Globally, songs and instrumental melodies are slower, higher, and use more stable pitches than speech

Yuto Ozaki, Adam Tierney, Peter Pfordresher, John Mcbride, Emmanouil Benetos, Polina Proutskova, Gakuto Chiba, Fang Liu, Nori Jacoby, Suzanne Purdy, et al.

Globally, songs and instrumental melodies are slower, higher, and use more stable pitches than speech [Stage 2 Registered Report]

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

[@]Correspondence: yozaki@sfc.keio.ac.jp and psavage@sfc.keio.ac.jp

NB: Order of authors other than first and last authors is based on the order in which they joined the project. See Author Contributions statement for detailed information.

Please note: This is a Stage 2 Registered Report peer-reviewed and recommended for publication by *Peer Community In Registered Reports* (PCI RR). For full open peer reviews, author replies, editorial recommendation, and list of PCI RR-friendly journals eligible for publication without further review, see https://doi.org/10.24072/pci.rr.100469. Please direct correspondence to yozaki@sfc.keio.ac.jp and psavage@sfc.keio.ac.jp.

URL to the preregistered Stage 1 protocol: https://osf.io/jdhtz

Recommended citation: Ozaki, Y., Tierney, A., Pfordresher, P. Q., McBride, J., Benetos, E., Proutskouva, P., Chiba, G., Liu, F., Jacoby, N., Purdy, S. C., Opondo, P., Fitch, W. T., Rocamora, M., Thorne, R., Nweke, F., Sadaphal, D., Sadaphal, P., Hadavi, S., Fujii, S., … Savage, P. E. (Accepted ["Recommended"]). Globally, songs and instrumental melodies are slower, higher, and use more stable pitches than speech [Stage 2 Registered Report]. *Peer Community In Registered Reports.* Preprint: https://doi.org/10.31234/osf.io/jr9x7

[1]Graduate School of Media and Governance, Keio University, Japan
[2]Department of Psychological Sciences, Birkbeck, University of London, UK
[3]Department of Psychology, University at Buffalo, State University of New York, USA
[4]Institute for Basic Science, South Korea
[5]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[6]Faculty of Environment and Information Studies, Keio University, Japan
[7]School of Psychology & Clinical Language Sciences, University of Reading, UK
[8]Max-Planck Institute for Empirical Aesthetics, Germany
[9]School of Psychology / Eisdell Moore Centre for Hearing and Balance Research, University of Auckland, New Zealand
[10]School of Arts, Music Discipline, University of KwaZulu-Natal, South Africa
[11]Department of Behavioral and Cognitive Biology / Department of Musicology, University of Vienna, Austria

[12]Music Cognition Lab, Department of Clinical Psychology, National Institute of Mental Health and Neuro Sciences, India
[13]Universidad de la República, Uruguay
[14]School of Music, Victoria University of Wellington, New Zealand
[15]Department of Creative Arts, University of Lagos, Nigeria
[16]Independent researcher, India
[17]Department of Human Neurosciences, Sapienza University of Rome, Italy
[18]Haponetay, Shimizu-cho, Hokkaido, Japan
[19]Independent researcher, Japan/Senegal
[20]Memorial University of Newfoundland, Canada
[21]Department of Music and Dance, Kenyatta University, Kenya
[22]Faculty of Music, University of Oxford, UK
[23]Aarhus Institute of Advanced Studies, Aarhus University, Denmark
[24]Institute of Biological and Medical Psychology, Department of Psychology, University of Bergen, Norway
[25]Centro de Linguística da Universidade do Porto, University of Porto, Portugal
[26]Adam Mickiewicz University, Faculty of Art Studies, Musicology Institute, Poland
[27]Max-Planck Institute for the Science of Human History, Germany
[28]Clinical Psychology, Pepperdine University, USA
[29]Department of Experimental Psychology, Institute of Psychology, University of São Paulo, Brazil
[30]Independent researcher, Taiwan
[31]School of International Liberal Studies, Waseda University, Japan
[32]Headmistress, The Royal Music Academy, Yangon, Myanmar
[33]Department of Cultural Policy, University of Hildesheim, Germany
[34]Director of MİRAS Centre for Cultural Sustainability, İstanbul, Turkey
[35]Sydney Environment Institute, University of Sydney, Australia
[36]International Research Center for Traditional Polyphony of the Tbilisi State Conservatoire, Georgia
[37]South-West University ""Neofit Rilski", Bulgaria
[38]Universidade de Aveiro, Portugal
[39]Dalarna University, Sweden
[40]Independent researcher, Austria
[41]Department of Music and Dance, University of Cape Coast, Ghana
[42]Department of Music, University of Ghana, Ghana
[43]Department of Ethnomusicology and Folklore, Indiana University, USA
[44]Department of Humanities and Arts, University of Los Lagos, Chile
[45]Millennium Nucleus on Musical and Sound Cultures (CMUS), Chile
[46]Traditional performer and culture bearer, Chile
[47]Kaunas University of Technology and Lithuanian Academy of Music and Theatre, Lithuania
[48]Music, Imaging and Neural Dynamics Lab, Northeastern University, USA
[49]Max Planck-NYU Center for Language, Music, and Emotion (CLaME), USA & Germany
[50]Txemim Puri Project - Puri Language Research, Vitalization and Teaching/ Recording and Preservation of Puri History and Culture, Brasil
[51]Independent researcher, Brazil
[52]University of Glasgow, UK
[53]Max Planck Institute for Psycholinguistics, Netherlands
[54]Rhode Island School of Design, USA
[55]Institute for English & American Studies (IEAS), Goethe University of Frankfurt am Main, Germany
[56]Music Technology Group, Universitat Pompeu Fabra, Spain
[57]Institute of Cognitive Science, University of Osnabrück, Germany
[58]Department of Psychology, University of Toronto Mississsauga, Canada
[59]Independent researcher, Tokyo, Japan
[60]Faculty of Fine Arts, Chiang Mai University, Thailand
[61]University of Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France

Abstract

What, if any, similarities and differences between music and speech are consistent across cultures? Both music and language are found in all known human societies and are argued to share evolutionary roots and cognitive resources, yet no studies have compared similarities and differences between song, speech, and instrumental music across languages on a global scale. In this Registered Report, we analyze a novel dataset of 300 high-quality annotated audio recordings representing matched sets of singing, recitation, conversational speech, and instrumental music from our 75[1] coauthors whose 55 1st/heritage languages span 21 language families to find strong evidence for cross-culturally consistent differences and similarities between music and language. Of our six pre-registered predictions, five were strongly supported: relative to speech, songs use 1) higher pitch, 2) slower temporal rate, and 3) more stable pitches, while both songs and speech used similar 4) pitch interval size, and 5) timbral brightness. Our 6th prediction that song and speech would show similar pitch declination was inconclusive, with exploratory analysis suggesting that songs tend to follow an arched contour while speech contours tend to decline overall but end with a slight rise. Because our non-representative language sample and unusual design involving coauthors as participants could affect our results, we also performed robustness analyses - including a parallel reanalysis of a previously published dataset of 418 song/speech recordings from 209 individuals whose 16 languages span 11 language families (Hilton & Moser et al., 2022, *Nature Human Behaviour*) - which confirmed that our conclusions are robust to these potential biases. Exploratory analyses identified additional features such as phrase length, intensity, and rhythmic/melodic regularity that also consistently distinguish song from speech, and suggest that such features also vary along a “musi-linguistic” continuum in a cross-culturally consistent manner when including instrumental melodies and recited lyrics. Further exploratory analysis suggests that pitch height is the only consistently sexually dimorphic feature (female singing/speaking is almost one octave higher than male on average), and that other factors such as musical training and recording context may also interact to influence the magnitude of song-speech differences. Our study provides strong empirical evidence for the existence of cross-cultural regularities in music and speech.

1. Introduction

Language and music are both found universally across cultures, yet in highly diverse forms (Evans & Levinson, 2009; Jacoby et al., 2020; Mehr et al., 2019; Savage 2019; Sammler, Under contract), leading many to speculate on their evolutionary functions and possible

[1]NB: 6 of the original 81 planned coauthors were unable to complete the recording and annotation process compared to our initially planned sample (compare the new Fig. 3 map with the originally planned Fig. S1 map). These six collaborators were excluded, following our exclusion criteria (S.1.2.2). 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 re-record themselves in the correct language as required by the recording Protocol (Appendix 1). 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. We aim to include their re-recorded sets if they can submit them in time to finalize the manuscript for publication. Updating these will not change the results of the Table 3 robustness check, as these collaborators were already not blind to our hypotheses, so they would be excluded from this analysis anyway. It is also unlikely to change the p-values in Table 2 calculated based on 73 recording sets. We commit to updating our analyses to reflect their new recordings if they can be submitted in time, regardless of if/how it impacts our conclusions.

coevolution (e.g., Darwin, 1871; Haiduk & Fitch, 2022; Mehr et al., 2021; Patel, 2008; Savage et al., 2021; Valentova et al., 2019). Yet such 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 (Albouy et al., 2020; Doelling et al., 2019; Morrill et al., 2015; Patel, 2008, 2011; Peretz, 2009; Rogalsky et al., 2011), there has been relatively less comparative analysis of *acoustic* attributes of music and language (e.g., Ding et al., 2017; Patel et al., 2006), 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 (Bickel, 2011; Brown, 1991; Brown and Jordiana, 2013; Savage et al., 2015). 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) (Mehr et al., 2019; Savage et al., 2015). 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; use of higher pitch when vocalizing to infants) (Haiduk & Fitch, 2022; Hilton et al., 2022; Ozaki et al., 2022; Patel, 2008; Tierney et al., 2011). 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 (Valentova et al., 2019).

Many hypotheses make predictions about cross-cultural similarities and differences between song and speech. For example, the social bonding hypothesis (Savage et al., 2021) predicts that song is more predictably regular than speech to facilitate synchronization and social bonding. In contrast, Tierney et al.'s (2011) motor constraint hypothesis predicts similarities in pitch interval size and melodic contour due to shared constraints on sung and spoken vocalization. Similarly, the sexual selection hypothesis (Valentova et al., 2019) predicts similarities between singing and speaking due to their redundant functions as 'backup signals' indicating similar underlying mate qualities (e.g., body size). Finally, 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 (List, 1971).

Culturally relativistic hypotheses appear to be dominant among ethnomusicologists. For example, in a Jan 13, 2022 email to the International Council for Traditional Music (ICTM) email list entitled "What is song?", ICTM 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. (2015) 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 [1968] 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" [Blacking, 1973] other than speech, with the distinction between speech and music being left to each culture's emic (insider, 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."*** (emphasis added)

Importantly, however, Savage et al.'s conclusion 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 (Hansen et al., 2020; Merrill & Larrouy-Maestri, 2017; Sharma et al., 2021; Vanden Bosch der Nederlanden et al., 2022). 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 (cf. Blasi et al., 2022). The available dataset closest to our study is Hilton, Moser, et al.'s (2022) 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 vs. infant-directed versions of song and of speech, and they did not directly compare singing vs. speaking. We performed exploratory analyses of their dataset (Ozaki et al., 2022), 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 [Durojaye et al., 2021]; bird song notes separated by micro-breaths [Roeske et al. 2020]), we found 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 (cf. Fig. S6). For example, Mertens' (2022) automated segmentation algorithm used by Hilton et al. (2022) mis-segmented two out of the first three words "by a lonely" from the English song used in our pilot analyses ("The Fields of Athenry"), over-segmenting "by" into "b-y", and under-segmenting "lonely" by failing to divide it into "lone-ly" (cf. 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, phrases) by the coauthor who recorded it in their own 1st/heritage language. Furthermore, because singing and speaking exist on a broader "musi-linguistic" spectrum including forms such as instrumental music and poetry recitation (Brown, 2000; Leongómez et al., 2022; Tsur and Gafni, 2022), we collected four types of recordings to capture variation across this spectrum: **1) singing**, **2) recitation** of the sung lyrics, **3) spoken description** of the song, and **4) instrumental** version of the sung melody (Fig. 1).

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 vs. natural speech by comparing the exact same lyrics when sung vs. 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 vs. song) versus a different system (speech vs. instrument).

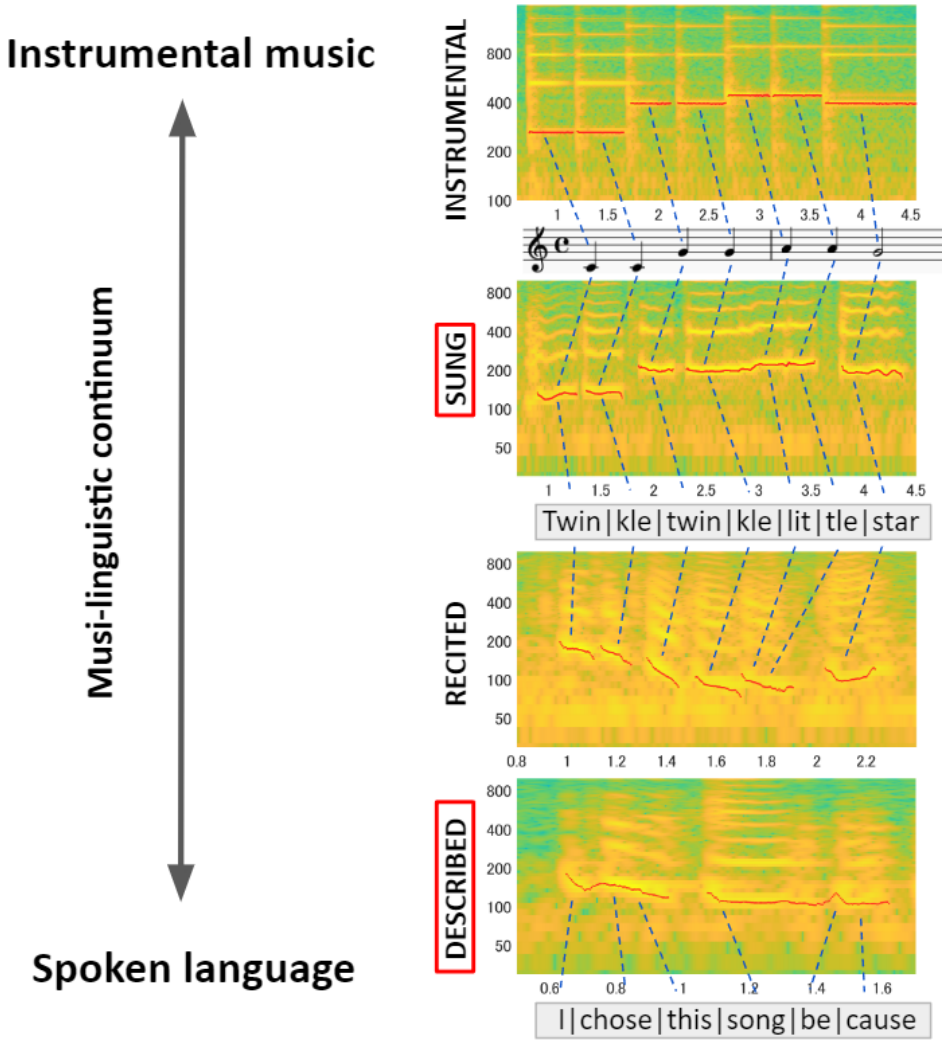


Figure 1. **Example excerpts of the four recording types collected in this study, arranged in a "musi-linguistic continuum" from instrumental music to spoken language.** Spectrograms (x-axis: time [seconds], y-axis: frequency [Hz]) of the four types of recordings are displayed on the right-hand side (excerpts of author Savage performing/describing "Twinkle Twinkle", 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.

1.1. 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 in

order 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 (Savage et al., 2021b). Such 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 (Patel, 2018; Savage et al., 2021; Hoeschele & Fitch, 2022). Rather than try to disambiguate such 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 (see Section S1.4), we settled on six features we believe we can reliably test for predicted similarities/differences: **1) pitch height, 2) temporal rate, 3) pitch stability, 4) timbral brightness, 5) pitch interval size**, and **6) pitch declination** (cf. Table 1). Detailed speculation on the possible mechanisms underlying potential similarities and differences are described in the Supplementary Discussion section (S2).

Table 1. Registered Report Design Planner. Includes six hypotheses (H1-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. 3)**. Recruitment was opportunistic based on collaborator networks aiming to maximize global diversity and achieve greater than 95% a priori power even if some data has to be excluded (see **Sec. S1.2** for inclusion/ exclusion criteria).	Meta-analysis framework **(Fig. 2)** 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 (2019) 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 α = .05; β = .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, 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 (*inter-onset interval (IOI) rate*)** instead of **pitch height (f_0)**					
	3) Song uses **more stable pitches** than speech	Same as H1, but for **pitch stability** (-$\|\Delta f_0\|$) 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 ±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.

1.2. 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 will be applied to the prediction of the statistical difference. Another type of null hypothesis is whether the effect size of the feature exceeds the smallest effect size of interest (SESOI) (Lakens, 2017). This hypothesis will be 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 (Brysbaert, 2019). 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, 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 will perform the within-participant 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 (cf. red boxes in Fig. 1; the comparisons with lyrics recitation and with instrumental versions will be saved for exploratory analyses). In addition, terms in the computed difference scores will be arranged so that for our predicted differences (H1-H3), a positive value indicates a difference in the predicted direction (cf. Fig. S3).

Evaluation of difference in the magnitude of each feature is performed with nonparametric relative effects (Brunner et al., 2018) which is also known as stochastic superiority (Vargha & Delaney, 1998) or probability-based measure of effect size (Ruscio, 2008). 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. 2). 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 (Brockwell & Gordon, 2001; Liu et al., 2018), and we frame our hypotheses as the inference of the mean parameter of Gaussian random-effects models which indicates the population effect size.

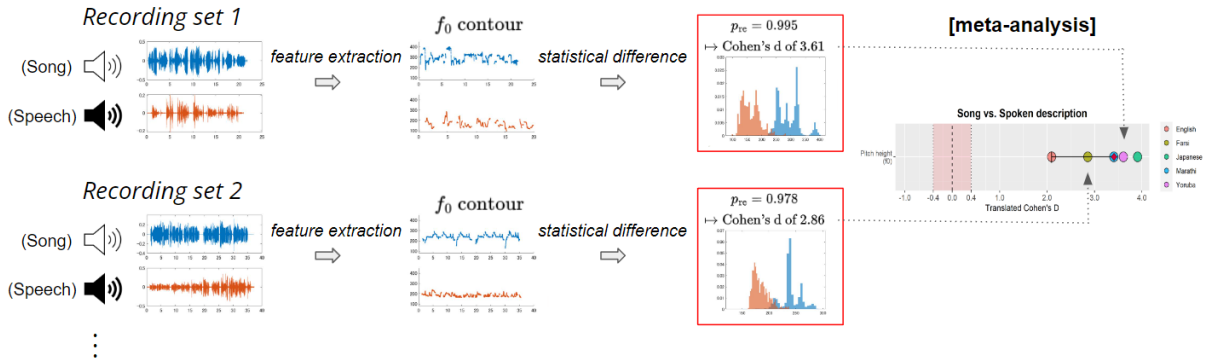


Figure 2. **Schematic overview of the analysis pipeline from raw audio recordings to the paired comparisons shown in Figure S2.** Recording sets 1 and 2 represent pilot data of singing and speaking in Yoruba and Farsi by coauthors Nweke and Hadavi, respectively. From each pair of song/spoken audio recordings by a given person, we quantify the difference using the effect size for each feature. p_{re} 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). In order 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 other recording sets to create the processed data for hypothesis testing shown in Figure S2, Different types of hypothesis testing are applied depending on the feature (i.e. hypothesis of difference and hypothesis of similarity).

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) (Brysbaert, 2019). We test six features, and thus test six null hypotheses.

Since we test multiple hypotheses, we will use the false discovery rate method with the Benjamini-Hochberg step-up procedure (Benjamini & Hochberg, 1995) to decide on the rejection of the null hypotheses. We define the alpha level as 0.05.

For the hypothesis testing of null effect size (H1-H3), we test whether the endpoints 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. (2018) and Wang & Tian (2018) to construct the confidence interval. For the hypothesis testing of equivalence (H4-H6), we first estimate the mean parameter (i.e. overall treatment effect) with the exact confidence interval (Liu et al., 2018; Wang & Tian, 2018) and the between-study variance with the DerSimonian-Laird estimator (DerSimonian & Laird, 1986). 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 (2005) shrinking alternative parameter space method to test whether the population mean is within the SESOI as specified above.

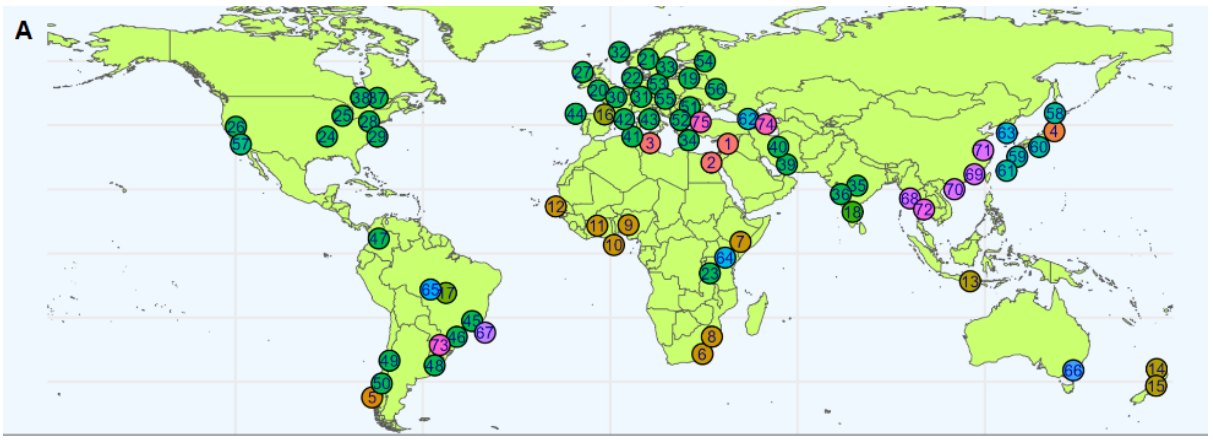
Our choice of an SESOI of d = 0.4 based on Brysbaert's (2019) recommendation after reviewing psychological studies is admittedly somewhat arbitrary. Future studies might be able to choose a different SESOI on a more principled basis based on the data and analyses we provide here, and the value of our 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, model selection based on Bayes factors) independent of the choice of prior distributions. We thus chose to rely on Brysbaert's recommended SESOI of d = 0.4 (and its equivalent relative effect of p_{re} = 0.61) in the absence of better alternatives.

Visual and aural inspection 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/spoken examples so one can hear what our proposed SESOI would sound like for a range of languages and features (Section S7 and Table S1; audio files also at https://osf.io/mzxc8/.

2. Methods [NB: The current manuscript is structured as a completed Stage 2 Registered Report, but this format can be modified if required based on additional reviewer feedback]

All details are written in the S1 Supplementary methods section. Here, we briefly introduce two key aspects: language sample and acoustic features.

We have recruited 75 collaborators from around the world, spanning the speakers of 21 language families (Fig. 3). All audio recordings analyzed are made by our group of 75 coauthors recording ourselves singing/speaking in our 1st/heritage languages. Collaborators were chosen by opportunistic sampling beginning from co-corresponding author Savage's network of researchers (cf. S1.2. for details).



