. Adult-directed; 32%), although confusions between Infant-directed speech and Adult-directed song were also prevalent (16%). We attempted a simplified post-hoc classification into the 2 categories of "speech" (Infant- and Adult-directed) and "song" (Infant- and Adult-directed). This model achieved an accuracy of 78.8% (one-sided p-value [Acc > NIR] < 0.001), where speech was slightly more likely to be misclassified as song (60% of errors) than the reverse. The AIC of the simplified model was 188.95, and so it is taken as the better model.

The model classifying the Ozaki-Savage dataset into the 4 categories specified by the authors achieved an accuracy of 73.9% (one-sided p-value [Acc > NIR] < 0.001) with an AIC of 78.6. The most common source of error was confusion between Described and Recited speech (61%). We attempted a simplified post-hoc classification into the 3 categories of "speech" (Described and Recited), "song," and "instrument." This model achieved an accuracy of 91.3% (one-sided p-value [Acc > NIR] < 0.001) and an AIC of 41.3. Interestingly, the most common source of error was confusion between song and instrumental music (67%). The simplified model was taken as the better model.

The model classifying the Phillips-Brown dataset into the 5 categories of song described in this article (Extended Data Table 1) only achieved an accuracy of 42.2% (one-sided p-value [Acc > NIR] = 0.24) and an AIC of 268.3. This is not surprising given that this analysis was restricted to song alone, with its relatively small range of silhouette values. The most common source of error was confusion between Melogenic song and Motivic song (59%)—which were the most prevalent song-types in the corpus—although Repetition Chant also caused some errors. Given the high NIR resulting from the very unbalanced groups, we did not attempt any post-hoc models for this dataset.

Model			BIC	AIC	Acc (%) [95% CI]	p [Acc > NIR]
Hilton-Mehr	Full	Adult-directed speech Infant-directed speech Adult-directed song Infant-directed song	586.3	564.2	43.6 [37.4, 50]	4.521e−7
	Reduced	Song Speech	200	189	78.8 [73.2, 83.7]	5.509e−13
Ozaki-Savage	Full	Described Recited Song Instrument	95.5	78.6	73.9 [61.9, 83.8]	1.687e−15
	Reduced	Speech Song Instrument	54	41.3	91.3 [82.0, 96.7]	5.302e−11
Phillips-Brown		Logogenic Repetition chant Melodious logogenic Motivic Melogenic	291.4	268.3	42.2 [32.4, 52.3]	0.24
Overall	Full	Speech Prosodic speech Song Instrument	539.9	515.7	73.8 [69.3, 77.9]	1.358e−6
	Reduced	Speech Song Instrument	309.4	291.2	86.1 [82.4, 89.2]	2.2e−16

Table 1. Predictive analysis results. *BIC* Bayesian Information Criterion, *AIC* Akaike Information Criterion, *Acc* Accuracy, *CI* Confidence Interval, *p* p-value, *NIR* no information rate.

The model classifying the entire dataset into 3 categories ("speech," "song," and "instrumental") achieved an accuracy of 86.1% (one-sided p-value [Acc > NIR] < 0.001), with an AIC of 291.2. The most common source of error was confusion between speech and song (93%), with misclassification of song as speech being more common than the reverse. When "prosodic speech" (Infant-directed and Recited) was introduced as a fourth category, the accuracy dropped to 73.8% (one-sided p-value [Acc > NIR] < 0.001) and the AIC rose to 515.7. The most common source of error was confusion between speech and prosodic speech (39%), although both were also confused with song. The larger AIC indicated that the added complexity of this model was not preferable. Overall, the 3-way classification of the samples into speech, song, and instrumental music gave the strongest classification accuracy, both within and across corpora, suggesting that these are fundamental categories of scaledness.

Discussion

In the current study, we developed a new means of measuring the scaledness of an acoustic sample, as separate both from Western-theoretic frameworks for tonality and from the "discreteness" of individual pitches. Using this metric of scaledness, we were able verify the longstanding notion that music is overall more scaled than speech. However, we did not validate the notion that prosodic speech is more scaled than conversational speech, or that tone-language speech is more scaled than speech produced using an intonation language. In contrast to these null differences for styles of speaking, we demonstrated large variation in the scaledness of music, with instrumental music being far more scaled than vocal music (even when controlling for the musical material), and melodious singing being somewhat more scaled than chanting. Overall, we observed that scaledness increased significantly from speech to song to instrumental music, demonstrating marked differences not only between vocal and instrumental music, but between vocal music and speech.

