ANTHROPOLOGY

# Globally, songs and instrumental melodies are slower and higher and use more stable pitches than speech: A Registered Report

Yuto Ozaki[1]*†, Adam Tierney[2], Peter Q. Pfordresher[3], John M. McBride[4], Emmanouil Benetos[5], Polina Proutskova[5], Gakuto Chiba[1], Fang Liu[6], Nori Jacoby[7], Suzanne C. Purdy[8,9], Patricia Opondo[10], W. Tecumseh Fitch[11], Shantala Hegde[13], Martín Rocamora[14,15], Rob Thorne[16], Florence Nweke[17,18], Dhwani P. Sadaphal[11], Parimal M. Sadaphal[19], Shafagh Hadavi[1], Shinya Fujii[20], Sangbuem Choo[1], Marin Naruse[21], Utae Ehara[22], Latyr Sy[23,24], Mark Lenini Parselelo[25,26], Manuel Anglada-Tort[27], Niels Chr. Hansen[28,29,30,31], Felix Haiduk[11,32], Ulvhild Færøvik[33], Violeta Magalhães[34,35,36], Wojciech Krzyżanowski[37], Olena Shcherbakova[38], Diana Hereld[39], Brenda Suyanne Barbosa[12], Marco Antonio Correa Varella[40], Mark van Tongeren[41], Polina Dessiatnitchenko[42], Su Zar Zar[43], Iyadh El Kahla[44], Olcay Muslu[45,46], Jakelin Troy[47], Teona Lomsadze[48,49], Dilyana Kurdova[50,51], Cristiano Tsope[52], Daniel Fredriksson[53], Aleksandar Arabadjiev[54], Jehoshaphat Philip Sarbah[55], Adwoa Arhine[56], Tadhg Ó Meachair[57], Javier Silva-Zurita[58,59], Ignacio Soto-Silva[58,59], Neddiel Elcie Muñoz Millalonco[60], Rytis Ambrazevičius[61], Psyche Loui[62], Andrea Ravignani[63,64,65], Yannick Jadoul[63,64], Pauline Larrouy-Maestri[66,67], Camila Bruder[66], Tutushamum Puri Teyxokawa[68], Urise Kuikuro[69], Rogerdison Natsitsabui[69], Nerea Bello Sagarzazu[70], Limor Raviv[64,71], Minyu Zeng[1,72], Shahaboddin Dabaghi Varnosfaderani[73,74], Juan Sebastián Gómez-Cañón[15], Kayla Kolff[75], Christina Vanden Bosch der Nederlanden[76], Meyha Chhatwal[76], Ryan Mark David[76], I. Putu Gede Setiawan[77], Great Lekakul[78], Vanessa Nina Borsan[1,79], Nozuko Nguqu[10], Patrick E. Savage[8,20]*



**Both music and language are found in all known human societies, yet no studies have compared similarities and differences between song, speech, and instrumental music on a global scale. In this Registered Report, we analyzed two global datasets: (i) 300 annotated audio recordings representing matched sets of traditional songs, recited lyrics, conversational speech, and instrumental melodies from our 75 coauthors speaking 55 languages; and (ii) 418 previously published adult-directed song and speech recordings from 209 individuals speaking 16 languages. Of our six preregistered predictions, five were strongly supported: Relative to speech, songs use (i) higher pitch, (ii) slower temporal rate, and (iii) more stable pitches, while both songs and speech used similar (iv) pitch interval size and (v) timbral brightness. Exploratory analyses suggest that features vary along a "musi-linguistic" continuum when including instrumental melodies and recited lyrics. Our study provides strong empirical evidence of cross-cultural regularities in music and speech.**

*Before submitting to* Science Advances *for further review, this Registered Report (Stage 2) was peer-reviewed and recommended for publication by Peer Community In Registered Reports (PCI-RR) (1). Stage 1 (review of the design and analysis) was also reviewed by PCI-RR (2). The authors have moved parts of the preregistered Introduction and Methods sections to fit* Science Advances *formatting requirements but have not changed their content from the version that was granted In Principle Acceptance by PCI-RR on 17 January 2023 (except where explicitly noted in the text).*

## INTRODUCTION

Language and music are both found universally across cultures, yet in highly diverse forms (*3*–*7*), leading many to speculate on their evolutionary functions and possible coevolution (*8*–*13*). However, this speculation still lacks empirical data to answer the question: What similarities and differences between music and language are shared cross-culturally? Although comparative research has revealed distinct and shared neural mechanisms for music and language (*9*, *14*–*19*), there has been relatively less comparative analysis of acoustic attributes of music and language (*20*, *21*) and even fewer that directly compare the two most widespread forms of music and language that use the same production mechanism: vocal music (song) and spoken language (speech).

Cross-cultural analyses have identified "statistical universals" shared by most of the world's musics and/or languages (*22*–*25*). In music, these include regular rhythms, discrete pitches, small melodic intervals, and a predominance of songs with words (rather than instrumental music or wordless songs) (*5*, *25*). However, non-signed languages also use the voice to produce words, and other proposed musical universals may also be shared with language (e.g., discrete pitch in tone languages, regular rhythms in "syllable-timed"/"stress-timed" languages, and use of higher pitch when vocalizing to infants) (*9*, *13*, *26*–*28*). Moreover, vocal parameters of speech and singing, such as fundamental frequency and vocal tract length as estimated from formant frequencies, are strongly intercorrelated in both men and women (*10*).

Many hypotheses make predictions about cross-cultural similarities and differences between song and speech. For example, the social bonding hypothesis (*11*) predicts that song is more predictably regular than speech to facilitate synchronization and social bonding. In contrast, the motor constraint hypothesis of Tierney *et al.* (*28*)

predicts similarities in pitch interval size and melodic contour due to shared constraints on sung and spoken vocalization. Similarly, the sexual selection hypothesis predicts similarities between singing and speaking due to their redundant functions as "backup signals" indicating similar underlying mate qualities (e.g., body size) (*10*). Last, culturally relativistic hypotheses instead predict neither regular cross-cultural similarities nor differences between song and speech but rather predict that relationships between song and speech are strongly culturally dependent without any universal regularities (*29*).

Culturally relativistic hypotheses appear to be dominant among ethnomusicologists. For example, in a 13 January 2022 email to the International Council for Traditional Music email list entitled "What is song?," International Council for Traditional Music Vice-President Don Niles requested definitions for "song" that might distinguish it from "speech" cross-culturally. Much debate ensued, but the closest to such a definition that appeared to emerge, was the following conclusion published by Savage *et al.* (*25*) based on a comparative analysis of 304 audio recordings of music from around the world:

"Although we found many statistical universals, absolute musical universals did not exist among the candidates we were able to test. The closest thing to an absolute universal was Lomax and Grauer's (*30*) definition of a song as a vocalization using "discrete pitches or regular rhythmic patterns or both," which applied to almost the entire sample, including instrumental music. However, three musical examples from Papua New Guinea containing combinations of friction blocks, swung slats, ribbon reeds, and moaning voices contained neither discrete pitches nor an isochronous beat. It should be noted that the editors of the Encyclopedia did not adopt a formal definition of music in choosing their selections. We thus assume that they followed the common practice in ethnomusicology of defining music as "humanly organized sound" (*31*) other than speech, with the distinction between speech and music being left to each culture's emic (insider and subjective) conceptions, rather than being defined objectively by outsiders. Thus, our analyses suggest that there is no absolutely universal and objective definition of music but that Lomax and Grauer's definition may offer a useful working definition to distinguish music from speech."

However, the conclusion of Savage *et al.* (*25*) was based only on an analysis of music; thus, the contrast with speech is speculative and not based on comparative data. Some studies have identified differences between speech and song in specific languages, such as song being slower and higher-pitched (*32*–*35*). However, a lack of annotated cross-cultural recordings of matched speaking and singing has hampered attempts to establish cross-cultural relationships between speech and song (*36*). The available dataset closest to our study is Hilton *et al.*'s (*26*) recordings sampled from 21 societies. Their dataset covers 11 language families, and each participant produced a set of adult-directed and infant-directed song and speech. However, their dataset was designed to independently compare adult-directed versus infant-directed versions of song and of speech, and they



[1]Graduate School of Media and Governance, Keio University, Fujisawa, Kanagawa, Japan. [2]Department of Psychological Sciences, Birkbeck, University of London, London, UK. [3]Department of Psychology, University at Buffalo, State University of New York, Buffalo, NY, USA. [4]Center for Algorithmic and Robotized Synthesis, Institute for Basic Science, Ulsan, South Korea. [5]School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK. [6]School of Psychology and Clinical Language Sciences, University of Reading, Reading, UK. [7]Computational Auditory Perception Group, Max-Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany. [8]School of Psychology, University of Auckland, Auckland, New Zealand. [9]Centre for Brain Research and Eisdell Moore Centre for Hearing and Balance Research, University of Auckland, Auckland, New Zealand. [10]School of Arts, Music Discipline, University of KwaZulu Natal, Durban, South Africa. [11]Department of Behavioral and Cognitive Biology, University of Vienna, Vienna, Austria. [12]Department of Musicology, University of Vienna, Vienna, Austria. [13]Music Cognition Lab, Department of Clinical Psychology, National Institute of Mental Health and Neuro Sciences, Bangalore, Karnataka, India. [14]Universidad de la República, Montevideo, Uruguay. [15]Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain. [16]School of Music, Victoria University of Wellington, Wellington, New Zealand. [17]Department of Creative Arts, University of Lagos, Lagos, Nigeria. [18]Department of Music, Mountain Top University, Ogun, Nigeria. [19]Independent Researcher, New Delhi, India. [20]Faculty of Environment and Information Studies, Keio University, Fujisawa, Kanagawa, Japan. [21]Faculty of Policy Management, Keio University, Fujisawa, Kanagawa, Japan. [22]Haponetay, Shimizucho, Hokkaido, Japan. [23]Independent researcher, Tokyo, Japan. [24]Independent researcher, Dakar, Sénégal. [25]Memorial University of Newfoundland, St. John's, NL, Canada. [26]Department of Music and Dance, Kenyatta University, Nairobi, Kenya. [27]Department of Psychology, Goldsmiths, University of London, London, UK. [28]Aarhus Institute of Advanced Studies, Aarhus University, Aarhus, Denmark. [29]Centre of Excellence in Music, Mind, Body and Brain, University of Jyväskylä, Jyväskylä, Finland. [30]Interacting Minds Centre, School of Culture and Society, Aarhus University, Aarhus, Denmark. [31]Royal Academy of Music Aarhus/Aalborg, Aarhus, Denmark. [32]Department of General Psychology, University of Padua, Padua, Italy. [33]Institute of Biological and Medical Psychology, Department of Psychology, University of Bergen, Bergen, Norway. [34]Centre of Linguistics of the University of Porto (CLUP), Porto, Portugal. [35]Faculty of Arts and Humanities of the University of Porto (FLUP), Porto, Portugal. [36]School of Education of the Polytechnic of Porto (ESE IPP), Porto, Portugal. [37]Adam Mickiewicz University, Faculty of Art Studies, Musicology Institute, Poznań, Poland. [38]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [39]Department of Psychiatry, UCLA Semel Institute for Neuroscience and Human Behavior, Los Angeles, CA, USA. [40]Department of Experimental Psychology, Institute of Psychology, University of São Paulo, São Paulo, Brazil. [41]Independent researcher, Taoyuan City, Taiwan. [42]School of International Liberal Studies, Waseda University, Tokyo, Japan. [43]Headmistress, The Royal Music Academy, Yangon, Myanmar. [44]Department of Cultural Policy, University of Hildesheim, Hildesheim, Germany. [45]Centre for the Study of Higher Education, University of Kent, Canterbury, UK. [46]MIRAS, Centre for Cultural Sustainability, Istanbul, Turkey. [47]Director, Indigenous Research, Office of the Deputy Vice-Chancellor (Research); Department of Linguistics, Faculty of Arts and Social Sciences, The University of Sydney, Camperdown, NSW, Australia. [48]International Research Center for Traditional Polyphony of the Tbilisi State Conservatoire, Tbilisi, Georgia. [49]Georgian Studies Fellow, University of Oxford, Oxford, UK. [50]South-West University Neofit Rilski, Blagoevgrad, Bulgaria. [51]Phoenix Perpeticum Foundation, Sofia, Bulgaria. [52]Universidade de Aveiro, Aveiro, Portugal. [53]Dalarna University, Falun, Sweden. [54]Department of Folk Music Research and Ethnomusicology, University of Music and Performing Arts–MDW, Wien, Austria. [55]Department of Music and Dance, University of Cape Coast, Cape Coast, Ghana. [56]Department of Music, University of Ghana, Accra, Ghana. [57]Department of Ethnomusicology and Folklore, Indiana University, Bloomington, IN, USA. [58]Department of Humanities and Arts, University of Los Lagos, Osorno, Chile. [59]Millennium Nucleus on Musical and Sound Cultures (CMUS NCS 2022-16), Santiago, Chile. [60]Traditional Performer and Culture Bearer, Castro, Chile. [61]Kaunas University of Technology, Kaunas, Lithuania. [62]Music, Imaging and Neural Dynamics Lab, Northeastern University, Boston, MA, USA. [63]Department of Human Neurosciences, Sapienza University of Rome, Rome, Italy. [64]Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands. [65]Center for Music in the Brain, Department of Clinical Medicine, Aarhus University, Aarhus, Denmark & The Royal Academy of Music Aarhus/Aalborg, Aarhus, Denmark. [66]Music Department, Max-Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany. [67]Max Planck—NYU Center for Language, Music, and Emotion (CLaME), New York, NY, USA. [68]Txemim Puri Project–Puri Language Research, Vitalization and Teaching/Recording and Preservation of Puri History and Culture, Rio de Janeiro, Brasil. [69]Independent Researcher, Brazil. [70]Department of Education, University of Glasgow, Glasgow, UK. [71]cSCAN, University of Glasgow, Glasgow, UK. [72]Rhode Island School of Design, Providence, RI, USA. [73]Institute for English and American Studies (IEAS), Goethe University of Frankfurt am Main, Frankfurt am Main, Germany. [74]Cognitive and Developmental Psychology Unit, Centre, for Cognitive Science, University of Kaiserslautern-Landau (RPTU), Kaiserslautern, Germany. [75]Institute of Cognitive Science, University of Osnabrück, Osnabrück, Germany. [76]Department of Psychology, University of Toronto Mississauga, Mississauga, ON, Canada. [77]Independent researcher, Tokyo, Japan. [78]Faculty of Fine Arts, Chiang Mai University, Chiang Mai, Thailand. [79]Université de Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France.
*Corresponding author. Email: yuto_ozaki@keio.jp (Y.O.); patrick.savage@auckland.ac.nz (P.E.S.)
†Order of authors other than first and last authors is based on the order in which they joined the project.

did not directly compare singing versus speaking. We performed exploratory analyses of their dataset (*37*) but found that since their dataset does not include manual annotations for acoustic units (e.g., note, syllable, sentence, phrase, etc.), it is challenging to analyze and compare key structural aspects such as pitch intervals, pitch contour shape, or note/syllable duration. While automatic segmentation can be effective for segmenting some musical instruments and animal songs [e.g., percussion instruments (*38*) and bird song notes separated by microbreaths (*39*)], we found that they did not provide satisfactory segmentation results compared to human manual annotation for the required task of segmenting continuous song/speech into discrete acoustic units such as notes or syllables (compare fig. S6). For example, Mertens' (*40*) automated segmentation algorithm used by Hilton *et al.* (*26*) mis-segmented two of the first three words "by a lonely" from the English song used in our pilot analyses ("The Fields of Athenry"), oversegmenting "by" into "b-y," and undersegmenting "lonely" by failing to divide it into "lone-ly" (compare fig. S6 for systematic comparison of annotation by automated methods and by humans speaking five different languages from our pilot data).

Our study overcomes these issues by creating a unique dataset of matched singing and speaking of diverse languages, with each recording manually segmented into acoustic units (e.g., syllables, notes, and phrases) by the coauthor who recorded it in their own first/heritage language. Furthermore, because singing and speaking exist on a broader "musi-linguistic" spectrum including forms such as instrumental music and poetry recitation (*41*–*43*), we collected four types of recordings to capture variation across this spectrum: (i) singing, (ii) recitation of the sung lyrics, (iii) spoken description of the song, and (iv) instrumental version of the sung melody (Fig. 1).