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					

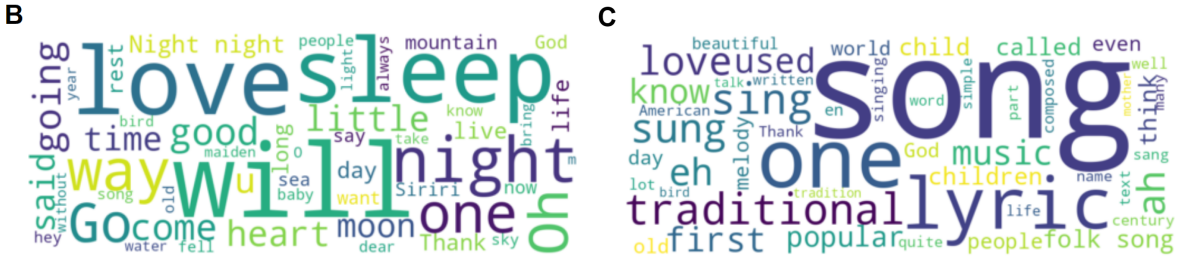


Figure 3. Map of the linguistic varieties spoken by our 75 coauthors as 1st/heritage languages (A). (NB: 6 of the original 81 planned coauthors were unable to complete the recording and annotation process compared to our initially planned sample; cf. Fig. S1 for the original map of 81 linguistic varieties). Each circle represents a coauthor singing and speaking in their 1st (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 (Hammarström et al., 2022), 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 (Dryer et al., 2013; https://wals.info/languoid). 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).

We compared the following six acoustic features between song and speech for our main confirmatory analyses:

1) Pitch height (fundamental frequency (f_0)) [*Hz*],
2) Temporal rate (inter-onset interval (IOI) rate) [*Hz*],
    - 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, 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, 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 (-|f_0|) [*cent/sec.*],
4) Timbral brightness (spectral centroid) [*Hz*],
5) Pitch interval size (f_0 ratio) [*cent*],
    - 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.

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 (cf. Fig. 2). Details can be found in S1.3. and S3.

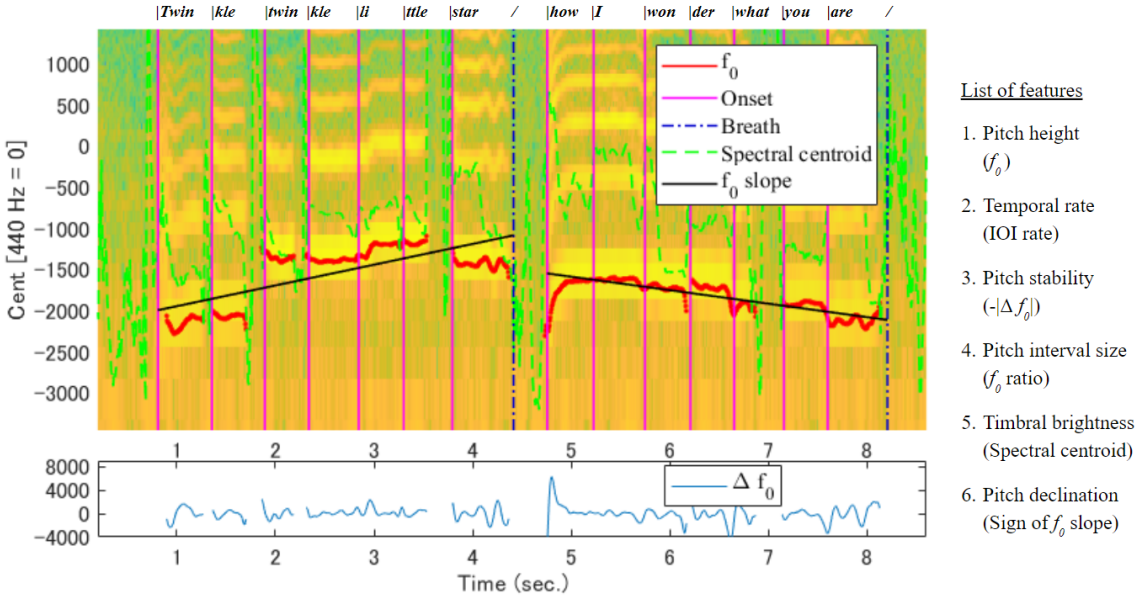


Figure 4. Schematic illustration of the six features analyzed for confirmatory analysis, using a recording of author Savage singing the first two phrases of "Twinkle Twinkle Little Star" as an example. Onset and breathing annotations are based on the segmented texts displayed on the top of the spectrogram. 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.

3. Changes to Stage 1 Registered Report protocol (Introduction and Method sections 1-2 plus Supplementary Materials)

We have left the content of Introduction and Method (Sections 1-2) and Supplementary Materials unchanged from the version granted In Principle Acceptance (accessible at https://osf.io/download/6387919ba98e5f286310370d/?version=4), following Registered Report procedures to avoid any possibility of adjusting hypotheses or analyses after knowing the results. However, we have moved the majority of the Method section to Supplementary Materials to make the main result and discussion easier to read. At the time we submitted the Stage 1 manuscript, we mainly reported our pilot data results included in the Method section, but now those results have been moved to Supplementary Information.

As a result, we have renumbered Section and Figure numbers and have updated cross-references to them. In addition, we have added a subsection title to the paragraph explaining exploratory features in the supplementary materials which should have been there. Minor typos have also been corrected accordingly.

Note that the map in the Methods section (Fig. 3) reflects the final 75 collaborators who provided audio recording data, not the original 81 collaborators shown in the original map (Fig. S1), as 6 collaborators were unable to provide recording data. We have also added a word cloud visualization of the translated content of the sung/spoken audio recordings to accompany this map.

4. Results

4.1. Confirmatory analysis

The results of the confirmatory hypothesis testing with 73 recording sets confirm 5 of our 6 predictions (Fig. 5 and Table 2; all $p < 1x10^{-5}$). 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=.57), with an estimated effect size of D = 0.42 slightly greater than our pre-specified "Smallest Effect Size of Interest" (SESOI) of D = 0.4. In section 4.2.7 we perform alternative exploratory analyses to understand possible reasons for this failed prediction.

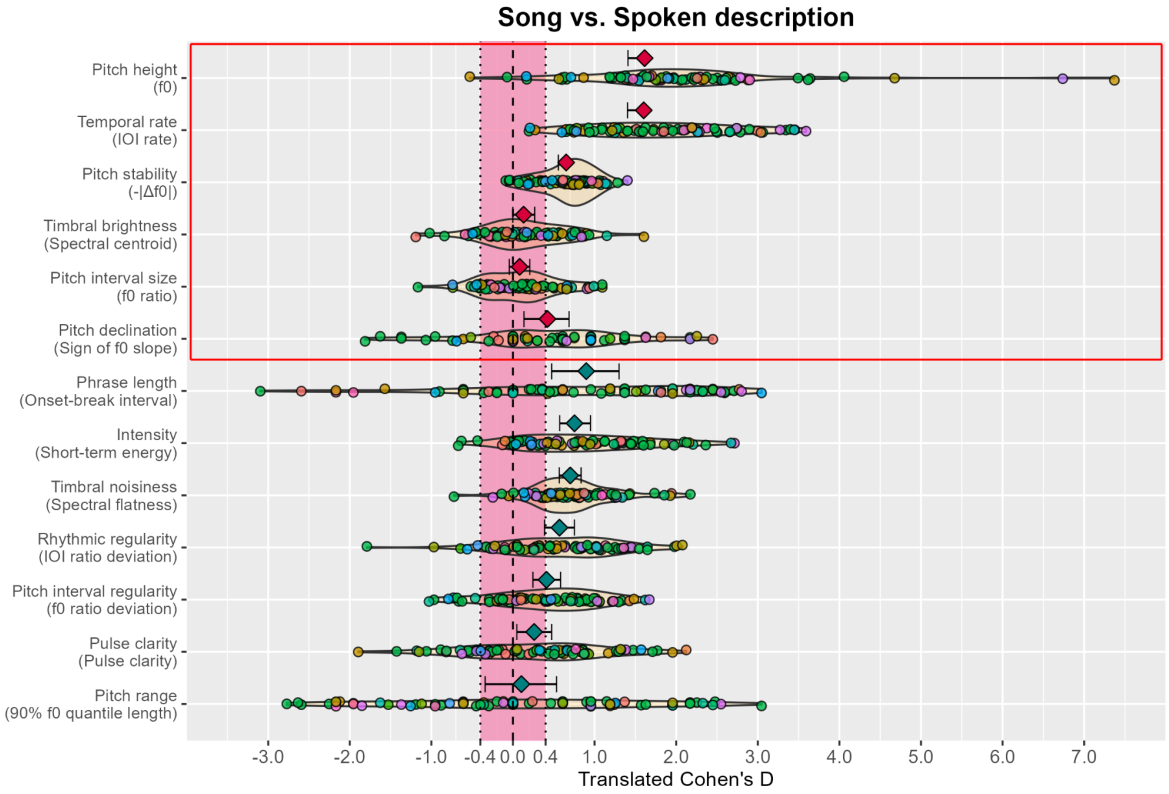


Figure 5. Plot of effect sizes showing differences of each feature between singing and spoken description of the 73 recording sets for the confirmatory analysis and 75 recording sets for the exploratory analysis. The plot includes 7 additional exploratory features, and the 6 features corresponding to 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 are measured per recording pair. Readers can find further information on how to interpret the figure in the caption of Figure S2 and Figure S9. Note that the colors of data points indicate language families, which are coded the same as in Figure 3, and violin plots are added to this figure compared to Figure S2.

Hypothesis	Feature	Test	Combined ES	CI (α = 0.05/6)	p-value
1) Song uses higher pitch than speech	f_0	One-tailed confidence interval of the combined effect size	1.61	1.41, n/a	***< 1.0x10^{-8}**
2) Song is slower than speech	IOI rate		1.60	1.40, n/a	***< 1.0x10^{-8}**
3) Song uses more stable pitches than speech	$-\lvert\Delta f_0\rvert$		0.65	0.56, n/a	***< 1.0x10^{-8}**
4) Song and speech use similar timbral brightness	Spectral centroid	Equivalence test for the combined effect size	0.13	-0.0046, 0.27	***5.2x10^{-6}**
5) Song and speech use similar sized pitch intervals	f_0 ratio		0.082	-0.044, 0.21	***< 1.0x10^{-8}**
6) Song and speech use similar pitch contours	Sign of f_0 slope		0.42	0.13, 0.69	.57

Table 2. Results of the confirmatory analysis. The effect sizes reported in the table are Cohen's *d* transformed from relative effects for ease of interpretation, but the hypothesis tests were conducted with relative effects. The CIs are either one-tailed or two-tailed, depending on the aim of the test. Note the equivalence test uses statistics different from the above meta-analysis CIs to verify equivalence hypotheses. Asterisks in p-values indicate that the null hypothesis is rejected.

Our robustness checks 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 3). This result suggests our unusual "participants as coauthors" model did not influence our confirmatory analyses. In addition, the other robustness check suggests that the measured effect sizes do not have language family-specific variance (Table 4), which supports the appropriateness of the use of simple random-effect models in the analyses.

Hypothesis	Feature	Test	Combined ES	CI (α = 0.05/6)	p-value
1) Song uses higher pitch than speech	f_0	One-tailed confidence interval of the combined effect size	1.73	1.46, n/a	***< 1.0x10^{-8}**
2) Song is slower than speech	IOI rate		1.64	1.40, n/a	***< 1.0x10^{-8}**
3) Song uses more stable pitches than speech	$-\lvert\Delta f_0\rvert$		0.64	0.51, n/a	***< 1.0x10^{-8}**
4) Song and speech use similar timbral brightness	Spectral centroid	Equivalence test for the combined effect size	0.14	-0.028, 0.31	***3.3x10^{-4}**
5) Song and speech use similar sized pitch intervals	f_0 ratio		0.10	-0.067, 0.27	***3.5x10^{-5}**
6) Song and speech use similar pitch contours	Sign of f_0 slope		0.23	-0.11, 0.60	.12

Table 3. Results of the robustness check, which used data only from the collaborators who had not known the hypotheses when generating data (47 pairs of singing and spoken description recordings).

Hypothesis	AIC (standard)	AIC (multi-level)	Log likelihood (standard)	Log likelihood (multi-level)	Variance of the effects at language family
1) Song uses higher pitch than speech	**-87.08**	-85.08	45.54	45.54	< 1.0×10^{-8}
2) Song is slower than speech	**-111.64**	-109.73	57.82	57.86	1.86×10^{-3}
3) Song uses more stable pitches than speech	**-153.53**	-151.53	78.76	78.76	< 1.0×10^{-8}
4) Song and speech use similar timbral brightness	**-86.32**	-84.90	45.16	45.45	2.07×10^{-3}
5) Song and speech use similar sized pitch intervals	**-95.90**	-93.90	49.95	49.95	< 1.0×10^{-8}
6) Song and speech use similar pitch contours	**-7.24**	-5.48	5.62	5.74	2.29×10^{-3}

Table 4. Results of the robustness check comparing models taking into account dependency by language families. Superior AIC scores are highlighted in bold. Maximum likelihood estimation is used to fit the models. “standard” refers to standard random-effects models used in the confirmatory analyses, and “multi-level” refers to two-level random-effects models grouping data by language families. The right-most column shows the maximum likelihood estimate of the variance parameters appearing in the multi-level models. The log-likelihoods are almost identical between the two models, and multi-level models degenerate to standard random effects models (i.e. variance due to language family is negligible), which means grouping data by language family is redundant and simple random effects models are enough to model data.

4.2. Exploratory analysis

4.2.1. More acoustic features

We specified six features for our confirmatory analyses, but human music and speech can be characterized by additional acoustic features. We include 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 (cf. section S6). Although we do not formally construct and test hypotheses for this analysis, Figure 5 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, higher sound pressure, and have less vocal noise than speech. Note that as described in 1.2, the order of comparison is arranged so that difference is expressed as a positive value, so that difference in timbral noisiness is calculated as noisiness of spoken description relative to song.

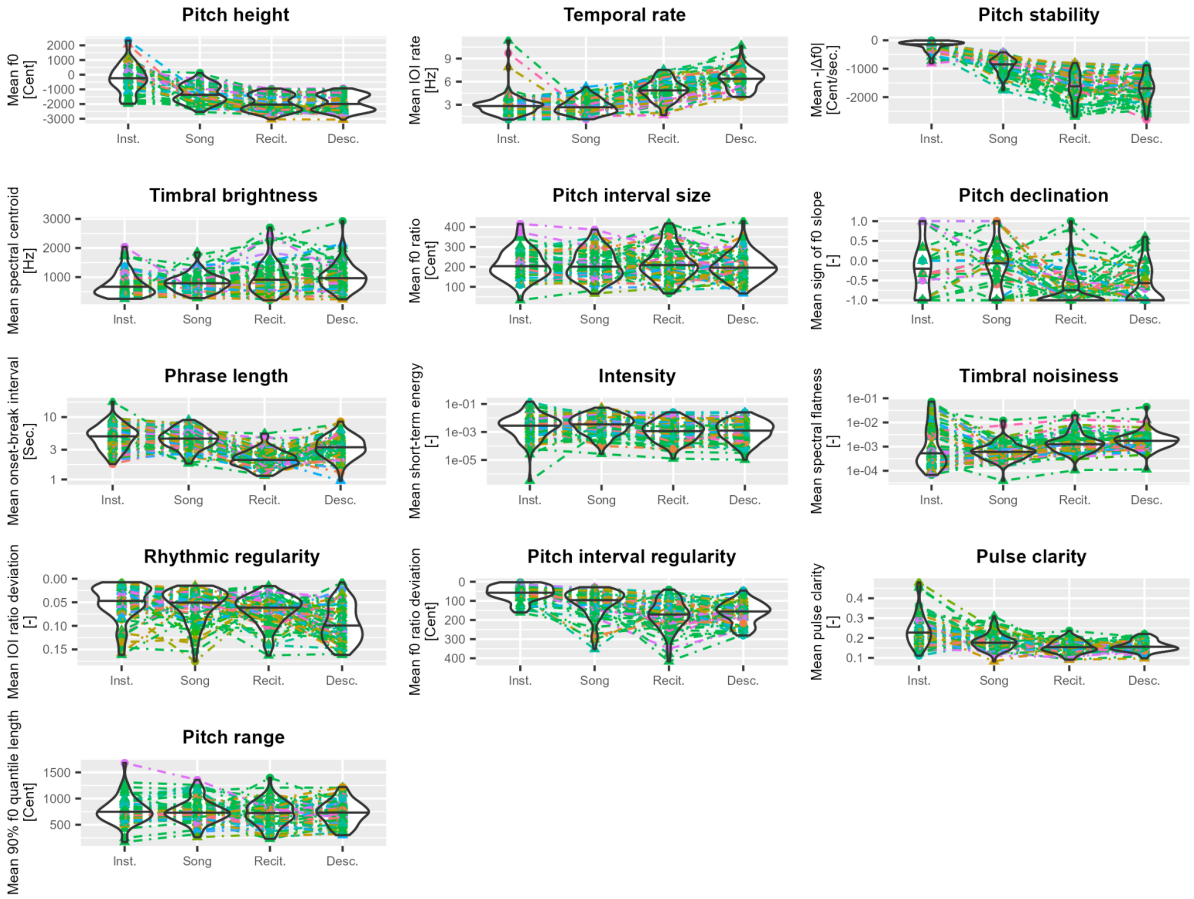


Figure 6. Alternative visualization of Figure 5 showing mean values of each feature rather than paired differences but with all recording types. Note that the colors of data points indicate language families, which are coded the same as in Figure 3. The horizontal lines in the violin plots indicate the median.

4.2.2. Music-language continuum: including instrumental/recited lyrics

Exploratory analyses that include comparisons with lyrics recitation and instrumental recordings (cf. Fig. S13 and Fig. 6) suggest that 1) comparing singing vs. lyrics recitation shows qualitatively the same results as for singing vs. spoken description in terms of how confidence intervals intersect with the null point and the equivalence region; 2) comparing instrumental vs. speech (both spoken description/lyrics recitation) reveals larger differences in pitch height, temporal rate, and pitch stability than found with song vs. speech; 3) features shown to be similar between song vs. speech (e.g., timbral brightness and pitch interval size) show differences when comparing instrumental vs. speech; 4) few major differences are observed between lyrics recitation and spoken description, except that recitation tends be slower and use shorter phrases; 5) the instrumental generally has a more extreme (larger/smaller) magnitude than singing for each feature except for temporal rate; and 6) pitch height, temporal rate, and pitch stability display a noticeable constantly increasing (or decreasing) continuum from spoken description to instrumental.

A similar trend is also found in additional differentiating features discussed in 4.2.1 (i.e., phrase length, timbral noisiness, and loudness). We also performed a nonparametric trend test (cf., Table S2) 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.

4.2.3. 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 ~1,000 cents [almost one octave] consistently for song, spoken description, and recited lyrics). However, there is 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 (Chen et al., 2022; Feinberg et al., 2018; Puts et al., 2006; 2016; Valentova et al., 2019).

4.2.4. Analysis by linguistic factors: nPVI

We employed nPVI (Patel & Daniele, 2003) to examine the degree of variation in inter-onset intervals and onset-break intervals (cf. S3.2. & S8.) 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 (Ling et al., 2000), although our duration is defined by the sequence of onset (cf. S1.1.) and break annotations (cf. S8.) which 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 (cf. Fig. S20). However, we observed that

(1) nPVIs of song and spoken description do not seem to create distinct clusters among our recordings (whether into “syllable-timed”, “stress-timed”, or any other categories), (2) nPVIs of song and spoken description do 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), and (3) nPVIs of spoken description tend to be slightly larger than song and instrumental. The third result suggests 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 (Savage et al., 2015). In addition, though linguists use various features (Grabe & Low, 2002) to carefully characterize the rhythm of speech, the first two observations suggest 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.

4.2.5. Reliability of annotation process: Inter-rater reliability of onset annotations

We analyzed the inter-rater reliability of onset annotations to check how large individual varieties are in the annotation. As stipulated in S1.7.7, Savage created onset annotations to the first 10 seconds of randomly chosen 8 pairs of song and spoken description recordings. In this 10-second 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 YO and confirmed by the coauthor who made the recording. Therefore, the annotation by PES follows the same segmentation as the annotation by YO, but can differ in the exact timing for which each segmentation is judged to begin. We measured intra-class correlations (ICCs) of onset times with two-way random-effects models measuring absolute agreement. As a result, all annotations show strong ICCs (> .99), which indicates 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 is also conducted (cf., Fig. S21). In the case of singing, 90% of onset time differences are within 0.083 seconds. Similarly, in the case of spoken description, 90% of onset time differences are within 0.055 seconds. In other words, Ozaki’s manual onset annotations that form 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.

4.2.6. Exploring recording representativeness and automated scalability: Comparison with alternative speech-song dataset (Hilton et al., 2022)

As stated in S1.7.8, we performed two exploratory analyses using automated methods to investigate (1) the reproducibility of our findings with another corpus and (2) the applicability of automated methods to substitute data extraction processes involving manual work. We analyzed the recordings of adult-directed singing and speech of Hilton et al.’s (2022) 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 5 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 nuclei segmentation method by de Jong & Wempe (2009) failed to capture the major differences in temporal rate that we identified using manual segmentation in Fig. 5. Instead, if we had used this automated method, 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, 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 pYIN (Mauch & Dixon, 2014) 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. The result suggests that (1) the same statistical significance can be obtained from Hilton et al.'s data though overall effect sizes tend to be weakened, and (2) combined effect sizes based on pYIN with full-length duration only show negligible differences from the original analysis involving manual work despite the drastic 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 sample (translated Cohen's d=0.30) are small enough to be within our defined equivalence region ($|d|<0.4$) if we had predicted it to be equivalent, but it is 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 ($p < .005$). Similar to Fig. 6, mean values of each feature per recording can be found in the supplementary information (Fig. S17-S19).

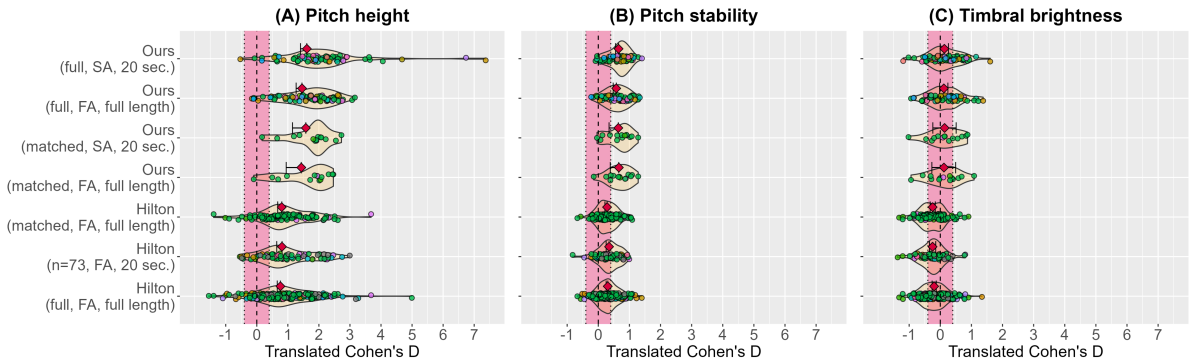


Figure 7. Re-running the analyses on four different samples using different fundamental frequency extraction methods: 1) our full sample (matched song and speech recordings from our 75 coauthors); 2) Hilton et al.'s (2022) full sample (matched song and speech recordings from 209 individuals); 3) a sub–sample of our 14 coauthors singing/speaking in English, Spanish, Mandarin, Kannada, and Polish), and 4) a sub-sample of Hilton et al.'s 122 participants also singing/speaking in English, Spanish, Mandarin, Kannada, and Polish). "SA" means that f_0s are extracted in a semi-automated manner (cf. S3.1), while "FA" means they were exactly in a fully automated manner (using the pYIN algorithm). Semi-automated analyses could only be performed on 20s excerpts of our recordings annotated by the coauthor who recorded them, while automated analyses could be applied to the full samples. In order to make the comparison with our results more interpretable, we have also added the analysis of Hilton's data using the same number of song-speech recording pairs with us (i.e., randomly selected 74 pairs of recordings), extracting features from the first 20 seconds. 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 visualization follows the same convention as in Figure 5 and Figure 8. However, Hilton et al.'s (2022) dataset contains languages that are not in our dataset. Therefore, slightly different color mapping was applied (cf. Fig. S16). Note that some large effect sizes (D > 3.5) in the pitch height of our original analysis (i.e., full-SA-20 sec.) 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_0s of pYIN are very high in spoken description or very low in singing, relative effects become smaller than semi-automated methods that remove such errors.