Low variation in scaledness across manners of speaking

Our focus on scaledness, rather than the more standard concept of "discreteness," clarifies an important point about the acoustic nature of speech: speech shows pitch discreteness at the level of individual syllables, but not at the level of phrasal melodies. There is good evidence that speech is discrete at the level of individual syllables[6,12]. While contoured syllables exist as well (e.g., rising tones, falling tones), there are also many relatively flat tones in speech. However, this does not make speech "tonal" in the musical sense, since speech does not make use of a limited set of recurrent pitch-classes in the way that music does. Even so-called tonal languages that make use of lexical tones do not align these tones to a fixed scale. Hence, they lack tonality in the musical sense, as Chow and Brown have previously claimed[12]. When tone-language speech is set to music, musical tonality often supersedes lexical tone, even when explicit attempts are made to preserve the relative pitch of the tones[38,39].

Our results revealed that even the most prosodic or "tonal" forms of speaking were significantly less scaled, on average, than the most speech-like forms of music, namely logogenic and repetition chanting. These results are likely due to the fact that chants, but not speech, make use of a small set of recurring pitches—or even a single pitch in the case of a monotonic chant—and typically elongate syllables relative to speech. This use of pitch elongation and repetition creates at least one stable pitch-class, resulting in the most basic form of a scale. Overall, our results allow for a disambiguation of the two meanings of discreteness in the acoustics literature. They reveal that speech may show *pitch* discreteness (i.e., flatness), but not *pitch-class* discreteness (i.e., scaledness).

Our speech results fail to support Spencer's evolutionary hypothesis that music emerged through a prosodic accentuation of speech. Scaledness is not the same thing as prosody, but is instead a parallel feature to it and an alternative means of organizing pitch during vocal communication[31,40]. Speech prosody and tonal languages operate by modulating pitch registers (e.g., high vs. low) and pitch contours (e.g., rising vs. falling), but do not employ recurring sets of discrete pitches, which is a characteristic feature of much music. It is important to note that, while a number of models of the origins of music talk about a "prosodic" precursor to music[40–44], the prosody in this case refers to vocal emotion in a general sense and not to speech prosody from a fully-fledged speech capacity the way that Spencer conceived of it.

The fact that infant- and adult-directed speech are easily distinguishable to the ear may be due to more-immediate melodic and timbral differences (e.g., pitch height and range), or even to cognitive differences, such as semantic salience, rather than to their underlying tonal structure. In addition, we observed only small differences in scaledness between Recited and Described speech in the Ozaki-Savage corpus. We speculate that this difference might be due to the fact that reciting the lyrics of a well-known song might prime some participants to implicitly incorporate features of the musical melody into their recitation. Similar to our results, Ozaki et al.[34] observed "few major differences between lyrics recitation and spoken description, except that recitation tends to be slower and use shorter phrases." Indeed, another factor that may make speech appear music-like is rhythm. In the same way that music's tonal dimension makes use of pitch-class discreteness, music's rhythmic dimension makes use of discreteness at the level of duration values. Future work should aim to examine how the discreteness of pitch and that of rhythm are related to one another across a diverse collection of song types.

High variation in scaledness across manners of musicking

In contrast to the relatively null differences observed across the various manners of speaking—covering conversational speech, prosodic speech, and tone-language speech—we observed strong differences among the different manners of creating music. This was seen at two levels. The first and most significant was the large difference in scaledness between instrumental music and vocal music. This was seen in the comparison between the Instrumental and Song categories within the Ozaki-Savage corpus, which had the important benefit of controlling for the melodic material, including the scale used. Such a result is consistent with our previous study showing a far greater imprecision in vocal production compared to instrumental production[31], even when controlling for the musical material. The correlational analysis in Fig. 4 supported this finding by showing a significant negative correlation between scaledness and vocal imprecision in our song corpus. The second type of variation in the scaledness of music occurred within vocal music. In particular, we provided support for Sachs' distinction between a more melodious type of a singing and a more speech-like form of chanting, where the melogenic style was more scaled than the logogenic style. An interesting intermediate case emerged in form of the "melodious logogenic" style, which combines a speech-driven style of chanting with a stronger adherence to a musical scale. Overall, the variation of scaledness for music far surpassed that for speech, even within vocal musics that were largely word-based.