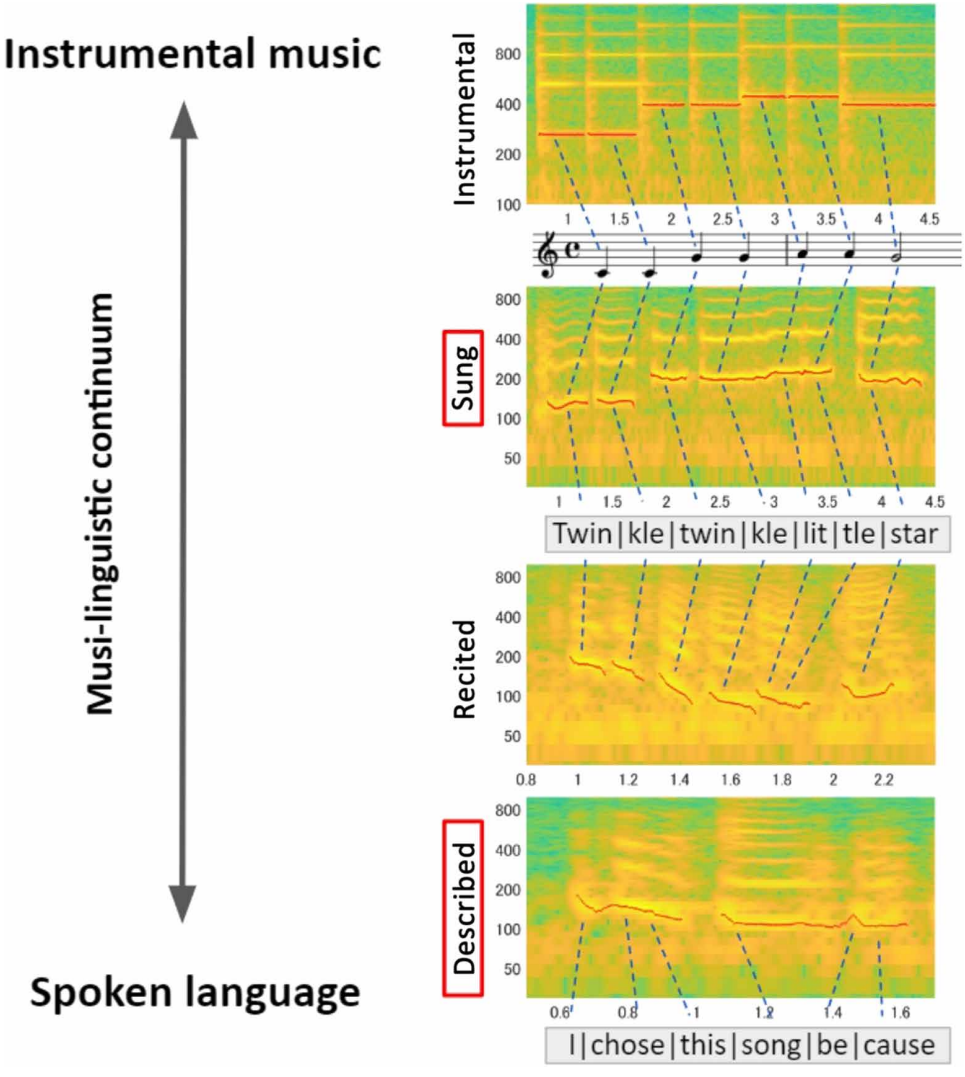



**Fig. 1. Example excerpts of the four recording types collected in this study arranged in a musi-linguistic continuum from instrumental music to spoken language.** Spectrograms [*x* axis, time (in seconds); *y* axis, frequency (in hertz)] of the four types of recordings are displayed on the right-hand side (excerpts of author Savage performing/describing "Twinkle Twinkle Little Star," using a piano for the instrumental version). Blue dashed lines show the schematic illustration of the mapping between the audio signal and acoustic units (here syllables/notes). For this Registered Report, we focus our confirmatory hypothesis only on comparisons between singing and spoken description (red rectangles), with recited and instrumental versions saved for post hoc exploratory analysis.

The spoken description represents a sample of naturalistic speech. In contrast, the lyrics recitation allows us to control for potential differences between the words and rhythmic structures used in song versus natural speech by comparing the exact same lyrics when sung versus spoken but, as a result, may be more analogous to poetry than to natural speech. The instrumental recording is included to capture the full musi-linguistic spectrum from instrumental music to spoken language, allowing us to determine how similar/different music and speech are when using the same effector system (speech versus song) versus a different system (speech versus instrument).

### Study aims and hypotheses

Our study aims to determine cross-cultural similarities and differences between speech and song. Many evolutionary hypotheses result in similar predicted similarities/differences between speech and song: for example, song may use more stable pitches than speech to signal desirability as a mate and/or to facilitate harmonized singing and by association bond groups together or signal their bonds to outside groups (*44*). These similarities and differences between song and speech could arise through a combination of purely cultural evolution, purely biological evolution, or some combination of gene-culture coevolution (*11*, *45*, *46*). Rather than try to disambiguate these ultimate theories, we focus on testing more proximate predictions about similarities and differences in the acoustic features of song and speech, which can then be used to develop more cross-culturally general ultimate theories in future research. Through literature review and pilot analysis (details provided in the "Pilot data analysis" section in the Supplementary Materials), we settled on six features that we believe we can reliably test for predicted similarities/differences: (i) pitch height, (ii) temporal rate, (iii) pitch stability, (iv) timbral brightness, (v) pitch interval size, and (vi) pitch declination (compare Table 1). Detailed speculation on the possible mechanisms underlying potential similarities and differences are described in the "Literature review of hypotheses and potential mechanisms" section in the Supplementary Materials.

## RESULTS

We have recruited 75 collaborators from around the world, spanning the speakers of 21 language families (Fig. 2) [Note: language classification follows the conventions of Glottolog and the World Atlas of Language Structures (*47*, *48*)]. Approximately 85% of our coauthors are first-language speakers of their recorded language (compare the "List of songs, instruments, and languages" section in the Supplementary Materials). Note that 6 of the original 81 planned coauthors were unable to complete the recording and annotation process compared to our initially planned sample (compare the Fig. 2 map with the originally planned fig. S1 map). These six collaborators were excluded, following our exclusion criteria (compare the "Exclusion criteria and data quality checks" section in the Supplementary Materials). Two collaborators (Thorne and Hereld) submitted recording sets with spoken descriptions in English instead of the language of their song (Te Reo Māori and Cherokee, respectively), and have not yet been able to rerecord themselves in the correct language as required by the "Recording protocol" (which can be found in the Supplementary Materials). Hereld's recording set is also an uncontrolled amalgam of recordings made for different settings. We have thus included Thorne and Hereld's recordings for the exploratory analyses but excluded them from the confirmatory analyses (i.e., 73 recording sets were used in the confirmatory analysis).

All audio recordings analyzed are made by our group of 75 coauthors recording ourselves singing/speaking in our first/heritage languages. Collaborators were chosen by opportunistic sampling beginning from cocorresponding author Savage's network of researchers (compare the "Language sample" section in the Supplementary Materials for details). Each coauthor made four recordings: (i) singing a traditional song chosen by the singer themself, (ii) reciting the song's lyrics, (iii) spoken description of the song's meaning, and (iv) instrumental version of the song's melody. The first 20 s of each recording was used for confirmatory analyses. Note that 28 instrumental recordings were made by clapping the rhythm of songs or using electronic instruments whose pitches are mechanically controlled. These recordings were excluded from analyses involving features related to pitch, such as pitch height. Although we asked coauthors to record traditional songs of their cultures, the chosen songs are not necessarily representative of the repertoires of their traditions. We did not collect standardized information about the function/context of songs, but the word clouds of lyrics translated to English (compare Fig. 2B) may provide an idea about what the songs are about, as do the English translations of the spoken descriptions (all available with other data at https://osf.io/mzxc8).

We compared the following six acoustic features (Fig. 3; compare the "Features" section in the Supplementary Materials for details) between song and speech for our main confirmatory analyses:

1) Pitch height [fundamental frequency ($f_0$)] (in hertz). $f_0$ is estimated with a custom tool in a semiautomated way like the annotation in the Erkomaishvili dataset (*49*), which used an interactive $f_0$ extraction tool (*50*).

2) Temporal rate [interonset interval (IOI) rate] (in hertz). The unit of IOI is seconds, and IOI rate is the reciprocal of IOI. Onset represents the perceptual center (P-center) of an acoustic unit (e.g., syllables, mora, and note), which represents the subjective moment when the sound is perceived to begin. The P-center can be interpreted to reflect the onset of linguistic units (e.g., syllable and mora) and musical units (e.g., note), with the segmentation of acoustic units determined by the person who made the recording. This measure includes the interval between a break and the onset immediately preceding the break. Breaks were defined as relatively long pauses between sounds. For vocal recordings, that would typically constitute when the participant would inhale.

3) Pitch stability ($-|\Delta f_0|$) (in cents per second).

4) Timbral brightness (spectral centroid) (in hertz).

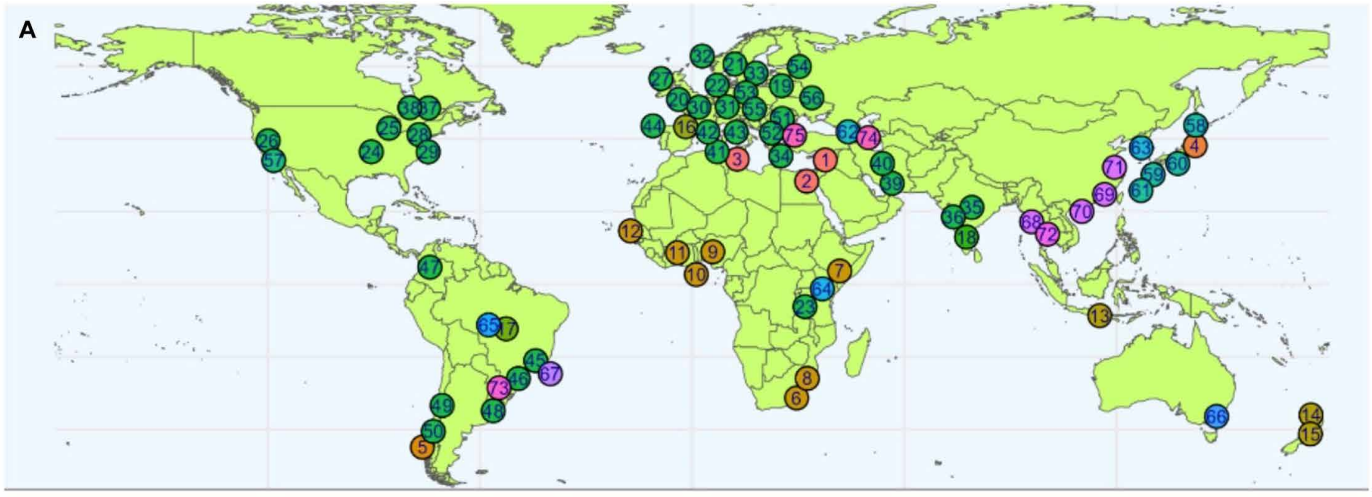
5) Pitch interval size ($f_0$ ratio) (in cents). Absolute value of pitch ratio converted to the cent scale.

6) Pitch declination (sign of $f_0$ slope) (dimensionless). Sign of the coefficient of robust linear regression fitted to the phrase-wise $f_0$ contour. A phrase is identified by the onset annotation after the break annotation (or the initial onset annotation for the first phrase) and the first break annotation following that.

For each feature, we compared its distribution in the song recording with its distribution in the spoken description by the same singer/speaker, converting their overall combined distributions into a single scalar measure of nonparametric standardized difference (compare Materials and Methods). Details can be found in the "Features" section in the Supplementary Materials. Temporal rate, pitch interval

**Table 1. Registered Report design planner includes six hypotheses (H1 to H6).**

| Question | Hypothesis | Sampling plan | Analysis plan | Rationale for deciding the test sensitivity | Interpretation given different outcomes | Theory that could be shown wrong by the outcomes | Actual outcome |
|---|---|---|---|---|---|---|---|
| Are any acoustic features reliably different between song and speech across cultures? | 1) Song uses higher pitch than speech | $n = 81$ pairs of audio recordings of song/speech, with each pair sung/spoken by the same person (Fig. 2). Recruitment was opportunistic based on collaborator networks aiming to maximize global diversity and achieve greater than 95% a priori power even if some data have to be excluded (see the "Language sample" section in the Supplementary Materials for inclusion/exclusion criteria). | Meta-analysis framework (compare Materials and Methods) calculates a paired effect size for pitch height ($f_0$) for each song/speech pair and tests whether the population effect size (relative effect $p_{re}$) is significantly larger than 0.5. | Power analysis estimate of minimum $n = 60$ pairs was based on converting Brysbaert's (*93*) suggested smallest effect size Of interest (SESOI) of Cohen's $D = 0.4$ to the corresponding $p_{re} = 0.61$. We control for multiple comparisons using false discovery rate (Benjamini-Hochberg step-up method; family-wise $\alpha = 0.05$; $\beta = 0.95$). | The null hypothesis of no difference in $f_0$ between sung and spoken pitch height is rejected if the population effect size is significantly larger than $p_{re} = 0.5$. Otherwise, we neither reject nor accept the hypothesis. | Our design cannot falsify specific ultimate theories (e.g., social bonding hypothesis and motor constraint hypothesis) but can falsify cultural relativistic theories that argue against general cross-cultural regularities in song-speech relationships. | All three hypothesized differences between song and speech (pitch height, temporal rate, and pitch stability) were confirmed |
| | 2) Song is slower than speech | Same as H1, but for temporal rate [interonset interval (IOI) rate] instead of pitch height ($f_0$) | | | | | |
| | 3) Song uses more stable pitches than speech | Same as H1, but for pitch stability ($-\lvert\Delta f_0\rvert$) instead of pitch height | | | | | |
| Are any acoustic features reliably shared between song and speech across cultures? | 4) Song and speech use similar timbral brightness | Same as H1. | Same as H1, except test whether the effect size for timbral brightness is significantly smaller than the SESOI. | Same as H1. | The null hypothesis of spectral centroid of singing being meaningfully lower or higher than speech is rejected if the population effect size is significantly within the SESOI ($0.39 < p_{re} < 0.61$, corresponding to $\pm 0.4$ of Cohen's *D*. Otherwise, we neither reject nor accept the hypothesis. | Same as H1. | The hypothesized similarities in timbral brightness and pitch interval size were confirmed |
| | 5) Song and speech use similar sized pitch intervals | Same as H4, but for pitch interval size ($f_0$ ratio) instead of timbral brightness. | | | | | |
| | 6) Song and speech use similar pitch contours | Same as H4, but for pitch declination (sign of $f_0$ slope) instead of timbral brightness. | | | | | The hypothesized similarity in pitch contour was neither rejected nor confirmed. |




| Afro-Asiatic | | Indo-European | | | | Nilotic | |
|---|---|---|---|---|---|---|---|
| 1 Modern Hebrew [Jerusalem] | | 19 Lithuanian | Baltic | 44 Portuguese [Porto] | | 64 Luo (dholuo) (Luo (Kenya and Tanzania)) | |
| 2 Modern Hebrew [Tel Aviv] | | 20 Gaeilge (Irish) | Celtic | 45 Portuguese [São Paulo] | | **Nuclear-Macro-Jê** | |
| 3 Tunisian Arabic | | 21 Danish | Germanic | 46 Portuguese [São Paulo] | | 65 Rikbaktsa | |
| **Ainu** | | 22 Dutch [Heemstede] | | 47 Spanish [Bogotá] | | **Pama-Nyungan** | |
| 4 Aynu (Hokkaido Ainu) | | 23 Dutch [Nairobi] | | 48 Spanish [Montevideo] | | 66 Ngarigu | |
| **Araucanian** | | 24 English [Indiana] | | 49 Spanish [Santiago] | | **Puri-Coroado** | |
| 5 Tsesungún (Huilliche) | | 25 English [Michigan] | | 50 Spanish [Osorno] | | 67 Puri Kwaytikindo (Puri) | |
| **Atlantic-Congo** | | 26 English [Nevada] | | 51 Bulgarian | Slavic | **Sino-Tibetan** | |
| 6 IsiXhosa (Xhosa) | Bantu | 27 English [Newry] | | 52 Macedonian | | 68 Myanmar (Burmese) | Burmese-Lolo |
| 7 Kiswahili (Swahili) | | 28 English [Pennsylvania] | | 53 Polish | | 69 Cantonese (Yue Chinese) | Chinese |
| 8 Ronga | | 29 English [Washington D.C.] | | 54 Russian | | 70 HainanHua (Min Nan Chinese) | |
| 9 Yoruba | Defoid | 30 Flemish (Dutch) | | 55 Slovenian | | 71 Mandarin Chinese | |
| 10 Fante (Akan) | Tano | 31 German | | 56 Ukrainian | | **Tai-Kadai** | |
| 11 Twi (Akan) | | 32 Norwegian | | **Iroquoian** | | 72 Thai | |
| 12 Wolof | Wolof | 33 Svenska (Swedish) | | 57 Cherokee | | **Tupian** | |
| **Austronesian** | | 34 Greek | Greek | **Japonic** | | 73 Mbyá-Guaraní | |
| 13 Balinese | | 35 Hindi | Indic | 58 Japanese [Hokkaido] | | **Turkic** | |
| 14 Te Reo Māori (Māori) [Auckland] | | 36 Marathi | | 59 Japanese [Hyogo] | | 74 North Azerbaijani | |
| 15 Te Reo Māori (Māori) [Wellington] | | 37 Punjabi (Eastern Panjabi) | | 60 Japanese [Tokyo] | | 75 Turkish | |
| **Basque** | | 38 Urdu | | 61 Northern Amami-Oshima | | | |
| 16 Euskara (Basque) [Errenteria] | | 39 Western Farsi [Isfahan] | Iranian | **Kartvelian** | | | |
| **Cariban** | | 40 Western Farsi [Tehran] | | 62 Georgian | | | |
| 17 Língua Kuikuro (Kuikúro-Kalapálo) | | 41 Catalan | Romance | **Koreanic** | | | |
| **Dravidian** | | 42 French | | 63 Korean | | | |
| 18 Kannada | | 43 Italian | | | | | |

B


C

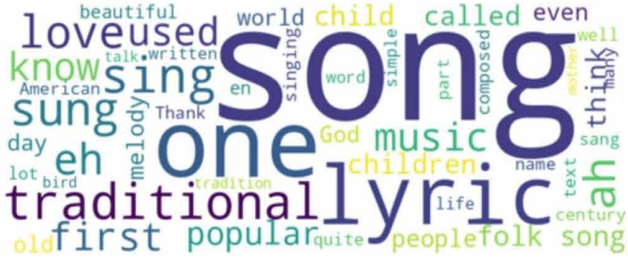



**Fig. 2. Visualization of the diversity of the primary sample of 300 audio recordings of singing/speaking/recitation/instrumental melodies.** Map of the linguistic varieties spoken by our 75 coauthors as first/heritage languages (**A**). (Note: 6 of the original 81 planned coauthors were unable to complete the recording and annotation process compared to our initially planned sample; compare fig. S1 for the original map of 81 linguistic varieties). Each circle represents a coauthor singing and speaking in their first (L1) or heritage language. The geographic coordinates represent their hometown where they learned that language. In cases when the language name preferred by that coauthor (ethnonym) differs from the L1 language name in the standardized classification in the Glottolog (*47*), the ethnonym is listed first, followed by the Glottolog name in round brackets. Language family classifications (in bold) are based on Glottolog. Square brackets indicate geographic locations for languages represented by more than one coauthor. Atlantic-Congo, Indo-European, and Sino-Tibetan languages are further grouped by genus defined by the World Atlas of Language Structures (*48*). The word clouds outline the most common textual content of English translations of the song lyrics (**B**) and spoken descriptions (**C**) provided by our 75 coauthors (larger text indicates words that appear more frequently).