4.2.7. Alternative analysis approaches for pitch declination (hypothesis 6)

The only one of our 6 predictions that was not confirmed was our prediction that song and speech would display similar pitch declination. However, we would like to point out that only 3 to 4 f_0 slopes (equal to the number of "phrases" or intervals from the first onset after a break and to the next break, cf. Fig. 4) are, on average, included in the 20s 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 30s recordings which contain 5 to 7 f_0 slopes for singing and spoken description on average, still the p-value was not small enough to reject the null hypothesis (p = .48, CI [.17, .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 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 which we use (e.g. DerSimonian-Laird estimator) since the standard deviations of effect sizes become zero, leading to computation undefined. We employed the same procedure used in our power analysis for such cases (cf. S4.2), but a more widely known practice would be zero-cell corrections used in binary outcome data analysis (Weber et al., 2020). Signs of f_0 slopes are dichotomous outcomes (i.e. positive or negative), and drawing upon zero-cell corrections, we artificially appended a plus and minus sign to each of the signs of f_0 slopes from singing and spoken description recordings when estimating standard errors of relative effects if needed (e.g. [-1, -1, -1] → [-1, -1, -1, 1, -1] for the case of 3 f_0 slopes). In zero-cell corrections, 0.5 is added to all cells of the 2×2 table. Our analysis is not based on count data, so we cannot exactly follow this correction. However, adding plus and minus signs to the outcome of both singing and spoken description recordings has a similar effect. In other words, our additional procedure is similar to zero-cell corrections but adding 1 instead of 0.5 to all cells. This additional analysis provided virtually identical results with the main analysis reported in 3.1 (p = .66, CI [.15, .71]), suggesting that the way to handle zero frequency f_0 slope sign data is not crucial.

Lastly, we also checked the average trend of f_0 contours segmented by onset and break annotations (cf. Figure 8). 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 even though the second half of songs tend to descend as predicted, the first half of songs tend to rise, in contrast to speech which 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 (cf. 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, cf. Figure 8).

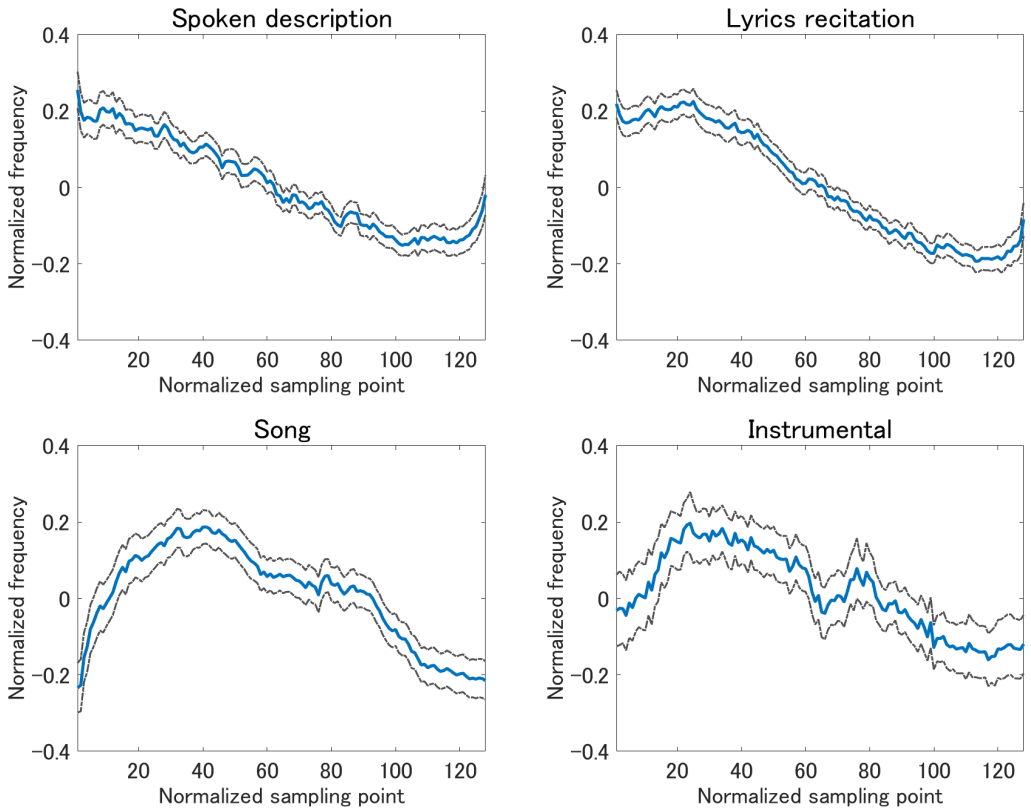


Figure 8. Averaged f_0 contours. f_0 contours extracted by the segments between onset and break were averaged to visualize the overall trend. The extracted f_0 contours were normalized to the length of 512 samples using interpolation by Fourier transform and resampling (Fraser, 1989; Schafer & Rabiner, 1973). The implementation by the MATLAB function interpft is used. Besides, the frequencies of extracted f_0 contours were standardized. Missing data from unvoiced segments of f0 contours were excluded. The blue lines represent averaged f_0 contours, and the black lines indicate 95% confidence intervals assuming the frequencies at each normalized sampling point were distributed normally. The average widths of confidence intervals of each category are .14 for instrumental, .097 for song, .060 for lyrics recitation, and .065 for spoken description.

Furthermore, the width of standard errors around the mean contour (cf. Figure. 8) 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. Indeed, musical melodies are considered to have multiple typical shapes (Adams, 1976), so the overall average contour is not necessarily representative of all samples.

4.2.8. Explanatory power of the features in song-speech classification

In order to probe the explanatory power of features on classifying acoustic signals into song and speech, we evaluated feature importance using permutation importance (Breiman, 2001) 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 (*Permutation Importance*, n.d.). Since how the feature contributes to solving the given task

differs in machine learning models, we employed three binary classification models to mitigate the bias from particular models: logistic regression with L2 regularization, SVM with RBF kernel, and naive Bayes with Laplace smoothing.

We computed permutation importance by randomly splitting 75 recording sets into the training set (n = 67) and test set (n = 8, 10% held-out) to fit the model and to evaluate the importance of features in the classification task, and repeated the same process 1024 times. The mean values of the feature, which are plotted in Figure 6, were used as data after normalization. The average of 1024 realizations of permutation importance values was reported here as the final output.

The result suggests at least temporal rate, pitch stability, and pitch declination are constantly weighed among these three models (cf. Fig. S22). All classifiers achieved average accuracy and F1 score higher than 90 (cf. Table S3). The importance of the other features depends on the models. For example, logistic regression gave the highest importance to pitch interval regularity as their 3rd most important feature. Naive Bayes chose rhythmic regularity as the 2nd most important feature, but this feature did not have a noticeable impact on SVM. On the other hand, it is consistent with the confirmatory analysis that pitch interval size and timbral brightness are evaluated as unimportant in discriminating between song and speech.

Interestingly, there are several cases that some features showing a strong difference within subjects were not evaluated as important in this analysis, including pitch height and intensity (cf. Fig. 5 and Fig. S22). Two reasons can be considered. One reason is relative largeness within the individual is not as informative in classifying acoustic signals collected from multiple individuals. In this case, between-subjects 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 (cf. Fig. S23-S24) show several features have medium to large size correlation (e.g., increase in pitch interval regularity with a decrease in temporal rate in singing with r = -.53). Therefore, there is a possibility that some features are evaluated as unimportant not because that feature is irrelevant to classify 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.

Inspection of the correlation matrices suggests 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. However, correlation is considered acting in the underestimation of permutation importance (Pereira et al., 2022). Therefore, at least the two features that consistently scored high among the three between-participant models and that confirmed our predicted within-participant differences - namely, temporal rate and pitch stability - capture important factors differentiating song and speech across cultures.

5. Discussion

5.1. Main confirmatory predictions and their robustness

Our analyses strongly support five out of our six predictions across an unprecedentedly diverse global sample of music/speech recordings: 1) song uses higher pitch than speech, 2) song is slower than speech, 3) song uses more stable pitches than speech, 4) song and speech use similar timbral brightness, and 5) song and speech use similar sized pitch intervals (Fig. 5). 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. 6).

While some of our findings were already expected from previous studies mainly focused on English and other Indo-European languages (Chang et al., 2022; Ding et al., 2017; Hansen et al., 2020; Merrill & Larrouy-Maestri, 2017; Sharma et al., 2021; see also S2.1 and Blasi et al., 2022), 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. Fig. 5 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 vs. spoken timbre or interval size were substantially larger than our designated "Smallest Effect Size Of Interest". The most consistent differences were found for temporal rate, as song was slower than speech for all 73 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. 8). 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 (Tierney et al., 2011; Savage et al., 2017). 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 confirm that our primary confirmatory results were not artefacts of our choice to record from a non-representative sample of coauthors. Specifically: 1) language families do not account for variances in the measured song-speech differences and similarities (Table 4), which means that these differences and similarities are cross-linguistically regular phenomena, and 2) analyzing only recordings from coauthors who made recordings prior to learning our hypotheses produced qualitatively identical conclusions (Table 3). Analysis of Hilton et al.'s (2022) dataset of field recordings also

supplemented our findings, producing qualitatively identical conclusions, regardless of the precise analysis methods or specific sample/sub-sample used (Fig. 7).

5.2. Implications from the exploratory analyses

Comparisons with lyrics recitation and instrumental recordings revealed 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. 6). 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; mapping music and language according to pitch, rhythm, and propositional/emotional functionality) may elucidate a more nuanced spectrum of musi-linguistic continuum (Brown, 2000; Leongómez et al., 2022; Hilton et al., 2022).

5.3. 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., 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 like flutes also are limited by breathing capacity), and some of which are not (e.g., chordophones like violins are limited by finger motor control). For example, though 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. In order to compare sung and instrumental features, all of our collaborators agreed to at least record themselves tapping the rhythm of their singing, but such 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 non-spoken forms of language (e.g., sign language, written language).

5.4. Comparison with alternative dataset (Hilton & Moser et al., 2022)

Interestingly, while the qualitative results using Hilton et al.'s dataset were identical, the magnitude of their song-speech differences were 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 (~600 cents [half an octave] on average vs. ~300 cents [quarter octave], respectively). These differences were consistent even when analyzed using matching sub-samples speaking the same languages and using the same fully automated analysis methods (Fig. 7), suggesting 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 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; cf. Fig. S11).Such background noise may reduce the observed differences between speech and song.

Another potential factor is musical experiences. Our coauthors were mostly recruited from academic societies studying music, and many also have substantial experience as performing musicians. Although the degree of musical experiences of Hilton et al.'s participants is not clear, the musical training of our participants is likely more extensive than a group of people randomly chosen from general populations. Such relatively greater musical training may have influenced the production of higher and more stable pitches in singing. In fact, we confirmed that there is no obvious difference in pitch stability of speech between ours and Hilton et al.'s dataset (2022), 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 recordings due to noise, our singing showcased the use of more heightened pitch (Fig. S17).

Interestingly, we also observed that our spoken recordings have slightly lower pitch height than Hilton et al.'s spoken recordings. Possible factors that may underlie this difference include age (Berg et al., 2017), body size (Pisanski, 2014), and possibly avoiding using low frequencies not to intimidate accompanied infants (Puts et al., 2006). 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 instructions to describe "a topic of their choice (for example…their daily routine)", and such task differences can also affect speaking pitch (Barsties, 2013). On the other hand, this result is unlikely to be due to the exposure of Western styles to participants, since the subset of Hilton's data including only English, Mandarin, Polish, Spanish, and Kannada speakers show 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, Albouy et al. (2023) also reanalysed Hilton et al.'s (2022) recordings using different but related methods that also emphasize pitch stability and temporal rate ("spectro-temporal modulations"). Albouy et al. transformed audio recordings to extract two-dimensional density features

(spectro-temporal modulations where one axis is temporal modulations [Hz] and the other is spectral modulations [cyc/kHz]) to characterize song and speech acoustically. Their finding is similar to our results that speech has higher density in the temporal modulation range of 5-10 Hz, which matches the syllable rate and amplitude modulation rate of speech investigated cross-culturally (Ding et al., 2017; Pellegrino et al., 2011; Poeppel & Assaneo, 2020), 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 2-5 cyc/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 listeners rely on spectral and temporal modulation information to judge whether the uttered vocalization is song or speech, which suggests spectro-temporal modulation is an acoustic cue differentiating song and speech. Although they have not reported other features such as pitch height, the convergence of our study and their study identifying the same features implies that temporal rate and pitch stability are robust features distinguishing song and speech across cultures.

5.5. Evolutionary and functional mechanisms

“Discrete pitches or regular rhythmic patterns” are often considered defining features of music that distinguish it from speech (cf. Fitch, 2006; and Savage et al. 2015 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. 5). 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 chain may be effective in capturing causal mechanisms underlying the manipulation of these parameters depending on communicative goals (e.g., Ma et al., 2019; Ozaki et al., 2023).

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 on 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. In fact, the temporal rate is the only feature showing almost no difference between singing and the instrumental (cf. 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 (Savage et al., 2021). 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 (cf. Lomax, 1968) 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 (Alipour & Scherer, 2007). 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 1k Hz (Suzuki & Takeshima, 2004), but generally, the frequency range of f_0s of human voice is below 1k Hz, so it is reasonable to heighten pitches to exploit higher loudness sensitivity, which may be helpful for creating bonding through acoustic communication extensively utilizing pitch control.

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 phrase, greater sound pressure, and less noisy sounds may ease the intelligibility of pitch information. On the contrary, similar timbral brightness, pitch interval size, and pitch range between song and speech may be due to motor and mechanistic constraints, like the difficulty of rapid transitioning to distanced pitch 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.

5.6. Inclusivity and global collaboration

Our use of a new "participants-as-coauthors" paradigm allowed us to discover new findings that would 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 surprisingly large differences for features such as temporal rate when analysed using their manual segmentations that we would have underestimated if we relied on automated segmentation (cf. combined effect size of translated Cohen's $d>1.5$ in Fig. 5 vs. $d<0.4$ in Fig. S15). This highlights that equitable collaboration is not merely an issue of social justice but also of scientific quality (Nature Editors, 2022; Urassa et al., 2021).

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 under-represented languages from our originally planned sample. Such under-represented 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 (e.g., Nicas, 2023). Of our coauthors representing Indigenous and under-represented 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. Indeed, 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, Hebrew; Troy & Barwick, 2020; Savage et al., 2015). 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 data. Many of these limitations have been addressed through our robustness analyses and converging results from our own and Albouy et al.'s (2023) reanalyses of Hilton et al.'s independent speech/song dataset described above. 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, 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.

6. Conclusion

Overall, our Registered Report comparing music and speech from our coauthors speaking diverse languages shows strong evidence for cross-cultural regularities in music and language amidst substantial global diversity. The features that we identified as differentiating music and speech along a "musilinguistic 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 (Fitch, 2006; Patel, 2008; Savage et al., 2021). 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 (Tierney et al., 2011; Trehub, 2015; Ozaki et al., 2023; Singh & Mehr, 2023). 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.

Data/code availability:

Analysis code: https://github.com/comp-music-lab/song-speech-analysis
Data*:* https://osf.io/mzxc8/

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 non-coauthor Ole Manyas are included as part of a separate ethical approval by the Kenyan National Commision for Scientice, Technology & Innovation to Parselelo (NACOSTI/P/23/24284).

Author contributions:

-Conceived the project: Savage, Ozaki, Tierney, Pfordresher, Benetos, McBride, Proutskova, Liu, Purdy, Opondo, Jacoby, Fitch
-Funding acquisition: Savage, Ozaki, Purdy, Benetos, Jacoby, Opondo, Fitch, Thorne, Pfordresher, Liu, Rocamora
-Project management: Savage, Ozaki
-Recruitment: Savage, Ozaki, Jacoby, Opondo, Pfordresher, Fitch, Barbosa
-Translation: Barbosa, Savage, Ozaki
-Audio recordings for pilot analyses: Ozaki, Hadavi, Nweke, P. Sadaphal, McBride
-Annotations for pilot analyses: Ozaki, Hadavi, Nweke, D. Sadaphal, Savage
-Recording and text transcription/segmentation of own singing/speaking/instrumental performance: all authors
-Detailed (millisecond-level) onset annotations: Ozaki (all data), Savage (inter-rater reliability subset)
-Checking/correcting onset annotations for own singing/speaking/instrumental performance: all authors
-Conducted analyses: Ozaki
-Made Fig. 3 word clouds: Gomez
-Drafting initial manuscript: Ozaki, Savage
-Editing manuscript: many (but not all) authors

Inclusivity statement:

We endeavored to follow best practices in cross-cultural collaborative research (Tan & Ostashewski, 2022; Savage, Jacoby, Margulis, et al., 2023), 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 Appendix 2). Each recording set analyzed comes from a named coauthor who speaks that language as their 1st or heritage language.

Conflict of interest disclosure:

The authors of this article declare that they have no financial conflict of interest with the content of this article. Patrick Savage is a Recommender at PCI Registered Reports.

Acknowledgments:

We thank Chris Chambers, Bob Slevc, and Nai Ding for helpful reviews, and Tomoko Tanaka for serving as the Research Assistant to securely monitor and check audio recordings. We thank Joseph Bulbulia for assistance in securing funding, and thank Ozaki's PhD committee members Akira Wakita and Nao Tokui and students from the Keio University CompMusic and NeuroMusic Labs for feedback on earlier drafts of the manuscript. We thank Aritz Irurtzun, Joel Maripil, Aeles Lrawbalrate, Morzaniel Iramari Aranariutheri, Tumi Uisu Paulo Matis, and Samira 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 Government, administered by the Royal Society Te Apārangi (Rutherford Discovery Fellowship 22-UOA-040 to Savage and Marsden Fast-Start Grant 22-UOA-052 to Savage, Purdy, Opondo, Jacoby, Benetos, and Fitch), and by the Japanese Government, administered by the Japan Society for the Promotion of Science (KAKENHI Grant-in-Aid #19KK0064 to Savage, Fujii, and Jacoby) and the Japan Science and Technology Agency (Grant Number JPMJSP2123 of Support for Pioneering Research Initiated by the Next Generation to Ozaki).

References:

Adams, C. R. (1976). Melodic Contour Typology. *Ethnomusicology*, 20(2), 179-215. doi:10.2307/851015

Albouy, P., Benjamin, L., Morillon, B., & Zatorre, R. J. (2020). Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science*, *367*(6481), 1043–1047. https://doi.org/10.1126/science.aaz3468

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

Alipour, F., & Scherer, R. C. (2007). On pressure-frequency relations in the excised larynx. *The Journal of the Acoustical Society of America*, *122*(4), 2296–2305. https://doi.org/10.1121/1.2772230

Anikin, A. (2020). The link between auditory salience and emotion intensity. *Cognition and Emotion*, *34*(6), 1246–1259. https://doi.org/10.1080/02699931.2020.1736992

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, *96*, 104159. https://doi.org/10.1016/j.jesp.2021.104159

Barbosa, P. A., Arantes, P., Meireles, A. R., & Vieira, J. M. (2005). Abstractness in speech-metronome synchronisation: P-centres as cyclic attractors. *Interspeech 2005*, 1441–1444. https://doi.org/10.21437/Interspeech.2005-512

Barnes, J. J., Davis, P., Oates, J., & Chapman, J. (2004). The relationship between professional operatic soprano voice and high range spectral energy. *The Journal of the Acoustical Society of America*, *116*(1), 530–538. https://doi.org/10.1121/1.1710505

Barsties, B. (2013). Einfluss verschiedener Methoden zur Bestimmung der mittleren Sprechstimmlage. *HNO*, *61*(7), 609–616. https://doi.org/10.1007/s00106-012-2665-0

Bârzan, H., Moca, V. V., Ichim, A.-M., & Muresan, R. C. (2021). Fractional Superlets. *2020 28th European Signal Processing Conference (EUSIPCO)*, 2220–2224. https://doi.org/10.23919/Eusipco47968.2020.9287873

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Berg, M., Fuchs, M., Wirkner, K., Loeffler, M., Engel, C., & Berger, T. (2017). The Speaking Voice in the General Population: Normative Data and Associations to Sociodemographic and Lifestyle Factors. *Journal of Voice*, *31*(2), 257.e13-257.e24. https://doi.org/10.1016/j.jvoice.2016.06.001

Bickel, B. (2011). Absolute and statistical universals. In *The Cambridge Encyclopedia of the Language Sciences*, P. C. Hogan, ed. (Cambridge University Press), pp. 77–79.

Blacking, J. (1973). *How musical is man?* University of Washington Press.

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, https://doi.org/10.1016/j.tics.2022.09.015

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370. https://doi.org/10.1007/BF02294361

Brockwell, S. E., & Gordon, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine*, *20*(6), 825–840. https://doi.org/10.1002/sim.650

Brown, D. E. (1991). *Human Universals*. New York: McGraw-Hill.

Brown, S. (2000). The Musilanguage Model of Music Evolution. In S. Brown, B. Merker, & C. Wallin (Eds.), *The Origins of Music* (pp. 271–300). The MIT Press.

Brown, S., & Jordania, J. (2013). Universals in the world's musics. *Psychology of Music*, *41*(2), 229–248. https://doi.org/10.1177/0305735611425896

Brown, S., Savage, P. E., Ko, A. M.-S., Stoneking, M., Ko, Y.-C., Loo, J.-H., & Trejaut, J. A. (2014). Correlations in the population structure of music, genes and language. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1774), 20132072. https://doi.org/10.1098/rspb.2013.2072

Brunner E., Bathke A. C., & Konietschke F. (2018). *Rank and pseudo-rank procedures for independent observations in factorial designs: Using R and SAS*. Springer. https://ci.nii.ac.jp/ncid/BB28708839

Bryant, G. A. (2021). The Evolution of Human Vocal Emotion. *Emotion Review*, *13*(1), 25–33. https://doi.org/10.1177/1754073920930791

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), 16. https://doi.org/10.5334/joc.72

Cannam, C., Landone, C., & Sandler, M. (2010). Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. *Proceedings of the 18th ACM International Conference on Multimedia*, 1467–1468. https://doi.org/10.1145/1873951.1874248

Chacón, J. E. (2020). The Modal Age of Statistics. *International Statistical Review*, *88*(1), 122–141. https://doi.org/10.1111/insr.12340

Chang, A., Teng, X., Assaneo, F., & Poeppel, D. (2022). *Amplitude modulation perceptually distinguishes music and speech*. *PsyArXiv* preprint: https://doi.org/10.31234/osf.io/juzrh

Chazal, F., Fasy, B., Lecci, F., Bertr, Michel, Aless, Rinaldo, R., & Wasserman, L. (2018). Robust Topological Inference: Distance To a Measure and Kernel Distance. *Journal of Machine Learning Research*, *18*(159), 1–40.