The prevalence of imprecise singing in our ethnographic corpus—and, as far as we can tell, the Hilton-Mehr corpus—suggests that the ability to measure tonal structure in song is likely attenuated by imprecise singing, which is ubiquitous across singing styles. Although we aimed to select excerpts of each sample that excluded extreme pitch drift or shift, it remains the case that the scaledness of song that we present here is not that of highly accurate singers in ideal conditions[45,46]. In addition, vocal ornamentation—through portamento, vibrato, and melisma[47]—is another factor that blurs differences between pitch-classes in sung music and that reduces its scaledness. However, even imprecise and ornamented singing is characterized by a degree of tonal structure that sets it apart from speech. The predictive analysis revealed that song was reliably distinguishable from speech, even though nearly all of the song samples used words.

Given that our results refute Spencer's hypothesis that musical tonality evolved from speech (either through prosody or lexical tone), how do we account for the evolutionary origins of scaledness? Do our results provide evidence for Sachs' dual-origins model? Our corpus contained too few songs without words for us to place vocable singing in the context of our 5-category scheme for song. While vocable singing is prevalent in *choral* traditions cross-culturally (e.g., African Pygmy singing, native American powwow singing), it seems to be less common in solo singing. Further work is necessary to examine scaledness in a large enough corpus of vocable songs to match the quantity of the word-based songs in our current corpus. Our prediction would be that vocable singing would tend to be more similar to melodious singing than to logogenic chanting, although imprecise singing could potentially confound that difference.

Limitations

Our analytic method for measuring scaledness is, at present, limited to monophonic acoustic samples, thereby precluding the analyses of choral singing, instrumental music, or group speech. It is hoped that future work will permit the analysis of group-level performance. Next, all of our silhouette values occurred in a relatively small range, even though silhouette values can in principle span from − 1 to 1. Values close to − 1 would indicate highly overlapping clustering, while values close to zero would indicate overlapping clusters. Because we used k-means clustering, which by design creates disjoint sets, our clusters could not be overlapping, making negative silhouette values impossible. Moreover, it is unlikely that k-means clustering would fail to find some degree of structure in the data, especially given that the maximum silhouette solution was chosen out of ten possible models, making values close to zero unlikely. Therefore, the functional range for our scaledness continuum spanned roughly from 0.5 to 1, and the speech samples reliably had silhouette values close to the low end. Further work should examine how the silhouette value of each of clustering solution varies per sample. Large variation would indicate that there are indeed obvious clusters, causing some solutions to be markedly better

than others. Small variation would indicate that the data are more linear and that the clustering algorithm is simply searching for minor differences to inform its solution.

Along the same lines, insight may be gained from recording the run-times of the clustering algorithm, where run-time may approximate the algorithm's difficulty in choosing clusters, and thus perhaps the linearity of the data. However, for run-times to be fairly compared across samples, the f0 lengths would first have to be standardized. Although we standardized f0 length within each corpus, the scaledness results were essentially the same as without standardization. Therefore, scaledness is not overly sensitive to the number of annotations, although extreme down-sampling—which would artificially introduce breaks in the f0 data—does cause problems. This is the main reason why we did not standardize f0 length *across* corpora, only within each corpus. The minimum length for the Phillips-Brown dataset was a fraction of that of the other two corpora.