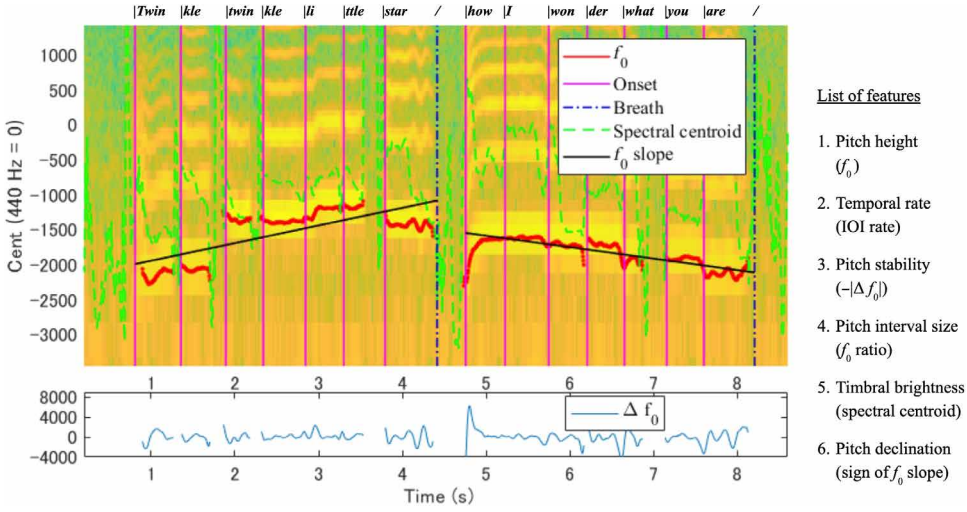



**Fig. 3. Schematic illustration of the six features analyzed for confirmatory analysis, using a recording of author Savage singing the first two phrases of "Twinkle Twinkle Little Star" as an example.** Onset and breathing annotations are based on the segmented texts displayed on the top of the spectrogram. The *y* axis is adjusted to emphasize the $f_0$ contour, so note that the spectral centroid information is not fully captured (e.g., high spectral centroid due to the consonant). The bottom figure shows pitch stability (rate of change of $f_0$ or derivative of the $f_0$ contour equivalently) of the sung $f_0$.

size, and pitch declination rely on the onset and break segmentations (roughly corresponding to note/syllable and phrase/breath boundaries, respectively). These segmentations were made manually by the coauthor who made the recording, as they are determined subjectively by the perception of the coauthors, as described above. First author Ozaki performed millisecond-level onset annotations of all recordings based on these segmentations, and coauthors checked the quality of annotations of their recordings (compare the "Recording and segmentation protocol" section in the Supplementary Materials).

## Confirmatory analysis

The results of the confirmatory hypothesis testing with 73 recording sets confirmed five of our six predictions (Fig. 4 and table S1; all $P < 1 \times 10^{-5}$). Specifically, relative to spoken descriptions, songs used significantly higher pitch (translated Cohen's $D = 1.6$), slower temporal rate ($D = 1.6$), and more stable pitches ($D = 0.7$), while both spoken descriptions and songs used significantly equivalent timbral brightness and pitch interval size (both $D < 0.15$). The one exception was pitch declination, which was not significantly equivalent between speech and song ($P = 0.57$), with an estimated effect size of $D = 0.42$ slightly greater than our prespecified "smallest effect size of interest" (SESOI) of $D = 0.4$. In the "Alternative analysis approaches for pitch declination (hypothesis 6)" section, we performed alternative exploratory analyses to understand possible reasons for this failed prediction.

Our robustness checks (compare the "Robustness analyses" section in the Supplementary Materials) confirmed that the tests with the recordings excluding collaborators who knew the hypotheses when generating data lead to the same decisions regarding the rejection of the null hypotheses (table S2). This result suggests that our unusual "participants as coauthors" model did not influence our confirmatory analyses. In addition, the other robustness check suggests that the measured effect sizes did not have language family-specific variance (table S3), which supports the appropriateness of the use of simple random-effect models in the analyses.

## Exploratory analysis

### *More acoustic features*

We specified six features for our confirmatory analyses, but human music and speech can be characterized by additional acoustic features. We included seven additional features to probe further similar and different aspects of music and speech, namely, rhythmic regularity, phrase length (duration between two breaths/breaks), pitch interval regularity, pitch range, intensity, pulse clarity, and timbral noisiness (compare the "Exploratory features" section in the Supplementary Materials). All 13 features were ones that we explored in our stage 1 pilot analyses based on previous analyses of acoustic features of music and speech (fig. S9). However, we chose to limit our confirmatory (preregistered) analyses to the six features that seemed most promising when considering both theoretical debate and pilot data to ensure sufficient statistical power to reliably test our hypotheses. For completeness, we also included the remaining seven features as exploratory analyses. Although we did not formally construct and test hypotheses for this analysis, Fig. 4 suggests that phrase length, intensity, and timbral noisiness may also inform differences between song and speech and pitch range can be another candidate for demonstrating similarities between song and speech. Specifically, songs appear to have longer intervals between breathing and higher sound pressure and have less vocal noise than speech. Note that the order of comparison was arranged so that difference is expressed as a positive value, so that difference in timbral noisiness was calculated as noisiness of spoken description relative to song (compare Materials and Methods).

### *Music-language continuum: Including instrumental melodies and recited lyrics*

Exploratory analyses that included comparisons with lyrics recitation and instrumental recordings (compare Fig. 5 and fig. S13) suggest that (i) comparing singing versus lyrics recitation showed

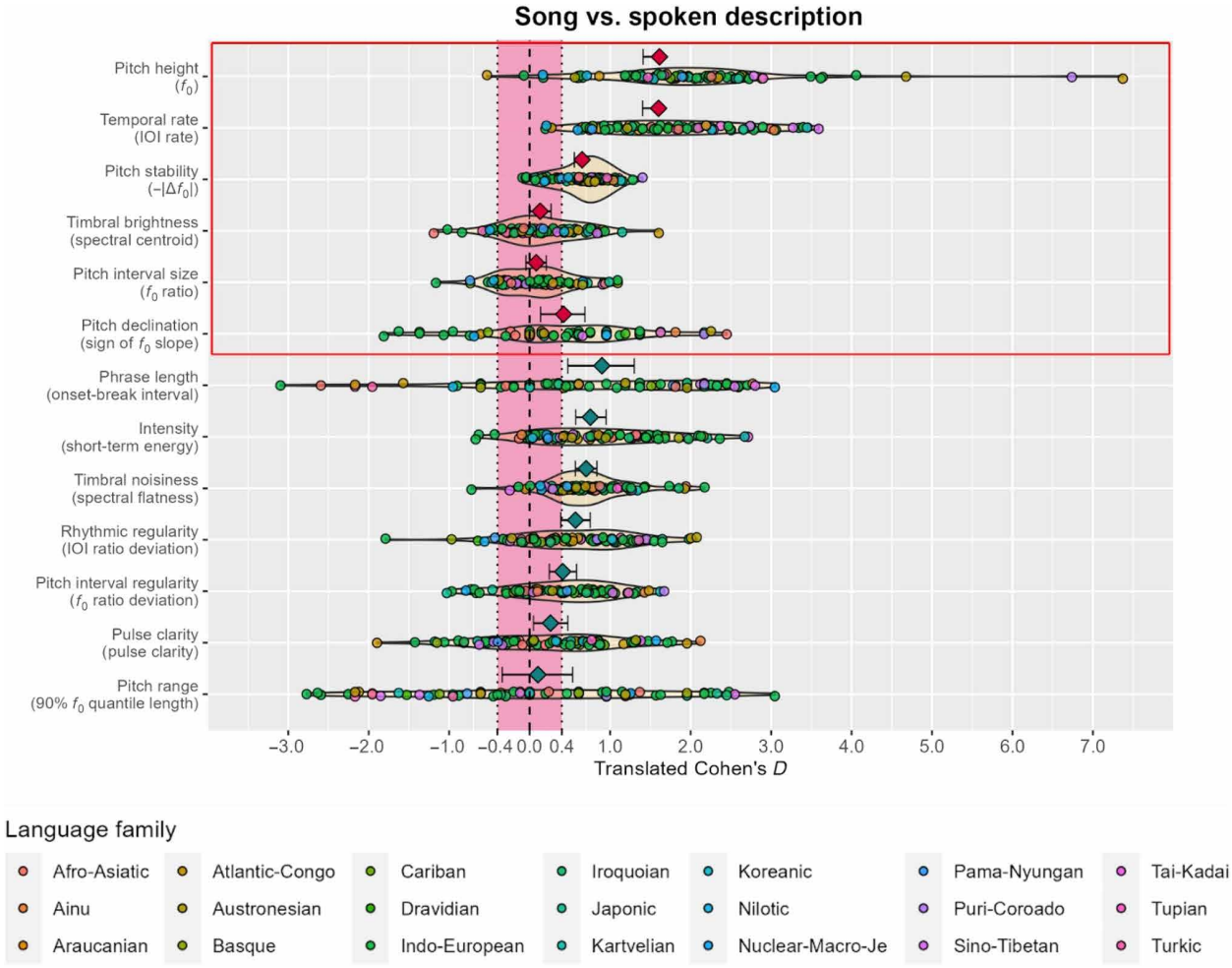



**Fig. 4. Plot of effect sizes showing differences of each feature between singing and spoken description of the 73 recording sets for the confirmatory analysis and 75 recording sets for the exploratory analysis.** The plot includes seven additional exploratory features, and the six features corresponding to the main confirmatory hypotheses are enclosed by the red rectangle. Confidence intervals are created using the same criteria in the confirmatory analysis (i.e., $\alpha = 0.05/6$). Each circle represents the effect size from each recording pair of singing and spoken description, and the set of effect sizes is measured per recording pair. Readers can find further information about how to interpret the figure in the caption of figs. S2 and S9. Note that the colors of data points indicate language families, which are coded the same as in Fig. 2, and violin plots are added to this figure compared to fig. S2.

qualitatively the same results as for singing versus spoken description in terms of how confidence intervals intersect with the null point and the equivalence region; (ii) comparing instrumental versus speech (both spoken description/lyrics recitation) revealed larger differences in pitch height, temporal rate, and pitch stability than found with song versus speech; (iii) features shown to be similar between song versus speech (e.g., timbral brightness and pitch interval size) showed differences when comparing instrumental versus speech; (iv) few major differences were observed between lyrics recitation and spoken description, except that recitation tended to be slower and use shorter phrases; (v) instrumental performances generally had a more extreme (larger/smaller) magnitude than singing for each feature except for temporal rate; and (vi) pitch height, temporal rate, and pitch stability displayed a noticeable constantly increasing (or decreasing) continuum from spoken description to instrumental.

A similar trend was also found in additional differentiating features discussed in the "More acoustic features" section (i.e., phrase length, timbral noisiness, and loudness). We also performed a nonparametric trend test (compare, table S4) to quantitatively assess the existence of trends, and the result suggests that features other than pitch interval size and pitch range display increasing/decreasing trends. These results tell us how acoustic characteristics are manipulated through the range of acoustic communication from spoken language to instrumental music.

### *Demographic factors: Sex differences in features*

Because we had a similar balance of female ($n = 34$) and male ($n = 41$) coauthors, we were able to perform exploratory analysis comparing male and female vocalizations (fig. S14). These analyses suggest that while there is some overlap in their distribution (e.g., some male speaking/singing was higher than some female speaking/singing), on average, female vocalizations were consistently higher-pitched than male vocalizations regardless of the language sung/spoken [by ~1000 cents (almost one octave) consistently for song, spoken description, and recited lyrics]. Specifically, the average frequencies of

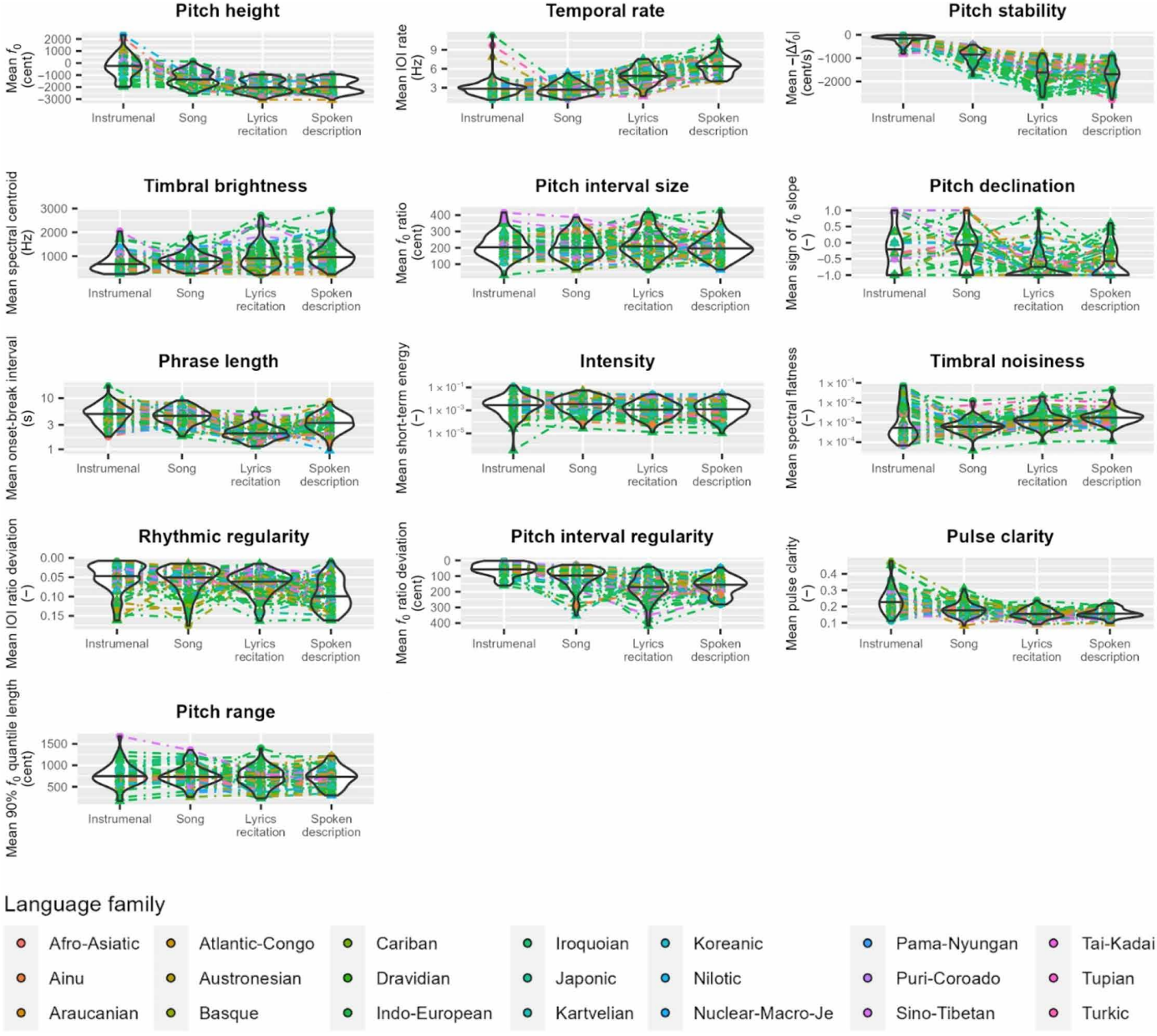



**Fig. 5. Mean values of acoustic features arranged along a "musi-linguistic continuum" from instrumental melodies to spoken descriptions.** This is an alternative visualization of the same sung/spoken data from Fig. 4, but showing mean values of each feature rather than paired differences, and now also including data for instrumental melodies and recited lyrics. The cent scale of $f_0$ is converted from Hertz, where 440 Hz corresponds to 0 cents and an octave interval equals 1200 cents. Note that the colors of data points indicate language families, which are coded the same as in Fig. 2. The horizontal lines in the violin plots indicate the median.

our data are as follows: male song, 161.3 Hz; male spoken description, 114.2 Hz; female song, 289.9 Hz; and female spoken description, 199.9 Hz. Incidentally, the data of Hilton *et al.* (*26*) also provide a similar result: male song, 152.6 Hz; male speech, 130.7 Hz; female song, 251.4 Hz; and female speech, 209.7 Hz. However, there was no apparent sexual dimorphism in vocal features other than pitch height (e.g., temporal rate, pitch stability, timbral brightness, etc.). Although this analysis is exploratory, this result is consistent with past research that often focuses on vocal pitch as a likely target of sexual selection (*10*, *51*–*55*).