Chaudhuri, P., & Marron, J. S. (1999). SiZer for Exploration of Structures in Curves. *Journal of the American Statistical Association*, *94*(447), 807–823. https://doi.org/10.1080/01621459.1999.10474186

Chen, S., Han, C., Wang, S., Liu, X., Wang, B., Wei, R., & Lei, X. (2022). Hearing the physical condition: The relationship between sexually dimorphic vocal traits and underlying physiology. *Frontiers in Psychology*, *13*. https://www.frontiersin.org/articles/10.3389/fpsyg.2022.983688

Chen, Y.-C., Genovese, C. R., Ho, S., & Wasserman, L. (2015). Optimal Ridge Detection using Coverage Risk. *Advances in Neural Information Processing Systems*, *28*. https://papers.nips.cc/paper/2015/hash/0aa1883c6411f7873cb83dacb17b0afc-Abstract.html

Chen, Y.-C., Genovese, C. R., & Wasserman, L. (2016). A comprehensive approach to mode clustering. *Electronic Journal of Statistics*, *10*(1), 210–241. https://doi.org/10.1214/15-EJS1102

Cheney, D. L., & Seyfarth, R. M. (2018). Flexible usage and social function in primate vocalizations. *Proceedings of the National Academy of Sciences*, *115*(9), 1974–1979. https://doi.org/10.1073/pnas.1717572115

Chow, I., Belyk, M., Tran, V., & Brown, S. (2015). Syllable synchronization and the P-center in Cantonese. *Journal of Phonetics*, *49*, 55–66. https://doi.org/10.1016/j.wocn.2014.10.006

Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(5), 603–619. https://doi.org/10.1109/34.1000236

Cooper, A. M., Whalen, D. H., & Fowler, C. A. (1986). P-centers are unaffected by phonetic categorization. *Perception & Psychophysics*, *39*(3), 187–196. https://doi.org/10.3758/BF03212490

Cox, C., Bergmann, C., Fowler, E., Keren-Portnoy, T., Roepstorff, A., Bryant, G., & Fusaroli, R. (2022). A systematic review and Bayesian meta-analysis of the acoustic features of infant-directed speech. *Nature Human Behaviour*, 1–20. https://doi.org/10.1038/s41562-022-01452-1

Cychosz, M., Cristia, A., Bergelson, E., Casillas, M., Baudet, G., Warlaumont, A. S., Scaff, C., Yankowitz, L., & Seidl, A. (2021). Vocal development in a large-scale crosslinguistic corpus. *Developmental Science*, *24*(5), e13090. https://doi.org/10.1111/desc.13090

Danielsen, A., Nymoen, K., Anderson, E., Câmara, G. S., Langerød, M. T., Thompson, M. R., & London, J. (2019). Where is the beat in that note? Effects of attack, duration, and frequency on the perceived timing of musical and quasi-musical sounds. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(3), 402–418. https://doi.org/10.1037/xhp0000611

Darwin, C. (1871). *The descent of man*. Watts & Co.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (2000). The role of computational models in neuropsychological investigations of language: Reply to Ruml and Caramazza (2000). *Psychological Review*, *107*, 635–645. https://doi.org/10.1037/0033-295X.107.3.635

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, *81*, 181–187. https://doi.org/10.1016/j.neubiorev.2017.02.011

Djurović, I., & Stanković, Lj. (2004). An algorithm for the Wigner distribution based instantaneous frequency estimation in a high noise environment. *Signal Processing*, *84*(3), 631–643. https://doi.org/10.1016/j.sigpro.2003.12.006

Doelling, K. B., Assaneo, M. F., Bevilacqua, D., Pesaran, B., & Poeppel, D. (2019). An oscillator model better predicts cortical entrainment to music. *Proceedings of the National Academy of Sciences*, *116*(20), 10113–10121. https://doi.org/10.1073/pnas.1816414116

Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, Accessed on 2022-10-03.)

Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, *473*(7345), 79–82. https://doi.org/10.1038/nature09923

Durojaye, C., Fink, L., Roeske, T., Wald-Fuhrmann, M., & Larrouy-Maestri, P. (2021). Perception of Nigerian Dùndún Talking Drum Performances as Speech-Like vs. Music-Like: The Role of Familiarity and Acoustic Cues. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.652673

Evans, N., and Levinson, S.C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behav. Brain Sci.* 32, 429–492.

Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., & Singh, A. (2014). Confidence sets for persistence diagrams. *The Annals of Statistics*, *42*(6), 2301–2339. https://doi.org/10.1214/14-AOS1252

Feinberg, D. R., Jones, B. C., & Armstrong, M. M. (2018). Sensory Exploitation, Sexual Dimorphism, and Human Voice Pitch. *Trends in Ecology & Evolution*, *33*(12), 901–903. https://doi.org/10.1016/j.tree.2018.09.007

Fitch, W. T. (2006). The biology and evolution of music: A comparative perspective. *Cognition*, *100*(1), 173–215. https://doi.org/10.1016/j.cognition.2005.11.009

Fraser, D. (1989). Interpolation by the FFT revisited-an experimental investigation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *37*(5), 665–675. https://doi.org/10.1109/29.17559

Freeberg, T. M., Dunbar, R. I. M., & Ord, T. J. (2012). Social complexity as a proximate and ultimate factor in communicative complexity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1597), 1785–1801. https://doi.org/10.1098/rstb.2011.0213

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., & Wasserman, L. (2014). Nonparametric ridge estimation. *The Annals of Statistics*, *42*(4), 1511–1545. https://doi.org/10.1214/14-AOS1218

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., & Wasserman, L. (2016). Non-parametric inference for density modes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *78*(1), 99–126.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the Rhythm Class Hypothesis. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 515–546). De Gruyter Mouton. https://doi.org/10.1515/9783110197105.2.515

Haiduk, F., Quigley, C., & Fitch, W. T. (2020). Song Is More Memorable Than Speech Prosody: Discrete Pitches Aid Auditory Working Memory. *Frontiers in Psychology*, *11*. https://www.frontiersin.org/article/10.3389/fpsyg.2020.586723

Haiduk, F., & Fitch, W. T. (2022). Understanding Design Features of Music and Language: The Choric/Dialogic Distinction. *Frontiers in Psychology*, *13*. https://www.frontiersin.org/article/10.3389/fpsyg.2022.786899

Hall, P., Sheather, S. J., Jones, M. C., & Marron, J. S. (1991). On Optimal Data-Based Bandwidth Selection in Kernel Density Estimation. *Biometrika*, *78*(2), 263–269. https://doi.org/10.2307/2337251

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2022). Glottolog 4.7. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.7398962 (Available online at http://glottolog.org, Accessed on 2023-05-14.)

Han, S. er, Sundararajan, J., Bowling, D. L., Lake, J., & Purves, D. (2011). Co-Variation of Tonality in the Music and Speech of Different Cultures. *PLOS ONE*, *6*(5), e20160. https://doi.org/10.1371/journal.pone.0020160

Hansen, J. H. L., Bokshi, M., & Khorram, S. (2020). Speech variability: A cross-language study on acoustic variations of speaking versus untrained singing. *The Journal of the Acoustical Society of America*, *148*(2), 829. https://doi.org/10.1121/10.0001526

Hilton, C. B., Moser, C. J., Bertolo, M., Lee-Rubin, H., Amir, D., Bainbridge, C. M., Simson, J., Knox, D., Glowacki, L., Alemu, E., Galbarczyk, A., Jasienska, G., Ross, C. T., Neff, M. B., Martin, A., Cirelli, L. K., Trehub, S. E., Song, J., Kim, M., … Mehr, S. A. (2022). Acoustic regularities in infant-directed speech and song across cultures. *Nature Human Behaviour*, 1–12. https://doi.org/10.1038/s41562-022-01410-x

Hoeschele, M., & Fitch, W. T. (2022). Cultural evolution: Conserved patterns of melodic evolution across musical cultures. *Current Biology*, *32*(6), R265–R267. https://doi.org/10.1016/j.cub.2022.01.080

Horn, M., & Dunnett, C. W. (2004). Power and sample size comparisons of stepwise FWE and FDR controlling test procedures in the normal many-one case. *Recent Developments in Multiple Comparison Procedures*, *47*, 48–65. https://doi.org/10.1214/lnms/1196285625

Howell, P. (1988). Prediction of P-center location from the distribution of energy in the amplitude envelope: I. *Perception & Psychophysics*, *43*(1), 90–93. https://doi.org/10.3758/BF03208978

Jackson, D., & Turner, R. (2017). Power analysis for random-effects meta-analysis. *Research Synthesis Methods*, *8*(3), 290–302. https://doi.org/10.1002/jrsm.1240

Jacoby, N., Margulis, E.H., Clayton, M., Hannon, E., Honing, H., Iversen, J., Klein, T.R., Mehr, S.A., Pearson, L., Peretz, I., Savage, P. E., et al. (2020). Cross-cultural work in music cognition: Methodologies, pitfalls, and practices. *Music Percept.* 37, 185–195.

Johnston, J. D. (1988). Transform coding of audio signals using perceptual noise criteria. *IEEE Journal on Selected Areas in Communications*, *6*(2), 314–323. https://doi.org/10.1109/49.608

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390. https://doi.org/10.3758/BRM.41.2.385

Jung, S.-H. (2005). Sample size for FDR-control in microarray data analysis. *Bioinformatics*, *21*(14), 3097–3104. https://doi.org/10.1093/bioinformatics/bti456

Ladd, D. R. (1984). Declination: A review and some hypotheses. *Phonology*, 1, 53-74.

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lartillot, O., Eerola, T., Toiviainen, P., & Fornari, J. (2008). Multi-feature modeling of pulse clarity: Design, validation and optimization. *Proc. of the 9th Int. Society for Music Information Retrieval Conf.*, 521–526.

Lartillot, O., Toiviainen, P., & Eerola, T. (2008). A Matlab Toolbox for Music Information Retrieval. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data Analysis, Machine Learning and Applications* (pp. 261–268). Springer. https://doi.org/10.1007/978-3-540-78246-9_31

Leongómez, J. D., Havlíček, J., & Roberts, S. C. (2022). Musicality in human vocal communication: An evolutionary perspective. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1841), 20200391. https://doi.org/10.1098/rstb.2020.0391

Lindblom, B., & Sundberg, J. (2007). The Human Voice in Speech and Singing. In T. D. Rossing (Ed.), *Springer Handbook of Acoustics* (pp. 669–712). Springer. https://doi.org/10.1007/978-0-387-30425-0_16

Ling, L. E., Grabe, E., & Nolan, F. (2000). Quantitative Characterizations of Speech Rhythm: Syllable-Timing in Singapore English. *Language and Speech*, *43*(4), 377–401. https://doi.org/10.1177/00238309000430040301

List, G. (1971). On the Non-Universality of Musical Perspectives. *Ethnomusicology*, *15*(3), 399–402.

Liu, S., Tian, L., Lee, S., & Xie, M. (2018). Exact inference on meta-analysis with generalized fixed-effects and random-effects models. *Biostatistics & Epidemiology*, *2*(1), 1–22. https://doi.org/10.1080/24709360.2017.1400714

Lomax, A., & Grauer, V. (1968). The Cantometric coding book. In A. Lomax (Ed.), *Folk song style and culture* (pp. 34–74). American Association for the Advancement of Science.

Ma, W., Fiveash, A., & Thompson, W. F. (2019). Spontaneous emergence of language-like and music-like vocalizations from an artificial protolanguage. *Semiotica*, *2019*(229), 1–23. https://doi.org/10.1515/sem-2018-0139

Matsumae, H., Ranacher, P., Savage, P. E., Blasi, D. E., Currie, T. E., Koganebuchi, K., Nishida, N., Sato, T., Tanabe, H., Tajima, A., Brown, S., Stoneking, M., Shimizu, K. K., Oota, H., & Bickel, B. (2021). Exploring correlations in genetic and cultural variation across language families in northeast Asia. *Science Advances*. https://doi.org/10.1126/sciadv.abd9223

Mauch, M., & Dixon, S. (2014). PYIN: A fundamental frequency estimator using probabilistic threshold distributions. *2014 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP), 659–663. https://doi.org/10.1109/ICASSP.2014.6853678

Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A. A., Hopkins, E. J., Howard, R. M., Hartshorne, J. K., Jennings, M. V., Simson, J., Bainbridge, C. M., Pinker, S., O'Donnell, T. J., Krasnow, M. M., & Glowacki, L. (2019). Universality and diversity in human song. *Science*, *366*(6468), eaax0868. https://doi.org/10.1126/science.aax0868

Mehr, S. A., Krasnow, M. M., Bryant, G. A., & Hagen, E. H. (2021). Origins of music in credible signaling. *Behavioral and Brain Sciences*, *44*. https://doi.org/10.1017/S0140525X20000345

Merrill, J., & Larrouy-Maestri, P. (2017). Vocal Features of Song and Speech: Insights from Schoenberg's Pierrot Lunaire. *Frontiers in Psychology*, *8*, 1108. https://doi.org/10.3389/fpsyg.2017.01108

Mertens, P. (2022). The Prosogram model for pitch stylization and its applications in intonation transcription. In J. Barnes & S. Shattuck-Hufnagel (Eds.), *Prosodic Theory and Practice* (pp. 259–286). MIT Press. https://mitpress.mit.edu/9780262543170/prosodic-theory-and-practice/

Moca, V. V., Bârzan, H., Nagy-Dăbâcan, A., & Mureșan, R. C. (2021). Time-frequency super-resolution with superlets. *Nature Communications*, *12*(1), 337. https://doi.org/10.1038/s41467-020-20539-9

Morrill, T. H., McAuley, J. D., Dilley, L. C., & Hambrick, D. Z. (2015). Individual differences in the perception of melodic contours and pitch-accent timing in speech: Support for domain-generality of pitch processing. *Journal of Experimental Psychology: General*, *144*, 730–736. https://doi.org/10.1037/xge0000081

Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centers (P-centers). *Psychological Review*, *83*(5), 405–408. https://doi.org/10.1037/0033-295X.83.5.405

Müller, M., Rosenzweig, S., Driedger, J., & Scherbaum, F. (2017, June 13). *Interactive Fundamental Frequency Estimation with Applications to Ethnomusicological Research*. Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio. https://www.aes.org/e-lib/browse.cfm?elib=18777

Natke, U., Donath, T. M., & Kalveram, K. Th. (2003). Control of voice fundamental frequency in speaking versus singing. *Journal of the Acoustical Society of America*, *113*(3), 1587–1593. https://doi.org/10.1121/1.1543928

Nature Editors. (2022). Nature addresses helicopter research and ethics dumping. *Nature*, *606*(7912), 7–7. https://doi.org/10.1038/d41586-022-01423-6

Nicas, J. (2023, March 25). The Amazon's Largest Isolated Tribe Is Dying. *The New York Times*. https://www.nytimes.com/2023/03/25/world/americas/brazil-amazon-indigenous-tribe.html

Nikolsky, A., Alekseyev, E., Alekseev, I., & Dyakonova, V. (2020). The Overlooked Tradition of "Personal Music" and Its Place in the Evolution of Music. *Frontiers in Psychology*, *10*. https://www.frontiersin.org/article/10.3389/fpsyg.2019.03051

Nordhoff, S., & Hammarstrom, H. (2011). Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources. *Proceedings of ISWC 2011*, 1–6.

Novitski, N., Tervaniemi, M., Huotilainen, M., & Näätänen, R. (2004). Frequency discrimination at different frequency levels as indexed by electrophysiological and

behavioral measures. *Cognitive Brain Research*, *20*(1), 26–36. https://doi.org/10.1016/j.cogbrainres.2003.12.011

Ozaki, Y., Sato, S., McBride, J.M., Pfordresher, P.Q., Tierney, A.T., Six, J., Fujii, S., and Savage, P.E. (2022). Automatic acoustic analyses quantify pitch discreteness within and between human music, speech, and bird song. *Proc. 10th Int. Folk Music Anal. Work.*

Ozaki, Y., Kuroyanagi, J., Chiba, G., McBride, J., Proutskova, P., Tierney, A. T., Pfordresher, P. Q., Benetos, E., Liu, F., & Savage, P. E. (2022). Similarities and differences in a cross-linguistic sample of song and speech recordings. *Proceedings of the 2022 Joint Conference on Language Evolution*, 569–572.

Ozaki, Y., de Heer Kloots, M., Ravignani, A., & Savage, P. E. (2023) Cultural evolution of music and language. In D. Sammler (Ed.), *Oxford Handbook of Language and Music*. Oxford University Press. Preprint: https://doi.org/10.31234/osf.io/s7apx

Passmore, S., Wood, A. L. C., Barbieri, C., Shilton, D., Daikoku, H., Atkinson, Q. D., & Savage, P. E. (Under review). Independent histories underlie global musical, linguistic, and genetic diversity.

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

Patel, A. D. (2011). Language, music, and the brain: A resource-sharing framework. In P. Rebuschat, M. Rohmeier, J. A. Hawkins, & I. Cross (Eds.), *Language and Music as Cognitive Systems* (p. 204-223). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199553426.003.0022

Patel, A. D. (2018). Music as a transformative technology of the mind: An update. In H. Honing (Ed.), *The origins of musicality* (pp. 113–126). MIT Press.

Patel, A. D., & Daniele, J. R. (2003). An empirical comparison of rhythm in language and music. *Cognition*, 87(1), 35–45.

Patel, A. D., Iversen, J. R., & Rosenberg, J. C. (2006). Comparing the rhythm and melody of speech and music: The case of British English and French. *The Journal of the Acoustical Society of America*, *119*(5), 3034–3047. https://doi.org/10.1121/1.2179657

Patel, A. D., & Rueden, C. von. (2021). Where they sing solo: Accounting for cross-cultural variation in collective music-making in theories of music evolution. *Behavioral and Brain Sciences*, *44*, e85. https://doi.org/10.1017/S0140525X20001089

Peeters, G. (2004). *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).

Pellegrino, F., Coupé, C., & Marsico, E. (2011). A Cross-Language Perspective on Speech Information Rate. *Language*, *87*(3), 539–558.

Pereira, J. P. B., Stroes, E. S. G., Zwinderman, A. H., & Levin, E. (2022). Covered Information Disentanglement: Model Transparency via Unbiased Permutation Importance. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(7), Article 7. https://doi.org/10.1609/aaai.v36i7.20769

Peretz, I. (2009). Music, Language and Modularity Framed in Action. *Psychologica Belgica*, *49*(2–3), Article 2–3. https://doi.org/10.5334/pb-49-2-3-157

Pemutatoin Importance (n.d.). ELI5, Retrieved May 10, 2023, from https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html

Pfordresher, P. Q., Brown, S., Meier, K. M., Belyk, M., & Liotti, M. (2010). Imprecise singing is widespread. *The Journal of the Acoustical Society of America*, *128*(4), 2182–2190. https://doi.org/10.1121/1.3478782

Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J. M., Röder, S., Andrews, P. W., Fink, B., DeBruine, L. M., Jones, B. C., & Feinberg, D. R. (2014). Vocal indicators of body size in men and women: A meta-analysis. *Animal Behaviour*, *95*, 89–99. https://doi.org/10.1016/j.anbehav.2014.06.011

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Reviews Neuroscience*, *21*(6), 322–334. https://doi.org/10.1038/s41583-020-0304-4

Pompino-Marschall, B. (1989). On the psychoacoustic nature of the P-center phenomenon. *Journal of Phonetics*, *17*(3), 175–192. https://doi.org/10.1016/S0095-4470(19)30428-0

Pounds, S., & Cheng, C. (2005). Sample size determination for the false discovery rate. *Bioinformatics*, *21*(23), 4263–4271. https://doi.org/10.1093/bioinformatics/bti699

Puts, D. A., Gaulin, S. J. C., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior*, *27*(4), 283–296. https://doi.org/10.1016/j.evolhumbehav.2005.11.003

Puts, D. A., Hill, A. K., Bailey, D. H., Walker, R. S., Rendall, D., Wheatley, J. R., Welling, L. L. M., Dawood, K., Cárdenas, R., Burriss, R. P., Jablonski, N. G., Shriver, M. D., Weiss, D., Lameira, A. R., Apicella, C. L., Owren, M. J., Barelli, C., Glenn, M. E., & Ramos-Fernandez, G. (2016). Sexual selection on male vocal fundamental frequency in humans and other anthropoids. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1829), 20152830. https://doi.org/10.1098/rspb.2015.2830

Raposo de Medeiros, B., Cabral, J. P., Meireles, A. R., & Baceti, A. A. (2021). A comparative study of fundamental frequency stability between speech and singing. *Speech Communication*, *128*, 15–23. https://doi.org/10.1016/j.specom.2021.02.003

Robledo, J. P., Hurtado, E., Prado, F., Román, D., & Cornejo, C. (2016). Music intervals in speech: Psychological disposition modulates ratio precision among interlocutors' nonlocal f0 production in real-time dyadic conversation. *Psychology of Music*, *44*(6), 1404–1418. https://doi.org/10.1177/0305735616634452

Roeske, T. C., Tchernichovski, O., Poeppel, D., & Jacoby, N. (2020). Categorical Rhythms Are Shared between Songbirds and Humans. *Current Biology*, 30(18), 3544-3555.e6. https://doi.org/10.1016/j.cub.2020.06.072

Rogalsky, C., Rong, F., Saberi, K., & Hickok, G. (2011). Functional Anatomy of Language and Music Perception: Temporal and Structural Factors Investigated Using Functional Magnetic Resonance Imaging. *Journal of Neuroscience*, *31*(10), 3843–3852. https://doi.org/10.1523/JNEUROSCI.4515-10.2011

Romano, J. P. (2005). Optimal testing of equivalence hypotheses. *The Annals of Statistics*, *33*(3), 1036–1047. https://doi.org/10.1214/009053605000000048

Rosenzweig, S., Scherbaum, F., Shugliashvili, D., Arifi-Müller, V., & Müller, M. (2020). Erkomaishvili Dataset: A Curated Corpus of Traditional Georgian Vocal Music for Computational Musicology. *Transactions of the International Society for Music Information Retrieval*, *3*(1), 31–41. https://doi.org/10.5334/tismir.44

Ross, D., Choi, J., & Purves, D. (2007). Musical intervals in speech. *Proceedings of the National Academy of Sciences*, *104*(23), 9852–9857. https://doi.org/10.1073/pnas.0703140104

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, *13*(1), 19–30. https://doi.org/10.1037/1082-989X.13.1.19

Sammler, D., Ed. (Under contract). *Oxford Handbook of Language and Music*. Oxford University Press.