The song dataset of Phillips-Brown was characterized by unbalanced groups, which made the 5-category classification of song-types more difficult than the classification of the other corpora. The low frequency of some song-types made training difficult and led to a very high NIR in the predictive analysis. This was unavoidable. The relative prevalence of singing styles varies across world regions, and some styles were simply more common than others in our global dataset. Despite the extremely unbalanced groups across the three corpora of the combined dataset (with a preponderance of song and very few instrumental samples), we still observed an accuracy rate of 85% in the predictive analysis. These results revealed that instrumental music is categorically distinct from both song and speech, since the models required relatively few tokens of instrumental music in order to learn to identify it. The results also revealed that, despite song and speech sharing their use of the vocal system for production, they were nonetheless distinguishable by their level of tonal structure. This held true even when songs contained words, as they did in the majority of our samples. This suggests that song is a means of musicalizing text through the imposition of a scale onto syllables, making song a joint function between speech and music (Fig. 1). Nonetheless, it is important to note that we only studied monophonic singing here and barely scratched the surface of the diverse world of instrumental music. There is thus much further work that needs to be done in applying this analysis to other samples. It should also be acknowledged that not all musics—vocal or instrumental—are primarily concerned with structuring pitch, and so methods for measuring the underlying timbral and metrical organizations should also be developed. The place of these traditions in the speech-music continuum requires further study.

The structure of the data presented another major limitation such that the non-normality and highly unequal variances complicated the multiple-comparisons testing. Although the Games-Howell test has been shown to maintain a satisfactory Type I error rate with unequal group sizes and variances, especially when the smallest group has the largest variance, its power drops markedly when performing all-pairs comparisons[48]. It may not be possible for a dataset of scaledness across the speech-music continuum to demonstrate normality or equal variance, given that speech seems to be minimally scaled, whereas music's scaledness varies widely. We suggest that the p-values presented in the analyses be interpreted alongside the results of the predictive analysis and the estimated differences in means (shown with 95% confidence intervals in Extended Data Fig. 3).

Conclusions

The combination of the descriptive and prescriptive analyses supports a 3-category organization of the scaledness continuum, spanning from speech to song to instrumental music. The largest divide was seen between vocal and instrumental production, even when the melodic material was controlled for, as in the Ozaki-Savage dataset. Overall, the results reveal that speech shows minimal scaledness no matter how it is uttered, and that music's scaledness varies widely depending on its manner of production.

Methods

Datasets

To test our hypotheses, we obtained fundamental frequency (f0) annotations for audio samples from three cross-cultural datasets:

- Hilton-Mehr: A set of 1004 f0 annotations generated using Jupyter Notebooks[49] to apply Mauch and Dixon's pYIN algorithm[50] to a subset of Hilton et al.'s[51] recordings of infant- and adult-directed speech and song. For this dataset, we generated f0 annotations using the implementation of pYIN in both Tony and Jupyter Notebook to ensure that the results would be comparable across datasets. Although the raw annotations produced in Jupyter Notebook had a lower resolution (i.e., f0 was recorded in 10-cent intervals) than those from Tony, the scaledness data (see the "silhouette" parameter below) were essentially the same. We used 5–15 s excerpts that contained one complete uttered phrase with minimal background noise, as produced by a single participant. This dataset comprised 225 samples of adult-directed speech, 205 of infant-directed speech, 285 samples of adult-directed song, and 289 of infant-directed song. The samples included 16 languages, spanning 11 language families. According to our best classification using the World Atlas of Language Structures (WALS) online[52], there were 155 samples of complex tonal languages, 90 of simple tonal languages, 739 of nontonal languages, and 20 samples where tonality was unknown.
- Ozaki-Savage: A set of 276 f0 annotations from Ozaki et al.'s[53] Stage 2 registered report, which the authors generated by applying Mauch and Dixon's pYIN algorithm to 76 recordings of descriptive speech, 76 of lyric recitation, 76 of singing, and 48 of instrumental music. These were matched samples such that each performer was asked to provide a recording of each type. The vocal samples included 55 languages, spanning 21 language families. According to our best classification using the WALS online, there were 7 vocal samples of each type that were uttered in complex tonal languages, 11 in simple tonal languages, 49 in nontonal languages, and 9 where tonality was unknown. There were fewer instrumental samples of each type, but these were not included in our language tonality analysis.

- Phillips-Brown: A set of 416 f0 annotations generated by applying Tony's implementation[54] of Mauch and Dixon's pYIN algorithm to recordings from our global dataset of traditional songs[31]. This dataset comprised excerpts of solo monophonic songs from indigenous and traditional cultures, spanning 10 musical-style regions, as based on Lomax's[47] global classification of singing style. The distribution of languages in these recordings is not known.