### *Analysis by linguistic factors: Normalized pairwise variability index*

We used normalized pairwise variability index (nPVI) (*56*) to examine the degree of variation in IOIs and onset-break intervals (compare the "Temporal rate" and "Break annotation" sections in the Supplementary Materials) of our song and speech recordings. nPVI provides large values if adjacent intervals differ in duration on average and vice versa. Thus, nPVI can capture durational contrasts between successive elements. It was originally developed to characterize vowel duration of stress-timed and syllable-timed languages (*57*),

although our duration is defined by the sequence of onset (compare the "Recording and segmentation protocol" section in the Supplementary Materials) and break annotations (compare the "Break annotation" section in the Supplementary Materials) that are neither the same as vowel duration nor vocalic intervals. In this exploratory analysis, we mapped nPVIs of song and spoken description recordings of each collaborator on a two-dimensional space to explore potential patterns and also visualized the density of nPVIs per recording type (compare fig. S20). However, we observed that (i) nPVIs of song and spoken description did not seem to create distinct clusters among our recordings (whether into syllable-timed, stress-timed, or any other categories); (ii) nPVIs tended to increase along the musi-linguistic continuum, progressing from instrumental to spoken description; and (iii) nPVIs of song and spoken description did not have a clear correlation (Pearson's $r = 0.087$), while nPVIs of song and instrumental recording do show a substantial correlation (Pearson's $r = 0.52$). The first result does not necessarily imply that nPVIs are not helpful in classifying recordings into rhythm categories. There is a possibility that languages are actually well separated by rhythm classes (e.g., stress-timed, syllable-timed, and mora-timed) in fig. S20, although we could not find information about rhythm classes of all languages in our recordings. The first result suggests that data-driven discovery of rhythm categories is challenging with nPVIs for our data, although evaluating its capability to predict rhythm categories needs a different analysis. The second result suggests that durational contrast of speech is more variable compared to singing and instrumental, which is consistent with past work showing that music tends to have limited durational variability worldwide (*25*). Last, although linguists use various features (*58*) to carefully characterize the rhythm of speech, the third result suggests that song rhythm is potentially independent of speech rhythm even when produced by the same speaker in the same language, which suggests that temporal control of song and speech may obey different communicative principles.

### *Reliability of annotation process: Interrater reliability of onset annotations*

We analyzed the interrater reliability of onset annotations to check how large individual variation is in the annotation. Savage created onset annotations to the first 10 s of randomly chosen eight pairs of song and spoken description recordings (compare the "Reliability of annotation process" section in the Supplementary Materials). In this 10-s annotation, Savage created onset annotations using the same segmented text as Ozaki (the text provided by the coauthor who made the recording) but was blinded from the actual annotation created by Ozaki and confirmed by the coauthor who made the recording. Therefore, the annotation by Savage follows the same segmentation as the annotation by Ozaki but can differ in the exact timing for which each segmentation is judged to begin. We measured intraclass correlations of onset times with two-way random-effects models measuring absolute agreement. As a result, all annotations showed that strong intraclass correlations (>0.99), which indicates that who performs the annotation may not matter as long as they strictly follow the segmentation indicated in segmented texts. Alternative exploratory analysis inspecting the distribution of differences in onset times was also conducted (compare, fig. S21). In the case of singing, 90% of onset time differences were within 0.083 s. Similarly, in the case of spoken description, 90% of onset time differences were within 0.055 s. That is, Ozaki's manual onset annotations that formed a core part of our dataset have been confirmed by the coauthor who produced each recording and by Savage's independent blind codings to be highly accurate and reliable.

### *Exploring recording representativeness and automated scalability: Comparison with alternative speech-song dataset*

We performed two exploratory analyses using automated methods to investigate (i) the reproducibility of our findings with another corpus and (ii) the applicability of automated methods to substitute data extraction processes involving manual work (Fig. 6; compare



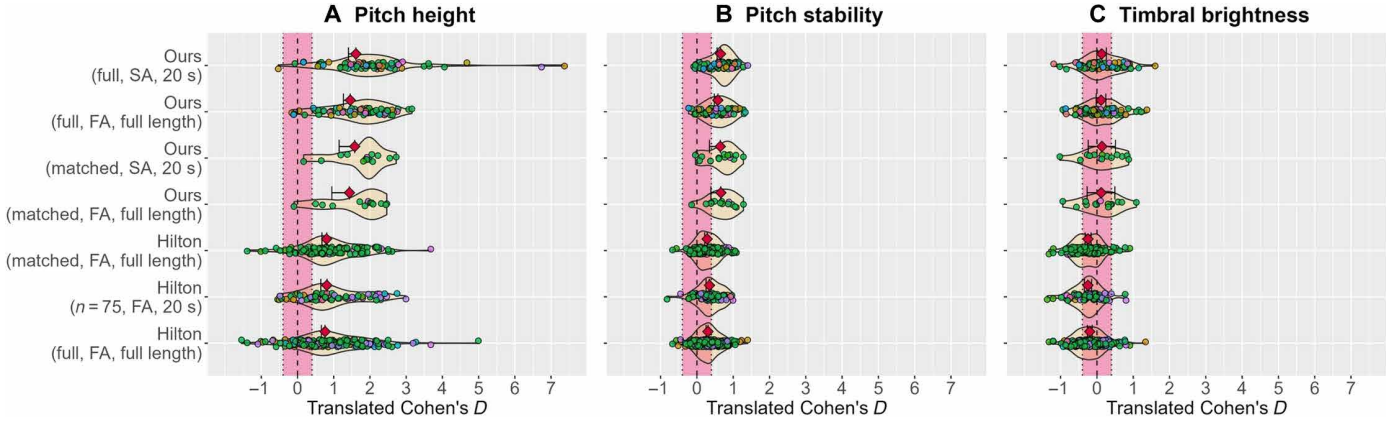



**Fig. 6. Rerunning the analyses on four different samples using different fundamental frequency extraction methods.** Three features could be directly compared between the different samples: pitch height (**A**), pitch stability (**B**), and timbral brightness (**C**). The following samples were compared: (i) Our full sample (matched song and speech recordings from our 75 coauthors); (ii) Hilton *et al.*'s (*26*) full sample (matched song and speech recordings from 209 individuals); (iii) a subsample of our 14 coauthors singing/speaking in English, Spanish, Mandarin, Kannada, and Polish; and (iv) a subsample of Hilton *et al.*'s (*26*) 122 participants also singing/speaking in English, Spanish, Mandarin, Kannada, and Polish). "SA" means that $f_0$ values are extracted in a semiautomated manner (compare the "Pitch height" section in the Supplementary Materials), while "FA" means they were exactly in a fully automated manner (using the pYIN algorithm). The visualization follows the same convention as in Figs. 4 and 5. However, Hilton *et al.*'s (*26*) dataset contains languages that are not in our dataset. Therefore, slightly different color mapping was applied (compare fig. S16). Note that some large effect sizes ($D > 3.5$) in the pitch height of our original analysis (i.e., full, SA, 20 s) are not observed in the automated analysis (i.e., full, FA, full length). This is due to estimation errors in the automated analyses. When erroneous $f_0$ values of pYIN are very high in spoken description or very low in singing, relative effects become smaller than semiautomated methods that remove these errors.

the "Exploring recording representativeness and automated scalability" section in the Supplementary Materials). We analyzed the recordings of adult-directed singing and speech of Hilton *et al.*'s (*26*) dataset. We especially analyzed both the full set of their data and the subset of their data representing languages also present in our own dataset—English, Spanish, Mandarin, Kannada, and Polish—to perform a matched comparison with our language varieties. However, in their dataset, not all individuals made a complete set of recordings (infant/adult-directed song/speech), and we analyzed recording sets containing matching adult-directed song and adult-directed speech recordings, which resulted in 209 individuals for the full data (i.e., individuals from full 21 societies/16 languages) and 122 individuals for the above subset of five languages.

Our data extraction processes involving manual work are fundamental frequency extraction, sound onset annotation, and sound break annotation, and we automated fundamental frequency extraction since reliable fundamental frequency estimators applicable to both song and speech signals are readily available. On the other hand, reliable automated onset and break annotation for both song and speech is still challenging. For example, we observed that a widely used syllable nucleus segmentation method (*59*) failed to capture the major differences in temporal rate that we identified using manual segmentation in Fig. 4. Instead, if we had used this automated method, then we would have mistakenly concluded that there is no meaningful difference in IOI rates of singing and speech (fig. S15). Therefore, as described in our stage 1 protocol (compare the "Exploring recording representativeness and automated scalability" section in the Supplementary Materials), we only focused on the automation of $f_0$ extraction that could provide reliable results even using purely automated methods without requiring manual annotations.

We chose the probabilistic YIN (pYIN) (*60*) $f_0$ extraction algorithm for this analysis. In addition, we analyzed full-length recordings by taking advantage of the efficiency of automated methods. Note that our timbral brightness analysis is already fully automated, so we use the same analysis procedure for this feature. Semiautomated analyses could only be performed on 20-s excerpts of our recordings annotated by the coauthor who recorded them, while automated analyses could be applied to the full samples. To make the comparison with our results more interpretable, we have also added the analysis of Hilton *et al.*'s (*26*) data using the same number of song-speech recording pairs as ours (i.e., randomly selected 75 pairs of recordings), extracting features from the first 20 s. Since temporal rate, pitch interval size, and pitch declination analyses require onset and break annotations, we focused on pitch height, pitch stability, and timbral brightness.

The result suggests that (i) the same statistical significance could be obtained from Hilton *et al.*'s (*26*) data although overall effect sizes tend to be weakened and (ii) combined effect sizes based on pYIN with full-length duration only showed negligible differences from the original analysis involving manual work despite the marked difference in the measurement of some effect sizes (i.e., no effect sizes larger than 3.5 in the automated analysis of the pitch height of our data). Note that the differences in pitch stability in Hilton *et al.*'s (*26*) sample (translated Cohen's $D = 0.3$) are small enough to be within our defined equivalence region ($|D| < 0.4$) if we had predicted it not only to be equivalent but also significantly greater than the null hypothesis of no difference (translated Cohen's $D = 0$ corresponding to relative effect of 0.5), as we predicted ($P < 0.005$). Similar to Fig. 5, mean values of each feature per recording can be found in the Supplementary Materials (figs. S17 to S19).

***Alternative analysis approaches for pitch declination (hypothesis 6)***

The only one of our six predictions that was not confirmed was our prediction that song and speech would display similar pitch declination. However, only three to four $f_0$ slopes (equal to the number of "phrases" or intervals from the first onset after a break and to the next break; compare Fig. 3) are, on average, included in the 20-s length recording of singing and spoken description, respectively, and so it is possible that this failed prediction could be due to the relatively more limited amount of data available for this feature. Therefore, we additionally checked the validity of the result of this hypothesis test using a longer duration to extract more signs of $f_0$ slopes to evaluate effect sizes. Although we performed exploratory reanalysis using 30-s recordings that contain five to seven $f_0$ slopes for singing and spoken description on average, still, the $P$ value was not small enough to reject the null hypothesis ($P = 0.48$; confidence interval, 0.17 to 0.60).

Note that we are judging the declination in an $f_0$ contour by looking at the sign of the slope of linear regression (i.e., the sign is negative means declination). Therefore, even if the $f_0$ contour is an arch shape, which means that it has a descending contour at the end part, it can be judged as no declination if the linear regression shows a positive slope. Therefore, the declination here means if the $f_0$ contour has a descending trend overall and not necessarily if the phrase is ending in a downward direction.

We report here an additional analysis based on a different approach for handling the case when signs of $f_0$ slopes are not directly analyzable. Some singing and spoken description recording pairs only contained negative signs (i.e., descending trend prosody). This is undesirable for inverse variance–weighted based meta-analysis methods that we used (e.g., DerSimonian-Laird estimator) since the SDs of effect sizes become zero, leading to computation undefined. We used the same procedure used in our power analysis for these cases (compare the "Power analysis" section in the Supplementary Materials), but a more widely known practice would be zero-cell corrections used in binary outcome data analysis (*61*) (compare the "Applying zero-cell correction to the signs of $f_0$ slopes" section in the Supplementary Materials). This additional analysis provided virtually identical results with the main analysis reported in 3.1 ($P = 0.66$; confidence interval, 0.15 to 0.71), suggesting that the way to handle zero-frequency $f_0$ slope sign data is not crucial.

Last, we also checked the average trend of $f_0$ contours segmented by onset and break annotations (compare Fig. 7). The averaged $f_0$ contour of spoken description recordings clearly exhibits a predominantly descending trend, albeit with a slight rise at the end. In contrast, the averaged $f_0$ contour of songs is close to an arch shape, so that although the second half of songs tend to descend as predicted, the first half of songs tend to rise, in contrast to speech that tends to mostly descend throughout the course of a breath. Thus, on average, spoken pitch contours tend to descend more than sung pitch contours, explaining our failure to confirm our prediction that their contours would display similar pitch declination (compare Fig. 5). We also noticed that vocalizers sometimes end their utterance by raising pitch in their spoken description recordings (and lyrics recitation as well), causing a slight rise at the end of the averaged $f_0$ contour of spoken description (and lyrics recitation; compare Fig. 7).

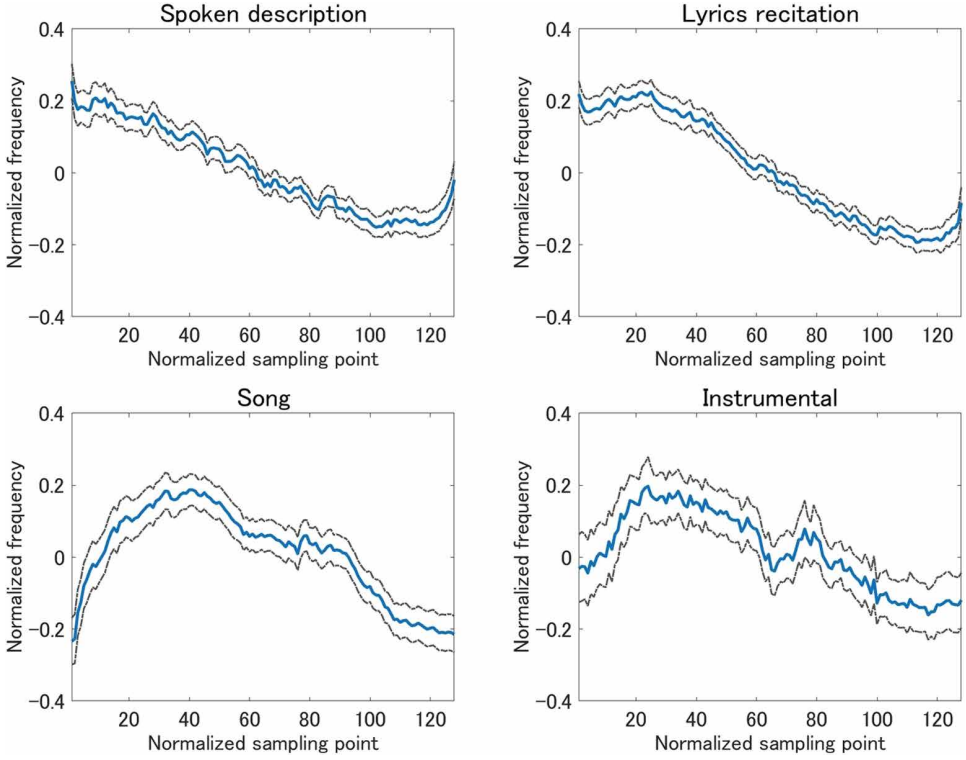



**Fig. 7. Averaged $f_0$ contours.** $f_0$ contours extracted by the segments between onset and break were averaged to visualize the overall trend. The length of contours is normalized to 128 samples. The average widths of confidence intervals of each category are 0.14 for instrumental, 0.097 for song, 0.060 for lyrics recitation, and 0.065 for spoken description. The details of the computation is provided in the "Computation of average $f_0$ contours of Fig. 7" section in the Supplementary Materials.

Furthermore, the width of SEs around the mean contour (compare Fig. 7) suggests that spoken description and lyrics recitation have more homogeneous variations of contours than song and instrumental. This difference may corroborate that music actually makes more use of the manipulation of the pitch in communication. Musical melodies are considered to have multiple typical shapes (*62*, *63*), so the overall average contour is not necessarily representative of all samples.

### *Explanatory power of the features in song-speech classification*

To probe the explanatory power of features in classifying acoustic signals into song and speech, we evaluated feature importance using permutation importance (*64*) with three simple machine learning models. Permutation importance informs the influence on the machine learning model by a particular variable by randomly shuffling the data of the variable (e.g., imagine a data matrix that row corresponds to observations and column corresponds to variables, and the data in a particular column are shuffled). Here, we use the permutation importance, which is the version implemented in Python's ELI5 package (*65*). Since how the feature contributes to solving the given task differs in machine learning models, we used three binary classification models to mitigate the bias from particular models: logistic regression with L2 regularization, support vector machines with radial basis function kernel, and naive Bayes with Laplace smoothing. The details of the computation are provided in the "Computation of permutation importance" section in the Supplementary Materials.