Savage, P. E. (2019). Universals. In J. L. Sturman (Ed.), *The SAGE International Encyclopedia of Music and Culture* (p. 2282–2285). Thousand Oaks: SAGE Publications. http://doi.org/10.4135/9781483317731.n759

Savage, P. E., Brown, S., Sakai, E., & Currie, T. E. (2015). Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences*, *112*(29), 8987–8992. https://doi.org/10.1073/pnas.1414495112

Savage, P. E., Loui, P., Tarr, B., Schachner, A., Glowacki, L., Mithen, S., & Fitch, W. T. (2021). Music as a coevolved system for social bonding. *Behavioral and Brain Sciences*, *44*. https://doi.org/10.1017/S0140525X20000333

Savage, P. E., Loui, P., Tarr, B., Schachner, A., Glowacki, L., Mithen, S., & Fitch, W. T. (2021b). Authors' response: Toward inclusive theories of the evolution of musicality. *Behavioral and Brain Sciences*, 44(e121), 132–140. https://doi.org/10.1017/S0140525X21000042

Savage, P.E., Jacoby, N., Margulis, E.H., Daikoku, H., Anglada-Tort, M., Castelo-Branco, S.E.-S., Nweke, F.E., Fujii, S., Hegde, S., Chuan-Peng, H., Opondo, P., et al. (2023). Building sustainable global collaborative networks: Recommendations from music studies and the social sciences. In E. H. Margulis, D. Loughridge, and P. Loui (Eds.), *The science-music borderlands: Reckoning with the past, imagining the future* (347-365). MIT Press. https://doi.org/10.7551/mitpress/14186.003.0032

Savage, P. E., Matsumae, H., Oota, H., Stoneking, M., Currie, T. E., Tajima, A., Gillan, M., & Brown, S. (2015). How "circumpolar" is Ainu music? Musical and genetic perspectives on the history of the Japanese archipelago. Ethnomusicology Forum, 24(3), 443–467. https://doi.org/10.1080/17411912.2015.1084236

Schafer, R. W., & Rabiner, L. R. (1973). A digital signal processing approach to interpolation. *Proceedings of the IEEE*, *61*(6), 692–702. https://doi.org/10.1109/PROC.1973.9150

Schamberg, I., Wittig, R. M., & Crockford, C. (2018). Call type signals caller goal: A new take on ultimate and proximate influences in vocal production. *Biological Reviews*, *93*(4), 2071–2082. https://doi.org/10.1111/brv.12437

Schwartz, D. A., Howe, C. Q., & Purves, D. (2003). The Statistical Structure of Human Speech Sounds Predicts Musical Universals. *Journal of Neuroscience*, *23*(18), 7160–7168. https://doi.org/10.1523/JNEUROSCI.23-18-07160.2003

Scott, S. K. (1998). The point of P-centres. *Psychological Research*, *61*(1), 4–11. https://doi.org/10.1007/PL00008162

Sera, F., Armstrong, B., Blangiardo, M., & Gasparrini, A. (2019). An extended mixed-effects framework for meta-analysis. *Statistics in Medicine*, *38*(29), 5429–5444. https://doi.org/10.1002/sim.8362

Shao, X., & Ma, C. (2003). A general approach to derivative calculation using wavelet transform. *Chemometrics and Intelligent Laboratory Systems*, *69*(1), 157–165. https://doi.org/10.1016/j.chemolab.2003.08.001

Sharma, B., Gao, X., Vijayan, K., Tian, X., & Li, H. (2021). NHSS: A speech and singing parallel database. *Speech Communication*, *133*, 9–22. https://doi.org/10.1016/j.specom.2021.07.002

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall. https://doi.org/10.1201/9781315140919

Singh, M., & Mehr, S. A. (2023). Universality, domain-specificity and development of psychological responses to music. *Nature Reviews Psychology*, 1–14. https://doi.org/10.1038/s44159-023-00182-z

Slifka, J. (2006). Respiratory system pressures at the start of an utterance. In P. Divenyi, S. Greenberg, & G. Meyer (Eds.), *Dynamics of speech production and perception* (pp. 45-57). Amsterdam: IOS Press.

Sommerfeld, M., Heo, G., Kim, P., Rush, S. T., & Marron, J. S. (2017). Bump hunting by topological data analysis. *Stat*, *6*(1), 462–471. https://doi.org/10.1002/sta4.167

Stegemöller, E. L., Skoe, E., Nicol, T., Warrier, C. M., & Kraus, N. (2008). Music Training and Vocal Production of Speech and Song. *Music Perception*, *25*(5), 419–428. https://doi.org/10.1525/mp.2008.25.5.419

Stone, R. E. (Ed), Cleveland, T. F., & Sundberg, J. (1999). Formant frequencies in country singers' speech and singing. *Journal of Voice*, *13*(2), 161–167. https://doi.org/10.1016/S0892-1997(99)80020-4

Sundberg, J. (2001). Level and Center Frequency of the Singer's Formant. *Journal of Voice*, *15*(2), 176–186. https://doi.org/10.1016/S0892-1997(01)00019-4

Suzuki, Y., & Takeshima, H. (2004). Equal-loudness-level contours for pure tones. *The Journal of the Acoustical Society of America*, *116*(2), 918–933. https://doi.org/10.1121/1.1763601

Tan, S. B., & Ostashewski, M. (Eds.). (2022). *DIALOGUES: Towards decolonizing music and dance studies*. International Council for Traditional Music. https://ictmdialogues.org/

Tan, Z.-H., Sarkar, A. kr., & Dehak, N. (2020). rVAD: An unsupervised segment-based robust voice activity detection method. *Computer Speech & Language*, *59*, 1–21. https://doi.org/10.1016/j.csl.2019.06.005

Thompson, B. (2014). Discrimination between singing and speech in real-world audio. *2014 IEEE Spoken Language Technology Workshop (SLT)*, 407–412. https://doi.org/10.1109/SLT.2014.7078609

Tierney, A. T., Russo, F. A., & Patel, A. D. (2011). The motor origins of human and avian song structure. *Proceedings of the National Academy of Sciences*, *108*(37), 15510–15515. https://doi.org/10.1073/pnas.1103882108

Trehub, S. E., Unyk, A. M., Kamenetsky, S. B., Hill, D. S., Trainor, L. J., Henderson, J. L., & Saraza, M. (1997). Mothers' and fathers' singing to infants. *Developmental Psychology*, *33*(3), 500–507. https://doi.org/10.1037/0012-1649.33.3.500

Troy, J., & Barwick, L. (2020). Claiming the 'Song of the Women of the Menero Tribe.' *Musicology Australia*, 42(2), 85–107. https://doi.org/10.1080/08145857.2020.1945254

Tsur, R., & Gafni, C. (2022). *Sound–Emotion Interaction in Poetry: Rhythm, Phonemes, Voice Quality.* John Benjamins.

Urassa, M., Lawson, D. W., Wamoyi, J., Gurmu, E., Gibson, M. A., Madhivanan, P., & Placek, C. (2021). Cross-cultural research must prioritize equitable collaboration. *Nature Human Behaviour*, *5*(6), Article 6. https://doi.org/10.1038/s41562-021-01076-x

Valentova, J. V., Tureček, P., Varella, M. A. C., Šebesta, P., Mendes, F. D. C., Pereira, K. J., Kubicová, L., Stolařová, P., and Havlíček, J. (2019). Vocal parameters of speech and singing covary and are related to vocal attractiveness, body measures, and sociosexuality: A cross-cultural study. *Frontiers in Psychology, 10*, 2029. https://doi.org/10.3389/fpsyg.2019.02029

Vanden Bosch der Nederlanden, C.M., Qi, X., Sequeira, S., Seth, P., Grahn, J.A., Joanisse, M.F., and Hannon, E.E. (2022). Developmental changes in the categorization of speech and song. *Developmental Science*, e13346. https://doi.org/10.1111/desc.13346

Vargha, A., & Delaney, H. D. (1998). The Kruskal-Wallis Test and Stochastic Homogeneity. *Journal of Educational and Behavioral Statistics*, *23*(2), 170–192. https://doi.org/10.3102/10769986023002170

Verhoef, T., & Ravignani, A. (2021). Melodic Universals Emerge or Are Sustained Through Cultural Evolution. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/article/10.3389/fpsyg.2021.668300

Villing, R. (2010). *Hearing the Moment: Measures and Models of the Perceptual Centre* [Phd, National University of Ireland Maynooth]. https://mural.maynoothuniversity.ie/2284/

Vos, J., & Rasch, R. (1981). The perceptual onset of musical tones. *Perception & Psychophysics*, *29*(4), 323–335. https://doi.org/10.3758/BF03207341

Wang, Y., & Tian, L. (2018). An efficient numerical algorithm for exact inference in meta analysis. *Journal of Statistical Computation and Simulation*, *88*(4), 646–656. https://doi.org/10.1080/00949655.2017.1402331

Watanabe, S. (2018). *Mathematical Theory of Bayesian Statistics*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315373010

Weber, F., Knapp, G., Ickstadt, K., Kundt, G., & Glass, Ä. (2020). Zero-cell corrections in random-effects meta-analyses. *Research Synthesis Methods*, *11*(6), 913–919. https://doi.org/10.1002/jrsm.1460

Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, *116*(4), 1195–1200. https://doi.org/10.1073/pnas.1814092116

Wood, A., Kirby, K. R., Ember, C., Silbert, S., Passmore, S., Daikoku, H., Mcbride, J., Paulay, F., Flory, M., Szinger, J., D'Arcangelo, G., Bradley, K. K., Guarino, M. F., Atayeva, M., Rifkin, J., Baron, V., Hajli, M. E., Szinger, M., & Savage, P. E. (2022). *The Global Jukebox: A public database of performing arts and culture*. *PLOS ONE 17(11), e0275469. https://doi.org/10.1371/journal.pone.0275469*

Zhang, R., & Ghanem, R. (2021). Normal-Bundle Bootstrap. *SIAM Journal on Mathematics of Data Science*, *3*(2), 573–592. https://doi.org/10.1137/20M1356002

Stage 1 Supplementary Materials

S1. Supplementary Methods

S1.1. Recording and segmentation protocol

In order to keep the quality and consistency of the recordings, we created a detailed recording protocol for coauthors to follow when recording (Appendix 1). The protocol gives detailed instructions for things like how to interpret the instructions to choose a “traditional song in their 1st or heritage language” for cases where they are multilingual; logistics such as recording duration (minimum 30s, maximum 5 minutes for the song and the spoken description), file format, and how to deliver recordings to a secure email account monitored by a Research Assistant who is not a coauthor on the manuscript. All recordings are made by the coauthor themselves singing/ speaking/ playing instruments.

In addition to the recordings, we also collect the texts of recordings which are segmented into acoustic units (e.g., notes, syllables) according to their perceptual center (P-center) (Danielsen et al., 2019; Howell, 1988; Morton et al., 1976; Pompino-Marschall, 1989; Scott, 1998; Vos & Rasch, 1981). Here, the P-center is defined as the moment sound is perceived to begin, and the P-center is considered to be able to capture the perceptual experience of rhythm (Scott, 1998; Villing, 2010). The segmentation by the P-center is expected to reflect the vocalizer’s perception of the beginning of acoustic units. Here, we use acoustic units as a general term that a listener perceives as a unit of sound sequences such as syllables and notes. However, some languages have their own linguistic unit (e.g. mora in Japanese) and music as well (Fushi 節 in Japanese traditional folk songs). It is challenging to identify the beginnings of acoustic units for different domains (e.g., language and music), musical traditions, and languages comprising different phonemic and suprasegmental properties. For example, the location of the P-center in speech is known to be dependent on various factors such as the duration of phonemic elements (e.g. vowel, consonant) and the type of the syllable-initial consonant (Barbosa et al., 2005; Chow et al., 2015; Cooper et al., 1986; Villing, 2010). Therefore, rather than building an objective definition of sound onset, we ask each participant to reflect on their interpretation of acoustic units of their song and speech focusing on the P-center. Segmented texts are used to create onset and breath annotations with SonicVisualizer software (Cannam et al., 2010; https://www.sonicvisualiser.org/) which will be the base of some features. SonicVisualizer was chosen because it provides a simple interface to add a click sound to the desired time location of the audio to reflect the P-center. Those annotations will be created by the first author (Ozaki) because the time required to train and ask each collaborator to create these annotations would not allow us to recruit enough collaborators for a well-powered analysis.

In order to maximize efficiency and quality in our manual annotations, we adopt the following 3-step process:

1) Each coauthor sends a text file segmenting their recorded song/speech into acoustic units and breathing breaks (see Appendix 1 for examples).

2) The first author (Ozaki) creates detailed millisecond-level annotations of the audio recording files based on these segmented texts. (This is the most time-consuming part of the process).
3) Each coauthor then checks Ozaki's annotations (by listening to the recording with "clicks" added to each acoustic unit) and corrects them and/or has Ozaki correct them as needed until the coauthor is satisfied with the accuracy of the annotation.

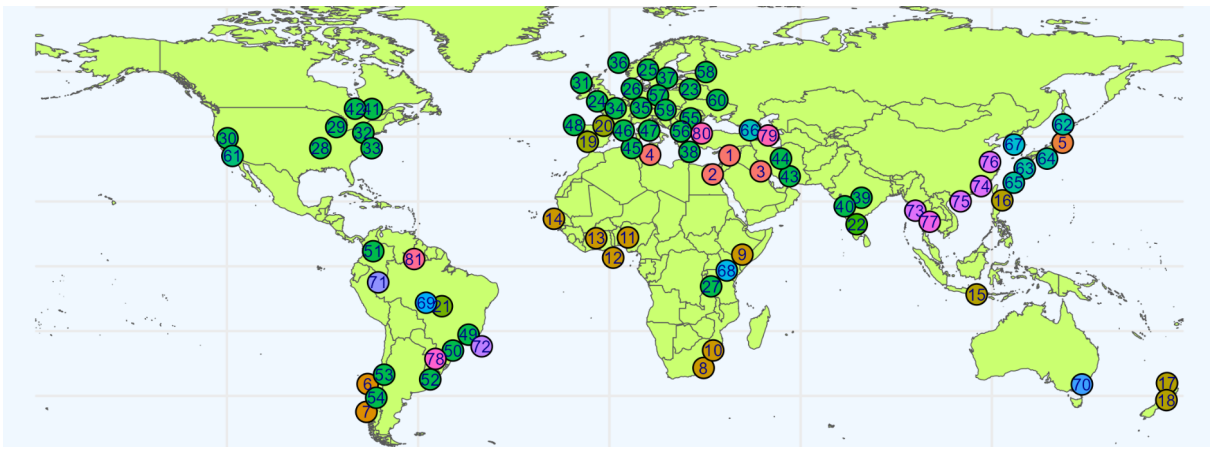
S1.2. Language sample

S.1.2.1. Inclusion criteria

All audio recordings analyzed are made by our group of 81 coauthors recording ourselves singing/speaking in our 1st/heritage languages, which span 23 language families (Fig. S1). Coauthors were chosen by opportunistic sampling beginning from co-corresponding author Savage's network of researchers, a public call to the email list of the International Council for Traditional Music (July 15 2022 to ictm-l@ictmusic.org; cf. Appendix 3), and recruitment at various conferences/symposia (International Council for Traditional Music, July 2022, Portugal; Joint Conference on Language Evolution, Sep 2022, Japan; Interdisciplinary Debates on the Empirical Aesthetics of Music series, Dec 2021, online; Social Bridges, Jan 2022, online; European Society for Cognitive Psychology, Feb 2022; AI Music Creativity, Sep 2022, online), with additional snowball recruitment from some collaborators using their own networks. Most authors are multilingual speakers who can speak English, though a few are multilingual in other languages (e.g., Portuguese, Japanese) with translations to and from English done by other coauthors as needed.

The set of linguistic varieties in this study represents a considerable portion of the world cross-linguistic variability in the main aspects that could conceivably play a role in shaping speech-song similarities/variabilities across languages (Dryer et al., 2013; https://wals.info/languoid):

- Head-complement order: languages with basic head-complement order (e.g. English), languages with basic complement-head order (e.g. Bengali)
- Vowel inventory size: moderate (e.g. Japanese), large (e.g. German)
- Consonant inventory size: small (e.g. Ainu), moderately small (e.g. Guaraní), average (e.g. Greek), moderately large (e.g. Swahili), large (e.g. Ronga)
- Consonant/vowel ratio: low (e.g. French), moderately low (e.g. Korean), average (e.g. Spanish), moderately high (e.g. Lithuanian), high (e.g. Russian)
- Potential syllable structures: simple (e.g. Yoruba), moderately complex (e.g. Catalan), complex (e.g. Kannada)
- Word-prosodic systems: stress-accent systems (e.g. Italian), pitch-accent systems (e.g. Swedish), tonal systems (e.g. Cantonese)
- Stress location: initial (e.g. Irish), postinitial (e.g. Basque), ante-penultimate (e.g. Georgian), penultimate (e.g. Polish), final (e.g. Balinese)
- Rhythm type: iambic (e.g. Mapudungun), trochaic (e.g. Hebrew)
- Complexity of tone systems: simple (e.g. Cherokee), complex (e.g. Thai)



Afro-Asiatic		Cariban		45 Catalan	*Romance*	Koreanic	
1 Modern Hebrew [Jerusalem]		21 Língua Kuikuro (Kuikúro-Kalapálo)		46 French		67 Korean	
2 Modern Hebrew [Tel Aviv]		Dravidian		47 Italian		Nilotic	
3 Palestinian Arabic (South Levantine Arabic)		22 Kannada		48 Portuguese [Porto]		68 Luo (dholuo) (Luo (Kenya and Tanzania))	
4 Tunisian Arabic		Indo-European		49 Portuguese [São Paulo]		Nuclear-Macro-Jê	
Ainu		23 Lithuanian	*Baltic*	50 Portuguese [São Paulo]		69 Rikbaktsa	
5 Aynu (Hokkaido Ainu)		24 Gaeilge (Irish)	*Celtic*	51 Spanish [Bogotá]		Pama-Nyungan	
Araucanian		25 Danish	*Germanic*	52 Spanish [Montevideo]		70 Ngarigu	
6 Mapudungun		26 Dutch [Heemstede]		53 Spanish [Santiago]		Pano-Tacanan	
7 Tsesungún (Huilliche)		27 Dutch [Nairobi]		54 Spanish [Osorno]		71 Matís	
Atlantic-Congo		28 English [Indiana]		55 Bulgarian	*Slavic*	Puri-Coroado	
8 IsiXhosa (Xhosa)	*Bantu*	29 English [Michigan]		56 Macedonian		72 Puri Kwaytikindo (Puri)	
9 Kiswahili (Swahili)		30 English [Nevada]		57 Polish		Sino-Tibetan	
10 Ronga		31 English [Newry]		58 Russian		73 Myanmar (Burmese)	*Burmese-Lolo*
11 Yoruba	*Defoid*	32 English [Pennsylvania]		59 Slovenian		74 Cantonese (Yue Chinese)	*Chinese*
12 Fante (Akan)	*Tano*	33 English [Washington D.C.]		60 Ukrainian		75 HainanHua (Min Nan Chinese)	
13 Twi (Akan)		34 Flemish (Dutch)		Iroquoian		76 Mandarin Chinese	
14 Wolof	*Wolof*	35 German		61 Cherokee		Tai-Kadai	
Austronesian		36 Norwegian		Japonic		77 Thai	
15 Balinese		37 Svenska (Swedish)		62 Japanese [Hokkaido]		Tupian	
16 Rukai		38 Greek	*Greek*	63 Japanese [Hyogo]		78 Mbyá-Guaraní	
17 Te Reo Māori (Māori) [Auckland]		39 Hindi	*Indic*	64 Japanese [Tokyo]		Turkic	
18 Te Reo Māori (Māori) [Wellington]		40 Marathi		65 Northern Amami-Oshima		79 North Azerbaijani	
Basque		41 Punjabi (Eastern Panjabi)		Kartvelian		80 Turkish	
19 Euskara (Basque) [Errenteria]		42 Urdu		66 Georgian		Yanomamic	
20 Euskara (Basque) [Hondarribia]		43 Western Farsi [Isfahan]	*Iranian*			81 Yanomami (Yanomám)	
		44 Western Farsi [Tehran]					

Figure S1. Map of the linguistic varieties spoken by our 81 coauthors as 1st/heritage languages. Each circle represents a coauthor singing and speaking in their 1st (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 (Hammarström et al., 2022), 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 (Dryer et al., 2013; https://wals.info/languoid).

S.1.2.2. Exclusion criteria and data quality checks

If coauthors choose to withdraw their collaboration agreement at any point prior to formal acceptance after peer review, their recording set will be excluded (cf. Appendix 2). If their recording quality is too poor to reliably extract features, or if they fail to meet the formatting requirements in the protocol we will ask them to resubmit a corrected recording set. In order to keep ourselves as blind as possible to the data prior to In Principle Acceptance and analysis, we ask coauthors to send only their segmented texts, not their audio recordings, to coauthors Ozaki & Savage to conduct formatting checks (e.g., ensuring that coauthors had understood the instructions to make all recordings in the same language and to segment their sung/spoken texts into acoustic units), so that we will not need to access the audio recordings until after In Principle Acceptance.

After we had already begun this process, we decided to add an additional layer of formatting and data quality checks by hiring a Research Assistant (RA) who is not a coauthor to create and securely monitor an external email account where authors could send their audio recordings. This allows us to prevent data loss (e.g., collaborators losing computers or accidentally deleting files), as well as allowing us to have the RA confirm that recording quality was acceptable, recordings met minimum length requirements, etc. The RA will not share the account password needed to access these recordings with us until we have received In Principle Acceptance.

S1.3. Features

We will compare the following six features between song and speech for our main confirmatory analyses:

1) Pitch height (fundamental frequency (f_0)) [*Hz*],
2) Temporal rate (inter-onset interval (IOI) rate) [*Hz*],
3) Pitch stability (-|f_0|) [*cent/sec.*],
4) Timbral brightness (spectral centroid) [*Hz*],
5) Pitch interval size (f_0 ratio) [*cent*],
    - 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.

For each feature, we will compare 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 (cf. Fig. 2).

We selected these features by reviewing what past studies focused on for the analysis of song-speech comparison and prominently observed features in music (e.g. Fitch, 2006; Hansen et al., 2020; Hilton et al., 2022; Savage et al., 2015; Sharma et al., 2021, see the Supplementary Discussion section S2 for a more comprehensive literature review). Here, f_0, rate of change of f_0, and spectral centroid are extracted purely from acoustic signals, while IOI rate is based purely on manual annotations. Pitch interval size and pitch declination analyses combine a mixture of automated and manual methods (i.e. extracted f_0 data combined with onset/breath annotations). The details of each feature can be found in the supplementary materials. Note that some theoretically relevant features we explored in our pilot analyses (especially the “regular rhythmic patterns” from Lomax & Grauer’s definition of song quoted in the introduction) proved difficult to quantify using existing metrics and thus are not included in our six candidate features (cf. Fig. S9 for pilot data and discussion for potential proxies that we found unsatisfactory such as “IOI ratio deviation” and “pulse clarity”).

S1.4. Pilot data analysis

We collected recordings from five coauthors for pilot data analysis[2] Each speaks a different 1st language: English, Japanese, Farsi, Marathi, and Yoruba. Figure S2 uses the analysis framework shown in Fig. 2 to calculate relative effect sizes for all five recording sets for all six hypothesized features. Note that our inferential statistical analysis uses the relative effects, but we translate these to Cohen's d in Fig. S2 for ease of interpretability, but technically our analysis is not the same as directly measuring Cohen's d of the data.

The primary purpose of the pilot analysis is to demonstrate feasibility and proof of concept, but we also used it to help decide on our final set of six features to focus on for our confirmatory analyses (Fig. S2). A full pilot analysis including additional features that we decided not to test is shown in Fig. S9. However, while some of our hypotheses appear to be strongly supported by our pilot data (e.g., song consistently appears much higher and much slower than speech, and timbral brightness appears consistently similar), others seem more ambiguous (e.g., pitch stability and pitch interval size show similar, weak trends although we predict pitch stability to differ but pitch interval size not to differ). In these cases, we prioritized our theoretical predictions over the pilot data trends, as effect sizes estimated from pilot data are not considered reliable (Brysbaert, 2019), while ample theory predicts that song should use more stable pitches than speech (e.g., Fitch, 2006) but sung and spoken pitch interval size should be similar (e.g., Tierney et al., 2010). However, we will be less surprised if our predictions for pitch stability and pitch interval size are falsified than if our predictions for pitch height and temporal rate are. Summary statistics visualizing the data underlying Fig. S2 in a finer-grained way are shown in Figure S3.