Song classification

For the Phillips-Brown dataset, the second author devised a multidimensional classification scheme of song-types in the full dataset, as based on Sachs'[7] primary classification of song into "logogenic" and "melogenic" varieties. The end result was a 5-category scheme, as shown in Extended Data Table 1. In between the two extremes of logogenic and melogenic are 3 newly-proposed categories. "Motivic song" is similar to melogenic, except that its rhythmic properties tend to be heterometric (rather than isometric), and so its phrases tend to be of variable lengths, interspersed with non-metric pauses. "Melodious logogenic" is similar to logogenic as a type of cantillation of text, except that it tends to be far more melodious (rather than monotonic), making extensive use of movement across scale pitches. It generally includes melismatic ornamentation in a way that pure logogenic does not and somewhat resembles a Classical opera "recitative" section. The last category of "repetition chant" typically consists of a single repeated phrase using 2–3 pitch-classes. Note that the classification scheme for song in Extended Data Table 1 is multidimensional even though we are only focusing on the scaledness feature in the current analysis.

After this 5-category scheme was proposed by the second author, the first author listened to the full dataset and carried out an independent rating of all of the songs. The two raters agreed on 60% of the categorizations as first choices; with second choices included, the agreement was 80%. All disputes were resolved by joint listening to the songs and coming to an agreement on every song's classification. This resulted in a breakdown of the original 416 songs into 16 logogenic songs, 45 melodious logogenic songs, 42 repetition chants, 154 motivic songs, and 159 melogenic songs).

Acoustic analysis

For each dataset, the file with the fewest f0 annotations was found, and its length was set as the baseline f0 length for that dataset. For each sample in that dataset, the f0 pitch-trace was converted from Hz to cents. Spurious annotations (more than 2 standard deviations beyond the median) were removed, and the data were down-sampled into evenly-spaced intervals such that the length of a sample would match the baseline. The cents data—originally plotted against time to form "melographs," or visual representations of the melody—were then sorted from lowest to highest pitch to create a "scalograph," which is an ordinal visual representation of the scale (Extended Data Fig. 1).

In order to quantify the "scaledness" of each sample, we applied automatic k-means clustering to the pitch profile in the scalograph using the Python package *sklearn*[55], allowing for 2–11 clusters per sample. We used the parameter of "silhouette" to approximate a sample's scaledness because it has components that intuitively map onto tonal discreteness: the flatness of a cluster (i.e., the extent to which the pitches within a class are similar) is measured by the intra-cluster distance (a), whereas the steppiness between clusters (i.e., the extent to which the pitch-classes are spaced apart) is measured by the nearest-cluster distance (b). A cluster's silhouette score is given by $\frac{(b-a)}{max(a,b)}$. The overall silhouette value of a model is given by the mean silhouette score of all of the component clusters from the k-means analysis. We chose the maximum silhouette value (of all models from 2 to 11 clusters) as representative of the overall scaledness of a sample. It is important to note that our measure of scaledness is agnostic with regard to "tonality" in the Western musicological sense of a tonic pitch and a hierarchical organization of scale tones. It merely describes the distinguishability of pitch-classes in a sample.

Descriptive statistical analysis

We first tested whether scaledness values differed significantly across the sample types of each dataset. Due to the unbalanced group sizes of the Phillips-Brown dataset, we used a one-way analysis of variance (ANOVA) with type-III sum of squares, followed by multiple comparisons testing using Tukey's "honest significant difference" (HSD) method, with a family-wise confidence level of 0.95. Due to the unequal variances and unbalanced groups of the Hilton-Mehr and Ozaki-Savage datasets, we used a one-way ANOVA with type-III sum of squares and a White-Huber (H3) heteroscedasticity-corrected covariance matrix, followed by multiple comparisons testing using a Games-Howell test, with a family-wise confidence level of 0.95[48,56]. Tests were carried out using the R packages *car*[57], *rstatix*[58] and *PMCMRplus*[59].