The result suggests that temporal rate, pitch stability, and pitch declination are constantly weighed among these three models (compare fig. S22). Several features showing a strong difference within participants were not evaluated as important in this analysis, including pitch height and intensity (compare Fig. 4 and fig. S22). Two reasons can be considered. One reason is that the difference in features (e.g., pitch height) between song and speech produced by the same person is not informative in classifying acoustic signals collected from multiple individuals. In this case, between-participants consistent differences would be more informative. Another scenario is that there is an overlap in information among features. Correlation matrices of the features within song and speech (compare figs. S23 and S24) show that several features have medium to large correlation (e.g., increase in pitch interval regularity with a decrease in temporal rate in singing with $r = -0.53$). Therefore, there is a possibility that some features are evaluated as unimportant, not because that feature is irrelevant to classifying song and speech but because the information in that feature overlaps with other features. This comes from the limitation of permutation importance that this measurement does not take into account correlation among features. Correlation is considered to cause the underestimation of permutation importance (*66*).

Inspection of the correlation matrices suggests that complex interactions exist among features. Although what is captured in correlation matrices is a linear dependency between two variables, nonlinear dependency among features or dependency among more than two variables can also happen in vocal sound production. Further study is necessary to accurately disentangle the importance of the features from complex interactions. However, the current analysis indicates that there are two features, namely, temporal rate and pitch stability,

that consistently scored high among the three between-participants models and confirmed our predicted within-participants differences. This coincidence suggests that temporal rate and pitch stability may capture important factors differentiating song and speech across cultures.

## DISCUSSION

### Main confirmatory predictions and their robustness

Our analyses strongly support five of our six predictions across an unprecedentedly diverse global sample of music/speech recordings: (i) Song uses higher pitch than speech, (ii) song is slower than speech, (iii) song uses more stable pitches than speech, (iv) song and speech use similar timbral brightness, and (v) song and speech use similar sized pitch intervals (Fig. 4). Furthermore, the first three features display a shift of distribution along the musi-linguistic continuum, with instrumental melodies tending to use even higher and more stable pitches than song and lyric recitation tending to fall in between conversational speech and song (Fig. 5).

While some of our findings were already expected from previous studies mainly focused on English and other Indo-European languages (*21*, *32*–*34*, *67*) [see also the "Literature review of hypotheses and potential mechanisms" section in the Supplementary Materials and (*36*)], our results provide the strongest evidence to date for the existence of "statistically universal" relationships between music and speech across the globe. However, none of these features can be considered an "absolute" universal that always applies to all music/speech. Figure 4 shows many exceptions for four of the five features: for example, Parselelo (Kiswahili speaker) sang with a lower pitch than he spoke, and Ozaki (Japanese speaker) used slightly more stable pitches when speaking than singing, while many recording sets had examples where differences in sung versus spoken timbre or interval size were substantially larger than our designated SESOI. The most consistent differences are found for temporal rate, as song was slower than speech for all recording sets in our sample. However, additional exploratory recordings have revealed examples where song can be faster than speech [e.g., Savage performing Eminem's rap from "Forgot About Dre" (https://osf.io/ba3ht); Parselelo's recording of traditional Moran singing by Ole Manyas, a member of Parselelo's ancestral Maasai community (https://osf.io/mfsjz)].

Our sixth prediction—that song and speech use similar pitch contours—remained inconclusive. Instead of our predicted similarities, our exploratory analyses suggest that, while both song and speech contours tend to decline toward the end of a breath, they tend to do so in different ways: song first rising before falling to end near the same height as the beginning, speech first descending before briefly rising at the end (Fig. 7). Our prediction was based in part on past studies by some of us finding similar pitch contours in human and bird song, which we argued supported a motor constraint hypothesis (*28*, *68*). However, our current results suggest that motor constraints alone may not be enough to explain similarities and differences between human speech, human song, and animal song and that future studies directly comparing all three domains will be needed.

Our robustness checks confirmed that our primary confirmatory results were not artifacts of our choice to record from a nonrepresentative sample of coauthors. Specifically, (i) language families did not account for variances in the measured song-speech differences and similarities (table S3), which means that these differences and similarities are cross-linguistically regular phenomena; and (ii) analyzing only recordings from coauthors who made recordings before learning our hypotheses produced qualitatively identical conclusions (table S2). Analysis of Hilton *et al.*'s (*26*) dataset of field recordings also supplemented our findings, producing qualitatively identical conclusions, regardless of the precise analysis methods or specific sample/subsample used (Fig. 6).

### Inclusivity and global collaboration

Our use of a "participants-as-coauthors" paradigm allowed us to discover findings that might not have been possible otherwise. For example, collaboration with native/heritage speakers who recorded and annotated their own speaking/singing relying on their own Indigenous/local knowledge of their language and culture allowed us to achieve annotations faithful to their perception of vocal/instrumental sound production that we could not have achieved using automated algorithms (particularly given that there were no apparent consistent criteria about what exactly constitutes acoustic units among our participants). This resulted in our identifying unexpectedly large differences for features such as temporal rate when analyzed using their manual segmentations that we would have substantially underestimated if we relied on automated segmentation (compare combined effect size of translated Cohen's $D > 1.5$ in Fig. 4 versus $D < 0.4$ in fig. S15). This highlights that equitable collaboration is not only an issue of social justice but also an issue of scientific quality (*69*, *70*).

On the other hand, this paradigm also created challenges and limitations. For example, 6 of our original 81 collaborators were unable to complete their recordings/annotations, and these were disproportionately from Indigenous and underrepresented languages from our originally planned sample. These underrepresented community members tend to be disproportionately burdened with requests for representation, and some also faced additional barriers including difficulty communicating via translation, loss of internet access, and urgent crises in their communities (*71*). Of our coauthors representing Indigenous and underrepresented languages who did complete their recordings and annotations, several were not native speakers, and so their acoustic features may not necessarily reflect the way they would have been spoken by native speakers. Several of our coauthors have been involved in reviving their languages and musical cultures despite past and/or continuing threats of extinction (e.g., Ngarigu, Aynu, and Hebrew) (*72*, *73*). By including their contributions as singers, speakers, and coauthors, we also hope to contribute to their linguistic and musical revival efforts.

Our requirement that all participant data come from coauthors, and vice versa, led to more severe sampling biases than traditional studies, as reflected in our discussion of our data showing higher, more stable-pitched singing than found in Hilton *et al.*'s (*26*) data. Many of these limitations have been addressed through our robustness analyses and converging results from our own and Albouy *et al.*'s (*74*) reanalyses of Hilton *et al.*'s (*26*) independent speech/song dataset described below. However, while our exploratory analyses revealed strong sex differences in pitch height that may reflect sexual selection, most demographic factors that may affect individual differences or cultural differences in music-speech relationships (e.g., musical training, age, and bilingualism) will require more comprehensive study with larger samples in the future. Because a key limitation of our participants-as-coauthors paradigm is sample size (as manual annotations are time-consuming and coauthor

recruitment is more time-intensive than participant recruitment), this model may not be feasible for future larger-scale analyses. Instead, other paradigms such as targeted recruitment of individuals speaking selected languages or mixed approaches combining manual and automated analyses may be needed.

### Implications from the exploratory analyses

Comparisons with lyrics recitation and instrumental recordings revealed that the relationship between music and language can noticeably change depending on the type of acoustic signal. In general, many features followed the predicted "musi-linguistic continuum" with instrumental music and spoken conversation most extreme (e.g., most/least stable pitches respectively), with song and lyric recitation occupying intermediate positions (Fig. 5). However, for temporal rate, songs were more extreme (slower) than instrumental music, while for phrase length, lyric recitation was more extreme (shorter) than spoken conversation. Increasing variations of acoustic signals and designing the continuum with multiple dimensions (e.g., by adding further categories such as infant-directed song/speech or speech intended for stage acting and by mapping music and language according to pitch, rhythm, and propositional/emotional functionality) may elucidate a more nuanced spectrum of musi-linguistic continuum (*26*, *41*, *42*).

Our nPVI analysis did not show correlations between music and speech, as some past studies found (*56*). Perhaps, a more nuanced comparison could be realized by analyzing the relationship between syllabic stress and metrically strong beats (*75*). However, extending the concept of stress and meter to music and languages cross-culturally and cross-linguistically is beyond the scope of the current study. Another possible approach could be the use of vocalic intervals, as originally analyzed by Patel and Daniele (*56*), rather than the IOIs we used. A vocalic interval consists of a vowel or sequence of vowels, regardless of whether they belong to the same syllable (*56*, *76*). This approach could be easier to implement since it eliminates the need for a detailed classification of vowels in the language; instead, we could group intervals based on vowel-like sounds.

### Limitations on generality

A limitation of our study is that because our paradigm was focused on isolating melodic and lyrical components of song, the instrumental melodies we analyzed are not representative of all instrumental music but only instrumental performance of melodies intended to be sung. It is thus possible that instrumental music intended for other contexts may display different trends (e.g., instrumental music to accompany dancing might be faster). Different instruments are also subject to different production constraints, some of which may be shared with singing and speech (e.g., aerophones such as flutes are also limited by breathing capacity), and some of which are not (e.g., chordophones such as violins are limited by finger motor control). For example, although most of our instrumental recordings followed the same rhythmic pattern of the sung melody, Dessiatnitchenko's instrumental performance on the Azerbaijani tar was several times faster than her sung version because the tar requires the performer to repeatedly strum the same note many times to produce the equivalent of a single long sustained note when singing (listen to her instrumental recording at https://osf.io/uj3dn).

Another limitation of our instrumental results is that while none of our collaborators reported any difficulty or unnaturalness in recording a song and then recording a recited version of the same lyrics, many found it unnatural to perform an instrumental version of the sung melody. For example, while the Aynu of Japan do use pitched instruments such as the tonkori, they are traditionally never used to mimic vocal melodies. To compare sung and instrumental features, all of our collaborators agreed to at least record themselves tapping the rhythm of their singing, but these recordings without comparable pitch information ($n = 28$ recordings) had to be excluded from our exploratory analysis of pitch features, and even their rhythmic features may not necessarily be representative of the kinds of rhythms that might be found in purely instrumental music. Likewise, the conversational speech recorded here is not necessarily representative of nonspoken forms of language (e.g., sign language and written language).

### Comparison with alternative dataset

Interestingly, while the qualitative results using Hilton *et al.*'s (*26*) dataset were identical, the magnitude of their song-speech differences was noticeably smaller. For example, while song was substantially higher-pitched than speech in both datasets, the differences were approximately twice as large in our dataset as in Hilton *et al.*'s (*26*) [~600 cents (half an octave) on average versus ~300 cents (quarter octave), respectively]. These differences were consistent even when analyzed using matching subsamples speaking the same languages and using the same fully automated analysis methods (Fig. 6), suggesting that they are not due to differences in the sample of languages or analysis methods we chose.

Instead, we speculate that these differences may be related to differences in recording context and participant recruitment. While our recordings were made by each coauthor recording themselves in a quiet, isolated environment, Hilton *et al.*'s (*26*) recordings were field recordings designed to capture differences between infant-directed and adult-directed vocalizations and thus contain various background sounds other than the vocalizer's speaking/singing (especially high-pitched vocalizations by their accompanying infants; compare fig. S11). This background noise may reduce the observed differences between speech and song.

Another potential factor is musical experience. Our coauthors were mostly recruited from academic societies studying music, and many have also substantial experience as performing musicians. Although the degree of musical experiences of Hilton *et al.*'s (*26*) participants is not clear, the musical training of our participants is likely more extensive than a group of people randomly chosen from general populations. This relatively greater musical training may have influenced the production of higher and more stable pitches in singing. We confirmed that there is no obvious difference in pitch stability of speech between ours and Hilton *et al.*'s (*26*) dataset, but our singing recordings have higher stability than theirs (fig. S18). Similarly, even if pitch estimation errors due to background noise erroneously inflated estimated $f_0$ of Hilton *et al.*'s (*26*) recordings due to noise, our singing showcased the use of more heightened pitch (fig. S17).

We also observed that our spoken recordings have slightly lower pitch height than Hilton *et al.*'s (*26*) spoken recordings. Possible factors that may underlie this difference include age (*77*), body size (*78*), and possibly avoiding using low frequencies not to intimidate accompanied infants (*54*). Our instructions to "describe the song you chose (why you chose it, what you like about it, what the song is about, etc.)" are also different from Hilton *et al.*'s (*26*) instructions to describe "a topic of their choice (for example… their daily routine)," and these

task differences can also affect speaking pitch (*79*). On the other hand, this result is unlikely to be due to the exposure of Western styles to participants, since the subset of Hilton *et al.*'s (*26*) data including only English, Mandarin, Polish, Spanish, and Kannada speakers shows almost the same result as one with their full data including participants from societies less influenced by Western cultures.

After our stage 1 Registered Report protocol received In Principle Acceptance, two independent studies also compared global datasets of singing and speaking, coming to similar conclusions as us. First, Albouy *et al.* (*74*) also reanalyzed Hilton *et al.*'s (*26*) recordings using different but related methods that also emphasize pitch stability and temporal rate ("spectrotemporal modulations"). Albouy *et al.* (*74*) transformed audio recordings to extract two-dimensional density features [spectrotemporal modulations where one axis is temporal modulations (in hertz) and the other is spectral modulations (in cycles per kilohertz)] to characterize song and speech acoustically. Their finding is similar to our results that speech has higher density in the temporal modulation range of 5 to 10 Hz, which matches the syllable rate and amplitude modulation rate of speech investigated cross-culturally (*20*, *80*, *81*), on the low spectral modulation range [rate of change in amplitude due to vocal sound production including the initiation of utterances and the transition from consonants to vowels, which is an automated proxy of our measurement of temporal rate via manually annotated acoustic unit (e.g., syllable/mora/note) durations], and song has higher density in the spectral modulation range of 2 to 5 cycles/kHz on the low temporal modulation range (prominent energy in upper harmonics without fast amplitude change, potentially related to pitch stability). Their behavioral experiment further confirmed that listeners rely on spectral and temporal modulation information to judge whether the uttered vocalization is song or speech, which suggests that spectrotemporal modulation is an acoustic cue differentiating song and speech.

Next, Anikin *et al.* (*82*) curated a different global recording dataset, including not only song and speech but also various nonverbal vocalizations (e.g., laughs, cries, and screams). Their analyses using spectrotemporal modulations also confirmed lower pitch in speech and steadier notes in singing. The convergent findings of our study and their studies identifying the same features imply that pitch height, temporal rate, and pitch stability are robust features distinguishing song and speech across cultures.

### Evolutionary and functional mechanisms

"Discrete pitches or regular rhythmic patterns" are often considered defining features of music that distinguish it from speech [(*83*) and (*25*) block quote in the introduction], and our analyses confirmed this using a diverse cross-cultural sample. At the same time, we were surprised to find that the two features that differed most between song and speech were not pitch stability and rhythmic regularity but rather pitch height and temporal rate (Fig. 4). Pitch stability was the feature differing most between instrumental music and spoken description, but sung pitches were substantially less stable than instrumental ones. Given that the voice is the oldest and most universal instrument, we suggest that future theories of the evolution of musicality should focus more on explaining the differences we have identified in temporal rate and pitch height. In this vein, experimental approaches such as transmission chains may be effective in capturing causal mechanisms underlying the manipulation of these parameters depending on communicative goals (*7*, *84*).

On the other hand, while pitch height showed larger differences between speech and song than pitch stability when comparing within the same individual, our exploratory analysis evaluating feature importance in song-speech classification showed that pitch stability was more useful than pitch height comparing song and speech between individuals. This is consistent with our intuition that song pitch can be artificially lowered in pitch and speech artificially raised in pitch without changing our categorical perception of them as song or speech. Future controlled perceptual experiments independently manipulating each feature may provide more insight into how these acoustic features are processed in our brains.

While our results do not directly provide evidence for the evolutionary mechanisms underlying differences between song and speech, we speculate that temporal rate may be a key feature underlying many observed differences. The temporal rate is the only feature showing almost no difference between singing and the instrumental melody (compare fig. S13). While slower singing reduces the amount of linguistic information that can be conveyed in the lyrics in a fixed amount of time, it gives singers more time to stabilize the pitch (which often takes some time to reach a stable plateau when singing), and the slower and more stable pitches may facilitate synchronization, harmonization, and ultimately bonding between multiple individuals (*11*). However, to ensure comparability between song and speech, we only asked participants to record themselves singing solo, even when songs are usually sung in groups in their culture, so future direct comparison of potential acoustic differences between solo and group vocalizations (*85*) may be needed to investigate potential relationships between our acoustic features and group synchronization/harmonization.

Furthermore, slow vocalization may also interact with high pitch vocalization since it needs deeper breaths to support sustained pitches, which may lead to an increase in subglottal pressure and accompanying higher pitch (*86*). The use of higher pitches in singing may also contribute to more effective communication of pitch information. Sensitivity to loudness for pure tones almost monotonically increases up to 1 kHz (*87*), but, generally, the frequency range of $f_0$ values of human voice is below 1 kHz, so it is reasonable to heighten pitches to exploit higher loudness sensitivity, which may be helpful for creating bonding through acoustic communication extensively using pitch control. Furthermore, in speech, we recognize phonemes by the shape of formants, which characterizes how upper harmonic content is emphasized or attenuated. In speech, the frequency content conveying information is not fundamental frequency but harmonics, whereas in music, it is the lower fundamental frequencies that contain the crucial melodic content (*9*). We speculate that this difference in emphasis on formants versus fundamental frequency may underlie the difference in pitch height between speech and music we have identified.