[2] Coauthors who contributed pilot data also recorded separate recording sets to be used in the main confirmatory analysis to ensure our main analyses are not biased by reusing pilot data.

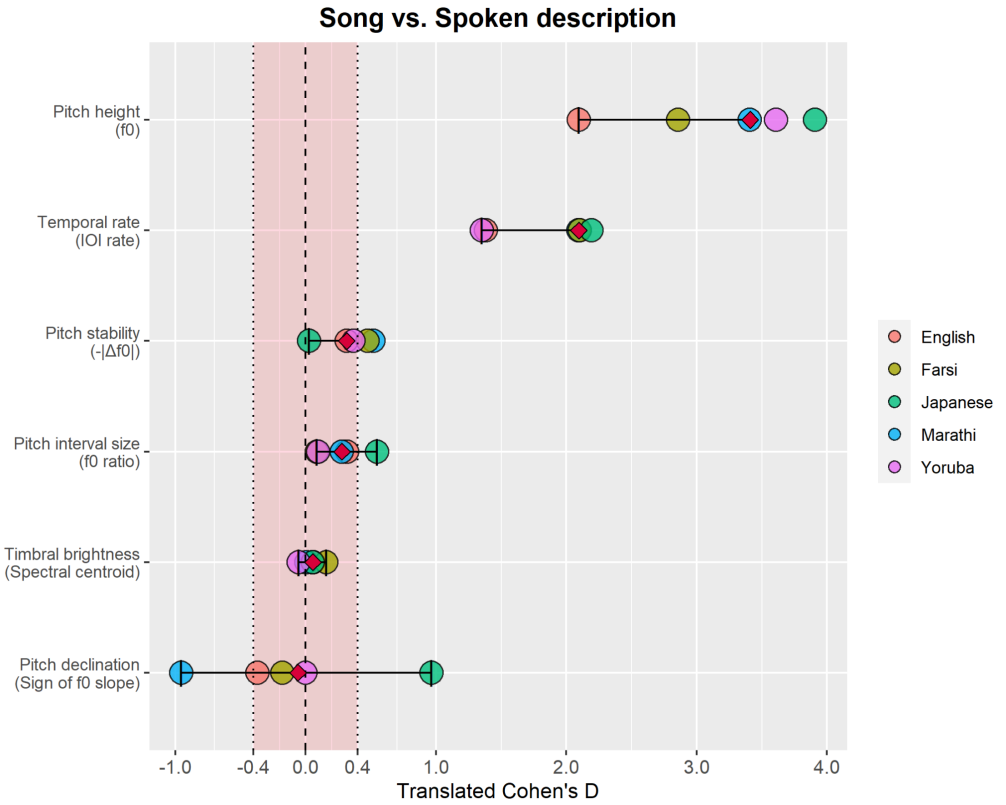


Figure S2. Pilot data showing similarities/differences between song and speech for each of the six hypothesized features across speakers of five languages (coauthors McBride, Hadavi, Ozaki, D. Sadaphal, and Nweke) Red diamonds indicate the population mean and black bars are confidence intervals estimated by the meta-analysis method. Although we use false discovery rate to adjust the alpha-level, these intervals are constructed based on Bonferroni corrected alpha (i.e. 0.05/6). Whether the confidence interval is one-sided or two-sided is determined by the type of the hypothesis. Positive effect sizes indicates song having a higher value than speech, with the exception of "temporal rate", whose sign is reversed for ease of visualization (i.e., the data suggest that speech is faster than song. The effect size is originally measured by relative effect, and that result is transformed into Cohen's d for interpretability. The red shaded area surrounded by vertical lines at ±0.4 indicate the "smallest effect size of interest" (SESOI) suggested by Brysbaert (2019). See Fig. 2 for a schematic of how each effect size is calculated from each pair of sung/spoken recordings.

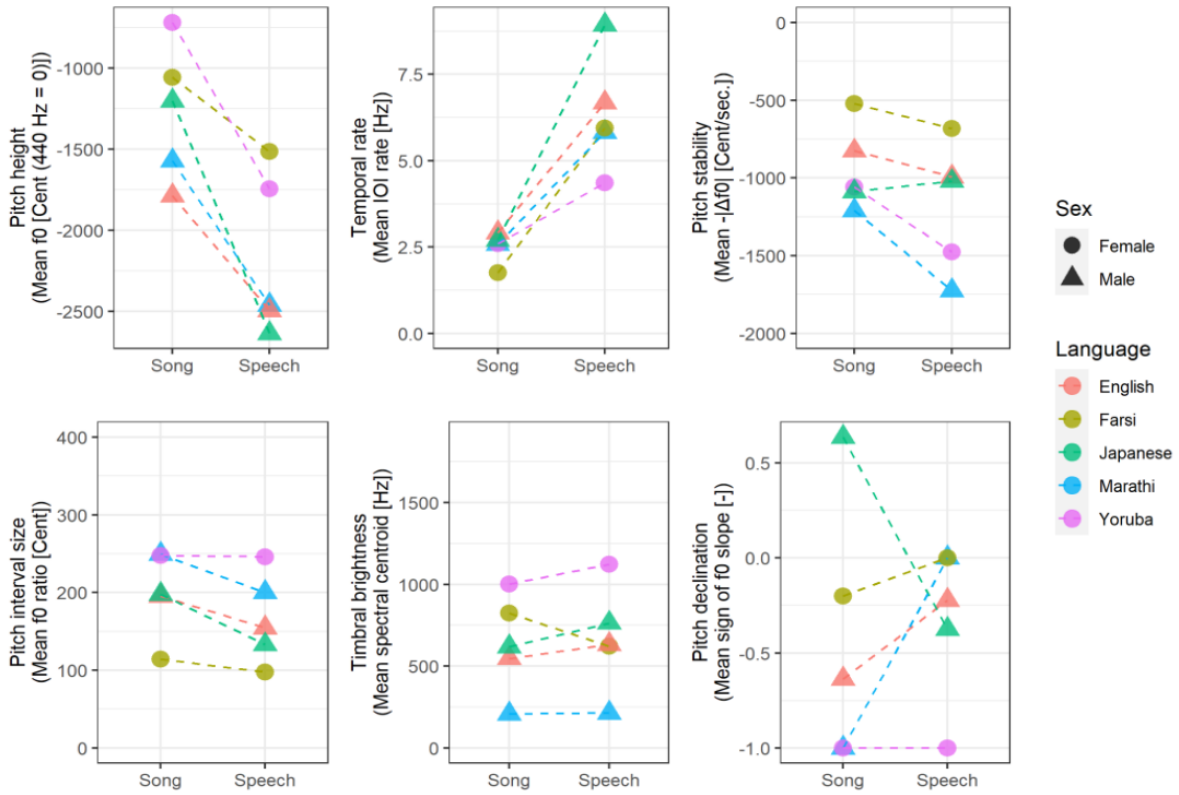


Figure S3. **Alternative visualization of Figure S2 showing mean values of each feature of song and speech, rather than paired differences.** "Speech" indicates spoken description (not lyric recitation). This figure allows us to visualize some trends not viewable from Figure S2, such as absolute values of each feature. For example, male voices all tend to be lower-pitched than female, but regardless of sex all singers use higher pitch for singing than speaking. (See Fig. S8 for an alternate version including exploratory analyses comparing instrumental and recited versions.)

In addition to the above main pilot analysis, we conducted two additional pilot analyses to validate our choice of duration of recording and annotation procedure. First, we investigated how estimated effect sizes vary with length of recording excerpt analyzed (Fig. S4). We concluded that 20 seconds approximately optimizes the tradeoff between accuracy of effect size estimation and the substantial time required to manually annotate onsets (roughly 10-40 minutes per 10 seconds of recording, with spoken description often taking several times longer to annotate than sung, instrumental, or recited versions).

Second, we had each of the five coauthors who annotated pilot data for their own language re-annotate a 10-second excerpt of their own recording (to determine intra-rater reliability) and then also annotate a 10-second excerpt of recordings in all other languages (to determine inter-rater reliability). They first did this once without any segmented text provided, and then corrected this after being provided with segmented texts. We then compared all these recordings against automated algorithms widely used in speech analysis (de Jong & Wempe, 2009; Mertens, 2022) to determine reliability of automated methods (Fig. S6).

The results of human-human comparisons were somewhat ambiguous, but overall suggested that (1) between-annotator differences in onset and break annotation are negligible even for different languages (provided they are provided with segmented texts),

(2) within-annotators randomness of annotation is also negligible as well, and (3) effect sizes based on the annotation provided by automated methods can be significantly different from human annotations. Note that Fig. S6 only compares temporal rate and pitch interval size, since most other features did not require manual annotations, while pitch declination was not analyzed because the 10-second excerpts were too short to have enough phrases to evaluate. Although our validation suggests the value of manual annotation, it would be desirable to increase its efficiency in future via options such as semi-automated methods or crowd-sourcing (though there will likely be tradeoffs between data quality and quantity; cf. Cychosz et al., 2021).

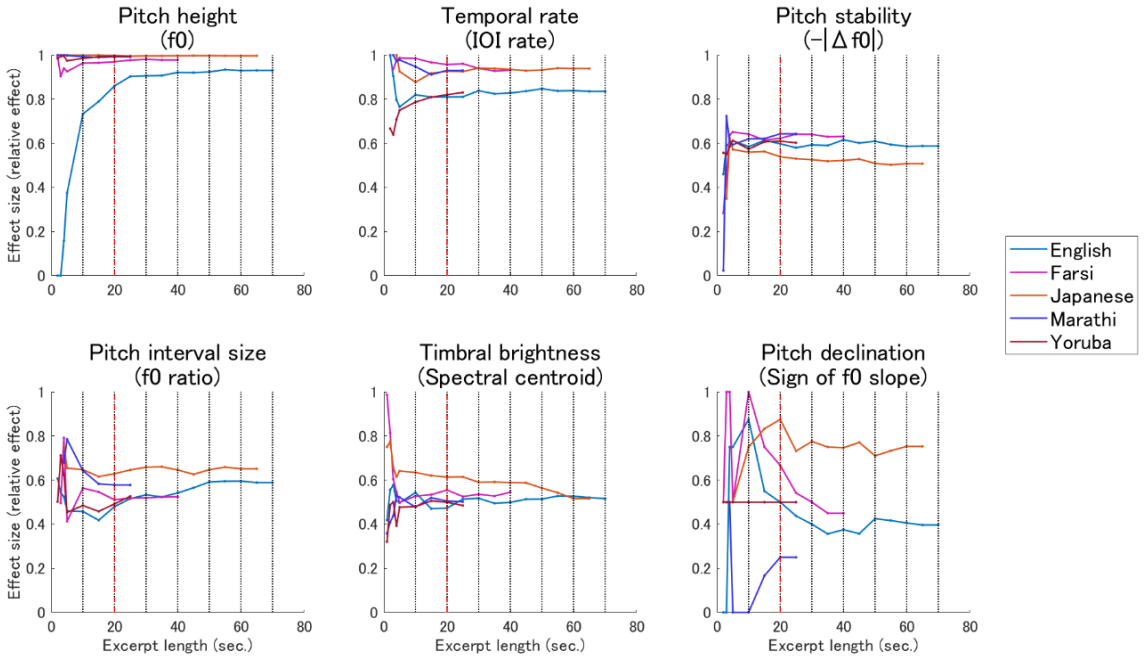


Figure S4. **Relationship between the duration of recording excerpt analyzed and estimated effect size for the 6 features and 5 sets of pilot recordings analyzed in Fig.S2.** Since the length of the pilot recordings ranged from under 30s to over 70s, plots are truncated at the point when there is no longer enough matching sung and spoken audio recording for that language (e.g., 25s for Marathi and Yoruba, 70s for English). The red vertical dashed line at 20s indicates the length we concluded approximately optimizes the tradeoff between accuracy of effect size estimation and the substantial time required to manually annotate onsets.

S1.5. Power analysis

We performed a power analysis to plan the number of recording sets (corresponding to the number of studies in meta-analysis) necessary to infer the statistical significance of the specified analyses. Because our pilot data consisting of only 5 recording sets is too small to empirically derive reliable effect size estimates, our power analyses used an SESOI corresponding to d = .4 (see Anvari & Lakens, 2021; Brysbaert, 2019 for the use of SESOI for sample size planning). However, there is one nuisance parameter in the model (i.e. between-study variance) necessary to specify for the power analysis, and we set this value with the estimate from the pilot data as a workaround.

Although we are planning to use the Benjamini-Hochberg step-up procedure (Benjamini & Hochberg, 1995) in our hypothesis testing, since the actual critical value depends on the p-value we will observe, it is challenging to specify sample size based on the false discovery rate especially when using nonparametric statistics, though some methods are available for parametric models (Jung, 2005; Pounds & Cheng, 2005). Therefore, we use the family-wise error rate for setting the alpha level for sample size planning as a proxy. Although it is known that when all null hypotheses are true, the false discovery rate becomes equal to the family-wise error rate (Benjamini & Hochberg, 1995), and the required sample size does not differ significantly between false discovery rate methods and stepwise family-wise error control methods in certain cases (Horn & Dunnett, 2004), our case may not necessarily match these conditions. Therefore our sample size estimate will be equal to or more than the size required for specified power assuming the alpha level determined by Bonferroni correction to set a stricter critical value.

We define the alpha level as 0.05 divided by six which is a family-wise error control by Bonferroni correction, and the statistical power as 0.95 for our sample size planning. Our statistical model is Gaussian random-effect models as explained in 1.2 Analysis plan.

Our power analysis estimated that n=60 recording sets is estimated as the minimum required sample size to achieve the above type I and type II error control levels when testing our six null hypotheses (see Supplementary Materials S3.2 for details). The features other than the sign of f_0 slope (i.e. f_0, IOI rate, rate of change of f_0, f_0 ratio, and spectral centroid) were estimated to have a relatively low between-study (recording set) variance, so the required number of recording sets computed for each feature is estimated to be lower than 10. However, as shown in Fig. S2, the sign of f_0 slope has a large between-study variance, and that resulted in 60 recording pairs being needed.

Please note that our power analysis does not take into account the specific languages used. While it would be ideal to have models that capture how languages (and other factors such as sex, age, etc.) influence the song-speech difference, we do not have enough empirical data or prior studies to build such models at this moment. Hence we simply treat each recording data without such factors, controlling for language family relationships separately in our robustness analyses. Future studies may be able to better incorporate such factors in a power analysis based on the data our study will provide.

S1.6. Robustness analyses

S1.6.1. Exclusion of data generated after knowing the hypotheses

One distinctive aspect of this study is that the authors ourselves generate the data for the analysis. Traditionally, personnel who provide data are blinded from the hypotheses to avoid biases where researchers (consciously or unconsciously) collect data to match their predictions. Here, we attempt to control for bias by withholding from analysis of audio data until we confirm the in-principle acceptance of this manuscript. We collect most recordings in a way that coauthors do not have access to each others' audio recordings until In Principle Acceptance (IPA) of this Registered Report, so that hypothesis formation and analysis methodology are specified a priori before accessing and analyzing the audio recordings. Still, some data are generated from the core team who planned and conducted the pilot analyses and thus already knew most hypotheses before we decided this issue needed to

be controlled for. Data from these authors may possibly include some biases due to knowing the details of the study (e.g., we may have consciously or unconsciously sung higher or spoke lower than we normally would to match our prediction that song would use higher pitch than speech). Therefore, we will test the robustness of our confirmatory analysis results by re-running the same analyses after excluding recordings provided by coauthors who already knew the hypotheses when generating data. Our confirmatory analyses test the direction of effect sizes, so applying the same tests allows us to check if that holds with varying conditions. In case the results of this analysis and the original confirmatory analysis do not match, we will interpret our results as not robust (whether due to potential confirmation bias or to other sampling differences) and will thus not draw strong conclusions regarding our confirmatory hypotheses.

S1.6.2. Potential dependency caused by language family lineage

Another potential bias in our design is the unbalanced sample of languages due to our opportunistic sampling design. Related languages are more likely to share linguistic features due to common descent, and sometimes these features can co-evolve following lineage-specific processes so that the dependencies between the features are observable only in some families but absent in others (Dunn et al., 2011)[3]. Thus, it is possible that our sample of speakers/singers may not represent independent data points. While our study includes a much more diverse global sample of languages/songs than most previous studies, like them our sample is still biased towards Indo-European and other larger languages families, which might bias our analyses. To determine whether the choice of language varieties affects our confirmatory analyses, we will re-run the same confirmatory analyses using multi-level meta-analysis models (linear mixed-effects models; Sera et al., 2019) with each recording set nested in the language family. We will perform model comparison using the Akaike Information Criterion (AIC; Bozdogan, 1987) for the original random-effects model and the multi-level model. The model having the lower AIC explains the data better in terms of the maximum likelihood estimation and the number of parameters (Watanabe, 2018), although critical assessment of information criteria and model selection methods in light of domain knowledge is also important (Dell et al., 2000). If the choice of model technique qualitatively changes the results of our confirmatory hypothesis testing, we will conclude that our results depend on the assumption of the language dependency..

S1.7. Exploratory analysis to inform future research

We are interested in a number of different questions that we cannot include in our main confirmatory analyses due to issues such as statistical power and presence of background noise. However, we plan to explore questions such as the following through post-hoc exploratory analyses, which could then be used to inform confirmatory analyses in future research:

---

[3] There is also some potential that musical and linguistic features may be related, although past analyses of such relationships between musical features and linguistic lineages have found relatively weak correlations (Brown et al., 2014; Matsumae et al., 2021; Passmore et al., Under review).

S1.7.1. More acoustic features:

We will also explore other features in addition to the specified five features to investigate what aspects of song and speech are similar and different. Supplementary Figure S9 shows the analysis using additional features.

S1.7.2. Relative differences between features:

Our confirmatory analysis will formally test whether a given feature is different or similar between song and speech, but will not directly test whether some features are more or less good than others at distinguishing between song and speech across cultures. To explore this question, we will rank the magnitude of effect sizes to investigate the most differentiating features and most similar features among the pairs of song and speech.

S1.7.3. Music-language continuum:

To investigate how music-language relationships vary beyond just song and spoken description, we will conduct similar analyses to our main analyses but adding in the other recording types shown in Fig. 1 made using instrumental music and recited song lyrics.

S1.7.4. Demographic factors:

Most collaborators also volunteered optional demographic information (age and gender), which may affect song/speech acoustics. Indeed, Fig. S3 suggests that pitch height differences between males and females are even larger than differences between song and speech. We will explore such effects for all relevant features.

S1.7.5. Linguistic factors:

We will also investigate whether typological linguistic features affect song-speech relationships (e.g., tonal vs. non-tonal languages; word orders such as Subject-Verb-Object vs. Subject-Object-Verb languages; “syllable-timed” vs. “stress-timed” languages and related measurements of rhythmic variability (nPVI; cf. Patel & Daniele, 2003), etc.

S1.7.6. Other factors:

In future studies, we also aim to investigate additional factors that may shape global diversity in music/language beyond those we can currently analyze. Such factors include things such as:

-functional context (e.g., different musical genres, different speaking contexts)
-musical/linguistic experience (e.g., musical training, mono/multilingualism)
-neurobiological differences (e.g., comparing participants with/without aphasia or amusia)

S1.7.7. Reliability of annotation process:

Each of Ozaki’s annotations will be based on segmented text provided by the coauthor who recorded it, and Ozaki’s annotations will be checked and corrected by the same coauthor, which should ensure high reliability and validity of the annotations. However, in order to objectively assess reliability, we will repeat the inter-rater reliability analyses shown in Fig. S6 on a subset of the full dataset annotated independently by Savage without access to Ozaki’s annotations. Like Fig. S6, these analyses will focus on comparing 10s excerpts of song and spoken descriptions, randomly selected from 10% of all recording sets (i.e., 8 out of the 81 coauthors, assuming no coauthors withdraw). Ozaki’s annotations corrected by the

original recorder will be used as the “Reference” datapoint as in Fig. S6, and Savage’s annotations (also corrected by the original recorder) will correspond to the “Another annotator” datapoints in Fig. S6. Note however that we predict that Savage’s corrected annotations will be more analogous to the “Reannotation” data points in Fig. S6, since in a sense our method of involving the original annotator in checking/correcting annotations is analogous to them reannotating themselves in the pilot study.

S1.7.8. Exploring recording representativeness and automated scalability:

Because our opportunistic sample of coauthors and their subjectively selected “traditional” songs are not necessarily representative of other speakers of their languages, we will replicate our analyses with Hilton, Moser et al.’s (2022) existing dataset, focusing on the subset of languages that can be directly compared. This subset of languages will consist of 5 languages (English, Spanish, Mandarin, Kannada, Polish) represented by matched adult-directed song and speech recordings by ~240 participants (cf. Hilton et al. Table 1).

Because our main analysis method requires time-intensive manual or semi-manual annotation involving the recorded individual that will not be feasible to apply to Hilton et al.’s dataset, we will instead rely for our reanalysis of Hilton et al.’s data on purely automated features. We will then re-analyze our own data using these same purely automated features. This will allow us to explore both the scalability of our own time-intensive method using automated methods, and directly compare the results from our own dataset and Hilton et al.’s using identical methods.

Fig. S10 demonstrate this comparison using pilot data for one feature (pitch height) based on a subset of Hilton et al.’s data that we previously manually annotated (Ozaki et al., 2022), allowing us to simultaneously compare differences in our sample vs. Hilton et al.’s sample and automated vs. semi-automated methods. Even though this analysis focuses on a feature expected to be one of the least susceptible to recording noise (pitch height), our pilot analyses found that these were mildly sensitive to background noise, such that purely automated analyses resulted in systematic underestimates of the true effect size as measured by higher-quality semi-automated methods (Fig. S10). While our recording protocol (Appendix 2) ensures minimal background noise, Hilton et al.’s field recordings were made to study infant-directed vocalizations and often contain background noises of crying babies as well as other sounds (e.g., automobile/animal sounds; cf. Fig. S11), which may mask potential differences and make them not necessarily directly comparable with our results. This supports the need to compare our results with Hilton et al.’s using both fully-automated and semi-automated extracted features to isolate differences that may be due to sample representativeness and differences that may be due to the use of automated vs. semi-automated methods.

S2. Supplementary discussion of hypotheses and potential mechanisms

This section outlines the literature review on the comparative analyses of music and language, with special emphasis on relevant hypotheses regarding their evolutionary origins. This section introduces possible mechanisms underlying differences and similarities between song and speech. We have include this text here for completeness but placed it in the Suplementary Material rather than in the “Study aims and hypotheses” section of the main text because, while relevant to our hypotheses, most are not directly testable in our proposed design.