Given that we did not standardize sample length across the datasets, we mainly investigated differences across the categories *within* each dataset, not the combined dataset. However, since the scaledness results seemed relatively insensitive to sample length (see "Discussion" section), we tested for significant differences in scaledness across the overarching categories of Speech (n = 582), Song (n = 1066), and Instrumental music (n = 48) in the combined dataset. We also broke down the Speech category into "Speech" (Adult-directed from Hilton-Mehr and Descriptive from Ozaki-Savage, n = 301) and "Prosodic speech" (Infant-directed from Hilton-Mehr and Recited from Ozaki-Savage, n = 281), and compared these categories to Song (Ozaki-Savage and Phillips-Brown) and Instrumental music (Ozaki-Savage) in the combined dataset. Because the overall dataset had unequal variance and unbalanced groups, we followed the same procedure as outlined for the Ozaki-Savage dataset above.

For the Hilton-Mehr and Ozaki-Savage datasets, we also tested for significant differences in scaledness across the various categories of language tonality in the vocal samples (excluding instrumental music from this analysis), using fully crossed two-way ANOVAs with type-III sum of squares (to account for the unbalanced group sizes) and multiple comparisons testing using Tukey's HSD, with a family-wise confidence level of 0.95.

For the Phillips-Brown dataset, we also examined the correlations between scaledness and some previously-measured properties of the samples, namely their vocal imprecision[31] and the number and average spacing of scale tones/intervals per sample[10].

Predictive analysis

To further test whether scaledness is a good marker for distinguishing categories across the speech-song continuum, we used the R package *xgboost*[60] to carry out multiclass logistic classification. For each model, we first performed soft classification, which produced predicted classification probabilities for each test observation. We used these values to calculate the log-likelihood and Akaike Information Criteria (AIC) of each model. We then hardened the classification predictions to compare them to the ground-truth labels (i.e., the classifications indicated by the authors of the studies) to obtain an overall accuracy score. To determine the significance of the classification accuracy, we compared the model accuracy to the "no information rate" (NIR), which is the accuracy of a naive model that would only predict the majority class. We compared across models for each dataset using the AIC, where a lower value of AIC is taken as the better model.

We used scaledness as the sole predictor in the following models: (1) classifying *each* dataset by *its own* categories; (2) classifying *each* dataset into a simplified model by inspecting the confusion matrix; (3) classifying the combined dataset into the overarching categories of Speech, Song, and Instrumental music; and (4) classifying the combined dataset into the overarching categories of Speech, Prosodic Speech, Song, and Instrumental music. Each model was performed with a 75/25 train/test partition, split within the levels of the outcome variable using the R package *caret*[61] to balance across the uneven class distributions within the splits. For each model, multiclass negative log-likelihood was used as the evaluation metric, and training was performed with a maximum boosting iteration of 10,000 and an early stopping point of 10 rounds, after which validation training would end if performance did not improve.

Data availability

All materials for this study are available at the following public Open Science Framework repository: https://osf.io/3j9wr/.