The exploratory analysis of additional features can also be interpreted from the same viewpoint that extra potential differentiating features also function to enhance the saliency of pitch information: Use of longer acoustic phrases, greater sound pressure, and less noisy sounds may ease the intelligibility of pitch information. This increased loudness and salience might also support evolutionary propositions that music evolved as a mnemonic device (*88*) or as a night-time, long-distance communication device (*89*). The lyrics of the chosen songs frequently mention "night," "moon," "sleep," and "love," which may further support the nocturnal hypothesis (*89*). On the other hand, similar timbral brightness, pitch interval size,

and pitch range between song and speech may be due to motor and mechanistic constraints, similar to the difficulty of rapid transitioning to distant pitches caused by the limiting control capacity of tension in the vocal folds. Since utilization of pitch can also be found in language (e.g., tonal languages; increasing the pitch of the final word in an interrogative sentence in today's English and Japanese), inclusively probing what we can communicate with pitch in human acoustic communication may give insights into the fundamental nature of songs.

Overall, our Registered Report comparing music and speech from our coauthors speaking diverse languages shows strong evidence for cross-cultural regularities in music and speech amidst substantial global diversity. The features that we identified as differentiating music and speech along a musi-linguistic continuum—particularly pitch height, temporal rate, and pitch stability—may represent promising candidates for future analyses of the (co)evolution of biological capacities for music and language (*9*, *11*, *83*). Meanwhile, the features we identified as shared between speech and song—particularly timbral brightness and pitch interval size—represent promising candidates for understanding domain-general constraints on vocalization that may shape the cultural evolution of music and language (*7*, *28*, *90*, *91*). Together, these cross-cultural similarities and differences may help shed light on the cultural and biological evolution of two systems that make us human: music and language.

## MATERIALS AND METHODS

### Analysis plan

We test two types of hypotheses, corresponding to the hypothesis of difference and the hypothesis of similarity, respectively. Formally, one type of null hypothesis is whether the effect size of the difference between song and speech for a given feature is null. This hypothesis is applied to the prediction of the statistical difference. Another type of null hypothesis is whether the effect size of the feature exceeds the SESOI (*92*). This hypothesis is applied to the prediction of statistical similarity. In this study, we particularly rely on the SESOI of 0.4 suggested by the review of psychological research (*93*). There are various ways to quantify the statistical difference or similarity (e.g., Kullbak-Leibler divergence, Jensen-Shannon divergence, Earth mover's distance, energy distance, $L_n$ norm, and Kolmogorov-Smirnov statistic). Here, we focus on effect sizes to facilitate interpretation of the magnitudes of differences.

Since our main interest lies in the identification of which features demonstrate differences or similarities between song and speech, we perform the within-participants comparison of the six features between the pairs of singing and speech, using the spoken description rather than the lyric recitation as the proxy for speech (compare red boxes in Fig. 1; the comparisons with lyrics recitation and with instrumental versions are saved for exploratory analyses). In addition, terms in the computed difference scores are arranged so that for our predicted differences (H1 to H3), a positive value indicates a difference in the predicted direction (compare Fig. 8).

Evaluation of difference in the magnitude of each feature is performed with nonparametric relative effects (*94*), which is also known as stochastic superiority (*95*) or probability-based measure of effect size (*96*). This measure is a nonparametric two-sample statistics and allows us to investigate the statistical properties of a wide variety of data in a unified way.

We apply the meta-analysis framework to synthesize the effect size across recordings to make statistical inference for each hypothesis (Fig. 8). In this case, the study sample size corresponds to the number of data points of the feature in a recording, and the number of studies corresponds to the number of language varieties. We use Gaussian random-effects models (*97*, *98*) (compare the "Statistical models" section in the Supplementary Materials), and we frame our hypotheses as the inference of the mean parameter of Gaussian random-effects models, which indicates the population effect size.

Our null hypotheses for the features predicted showing difference is that the true effect size is zero (i.e., relative effects of 0.5). On the other hand, the null hypotheses for the feature predicted showing similarity is that the true effect size is lower or larger than smallest effect sizes of interest in psychology studies (i.e., relative effects of 0.39 and 0.61 corresponding to ±0.4 of Cohen's *D*) (*93*). We test six features and, thus, test six null hypotheses.

Since we test multiple hypotheses, we use the false discovery rate method with the Benjamini-Hochberg step-up procedure (*99*) to decide on the rejection of the null hypotheses. We define the α level as 0.05.

For the hypothesis testing of null effect size (H1 to H3), we test whether the end points of the confidence interval of the mean parameter of the Gaussian random-effects model are larger than 0.5. We use the exact confidence interval proposed by Liu *et al.* (*98*) and Wang and Tian (*100*) to construct the confidence interval. For the hypothesis testing of equivalence (H4 to H6), we first estimate the mean parameter (i.e., overall treatment effect) with the exact confidence interval (*98*, *100*) and the between-study variance with the DerSimonian-Laird estimator (*101*). Since Gaussian random-effects models can be considered Gaussian mixture models having the same mean parameter, the overall variance parameter can be obtained by averaging the sum of the estimated between-study variance and the within-study variance. Then, we plug the mean parameter and overall variance into Romano's (*102*) shrinking alternative parameter space method to test whether the population mean is within the SESOI as specified above.

Our choice of an SESOI of Cohen's $D = 0.4$ based on Brysbaert's (*93*) recommendation after reviewing psychological studies is admittedly somewhat arbitrary. Future studies might be able to choose a different SESOI on a more principled basis based on the data and analyses we provide here, and the value of our database for such hypothesis generation and exploration is an important benefit beyond the specific confirmatory analyses proposed. However, we currently are faced with a chicken-and-egg problem in that it is difficult to justify an a priori SESOI for analysis until we have undertaken the analysis. The same argument may hold for Bayesian approaches (e.g., highest-density regions, region of practical equivalence, and model selection based on Bayes factors) independent of the choice of prior distributions. We thus chose to rely on Brysbaert's recommended SESOI of Cohen's $D = 0.4$ (and its equivalent relative effect of $p_{re} = 0.61$) in the absence of better alternatives.

Visual and aural inspections of the distribution of pilot data (figs. S2 and S9; audio recordings can be heard at https://osf.io/mzxc8/) also suggest that it is a reasonable (albeit arbitrary) threshold given the variance observed across a range of different features and languages. To enable the reader/listener to assess what an SESOI might sound like, we have created versions of the pilot data artificially raising/lowering the temporal rate and pitch height of sung/

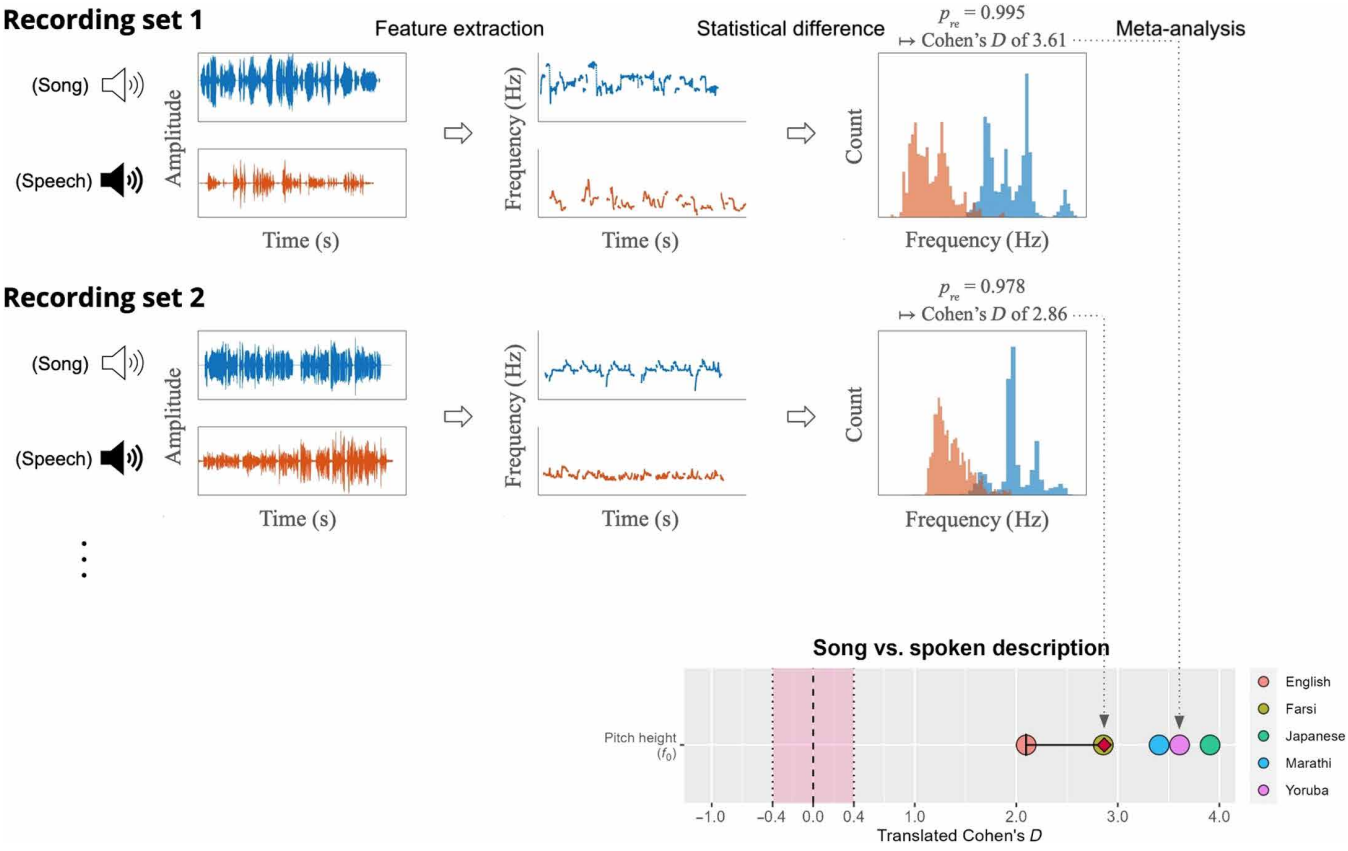



**Fig. 8. Schematic overview of the analysis pipeline from raw audio recordings to the paired comparisons.** This illustration is based on the pilot analysis of stage 1 (fig. S2), which served as a foundation for the subsequent main confirmatory analysis (Fig. 4). Recording sets 1 and 2 represent pilot data of singing and speaking in Yoruba and Farsi by coauthors F.N. and S.H., respectively. From each pair of song/spoken audio recordings by a given person, we quantify the difference using the effect size for each feature. *Pre* is the relative effect (converted to Cohen's *D* for ease of interpretability). In both cases, the distributions of sung and spoken pitch overlap slightly, but song is substantially higher on average (Cohen's $D > 2$). To synthesize the effect sizes collected from each recording pair to test our hypotheses, we apply meta-analyses by treating each recording pair as a study. This approach allows us to make an inference about the population effect size of features in song and speech samples. This example focuses on just one feature (pitch height) applied to just two recording sets, but the same framework is applied to the other five features and all recording sets in the actual analysis. Different types of hypothesis testing are applied depending on the feature (i.e., hypothesis of difference and hypothesis of similarity).

spoken examples so one can hear what our proposed SESOI would sound like for a range of languages and features [compare the "Manipulation of features to demonstrate our designated SESOI (Cohen's $D = 0.4$)" section in the Supplementary Materials and table S6; audio files also at https://osf.io/8mcev].

## Ethics

This research has been approved by the Keio University Shonan Fujisawa Campus's Research Ethics Committee (approval no. 449). The exploratory Maasai song/speech excerpts from noncoauthor Ole Manyas are included as part of a separate ethical approval by the Kenyan National Commision for Science, Technology and Innovation to Parselelo (NACOSTI/P/23/24284).

## Inclusivity statement

We endeavored to follow the best practices in cross-cultural collaborative research (*103*, *104*), such as involving collaborators from diverse backgrounds from the initial planning phases of a study and offering compensation via both financial (honoraria) and intellectual (coauthorship) mechanisms (see the "Collaboration agreement form" in the Supplementary Materials). Each recording set analyzed comes from a named coauthor who speaks that language as their first or heritage language.

## Supplementary Materials

**This PDF file includes:**

Supplementary Text
Figs. S1 to S24
Tables S1 to S6
References

## REFERENCES AND NOTES

1. C. Chambers, "Strong evidence for cross-cultural regularities in music and speech," *Peer Community Registered Report* **1**, (100469) (2023); https://rr.peercommunityin.org/articles/rec?id=469.
2. C. Chambers, "Exploring cross-cultural variation in speech and song," *Peer Community Registered Report* (2023); https://rr.peercommunityin.org/articles/rec?id=316.
3. N. Evans, S. C. Levinson, The myth of language universals: Language diversity and its importance for cognitive science. *Behav. Brain Sci.* **32**, 429–448 (2009).
4. P. E. Savage, Universals, in *The SAGE International Encyclopedia of Music and Culture* (SAGE Publications Inc., 2019), pp. 2283–2285; https://sk.sagepub.com/reference/the-sage-international-encyclopedia-of-music-and-culture/i21528.xml.