S2.1. Hypotheses for speech-song differences

We predict that the most distinguishing features will be those repeatedly reported in past studies, namely pitch height and temporal rate of sound production (Chang et al., 2022; Ding et al., 2017; Hansen et al., 2020; Merrill & Larrouy-Maestri, 2017; Sharma et al., 2021). Why have these features emerged specific to singing? From the viewpoint of the social bonding hypothesis, slower production rate may help multiple singers synchronize, facilitating "formation, strengthening, and maintenance of affiliative connections" (Savage et al., 2021). The social bonding hypothesis does not directly account for the use of high pitched voice; instead we speculate that this is related to the loudness perception of human auditory systems. It is known that the loudness sensitivity of human ears increases almost monotonically until 5k Hz. Furthermore, the magnitude of neural response to the frequency change by means of mismatch negativity also increases as the frequency range goes high in the range of 250 - 4000 Hz (Novitski et al., 2004). Therefore, heightening f_0 can be considered as conveying pitch information at a higher sensitive channel as possible. Also, in song and speech, melody is predominantly perceived via f_0, while timbre is predominantly perceived via the upper harmonics (Patel, 2008). Thus the tendency for music to emphasize melodic information and language to emphasize timbral information (Patel, 2008) may also explain a preference for higher sung pitch to optimize the frequency of the key melodic information. However, in addition to perceptual factors, higher pitch in singing may also be a consequence of the production mechanism required for sustaining the pitched voice, especially when keeping sub-glottal pressure at a high level to sustain phonation, which may facilitate raising pitch (Alipour & Scherer, 2007).

Interestingly, higher pitch and longer duration are identified as features contributing to saliency and perceived emotional intensity of sounds (but also other factors such as greater amplitude and higher spectral centroid, see Anikin (2020) for a more comprehensive list). This suggests our features predicted to show differences may originate in non-verbal emotional expression. In addition, the pattern of higher pitch height and slower sound production rate is also cross-culturally characteristic of infant-directed speech compared to adult-directed speech (Cox et al., 2022; Hilton et al., 2022). Along with other features in infant-directed speech, this difference is argued to play an important role in linguistic and social development (Cox et al., 2022).

Pitch discreteness is often considered a key feature of music (Brown and Jordiana, 2013; Fitch, 2006; Haiduk & Fitch, 2022; Savage et al., 2015; Ozaki et al., 2022; Vanden Bosch der Nederlanden et al., 2022). However, to our knowledge, there is no well-established way to analyze this property directly from acoustic signals. In this study, we measure pitch stability as a proxy of pitch discreteness. Our pitch stability measures how fast f_0 modulates, although we admit this may not fully account for the characteristics of pitch discreteness. For example, recent studies indicated pitch discreteness might relate to the ease of memorization (Haiduk et al., 2020; Verhoef & Ravignani, 2021), but our measurement does not directly take into account such effects. Based on the pilot analysis (Fig. S2), we confirmed that pitch stability can demonstrate the expected trend (i.e. more stable pitch in singing). The effect size can be medium (size corresponding to Cohen's d of 0.5) at best, but considering the limited capacity of human pitch control in singing (e.g. imprecise singing; Pfordresher et al. (2010)), it is plausible that pitch stability may not matter for the distinction between song and speech as much as pitch height and temporal rate. Still, we predict this

feature is worth testing for cross-cultural differences between song and speech, particularly given its prominence in previous debate (including Lomax an Grauer's definition of song cited in the introduction). In fact, several empirical studies documented that song usually produces more controlled f_0 than speech (Natke et al., 2003; Raposo de Medeiros et al., 2021; Stegemöller et al., 2008; Thompson, 2014).

In relation to the differentiation between song and speech, Ma et al. (2019) provided an intriguing simulation result of how a single vocal communication can diverge into a music-like signal and speech-like signal through transmission chain experiments. Their experiment was designed to test the musical protolanguage hypothesis (Brown, 2000) and found that music-like vocalization emerges when emotional functionality is weighted in the transmission and speech-like vocalization emerges when referential functionality is necessitated. This result may imply a scenario that singing behaviour emerged as one particular form of emotional vocal signals conveying internal states of the vocalizer, though its evolutionary theory has not particularly targeted music (Bryant, 2021). In fact, a melodic character of music is often considered to function in communicating mental states (Leongómez et al., 2022; Mehr et al., 2021) and infant-directed singing acts as the indication of emotional engagement (Trehub et al., 1997). Since our recordings are solo vocalizations however, our recordings may not display key features facilitating synchronization of multiple people such as regular and simple rhythmic patterns. Although this is out of scope of our study, it is intriguing to investigate whether this speculation also holds in the case of solo music traditions (Nikolsky et al., 2020; Patel & von Rueden, 2021).

S2.2. Hypotheses for speech-song similarities

We predict pitch interval size, timbre brightness and pitch declination will not show marked differences between song and speech. Amongst these three features, we introduce a novel way for assessing pitch interval size. Although there is a line of research studying musical intervals based on the limited notion of the interval defined with the Western twelve-tone equal-tempered scale (Ross et al., 2007; Schwartz et al., 2003; Stegemöller et al., 2008; but cf. Han et al., 2011; Robledo et al., 2016), our study treats interval more generally as a ratio of frequencies to characterize the interval of song and speech in a unified way.

Stone et al. (1999) reported that country singers use similar formant frequencies in both song and speech which is consistent with our pilot analysis (Figure S2), and they argued that the use of higher formant frequencies (e.g. singer's formant, see also Lindblom & Sundberg (2007)) in Western classical music tradition stemmed from the necessity of the singer's voice to be heard over a loud orchestral accompaniment. Similarly, Stegemöller et al. (2008) confirmed that speech and song have a similar spectral structure. Although we can find studies showing higher brightness in singing performed by professional singers (Barnes et al., 2004; Merrill & Larrouy-Maestri, 2017; Sharma et al., 2021; Sundberg, 2001), our dataset does not necessarily consist of recordings by professional musicians and as in the case of Stone et al. (1999) the prominent use of the high formant frequencies in singing may depend on musical style (but see Nikolsky et al., 2020) for the role of timbre played in personal music tradition). However, we would like to note that other aspects of timbre such as noisiness (spectral flatness) can potentially indicate the difference between song and speech (Durojaye et al., 2021).

Cross-species comparative studies identified that the shape of pitch contour is regulated by the voice production mechanism (Tierney et al., 2011; Savage et al., 2017). Since both humans and birds use respiratory air pressure to drive sound-producing oscillations in membranous tissues (Tierney et al., 2011), their pitch contours tend to result in descending towards the end of the phrase. Although previous studies only compared on pitch contours of human music (instrumental and vocal) and animal song, we predict the same pattern can be found in human speech since it still relies on the same motor mechanism of vocal production. More precisely, pitch declination is predicted to happen when subglottal pressure during exhalation can influence the speed of vocal fold vibration; the high pressure facilitates faster vocal fold vibration, and low pressure therefore makes the vibration relatively slower. Declarative speech is also subject to this mechanism (Ladd, 1984; Slifka, 2006).

S3. Features

The six features introduced in the main section are extracted as follows:

S3.1. Pitch height (f_0):

f_0 is estimated in a semi-automated way like the annotation in the Erkomaishvili dataset (Rosenzweig et al., 2020), which used an interactive f_0 extraction tool (Müller et al., 2017). We created a graphical user interface application with the following extraction process: 1) create the time-frequency representation of the audio signal using the fractional superlet transform (Bârzan et al., 2021; Moca et al., 2021); 2) a user specifies the set of points (beginning, end, upper and lower bound of frequency, and optional intermediate point(s) to be included in the contour) on the time-frequency plane to constraint the search region of f_0; 3) estimate an f_0 contour using the Viterbi algorithm (Djurović & Stanković, 2004). It is also possible to manually draw/delete/modify the contour if the f_0 is deemed not reliably estimated automatically due to severe interference by noise. The frequency resolution is 10 cents with 440 Hz = 0 (octave is 1200 cents), and the time resolution is 5 ms.

S3.2. Temporal rate (Inter-onset interval [IOI] rate):

Inter-onset interval rate is measured by first taking the difference between adjacent onset annotation times or onset and break annotation times and then taking that reciprocal. Our proxy for temporal rate is the inter-onset interval of consecutive P-centers (perceptual centers; Danielsen et al., 2019; Howell, 1988; Morton et al., 1976; Pompino-Marschall, 1989; Scott, 1998; Vos & Rasch, 1981), which is approximately similar to but not identical to the rate of linguistic and musical acoustic units (e.g. syllables, notes). Onset is a perceptual center determined by the person who made the recording.

S3.3. Pitch stability ($-|\Delta f_0|$):

The rate of change of f_0 is the negative absolute value of the numerical differentiation at each sampling point of the f_0 contour. The negative sign is used so that higher values indicate greater pitch stability. We use Shao & Ma's (2003) wavelet method with a first-order derivative of Gaussian to derive this because it is robust to noisy f_0 contours such as the ones in our pilot dat. We use 20 ms as the standard deviation parameter of the first-order derivative of Gaussian to smooth the noise, which corresponds to the scaling factor of the wavelet function.

S3.4. Pitch interval size:

Pitch interval is usually expressed as the ratio of pitch of two notes. We generalize this concept as follows. Firstly, segment an f_0 contour with the onset and break times. Secondly, take the outer product of the antecedent segmented f_0 contour and the reciprocal of the consequent f_0 contour. Here, rather than estimating a single representative pitch from each segment, we take exhaustive combinations of the ratio of f_0 values between adjacent segments and evaluate the interval as a distribution. This approach allows us to quantify intervals on both musical and linguistic acoustic signals. We calculate this outer product from each pair of adjacent segmented f_0 contours and aggregate all results as the pitch interval of the recording. However, one drawback of this method is the number of data points tends to become large due to taking outer products, though it can be mitigated by lengthening the sampling interval of f_0. Figure S5 shows a schematic overview of our approach.

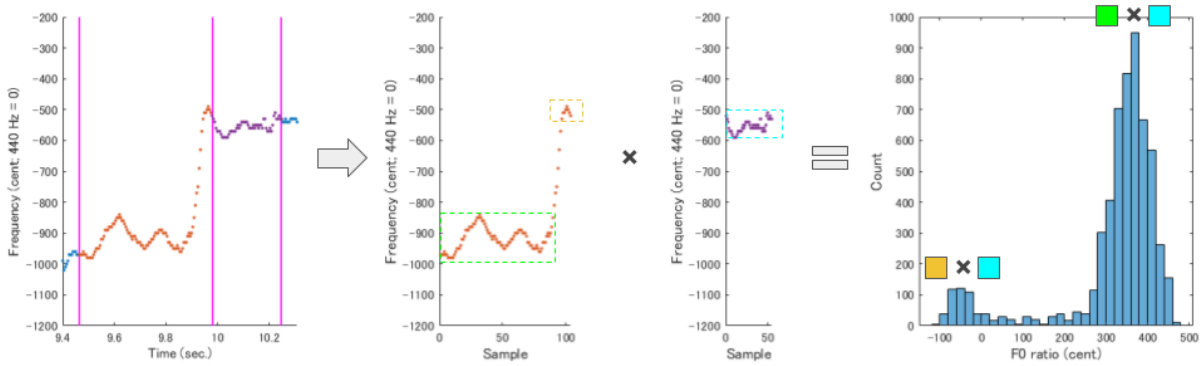


Figure S5. Process of computing f_0 ratios. The leftmost figure shows an f_0 contour which is segmented by three onset times. Then, the pitch ratio of the antecedent segmented f_0 contour (orange) and the consequent f_0 contour (purple) is calculated by taking exhaustive pairs of samples from two signals (104 samples × 55 samples in this example). The rightmost figure shows the obtained intervals by histogram which displays two peaks. The right-hand mode is the interval of ascending direction (around 370 cents) generated from the green rectangle part. The left-hand mode is the interval of descending direction (around -50 cents) generated from the orange rectangle part. Note that this example uses the cent scale rather than the frequency scale so that intervals can be calculated by subtraction.

S3.5. Timbral brightness (spectral centroid):

Spectral centroid is computed by obtaining a power spectrogram using 0.032 seconds Hanning window with 0.010 seconds hop size. The original sampling frequency of the signal is preserved. Please note silent segments during breathing/breaks are also included. However, the majority of the recordings contain a voice (or instrument), so the influence from silent segments should be minimal. Although we tried using an unsupervised voice activity detection algorithm by Tan et al. (2020), it was challenging to assess how much the failure of detection can impact the measurement of the effect size. The unsupervised algorithm was chosen to avoid the assumption of particular languages and domains as possible since we deal with a wide range of language varieties and audio signals of both music and language domains, which is usually beyond the scope of voice activity detection algorithms in general. Another limitation is that the measurement of spectral centroid can be affected by noise due to poor recording environment or equipment. However, our study focuses on the difference in terms of the relative effect in spectral centroid in two recordings (expected to be recorded

in the same environment/equipment/etc.), and we confirmed that the difference in spectral centroid itself is not markedly influenced by noise if the two recordings are affected by the same noise.

S3.6. Pitch declination (Sign of f_0 slope):

Pitch declination is estimated in the following steps. First, a phrase segment 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. Secondly, an f_0 contour is extracted from that segment. We treat f_0s as response variable data and correspondence times as dependent variable data. If there are frames where f_0 is not estimated, we discard that region. Finally, we fit a linear regression model with Huber loss and obtain the slope. If the pitch contour tends to have a descending trend at the end of the phrase, we expect the slope of the linear regression tends to be negative. MATLAB's fitlm() function was used to estimate the slope. Figure 3 illustrates linear models fitted to each phrase.

S4. Statistical models and power analysis

S4.1. Statistical models

The Gaussian random-effects model used in meta-analysis is (Brockwell & Gordon, 2001; Liu et al., 2018)

$$Y_i|\theta_i \sim \mathcal{N}(\theta_i, \sigma_i^2),\ \theta_i \sim \mathcal{N}(\mu_0, \tau^2),\ i = 1, \ldots, K$$

Y_i is the effect size (or summary statistics) from ith study, θ_i is the study-specific population effect size, σ_i^2 is the variance of ith effect size estimate (e.g. standard error of estimate) which is also called the within-study variance, μ_0 is the population effect size, τ^2 is the between-study variance, and K is the number of studies. In our study, Y_i is the relative effect and σ_i^2 is its variance estimator (Brunner et al., 2018). In addition, the term "studies" usually used in meta-analysis corresponds to recording sets. This model can also be written as

$$Y_i \sim N(\mu_0, \sigma_i^2 + \tau^2),\ i = 1, \ldots, K$$

S4.2. Power analysis

We first describe the procedure for sample size planning for the hypotheses testing differences (H1-3). In this case, hypothesis testing evaluates $H : \mu_0 = \mu_{\text{null}}$ vs. $K : \mu_0 > \mu_{\text{null}}$, which means that the null hypothesis assumes the population effect size is the same as no difference and the alternative hypothesis assumes the difference exists in the positive direction (one-sided). Since we use relative effects as our effect sizes, we define $\mu_{\text{null}} = 0.5$. As described in "S1.5 Power Analysis", we decided to use SESOI for sample size planning, meaning we assume that the population effect size is the same as SESOI. Therefore, we specify where $\mu_0 = \Phi(0.4/\sqrt{2}) \approx 0.6114$ $\Phi(\cdot)$ is the standard cumulative normal distribution.

The power of the Gaussian random-effects model is given by (Hedges & Pigott, 2001; Jackson & Turner, 2017)

$$\beta(\delta, \tau^2, \boldsymbol{\sigma}) = 1 + \Phi(-Z_\alpha - \delta/\sqrt{V_R}) - \Phi(Z_\alpha - \delta/\sqrt{V_R}) \quad (1)$$
$$V_R = \frac{1}{\sum_{i=1}^{K}(\sigma_i^2 + \tau^2)^{-1}}$$

, where Z_α satisfies $\Phi(Z_\alpha) = \alpha$ that α is the significance level of the test, and δ is non-centrality parameter defined as $\delta = \mu_0 - \mu_{\text{null}}$ which represents the gap between the parameter of the null hypothesis model and the population parameter.

In order to perform the power analysis, we first need to specify the nuisance parameter τ^2 (between-study variance) which is generally unknown. We use DerSimonian-Laird estimator (Dersimonian & Laird, 1986; Liu et al., 2018) to estimate τ^2 using pilot data. However, there is the issue that the within-study variance σ_i^2 of sign of f_0 slope of the Yoruba recordings became 0. This happened because the signs of f_0 slope of singing and spoken description are all -1, which means f_0 contours of all phrases show better fitting to a downward direction than the upward. Zero variance causes divergence (i.e., +∞) in the weighting used in the DerSimonian-Laird estimator. As a workaround, the hypothetical standard error of the relative effect is estimated by assuming at least one of the observations was +1 (i.e. one of the f_0 slopes fits the upward direction). Specifically, we first re-estimated the standard error of the relative effect with both patterns that one of the signs is +1 in either the singing or spoken description. Then we took the smaller variance estimate for the hypothetical standard error of this recording set.

Furthermore, we also need assumption for σ_i^2 to calculate the power and to estimate the necessary number of studies K since the power is the function of the non-centrality parameter, between-study variance, and within-study variances. We assume the within-study variance has a mean and plug in the average of the within-study variances from pilot data. Algorithmically, our procedure is

1. Estimate τ^2 and $\delta = \mu_0 - \mu_{\text{null}}$.
2. Calculate the average of the within study variance.

$$\bar{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}\sigma_n^2$$

   N is the number of pilot recording sets (i.e. N = 5) here.
3. Set $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_N\}$
4. Calculate the power using the equation (1)
5. If the calculated power is lower than the target power then,
   $\boldsymbol{\sigma} \leftarrow [\boldsymbol{\sigma}\ \bar{\sigma}]$ (append $\bar{\sigma}$ to the current $\boldsymbol{\sigma}$) and return to 4.
   Otherwise, take the number of elements of $\boldsymbol{\sigma}$ as the necessary number of studies.

For the power analysis of equivalence tests (H4-6), we first note that the Gaussian random-effects model is equivalent to a normal distribution since random-effects models are Gaussian mixture models having the same mean parameter among components, therefore

$$\begin{aligned} p(\mathbf{Y}|\boldsymbol{\sigma}, \tau^2, \mu_0) &= \frac{1}{K}\sum_{i=1}^{K}\mathcal{N}(Y_i|\mu_0, \sigma_i^2 + \tau^2) \\ &= \mathcal{N}(Y_i|\mu_0, \sigma_\tau^2),\ i = 1...K \end{aligned}$$

where

$$\sigma_\tau^2 = \frac{1}{K}\sum_{i=1}^{K}(\sigma_i^2 + \tau^2)$$

We use this reparameterized version for equivalence tests. We estimate the necessary number of studies K by simulating how many times the test can reject a null hypothesis under the alternative hypothesis being true out of the total number of tests. Specifically, the rejection criteria is (Romano, 2005)

$$K^{1/2}|\bar{Y}_K| \leq C(\alpha, \delta, \sigma_\tau)$$

where $C = C(\alpha, \delta, \sigma)$ satisfies

$$\Phi\left(\frac{C-\delta}{\sigma}\right) - \Phi\left(\frac{-C-\delta}{\sigma}\right) = \alpha$$

$\bar{Y}_K$ is sample estimate of the mean, and we use the estimated μ_0 instead of the simple average of effect sizes. Here, δ defines the boundary for equivalence testing, namely $H : |\theta| \geq \delta$ vs. $K : |\theta| < \delta$ that the boundary is symmetric at 0. We set the boundary parameter based on SESOI $\delta = \Psi(0.4/\sqrt{2}) - 0.5 \approx 0.1114$ that shifts the center of the relative effect to 0 from 0.5, and specify $\theta = 0$ assuming that the population effect sizes of the features to be tested are null. When running the simulation, we draw random samples as $Y_i \sim \mathcal{N}(\mu_0, \sigma_\tau^2)$ and increase the number of studies K gradually until the simulation satisfies the expected power under the specified significance level.

S5. Supplementary Figures

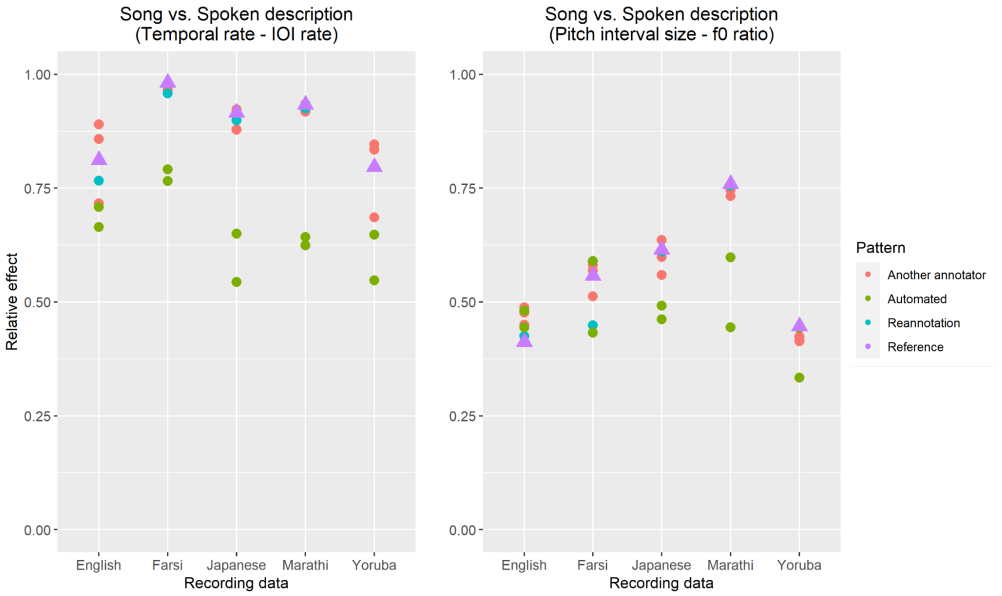


Figure S6. Within- and between-annotators randomness of onset annotations including automated methods (de Jong & Wempe, 2009; Mertens, 2022) discussed in Section S1.4 “Pilot data analysis”. 10-second excerpts were used. Reference is the result of the annotation by the person who originally made the recording.

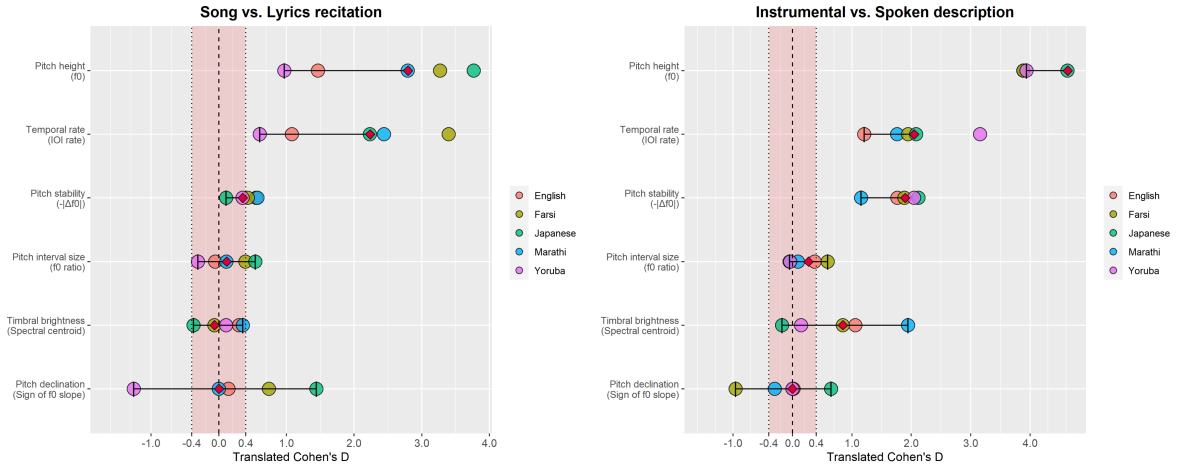


Figure S7. Effect sizes of each feature across five languages using the pilot data as in Figure S2 but with exploratory comparisons with recitation and instrumental recording types.