Received: 11 November 2024; Accepted: 19 May 2025
Published online: 01 July 2025

References

1. Bernstein, L. *The Unanswered Question: Six Talks at Harvard* (Harvard University Press, 1976).
2. Sloboda, J. A. *The Musical Mind: The Cognitive Psychology of Music* (Oxford University Press, 1986).
3. Swain, J. P. *Musical Languages* (WW Norton & Company, 1997).
4. Burns, E. M. Intervals, scales, and tuning. In *The Psychology of Music* (ed. Deutsch, D.) 215–264 (Elsevier, 1999).
5. Jeannin, M. Organizational structures in language and music. *World Music* **50**, 5–16 (2008).
6. Patel, A. D. *Music, Language, and the Brain* (Oxford University Press, 2008).
7. Sachs, C. *The Rise of Music in the Ancient World, East and West* (Dover, 1943).
8. Malm, W. P. *Music Cultures of the Pacific, the Near East, and Asia* 3rd edn. (Prentice-Hall, 1996).
9. Ellis, A. J. On the musical scales of various nations. *J. Soc. Arts* **33**, 485–527 (1885).
10. Brown, S., Phillips, E., Husein, K. & McBride, J. Musical scales optimize pitch spacing: A global analysis of traditional vocal music. *Humanities and Social Sciences Communications* **12**, 1–13. https://doi.org/10.1057/s41599-025-04881-1 (2025).
11. Savage, P. E., Brown, S., Sakai, E. & Currie, T. E. Statistical universals reveal the structures and functions of human music. *PNAS* **112**, 8987–8992 (2015).
12. Chow, I. & Brown, S. A musical approach to speech melody. *Front. Psychol.* **9**, 247 (2018).
13. Magdics, K. From the melody of speech to the melody of music. *Stud. Musicol. Acad. Sci. Hung.* **4**, 325–346 (1963).
14. Nooteboom, S. The prosody of speech: Melody and rhythm. *Handb. Phonetic Sci.* **5**, 640–673 (1997).
15. Zatorre, R. J. & Baum, S. R. Musical melody and speech intonation: Singing a different tune. *PLoS Biol.* **10**, e1001372 (2012).
16. Boas, F. *Primitive Art* (Dover Publications, 1927/2010).
17. Sborgi Lawson, F. R. Mousike or music? Using analysis to explore shifts in musical attention. *Ethnomusicol. Forum* **32**, 120–142 (2023).
18. Halle, J. & Lerdahl, F. A generative textsetting model. *Curr. Musicol.* **55**, 3–23 (1993).
19. Proto, T. Prosody, melody and rhythm in vocal music: The problem of textsetting in a linguistic perspective. *Ling. Netherlands* **32**, 116–129 (2015).
20. McPherson, L. Musical adaptation as phonological evidence: Case studies from textsetting, rhyme, and musical surrogates. *Lang. Ling. Compass* **13**, e12359 (2019).
21. Soderstrom, M. Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Dev. Rev.* **27**, 501–532 (2007).
22. Fernald, A. Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child Dev.* **60**, 1497–1510 (1989).
23. Fernald, A. & Simon, T. Expanded intonation contours in mothers' speech to newborns. *Dev. Psychol.* **20**, 104–113 (1984).
24. Fernald, A. & Mazzie, C. Prosody and focus in speech to infants and adults. *Dev. Psychol.* **27**, 209–221 (1991).
25. Papousek, M. Intuitive parenting: A hidden source of musical stimulation in infancy. In *Musical Beginnings: Origins and Development of Musical Competence* (eds Deliege, I. & Sloboda, J. A.) 88–112 (Oxford University Press, 1996).
26. Berry, M. & Brown, S. Acting in action: Prosodic analysis of character portrayal during acting. *J. Exp. Psychol. Gen.* **148**, 1407–1425 (2019).
27. Bryant, G. A. & Barrett, H. C. Recognizing intentions in infant-directed speech: Evidence for universals. *Psychol. Sci.* **18**, 746–751 (2007).
28. Vosoughi, S. & Roy, D. A longitudinal study of prosodic exaggeration in child-directed speech. In *Speech Prosody 2012* 194–197. https://doi.org/10.21437/SpeechProsody.2012-50 (ISCA, 2012).
29. Spencer, H. The origin and function of music. *Fraser's Mag.* **56**, 396–408 (1857).
30. Darwin, C. *The Descent of Man, and Selection in Relation to Sex* (Princeton University Press, 1871).
31. Phillips, E. & Brown, S. Vocal imprecision as a universal constraint on the structure of musical scales. *Sci. Rep.* **12**, 19820 (2022).