5. S. A. Mehr, M. Singh, D. Knox, D. M. Ketter, D. Pickens-Jones, S. Atwood, C. Lucas, N. Jacoby, A. A. Egner, E. J. Hopkins, R. M. Howard, J. K. Hartshorne, M. V. Jennings, J. Simson, C. M. Bainbridge, S. Pinker, T. J. O'Donnell, M. M. Krasnow, L. Glowacki, Universality and diversity in human song. *Science* **366**, eaax0868 (2019).
6. N. Jacoby, E. H. Margulis, M. Clayton, E. Hannon, H. Honing, J. Iversen, T. R. Klein, S. A. Mehr, L. Pearson, I. Peretz, M. Perlman, R. Polak, A. Ravignani, P. E. Savage, G. Steingo, C. J. Stevens, L. Trainor, S. Trehub, M. Veal, M. Wald-Fuhrmann, Cross-cultural work in music cognition: Challenges, insights, and recommendations. *Music Percept.* **37**, 185–195 (2020).
7. Y. Ozaki, M. de Heer Kloots, A. Ravignani, P. E. Savage, Cultural evolution of music and language. PsyArXiv 10.31234/osf.io/s7apx [Preprint] (2024). https://doi.org/10.31234/osf.io/s7apx.
8. C. Darwin, *The Descent of Man* (Watts & Co., 1871).
9. A. D. Patel, *Music, Language, and the Brain* (Oxford Univ. Press, 2008); https://oxford.universitypressscholarship.com/10.1093/acprof:oso/9780195123753.001.0001/acprof-9780195123753.
10. J. V. Valentova, P. Tureček, M. A. C. Varella, P. Šebesta, F. D. C. Mendes, K. J. Pereira, L. Kubicová, P. Stolařová, J. Havlíček, Vocal parameters of speech and singing covary and are related to vocal attractiveness, body measures, and sociosexuality: A cross-cultural study. *Front. Psychol.* **10**, 2029 (2019).
11. P. E. Savage, P. Loui, B. Tarr, A. Schachner, L. Glowacki, S. Mithen, W. T. Fitch, Music as a coevolved system for social bonding. *Behav. Brain Sci.* **44**, 1–22 (2021).
12. S. A. Mehr, M. M. Krasnow, G. A. Bryant, E. H. Hagen, Origins of music in credible signaling. *Behav. Brain Sci.* **44**, e60 (2021).
13. F. Haiduk, W. T. Fitch, Understanding design features of music and language: The choric/dialogic distinction. *Front. Psychol.* **13**, 786899 (2022).
14. I. Peretz, Music, language and modularity framed in action. *Psychol. Belg.* **49**, 157–175 (2009).
15. A. D. Patel, Language, music, and the brain: A resource-sharing framework, in *Language and Music as Cognitive Systems*, P. Rebuschat, M. Rohmeier, J. A. Hawkins, I. Cross, Eds. (Oxford Univ. Press, 2011), pp. 204–223; https://doi.org/10.1093/acprof:oso/9780199553426.003.0022.
16. C. Rogalsky, F. Rong, K. Saberi, G. Hickok, Functional anatomy of language and music perception: Temporal and structural factors investigated using functional magnetic resonance imaging. *J. Neurosci.* **31**, 3843–3852 (2011).
17. T. H. Morrill, J. D. McAuley, L. C. Dilley, D. Z. Hambrick, Individual differences in the perception of melodic contours and pitch-accent timing in speech: Support for domain-generality of pitch processing. *J. Exp. Psychol. Gen.* **144**, 730–736 (2015).
18. K. B. Doelling, M. F. Assaneo, D. Bevilacqua, B. Pesaran, D. Poeppel, An oscillator model better predicts cortical entrainment to music. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 10113–10121 (2019).
19. P. Albouy, L. Benjamin, B. Morillon, R. J. Zatorre, Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science* **367**, 1043–1047 (2020).
20. A. D. Patel, J. R. Iversen, J. C. Rosenberg, Comparing the rhythm and melody of speech and music: The case of British English and French. *J. Acoust. Soc. Am.* **119**, 3034–3047 (2006).
21. N. Ding, A. D. Patel, L. Chen, H. Butler, C. Luo, D. Poeppel, Temporal modulations in speech and music. *Neurosci. Biobehav. Rev.* **81**, 181–187 (2017).
22. D. E. Brown, *Human Universals* (McGraw-Hill, 1991).
23. B. Bickel, Absolute and statistical universals, in *The Cambridge Encyclopedia of the Language Sciences*, P. C. Hogan, Ed. (Cambridge Univ. Press, 2011), pp. 77–79.
24. S. Brown, J. Jordania, Universals in the world's musics. *Psychol. Music* **41**, 229–248 (2013).
25. P. E. Savage, S. Brown, E. Sakai, T. E. Currie, Statistical universals reveal the structures and functions of human music. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 8987–8992 (2015).
26. C. B. Hilton, C. J. Moser, M. Bertolo, H. Lee-Rubin, D. Amir, C. M. Bainbridge, J. Simson, D. Knox, L. Glowacki, E. Alemu, A. Galbarczyk, G. Jasienska, C. T. Ross, M. B. Neff, A. Martin, L. K. Cirelli, S. E. Trehub, J. Song, M. Kim, A. Schachner, T. A. Vardy, Q. D. Atkinson, A. Salenius, J. Andelin, J. Antfolk, P. Madhivanan, A. Siddaiah, C. D. Placek, G. D. Salali, S. Keestra, M. Singh, S. A. Collins, J. Q. Patton, C. Scaff, J. Stieglitz, S. C. Cutipa, C. Moya, R. R. Sagar, M. Anyawire, A. Mabulla, B. M. Wood, M. M. Krasnow, S. A. Mehr, Acoustic regularities in infant-directed speech and song across cultures. *Nat. Hum. Behav.* **6**, 1545–1556 (2022).
27. Y. Ozaki, S. Sato, J. Mcbride, P. Q. Pfordresher, A. T. Tierney, J. Six, S. Fujii, P. E. Savage, Automatic acoustic analyses quantify pitch discreteness within and between human music, speech, and birdsong, in *Proceedings of the 10th International Workshop on Folk Music Analysis*, A. Holzapfel, I. Ali-MacLachlan Eds. (KTH Royal Institute of Technology, Sweden, 2022), pp. 3–9. http://doi.org/10.5281/zenodo.7100288.
28. A. T. Tierney, F. A. Russo, A. D. Patel, The motor origins of human and avian song structure. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 15510–15515 (2011).
29. G. List, On the non-universality of musical perspectives. *Ethnomusicology* **15**, 399–402 (1971).
30. A. Lomax, V. Grauer, The Cantometric coding book, in *Folk Song Style and Culture*, A. Lomax, Ed. (American Association for the Advancement of Science, 1968), pp. 34–74.
31. J. Blacking, *How Musical Is Man?* (University of Washington Press, 1973).
32. J. H. L. Hansen, M. Bokshi, S. Khorram, Speech variability: A cross-language study on acoustic variations of speaking versus untrained singing. *J. Acoust. Soc. Am.* **148**, 829–844 (2020).
33. J. Merrill, P. Larrouy-Maestri, Vocal features of song and speech: Insights from Schoenberg's Pierrot lunaire. *Front. Psychol.* **8**, 1108 (2017).
34. B. Sharma, X. Gao, K. Vijayan, X. Tian, H. Li, NHSS: A speech and singing parallel database. *Speech Commun.* **133**, 9–22 (2021).
35. C. M. Vanden Bosch der Nederlanden, X. Qi, S. Sequeira, P. Seth, J. A. Grahn, M. F. Joanisse, E. E. Hannon, Developmental changes in the categorization of speech and song. *Dev. Sci.* **26**, e13346 (2023).
36. D. E. Blasi, J. Henrich, E. Adamou, D. Kemmerer, A. Majid, Over-reliance on English hinders cognitive science. *Trends Cogn. Sci.* **26**, 1153–1170 (2022).
37. Y. Ozaki, J. Kuroyanagi, G. Chiba, J. McBride, P. Proutskova, A. Tierney, E. Benetos, F. Liu, P. E. Savage, Similarities and differences in a cross-linguistic sample of song and speech recordings, in *Proceedings of the 2022 Joint Conference on Language Evolution* (Joint Conference on Language Evolution (JCoLE) Max Planck Institute for Psycholinguistics, 2022), pp. 569–572.
38. C. Durojaye, L. Fink, T. Roeske, M. Wald-Fuhrmann, P. Larrouy-Maestri, Perception of Nigerian dùndún talking drum performances as speech-like vs. music-like: The role of familiarity and acoustic cues. *Front. Psychol.* **12**, 652673 (2021).
39. T. C. Roeske, O. Tchernichovski, D. Poeppel, N. Jacoby, Categorical rhythms are shared between songbirds and humans. *Curr. Biol.* **30**, 3544–3555.e6 (2020).
40. P. Mertens, The Prosogram model for pitch stylization and its applications in intonation transcription, in *Prosodic Theory and Practice*, J. Barnes, S. Shattuck-Hufnagel, Eds. (The MIT Press, 2022), pp. 259–286; https://mitpress.mit.edu/9780262543170/prosodic-theory-and-practice/.
41. S. Brown, The musilanguage model of music evolution, in *The Origins of Music*, S. Brown, B. Merker, C. Wallin, Eds. (The MIT Press, 2000), pp. 271–300; https://direct.mit.edu/books/book/2109/chapter/56564/The-Musilanguage-Model-of-Music-Evolution.
42. J. D. Leongómez, J. Havlíček, S. C. Roberts, Musicality in human vocal communication: An evolutionary perspective. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **377**, 20200391 (2022).
43. R. Tsur, C. Gafni, *Sound–Emotion Interaction in Poetry: Rhythm, Phonemes, Voice Quality* (John Benjamins, 2022).
44. P. E. Savage, P. Loui, B. Tarr, A. Schachner, L. Glowacki, S. Mithen, W. T. Fitch, Authors' response: Toward inclusive theories of the evolution of musicality. *Behav. Brain Sci.* **44**, 132–140 (2021).
45. A. D. Patel, Music as a transformative technology of the mind: An update, in *The Origins of Musicality*, H. Honing, Ed. (The MIT Press, 2018); pp. 113–126; https://direct.mit.edu/books/book/4115/chapter/170183/Music-as-a-Transformative-Technology-of-the-Mind.
46. M. Hoeschele, W. T. Fitch, Cultural evolution: Conserved patterns of melodic evolution across musical cultures. *Curr. Biol.* **32**, R265–R267 (2022).
47. H. Hammarström, R. Forkel, M. Haspelmath, S. Bank, glottolog/glottolog: Glottolog database 4.7 (Leipzig: Max Planck Institute for Evolutionary Anthropology, 2022); https://doi.org/10.5281/zenodo.7398962.
48. M. S. Dryer, M. Haspelmath, *The World Atlas of Language Structures Online* (Max Planck Institute for Evolutionary Anthropology, 2013); http://wals.info.
49. S. Rosenzweig, F. Scherbaum, D. Shugliashvili, V. Arifi-Müller, M. Müller, Erkomaishvili dataset: A curated corpus of traditional georgian vocal music for computational musicology. *Trans. Int. Soc. Music Inf. Retr.* **3**, 31–41 (2020).
50. M. Müller, S. Rosenzweig, J. Driedger, F. Scherbaum, *Interactive Fundamental Frequency Estimation with Applications to Ethnomusicological Research* (Audio Engineering Society, 2017); www.aes.org/e-lib/browse.cfm?elib=18777.
51. N. Bannan, R. I. M. Dunbar, J. S. Bamford, The evolution of gender dimorphism in the human voice: The role of octave equivalence. PsyArXiv 10.31234/osf.io/f4j6b [Preprint] (2024). https://doi.org/10.31234/osf.io/f4j6b.
52. S. Chen, C. Han, S. Wang, X. Liu, B. Wang, R. Wei, X. Lei, Hearing the physical condition: The relationship between sexually dimorphic vocal traits and underlying physiology. *Front. Psychol.* **13**, 983688 (2022).
53. D. R. Feinberg, B. C. Jones, M. M. Armstrong, Sensory exploitation, sexual dimorphism, and human voice pitch. *Trends Ecol. Evol.* **33**, 901–903 (2018).
54. D. A. Puts, S. J. C. Gaulin, K. Verdolini, Dominance and the evolution of sexual dimorphism in human voice pitch. *Evol. Hum. Behav.* **27**, 283–296 (2006).
55. D. A. Puts, A. K. Hill, D. H. Bailey, R. S. Walker, D. Rendall, J. R. Wheatley, L. L. M. Welling, K. Dawood, R. Cárdenas, R. P. Burriss, N. G. Jablonski, M. D. Shriver, D. Weiss, A. R. Lameira, C. L. Apicella, M. J. Owren, C. Barelli, M. E. Glenn, G. Ramos-Fernandez, Sexual selection on male vocal fundamental frequency in humans and other anthropoids. *Proc. R. Soc. B Biol. Sci.* **283**, 20152830 (2016).
56. A. D. Patel, J. R. Daniele, An empirical comparison of rhythm in language and music. *Cognition* **87**, B35–B45 (2003).

57. L. E. Ling, E. Grabe, F. Nolan, Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English. *Lang. Speech* **43**, 377–401 (2000).
58. E. Grabe, E. L. Low, Durational variability in speech and the Rhythm Class Hypothesis, in *Laboratory Phonology 7*, C. Gussenhoven, N. Warner, Eds. (De Gruyter Mouton, 2002), pp. 515–546; www.degruyter.com/document/doi/10.1515/9783110197105.2.515/html?lang=en.
59. N. H. de Jong, T. Wempe, Praat script to detect syllable nuclei and measure speech rate automatically. *Behav. Res. Methods* **41**, 385–390 (2009).
60. M. Mauch, S. Dixon, PYIN: A fundamental frequency estimator using probabilistic threshold distributions, in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, 2014), pp. 659–663.
61. F. Weber, G. Knapp, K. Ickstadt, G. Kundt, Ä. Glass, Zero-cell corrections in random-effects meta-analyses. *Res. Synth. Methods* **11**, 913–919 (2020).
62. C. R. Adams, Melodic contour typology. *Ethnomusicology* **20**, 179–215 (1976).
63. D. Huron, The melodic arch in western folksongs. *Comput. Musicol.* **10**, 3–23 (1996).
64. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).
65. "Permutation importance," ELI5; https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html.
66. J. P. B. Pereira, E. S. G. Stroes, A. H. Zwinderman, E. Levin, Covered information disentanglement: Model transparency via unbiased permutation importance. *Proc. AAAI Conf. Artif. Intell.* **36**, 7984–7992 (2022).
67. A. Chang, X. Teng, F. Assaneo, D. Poeppel, Amplitude modulation perceptually distinguishes music and speech. PsyArXiv 10.31234/osf.io/juzrh [Preprint] (2022). https://doi.org/10.31234/osf.io/juzrh.
68. P. E. Savage, A. T. Tierney, A. D. Patel, Global music recordings support the motor constraint hypothesis for human and avian song contour. *Music Percept.* **34**, 327–334 (2017).
69. Nature addresses helicopter research and ethics dumping. *Nature* **606**, 7 (2022).
70. M. Urassa, D. W. Lawson, J. Wamoyi, E. Gurmu, M. A. Gibson, P. Madhivanan, C. Placek, Cross-cultural research must prioritize equitable collaboration. *Nat. Hum. Behav.* **5**, 668–671 (2021).
71. J. Nicas, The Amazon's Largest Isolated Tribe Is Dying, *The New York Times*, 25 March 2023; www.nytimes.com/2023/03/25/world/americas/brazil-amazon-indigenous-tribe.html.
72. J. Troy, L. Barwick, Claiming the 'Song of the Women of the Menero Tribe'. *Musicol. Aust.* **42**, 85–107 (2020).
73. P. E. Savage, H. Matsumae, H. Oota, M. Stoneking, T. E. Currie, A. Tajima, M. Gillan, S. Brown, How 'circumpolar' is Ainu music? Musical and genetic perspectives on the history of the Japanese archipelago. *Ethnomusicol. Forum* **24**, 443–467 (2015).
74. P. Albouy, S. A. Mehr, R. S. Hoyer, J. Ginzburg, R. J. Zatorre, Spectro-temporal acoustical markers differentiate speech from song across cultures. bioRxiv 2023.01.29.526133 [Preprint] (2023). https://doi.org/10.1101/2023.01.29.526133.
75. D. Temperley, Music and language. *Annu. Rev. Linguist.* **8**, 153–170 (2022).
76. F. Ramus, Acoustic correlates of linguistic rhythm: Perspectives. *Proc. Speech Prosody* **2002**, 115–120 (2002).
77. M. Berg, M. Fuchs, K. Wirkner, M. Loeffler, C. Engel, T. Berger, The speaking voice in the general population: Normative data and associations to sociodemographic and lifestyle factors. *J. Voice* **31**, 257.e13–257.e24 (2017).
78. K. Pisanski, P. J. Fraccaro, C. C. Tigue, J. J. M. O'Connor, S. Röder, P. W. Andrews, B. Fink, L. M. DeBruine, B. C. Jones, D. R. Feinberg, Vocal indicators of body size in men and women: A meta-analysis. *Anim. Behav.* **95**, 89–99 (2014).
79. B. Barsties, Einfluss verschiedener Methoden zur Bestimmung der mittleren Sprechstimmlage. *HNO* **61**, 609–616 (2013).
80. F. Pellegrino, C. Coupé, E. Marsico, Across-Language perspective on speech information rate. *Language* **87**, 539–558 (2011).
81. D. Poeppel, M. F. Assaneo, Speech rhythms and their neural foundations. *Nat. Rev. Neurosci.* **21**, 322–334 (2020).
82. A. Anikin, V. Canessa-Pollard, K. Pisanski, M. Massenet, D. Reby, Beyond speech: Exploring diversity in the human voice. *iScience* **26**, 108204 (2023).
83. W. T. Fitch, The biology and evolution of music: A comparative perspective. *Cognition* **100**, 173–215 (2006).
84. W. Ma, A. Fiveash, W. F. Thompson, Spontaneous emergence of language-like and music-like vocalizations from an artificial protolanguage. *Semiotica* **2019**, 1–23 (2019).
85. A. Lomax, *Folk Song Style and Culture* (American Association for the Advancement, 1968).
86. F. Alipour, R. C. Scherer, On pressure-frequency relations in the excised larynx. *J. Acoust. Soc. Am.* **122**, 2296–2305 (2007).
87. Y. Suzuki, H. Takeshima, Equal-loudness-level contours for pure tones. *J. Acoust. Soc. Am.* **116**, 918–933 (2004).
88. D. J. Levitin, Knowledge songs as an evolutionary adaptation to facilitate information transmission through music. *Behav. Brain Sci.* **44**, e105 (2021).
89. M. A. C. Varella, Nocturnal selective pressures on the evolution of human musicality as a missing piece of the adaptationist puzzle. *Front. Psychol.* **14**, 1215481 (2023).
90. S. E. Trehub, Cross-cultural convergence of musical features. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 8809–8810 (2015).
91. M. Singh, S. A. Mehr, Universality, domain-specificity and development of psychological responses to music. *Nat. Rev. Psychol.* **2**, 333–346 (2023).
92. D. Lakens, Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Soc. Psychol. Personal. Sci.* **8**, 355–362 (2017).
93. M. Brysbaert, How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *J. Cogn.* **2**, 16 (2019).
94. E. Brunner, A. C. Bathke, F. Konietschke, *Rank and Pseudo-Rank Procedures for Independent Observations in Factorial Designs: Using R and SAS* (Springer, 2018); https://ci.nii.ac.jp/ncid/BB28708839.
95. A. Vargha, H. D. Delaney, The Kruskal-Wallis test and stochastic homogeneity. *J. Educ. Behav. Stat.* **23**, 170–192 (1998).
96. J. Ruscio, A probability-based measure of effect size: Robustness to base rates and other factors. *Psychol. Methods* **13**, 19–30 (2008).
97. S. E. Brockwell, I. R. Gordon, A comparison of statistical methods for meta-analysis. *Stat. Med.* **20**, 825–840 (2001).
98. S. Liu, L. Tian, S. Lee, M.-g. Xie, Exact inference on meta-analysis with generalized fixed-effects and random-effects models. *Biostat. Epidemiol.* **2**, 1–22 (2018).
99. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
100. Y. Wang, L. Tian, An efficient numerical algorithm for exact inference in meta analysis. *J. Stat. Comput. Simul.* **88**, 646–656 (2018).
101. R. DerSimonian, N. Laird, Meta-analysis in clinical trials. *Control. Clin. Trials* **7**, 177–188 (1986).
102. J. P. Romano, Optimal testing of equivalence hypotheses. *Ann. Stat.* **33**, 1036–1047 (2005).
103. S. B. Tan, M. Ostashewski, Eds., *DIALOGUES: Towards Decolonizing Music and Dance Studies* (International Council for Traditional Music, 2022); https://ictmdialogues.org/.
104. P. E. Savage, N. Jacoby, E. H. Margulis, H. Daikoku, M. Anglada-Tort, S. E.-S. Castelo-Branco, F. E. Nweke, S. Fujii, S. Hegde, H. Chuan-Peng, J. Jabbour, C. Lew-Williams, D. Mangalagiu, R. McNamara, D. Müllensiefen, P. Opondo, A. D. Patel, H. Schippers, Building sustainable global collaborative networks: Recommendations from music studies and the social sciences, in *The Science-Music Borderlands: Reckoning with the Past, Imagining the Future*, E. H. Margulis, L. Loughridge, P. Loui, Eds. (The MIT Press, 2023), pp. 347–365; https://direct.mit.edu/books/oa-edited-volume/5578/chapter/4162120/Building-Sustainable-Global-Collaborative-Networks.
105. N. Novitski, M. Tervaniemi, M. Huotilainen, R. Näätänen, Frequency discrimination at different frequency levels as indexed by electrophysiological and behavioral measures. *Cogn. Brain Res.* **20**, 26–36 (2004).
106. A. Anikin, The link between auditory salience and emotion intensity. *Cogn. Emot.* **34**, 1246–1259 (2020).
107. C. Cox, C. Bergmann, E. Fowler, T. Keren-Portnoy, A. Roepstorff, G. Bryant, R. Fusaroli, A systematic review and Bayesian meta-analysis of the acoustic features of infant-directed speech. *Nat. Hum. Behav.* **7**, 114–133 (2023).
108. T. Verhoef, A. Ravignani, Melodic universals emerge or are sustained through cultural Evolution. *Front. Psychol.* **12**, 668300 (2021).
109. P. Q. Pfordresher, S. Brown, K. M. Meier, M. Belyk, M. Liotti, Imprecise singing is widespread. *J. Acoust. Soc. Am.* **128**, 2182–2190 (2010).
110. U. Natke, T. M. Donath, K. T. Kalveram, Control of voice fundamental frequency in speaking versus singing. *J. Acoust. Soc. Am.* **113**, 1587–1593 (2003).
111. B. Raposo de Medeiros, J. P. Cabral, A. R. Meireles, A. A. Baceti, A comparative study of fundamental frequency stability between speech and singing. *Speech Commun.* **128**, 15–23 (2021).
112. E. L. Stegemöller, E. Skoe, T. Nicol, C. M. Warrier, N. Kraus, Music training and vocal production of speech and song. *Music Percept.* **25**, 419–428 (2008).
113. B. Thompson, Discrimination between singing and speech in real-world audio, in *2014 IEEE Spoken Language Technology Workshop (SLT)* (IEEE, 2014), pp. 407–412.
114. G. A. Bryant, The evolution of human vocal emotion. *Emot. Rev.* **13**, 25–33 (2021).
115. S. E. Trehub, A. M. Unyk, S. B. Kamenetsky, D. S. Hill, L. J. Trainor, J. L. Henderson, M. Saraza, Mothers' and fathers' singing to infants. *Dev. Psychol.* **33**, 500–507 (1997).
116. A. Nikolsky, E. Alekseyev, I. Alekseev, V. Dyakonova, The overlooked tradition of "personal music" and its place in the evolution of music. *Front. Psychol.* **10**, 3051 (2020).
117. A. D. Patel, C. von Rueden, Where they sing solo: Accounting for cross-cultural variation in collective music-making in theories of music evolution. *Behav. Brain Sci.* **44**, e85 (2021).
118. D. Ross, J. Choi, D. Purves, Musical intervals in speech. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 9852–9857 (2007).
119. D. A. Schwartz, C. Q. Howe, D. Purves, The statistical structure of human speech sounds predicts musical universals. *J. Neurosci.* **23**, 7160–7168 (2003).

120. S. Han, J. Sundararajan, D. L. Bowling, J. Lake, D. Purves, Co-variation of tonality in the music and speech of different cultures. *PLOS ONE* **6**, e20160 (2011).
121. J. P. Robledo, E. Hurtado, F. Prado, D. Román, C. Cornejo, Music intervals in speech: Psychological disposition modulates ratio precision among interlocutors' nonlocal f0 production in real-time dyadic conversation. *Psychol. Music* **44**, 1404–1418 (2016).
122. R. E. Stone Jr., T. F. Cleveland, J. Sundberg, Formant frequencies in country singers' speech and singing. *J. Voice* **13**, 161–167 (1999).
123. B. Lindblom, J. Sundberg, The human voice in speech and singing, in *Springer Handbook of Acoustics*, T. D. Rossing, Ed. (Springer, 2007), pp. 669–712; https://doi.org/10.1007/978-0-387-30425-0_16.
124. J. J. Barnes, P. Davis, J. Oates, J. Chapman, The relationship between professional operatic soprano voice and high range spectral energy. *J. Acoust. Soc. Am.* **116**, 530–538 (2004).
125. J. Sundberg, Level and center frequency of the singer's formant. *J. Voice* **15**, 176–186 (2001).
126. D. R. Ladd, Declination: A review and some hypotheses. *Phonol. Yearb.* **1**, 53–74 (1984).
127. J. Slifka, Respiratory system pressures at the start of an utterance, in *Dynamics of Speech Production and Perception*, P. Divenyi, S. Greenbarg, G. Meyer, Eds. (IOS Press, 2006), pp. 45–57; https://ebooks.iospress.nl/volumearticle/392.
128. H. Bârzan, V. V. Moca, A.-M. Ichim, R. C. Muresan, Fractional superlets, in *2020 28th European Signal Processing Conference (EUSIPCO)* (IEEE, 2021), pp. 2220–2224.
129. V. V. Moca, H. Bârzan, A. Nagy-Dăbâcan, R. C. Mureşan, Time-frequency super-resolution with superlets. *Nat. Commun.* **12**, 337 (2021).
130. I. Djurović, L. Stanković, An algorithm for the Wigner distribution based instantaneous frequency estimation in a high noise environment. *Signal Process.* **84**, 631–643 (2004).
131. A. Danielsen, K. Nymoen, E. Anderson, G. S. Câmara, M. T. Langerød, M. R. Thompson, J. London, Where is the beat in that note? Effects of attack, duration, and frequency on the perceived timing of musical and quasi-musical sounds. *J. Exp. Psychol. Hum. Percept. Perform.* **45**, 402–418 (2019).
132. S. K. Scott, The point of P-centres. *Psychol. Res.* **61**, 4–11 (1998).
133. J. Vos, R. Rasch, The perceptual onset of musical tones. *Percept. Psychophys.* **29**, 323–335 (1981).
134. P. Howell, Prediction of P-center location from the distribution of energy in the amplitude envelope: I. *Percept. Psychophys.* **43**, 90–93 (1988).
135. J. Morton, S. Marcus, C. Frankish, Perceptual centers (P-centers). *Psychol. Rev.* **83**, 405–408 (1976).
136. B. Pompino-Marschall, On the psychoacoustic nature of the P-center phenomenon. *J. Phon.* **17**, 175–192 (1989).
137. X. Shao, C. Ma, A general approach to derivative calculation using wavelet transform. *Chemom. Intel. Lab. Syst.* **69**, 157–165 (2003).
138. Z.-H. Tan, A. k. Sarkar, N. Dehak, rVAD: An unsupervised segment-based robust voice activity detection method. *Comput. Speech Lang.* **59**, 1–21 (2020).
139. J. E. Chacón, The modal age of statistics. *Int. Stat. Rev.* **88**, 122–141 (2020).
140. P. Chaudhuri, J. S. Marron, SiZer for exploration of structures in curves. *J. Am. Stat. Assoc.* **94**, 807–823 (1999).
141. F. Chazal, B. Fasy, F. Lecci, B. Michel, A. Rinaldo, L. Wasserman, Robust topological inference: Distance to a measure and kernel distance. *J. Mach. Learn. Res.* **18**, 1–40 (2018).
142. Y.-C. Chen, C. R. Genovese, L. Wasserman, A comprehensive approach to mode clustering. *Electron. J. Stat.* **10**, 210–241 (2016).
143. D. Comaniciu, P. Meer, Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 603–619 (2002).
144. B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, A. Singh, Confidence sets for persistence diagrams. *Ann. Stat.* **42**, 2301–2339 (2014).
145. C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, L. Wasserman, Non-parametric inference for density modes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **78**, 99–126 (2016).
146. M. Sommerfeld, G. Heo, P. Kim, S. T. Rush, J. S. Marron, Bump hunting by topological data analysis. *Stat* **6**, 462–471 (2017).
147. R. Zhang, R. Ghanem, Normal-bundle bootstrap. *SIAM J. Math. Data Sci.* **3**, 573–592 (2021).
148. F. T. Pokorny, C. H. Ek, H. Kjellström, D. Kragic, "Topological constraints and kernel-based density estimation," *Advances in Neural Information Processing Systems 25, Workshop on Algebraic Topology and Machine Learning*, Nevada, USA, 8 December 2012.
149. G. Carlsson, Topology and data. *Bull. Am. Math. Soc.* **46**, 255–308 (2009).
150. B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, 1986).
151. P. Hall, S. J. Sheather, M. C. Jones, J. S. Marron, On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78**, 263–269 (1991).
152. Y.-C. Chen, C. R. Genovese, S. Ho, L. Wasserman, Optimal ridge detection using coverage risk, in *Advances in Neural Information Processing Systems* (Curran Associates Inc., 2015), vol. 28 pp. 1–9; https://papers.nips.cc/paper/2015/hash/0aa1883c6411f7873cb83dacb17b0afc-Abstract.html.
153. O. Lartillot, T. Eerola, P. Toiviainen, J. Fornari, Multi-feature modeling of pulse clarity: Design, validation and optimization, in *Proceedings of the 9th International Conference on Music Information Retrieval* (International Society for Music Information Retrieval, Philadelphia, PA, USA, 2008), pp. 521–526.
154. G. Peeters, "A large set of audio features for sound description (similarity and classification) in the Cuidado Project" (Technical Report, Institut de Recherche et Coordination Acoustique/Musique (IRCAM), 2004).
155. J. D. Johnston, Transform coding of audio signals using perceptual noise criteria. *IEEE J Sel Areas Commun* **6**, 314–323 (1988).
156. R. Villing, "Hearing the moment: Measures and models of the perceptual centre," thesis, National University of Ireland Maynooth (2010).
157. P. A. Barbosa, P. Arantes, A. R. Meireles, J. M. Vieira, Abstractness in speech-metronome synchronisation: P-centres as cyclic attractors, in *Interspeech 2005* (ISCA, 2005), pp. 1441–1444; www.isca-speech.org/archive/interspeech_2005/barbosa05_interspeech.html.
158. I. Chow, M. Belyk, V. Tran, S. Brown, Syllable synchronization and the P-center in Cantonese. *J. Phon.* **49**, 55–66 (2015).
159. A. M. Cooper, D. H. Whalen, C. A. Fowler, P-centers are unaffected by phonetic categorization. *Percept. Psychophys.* **39**, 187–196 (1986).
160. C. Cannam, C. Landone, M. Sandler, Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files, in *Proceedings of the 18th ACM International Conference on Multimedia* (Association for Computing Machinery, 2010), pp. 1467–1468; https://doi.org/10.1145/1873951.1874248.
161. M. Dunn, S. J. Greenhill, S. C. Levinson, R. D. Gray, Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* **473**, 79–82 (2011).
162. S. Brown, P. E. Savage, A. M.-S. Ko, M. Stoneking, Y.-C. Ko, J.-H. Loo, J. A. Trejaut, Correlations in the population structure of music, genes and language. *Proc. R. Soc. B Biol. Sci.* **281**, 20132072 (2014).
163. H. Matsumae, P. Ranacher, P. E. Savage, D. E. Blasi, T. E. Currie, K. Koganebuchi, N. Nishida, T. Sato, H. Tanabe, A. Tajima, S. Brown, M. Stoneking, K. K. Shimizu, H. Oota, B. Bickel, Exploring correlations in genetic and cultural variation across language families in northeast Asia. *Sci. Adv.* **7**, eabd9223 (2021).
164. S. Passmore, A. Wood, C. Barbieri, D. Shilton, H. Daikoku, Q. Atkinson, P. E. Savage, Global musical diversity is largely independent of linguistic and genetic histories. PsyArXiv 10.31234/osf.io/pty34 [Preprint] (2023). https://doi.org/10.31234/osf.io/pty34.
165. F. Sera, B. Armstrong, M. Blangiardo, A. Gasparrini, An extended mixed-effects framework for meta-analysis. *Stat. Med.* **38**, 5429–5444 (2019).
166. H. Bozdogan, Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52**, 345–370 (1987).
167. S. Watanabe, *Mathematical Theory of Bayesian Statistics* (Chapman and Hall/CRC, 2018).
168. G. S. Dell, M. F. Schwartz, N. Martin, E. M. Saffran, D. A. Gagnon, The role of computational models in neuropsychological investigations of language: Reply to Ruml and Caramazza (2000). *Psychol. Rev.* **107**, 635–645 (2000).
169. D. Fraser, Interpolation by the FFT revisited-an experimental investigation. *IEEE Trans. Acoust. Speech Signal Process.* **37**, 665–675 (1989).
170. R. W. Schafer, L. R. Rabiner, A digital signal processing approach to interpolation. *Proc. IEEE* **61**, 692–702 (1973).
171. M. Cychosz, A. Cristia, E. Bergelson, M. Casillas, G. Baudet, A. S. Warlaumont, C. Scaff, L. Yankowitz, A. Seidl, Vocal development in a large-scale crosslinguistic corpus. *Dev. Sci.* **24**, e13090 (2021).
172. F. Anvari, D. Lakens, Using anchor-based methods to determine the smallest effect size of interest. *J. Exp. Soc. Psychol.* **96**, 104159 (2021).
173. S.-H. Jung, Sample size for FDR-control in microarray data analysis. *Bioinformatics* **21**, 3097–3104 (2005).
174. S. Pounds, C. Cheng, Sample size determination for the false discovery rate. *Bioinformatics* **21**, 4263–4271 (2005).
175. M. Horn, C. W. Dunnett, Power and sample size comparisons of stepwise FWE and FDR controlling test procedures in the normal many-one case. *Lect. Notes-Monogr. Ser.* **47**, 48–65 (2004).
176. L. V. Hedges, T. D. Pigott, The power of statistical tests in meta-analysis. *Psychol. Methods* **6**, 203–217 (2001).
177. D. Jackson, R. Turner, Power analysis for random-effects meta-analysis. *Res. Synth. Methods* **8**, 290–302 (2017).


**Acknowledgments:** We thank C. Chambers, B. Slevc, and N. Ding for reviews and recommendations, which are publicly available along with the stage 1 and stage 2 Registered Report protocols (*1*, *2*). We also thank T. Tanaka for serving as the Research Assistant to securely monitor and check audio recordings. We thank J. Bulbulia for assistance in securing funding and Ozaki's PhD committee members A. Wakita and N. Tokui, students from the Keio University CompMusic and NeuroMusic Labs, M. Singh, and anonymous reviewers for feedback on earlier drafts of the manuscript. We thank A. Irurtzun, J. Maripil, A. Lrawbalrate, M. I. Aranariutheri,

T. U. P. Matis, and S. Farwaneh, who initially planned to be coauthors but were unable to complete the recording and annotation processes in time. **Funding:** This work is supported by funding from the New Zealand and Japanese governments, the European Research Council, the Yamaha Corporation, and Keio University as follows: Marsden Fast-Start Grant from the Royal Society Te Apārangi (MFP-UOA2236 to P.E.S., S.C.P., P.O., N.J., E.B., and W.T.F.), Rutherford Discovery Fellowship from the Royal Society Te Apārangi (RDF-UOA2202 to P.E.S.), KAKENHI Grant-in-Aid from the Japan Society for the Promotion of Science (#19KK0064 to P.E.S., S.F., and N.J.), Support for Pioneering Research Initiated by the Next Generation from the Japan Science and Technology Agency (JPMJSP2123 to Y.O. and G.C.), Collaborative Research Grant from the Yamaha Corporation (to P.E.S. and S.H.), Horizon 2020 Framework Programme (grant number 754513) and Horizon Europe European Research Council (grant number 101045747) to Hansen, European Research Council Starting Grant (ERC-StG-2015, CAASD, and 678733) to Liu, and Keio University International Journal Article Publication Fee Grant (type A) to P.E.S. for article processing charges. **Author contributions:** Conceived the project: P.E.S., Y.O., A.T., P.Q.P., E.B., J.M.M., P.P., F.L., S.C.P., P.O., N.J., and W.T.F. Funding acquisition: P.E.S., Y.O., S.C.P., E.B., N.J., P.O., W.T.F., R.T., P.Q.P., F.L., and M.R. Project management: P.E.S. and Y.O. Recruitment: P.E.S., Y.O., N.J., P.O., P.Q.P., W.T.F., and B.S.B. Translation: B.S.B., P.E.S., and Y.O. Audio recordings for pilot analyses: Y.O., S.H., F.N., P.M.S., and J.M.M. Annotations for pilot analyses: Y.O., S.H., F.N., D.P.S., and P.E.S. Recording and text transcription/segmentation of own singing/speaking/instrumental performance: all authors. Detailed (millisecond-level) onset annotations: Y.O. (all data) and P.E.S. (interrater reliability subset). Checking/correcting onset annotations for own singing/speaking/instrumental performance: all authors. Conducted analyses: Y.O. Made Fig. 2 word clouds: J.S.G.-C. Drafting initial manuscript: Y.O. and P.E.S. Editing manuscript: A.T., P.Q.P., J.M.M., E.B., P.P., G.C., F.L., N.J., S.C.P., P.O., W.T.F., S.He., M.R., F.N., D.P.S., S.Ha., S.F., S.C., M.N., M.L.P., M.A.-T., N.C.H., F.H., U.F., W.K., O.S., D.H., B.S.B., M.A.C.V., M.v.T., P.D., O.C., J.T., T.L., D.K., C.T., D.F., Al.A., J.P.S., Ad.A., T.O.M., J.S.-Z., I.S.-S., R.A., P.L., A.R., Y.J., P.L.-M., C.B., T.P.T., U.K., N.B.S., L.R., M.Z., S.D.V., J.S.G.-C., K.K., C.V.B.d.N., and V.N.B. **Competing interests:** P.E.S. is a recommender at Peer Community In Registered Reports. The authors declare that they have no other competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Analysis code and data (fixed permanent repository) is available at https://doi.org/10.5281/zenodo.10612357. Analysis code is available at https://github.com/comp-music-lab/song-speech-analysis. Data are available at https://osf.io/mzxc8/.

Submitted 15 November 2023
Accepted 19 April 2024
Published 15 May 2024
10.1126/sciadv.adm9797