32. Haiduk, F., Quigley, C. & Fitch, W. T. Song is more memorable than speech prosody: Discrete pitches aid auditory working memory. *Front. Psychol.* **11**, 3493 (2020).
33. Ozaki, Y. et al. *Automatic Acoustic Analyses Quantify Variation in Pitch Discreteness Within and Between Human Music, Speech, and Bird Song*. https://doi.org/10.31234/osf.io/7ywxm (2020).
34. Ozaki, Y. et al. Globally, songs and instrumental melodies are slower and higher and use more stable pitches than speech: A registered report. *Sci. Adv.* **10**, 9797 (2024).
35. Weiß, C. & Habryka, J. Chroma-based scale matching for audio tonality analysis. In *Proceedings of the 9th Conference on Interdisciplinary Musicology (CIM)* 168–173 (2014).
36. You, Y., Yang, H., Xu, H. & Zhou, Y. Music tonality detection based on Krumhansl-Schmuckler profile. In *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)* 85–88. https://doi.org/10.1109/ITAIC.2019.8785576 (2019).
37. Birajdar, G. K. & Patil, M. D. Speech/music classification using visual and spectral chromagram features. *J. Ambient Intell. Hum. Comput.* **11**, 329–347 (2020).
38. Schellenberg, M. H. *The Realization of Tone in Singing in Cantonese and Mandarin* (University of British Columbia, 2013).
39. Zhang, Q., Zhu, L. & Jiang, X. Can we sing the tones of a tonal language? The duration of Mandarin tones under music context. In *Speech Prosody 2024* 344–348. https://doi.org/10.21437/SpeechProsody.2024-70 (ISCA, 2024).
40. Brown, S. A joint prosodic origin of language and music. *Front. Psychol.* **8**, 1894 (2017).
41. Filippi, P. Emotional and interactional prosody across animal communication systems: A comparative approach to the emergence of language. *Front. Psychol.* **7**, 1393 (2016).
42. Fitch, W. T. *The Evolution of Language* (Cambridge University Press, 2010).
43. Mithen, S. J. *The Singing Neanderthals: The Origins of Music, Language, Mind, and Body* (Harvard University Press, 2006).
44. Rousseau, J. *Essay on the Origin of Languages. English Translation by JH Moran and A. Gode (1986)* (1781).
45. Mauch, M., Frieler, K. & Dixon, S. Intonation in unaccompanied singing: Accuracy, drift, and a model of reference pitch memory. *J. Acoust. Soc. Am.* **136**, 401–411 (2014).
46. Rosenzweig, S., Scherbaum, F. & Müller, M. Computer-assisted analysis of field recordings: A case study of Georgian funeral songs. *ACM J. Comput. Cult. Heritage* **16**, 1–16 (2022).
47. Lomax, A. *Folk Song Style and Culture* (Routledge, 2017).
48. Sauder, D. C. & DeMars, C. E. An updated recommendation for multiple comparisons. *Adv. Methods Pract. Psychol. Sci.* **2**, 26–44 (2019).
49. Kluyver, T. et al. Jupyter notebooks—A publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas* 87–90. https://doi.org/10.3233/978-1-61499-649-1-87 (2016).
50. Mauch, M. & Dixon, S. PYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 659–663. https://doi.org/10.1109/ICASSP.2014.6853678 (IEEE, 2014).
51. Hilton, C. B. et al. Acoustic regularities in infant-directed speech and song across cultures. *Nat. Hum. Behav.* **1**, 1–12. https://doi.org/10.1038/s41562-022-01410-x (2022).
52. Maddieson, I. *WALS Online. Zenodo*. https://doi.org/10.5281/zenodo.7385533 (2013).
53. Ozaki, Y. *et al. Globally, Songs and Instrumental Melodies are Slower, Higher, and Use More Stable Pitches than Speech [Stage 2 Registered Report]* (2023).
54. Mauch, M. *et al.* Computer-aided melody note transcription using the Tony software: Accuracy and efficiency. In *Proceedings of the First International Conference on Technologies for Music Notation and Representation* 8 (2015).
55. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
56. Lee, S. & Lee, D. K. What is the proper way to apply the multiple comparison test? *Korean J. Anesthesiol.* **71**, 353–360 (2018).
57. Fox, J. & Weisberg, S. *An R Companion to Applied Regression* (Sage, 2019).
58. Kassambara, A. *Rstatix: Pipe-Friendly Framework for Basic Statistical Tests* (2023).
59. Pohlert, T. *PMCMRplus: Calculate Pairwise Multiple Comparisons of Mean Rank Sums Extended* (2024).
60. Chen, T. & Guestrin, C. XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794. https://doi.org/10.1145/2939672.2939785 (2016).
61. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **28**, 1–26 (2008).

Acknowledgements

We thank Jinkyung (Sam) Kim for his assistance pre-processing the Hilton-Mehr corpus, John McBride for his analysis of pitch imprecision in the Phillips-Brown corpus, and Khalil Husein for his analysis of scale properties in the Phillips-Brown corpus. This work was supported by a grant to S.B. from Natural Sciences and Engineering Research Council (NSERC) of Canada (Grant Number RGPIN-2020-05718).

Author contributions

Both authors contributed to the study design. E.M.P. wrote the computational code, conducted the study, and analyzed the results. Both authors contributed to data interpretation, preparation of figures, and the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-03049-w.

Correspondence and requests for materials should be addressed to E.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.