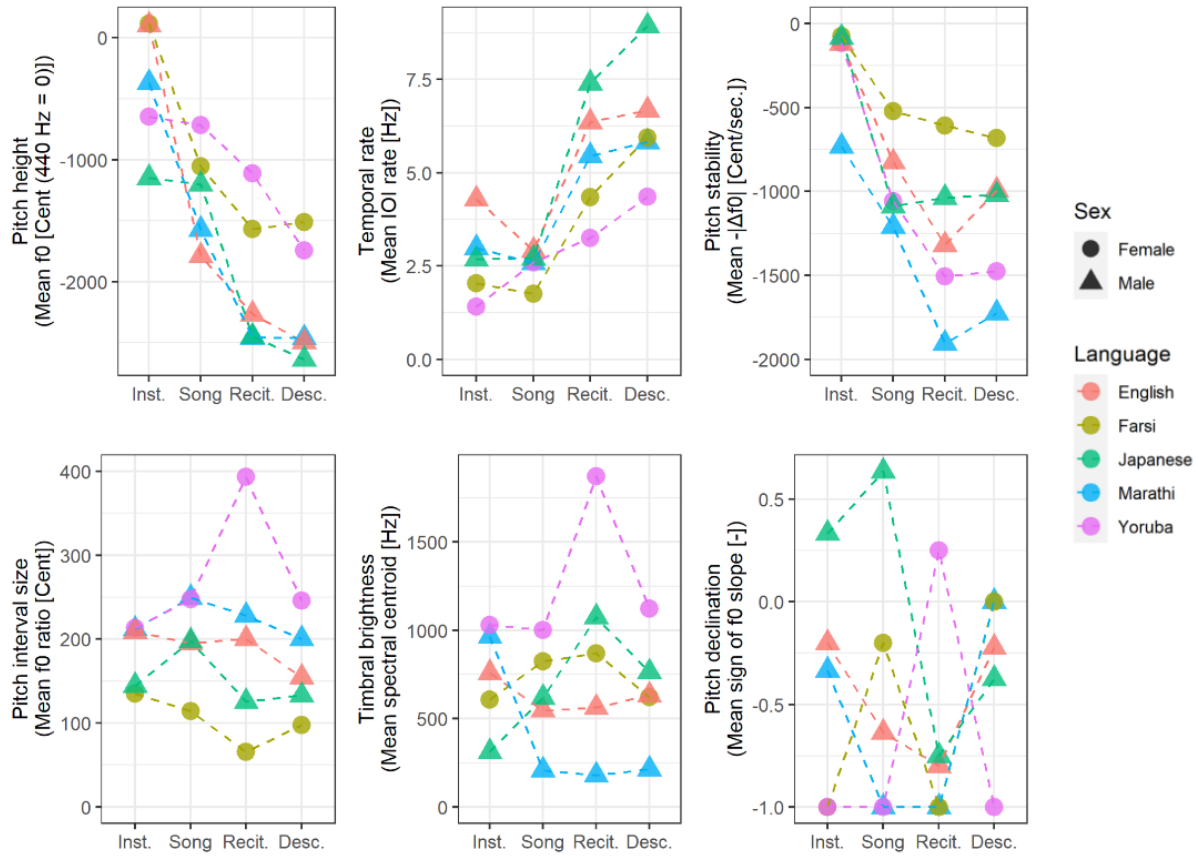


Figure S8. Mean values of each feature as in Figure S3 but with all recording types (including recitation and instrumental). “Desc.” means spoken description, “Recit.” means recited lyrics.

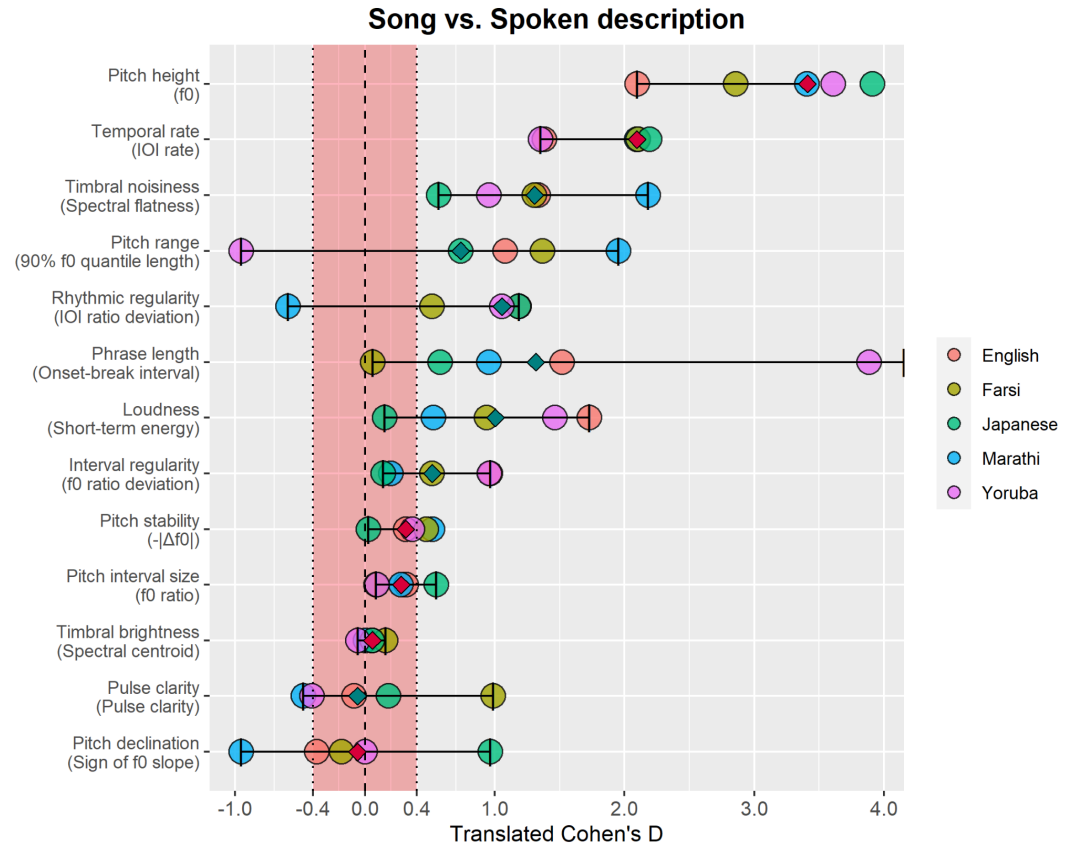


Figure S9. Effect sizes of each feature across five languages using the pilot data as in Figure S2 with additional exploratory features. Green-colored diamonds and two-sided confidence intervals are used for the features that hypotheses are not specified.

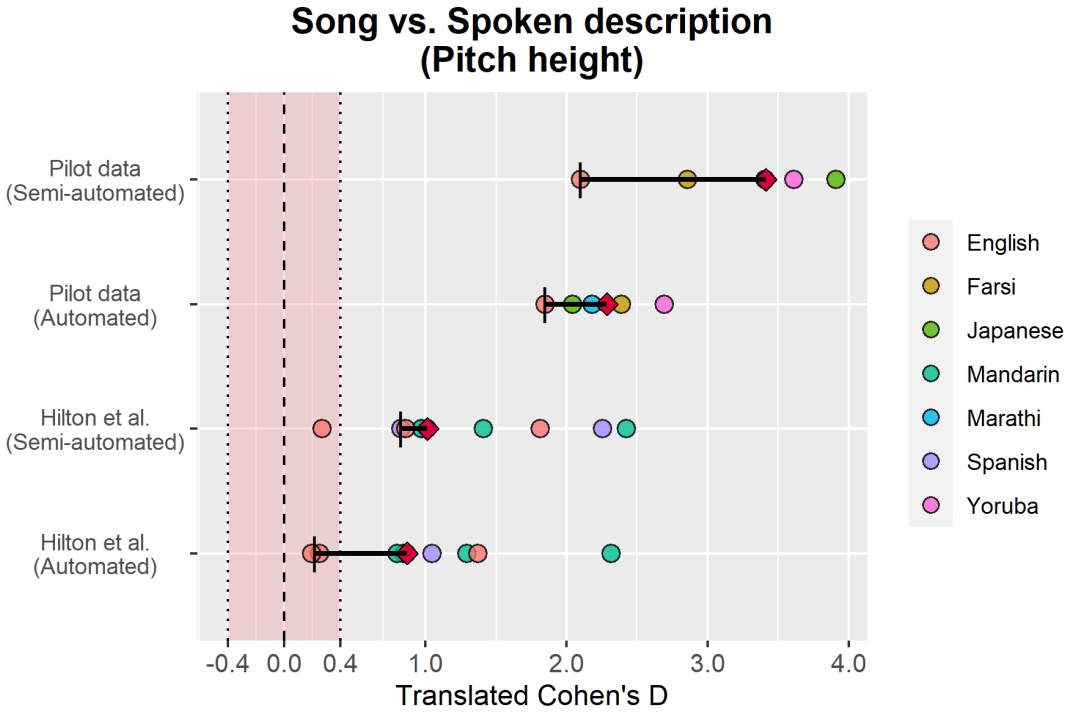


Figure S10. Pilot analysis of a subset of Hilton et al.'s (2022) data (pairs of adult-directed singing/speaking recordings from n=9 participants speaking English, Spanish, or Mandarin) focusing on pitch height. Ozaki et al., (2022) previously analyzed this subset for preliminary analyses using the same method described in S2.1 to avoid contamination by various noises included in audio (vocalization by babies, car noises, etc.), which allows us to explore issues such as whether such extraneous noises are likely to be a concern in our planned fully automated analysis of Hilton et al.'s full dataset (cf. Fig. S11). Although all four conditions demonstrate the predicted trend of song being consistently higher than speech, the effect size varies depending on the dataset and analysis method used (see Section S1.7.8. for discussion).

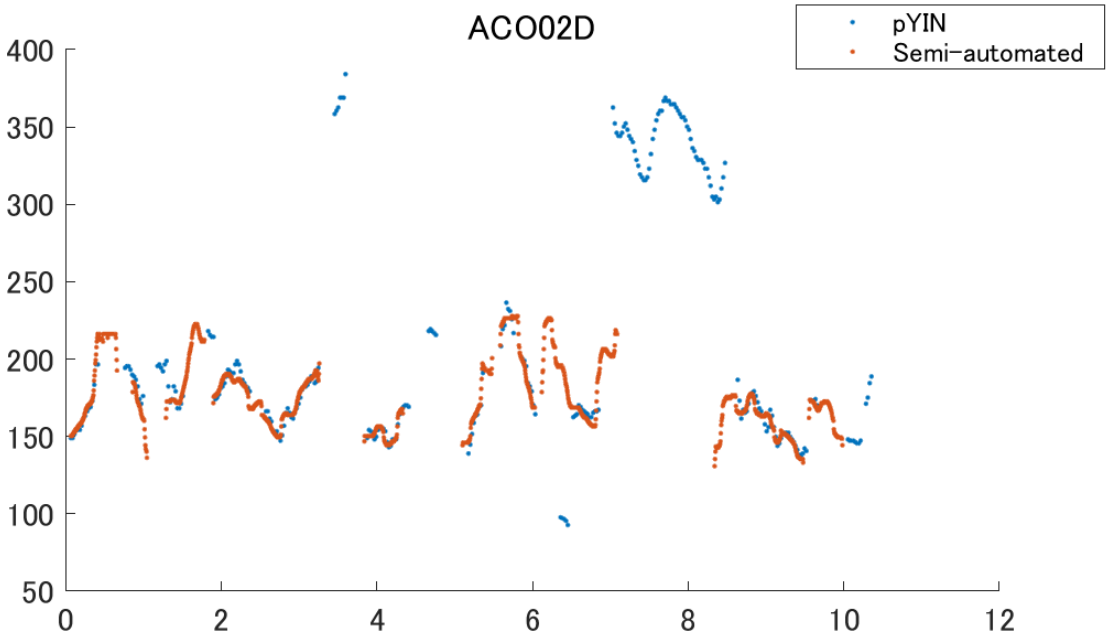


Figure S11. An example of fully-automated vs. semi-automated f0 extraction underlying the analyses in Fig. S7 for one of the field recordings from Hilton et al.'s dataset. AC002D = adult-directed speech [D] from individual #02 from the Spanish-speaking Afro-Colombian [ACO] sample). While the extracted f0 values are generally similar, the fully automated pYIN method sometimes has large leaps, particularly when there are external noises and the main recorded individual stops vocalizing to breathe (here the high-pitched blue contours at around 3.5 and 8 seconds correspond to the vocalizations of a nearby child while the recorded adult male takes a breath).

S6. Exploratory features

The summary of the additional features that will be examined in the exploratory analysis is as follows.

7) Rhythmic regularity (IOI ratio (Roeske et al., 2020) deviation) [*dimensionless*],
    - Absolute difference between the observed IOI ratios and the nearest mode estimated from the observed IOI ratios. If the perceived onsets constitute similar ratios over the recording, each data point (IOI ratio) would be concentrated around the mode thus small deviation from the most typical ratio would be expected. This idea is similar to measuring the variance of the within-cluster that modal clustering is used to create clusters. However, the

deviation of each data point from a cluster centroid is measured instead of variance.

- Various methods for density modes (equivalently zero-dimensional density ridges or degree zero homological features) have been recently proposed (Chacón, 2020; Chaudhuri & Marron, 1999; Chazal et al., 2018; Chen et al., 2016; Comaniciu & Meer, 2002; Fasy et al., 2014; Genovese et al., 2014; Genovese et al., 2016; Sommerfeld et al., 2017; Zhang & Ghanem, 2021). Here, we adopted techniques of topological data analysis. In particular, we use the mean-shift algorithm (Comaniciu & Meer, 2002) to detect the modes. Gaussian kernels are used and we choose to obtain a bandwidth parameter using Pokorny et al. (2012)'s method that selects a bandwidth from the range that the Betti number (number of modes in this case) is most stable (Carlsson, 2009; Pokorny et al., 2012). Note that this is not the only way and other criteria also exist (e.g. Genovese et al., 2016; Chazal et al., 2018) for the bandwidth selection from the viewpoint of topological features. The search space of bandwidth is set as $\sigma\{\log(n)/n\}$ as minimum following Genovese et al. (2016). The maximum bandwidth value is set as Silverman's rule-of-thumb (Silverman, 1986) since this bandwidth selection is usually considered oversmoothing (Hall et al., 1991), and this idea was previously also used for ridge detection analysis (Chen et al., 2015). Removing low density data points (outliers) to infer the persistent homology features is recommended (Chazal et al., 2018), so we set the threshold to eliminate data points that is $\{X_i : \hat{p}(X_i) < t\},\ t = \max(2, 0.01N)K(X;h)$ where $K(X;h)$ is a kernel density function with the bandwidth parameter h and $\hat{p}(X)$ is kernel density estimate using all data points. This threshold removes samples from density created by a few samples; equivalent to density less than 2 data points or less than 1% of the number of data points. Figure S12 illustrates our approach.

8) Phrase length (duration between two breaths/breaks) (onset-break interval) [*seconds*],
    - An interval between the first onset time after a break time (or the beginning onset time) and the first break time after the onset time, roughly corresponding to the length of a musical phrase or spoken utterance..
9) Pitch interval regularity (f_0 ratio deviation) [*cent*],
    - Like the IOI ratio deviation, the absolute difference between the observed f_0 ratios and the nearest mode. The method for calculating this feature is identical to the IOI ratio deviation, but for frequency rather than for time..
10) Pitch range (phrase-wise 90% f_0 quantile length) [*cent*],
    - The phrase is an interval as defined in 8) Phrase length. The sample quantile length of f_0 within each phrase is extracted.
11) Intensity (short–term energy) [*dimensionless*],
    - We measure the energy of the acoustic signal as a rough proxy of loudness although loudness is a perceptual phenomenon and these two are not necessarily equal. The short-term energy is the average of the power of the signal within a rectangular window whose length is 25 ms. We slide this window every 12.5 ms to collect the short-term energies of the recording. In order to avoid including the unvoiced segments, the energy is calculated from

the samples within IOIs or onset-break intervals. Since the relative effect is invariant with the order-preserving transformation, we do not apply a logarithm though the feature name is intensity. There are some limitations in this feature. One limitation is that recording is not strictly controlled. However, assuming the collaborator follows the protocol (e.g. keep the same distance between microphone and mouth/instrument and use the same recording device and recording environment across recordings), we assume the intensityof the recordings within each collaborator can be roughly compared. Another limitation is that the recording method is not unified across the collaborators. Therefore, even if there are the same level of differences in sound pressure level of singing and speech among the collaborators, the effect sizes to be calculated can be different. More precise control of recording conditions would be necessary for more accurate measurement of the difference in loudness in the future study.

12) Pulse clarity [*dimensionless*],
    - Pulse clarity is calculated using MIRToolbox V1.8.1 (Lartillot et al., 2008).
13) Timbre noisiness (spectral flatness (Johnston, 1988; Peeters, 2004)) [*dimensionless*]
    - Spectral flatness is measured at each acoustic unit, namely inter-onset intervals and onset-break intervals, as in Durojaye et al. (2021).

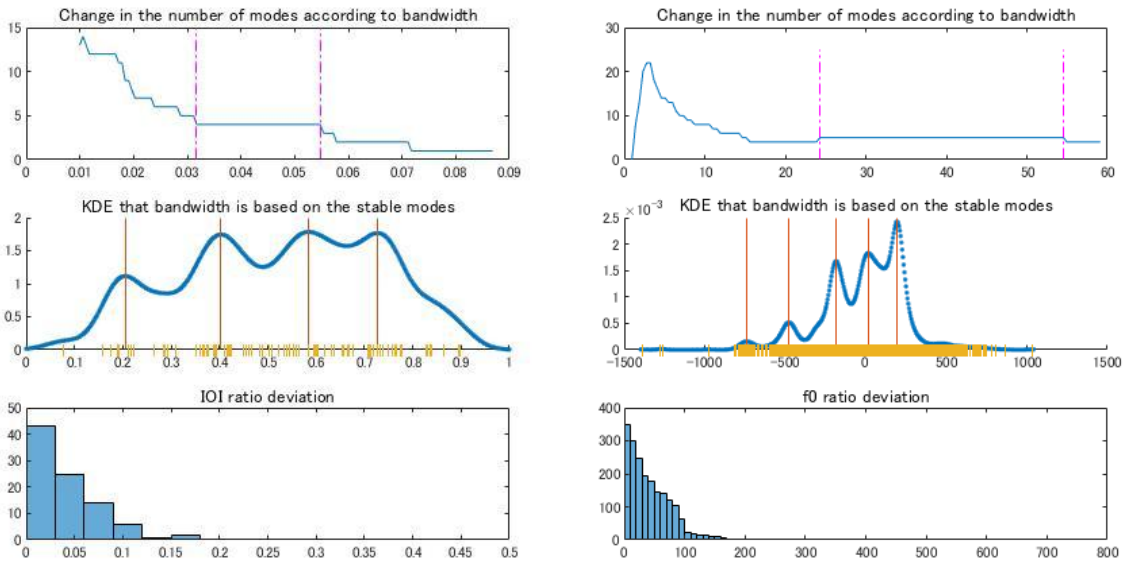


Figure S12. Illustration of the computation of IOI ratio deviation and f_0 ratio deviation. The interval between the magenta lines is the range of the bandwidth parameter that Betti number (number of modes) is most stable which we interpret as indicating the strong persistence of the topological features. Note that due to the removal of data points from the low density region, the number of modes does not simply monotonically decrease with the increase in the bandwidth parameter.

S7. Manipulation of features to demonstrate our designated SESOI (Cohen's D = 0.4).

Following Brysbaert's (2019) recommendation, we use the relative effect corresponding to 0.4 of Cohen's D as the SESOI for our hypothesis testing. Although the choice of 0.4 of Cohen's D is somewhat arbitrary, we empirically measured how much such differences

correspond to the physical attribute of audio using our pilot data focusing on pitch height and temporal rate. For each pair of singing and spoken description recording, we first measured the relative effect (3rd column: Relative effect (p_{re})). Then, we manipulated the corresponding feature of the song to result in a relative effect equal to 0.61 (corresponding to 0.4 of Cohen's D) and 0.5 (corresponding to no difference, 0.0 of Cohen's D). Specifically, we shifted down the entire f_0 for pitch height and slowed down the playback speed for temporal rate. The 4th and 5th columns show actual scale factors identified at each recording and feature. For example, the first row indicates the f_0 of the sung version needed to be shifted 730 cents downward to manipulate the difference in this feature between singing and spoken description to be as small as our proposed SESOI of Cohen's D = .4. Similarly, the sixth row indicates the IOIs of singing needed to be multiplied by 0.472 (i.e., each sung note sped up to be 47.2% as short as the original duration) to make no difference against the spoken description recording, meaning the playback speed of singing should be over 2x faster than the the original recording. Although there are only 5 recording pairs and this measurement does not directly provide the justification for using 0.4 of Cohen's D, we can see how the current SESOI threshold corresponds to the physical attribute of audio by comparing the 4th and 5th columns (106 cents for pitch height and factor of 0.091 for temporal rate in average), which to we authors seems reasonable borderlines for listeners to notice the change in audio content. The corresponding audio examples are available in our OSF repository (https://osf.io/mzxc8/files/osfstorage/638491c81daa6b1394759086).

Table S1. Overview of our pilot recordings with key features (pitch height [f0] and temporal rate [1/IOI]) manipulated to demonstrate what real examples of song and speech might sound like if they the differences were non-existent ("equivalence") or negligible (as small as our chosen SESOI [Smallest Effect Size Of Interest]).

Vocalizer	Feature	Relative effect (p_{re})	Manipulation to demonstrate SESOI (p_{re} = 0.611)	Manipulation to demonstrate equivalence (p_{re} = 0.5)
D. Sadaphal (Marathi)	f_0	0.992	-730 cents (i.e., pitch is transposed down such that sung pitch is more than half an octave lower than the original)	-860 cents
Nweke (Yoruba)	f_0	0.995	-930 cents	-1030 cents
McBride (English)	f_0	0.931	-650 cents	-770 cents
Hadavi (Farsi)	f_0	0.978	-430 cents	-480 cents
Ozaki (Japanese)	f_0	0.997	-1300 cents	-1430 cents
D. Sadaphal (Marathi)	IOI	0.931	x 0.544 (i.e., playback speed is increased by almost 2x such that the duration of each sung note is only 54.4% as fast as the original)	x 0.472
Nweke (Yoruba)	IOI	0.831	x 0.622	x 0.499

McBride (English)	IOI	0.836	x 0.530	x 0.415
Hadavi (Farsi)	IOI	0.932	x 0.396	x 0.324
Ozaki (Japanese)	IOI	0.939	x 0.393	x 0.320

Appendix 1 Recording protocol

We study how and why song and speech are similar or different throughout the world, and we need your help! We are recruiting collaborators speaking diverse languages who can record themselves singing one short (minimum 30 second) song excerpt, recitation of the same lyrics, spoken description of the song, and an instrumental version of the song's melody. In addition, we ask collaborators to include a transcribed text that segments your words according to the onset of the sound unit (e.g., syllable, note) that you feel reasonable. **The recording/transcription/segmentation process should take less than 2 hours.** (Later we will ask you to check sound recordings that we produce based on your segmented text, which may take up to 2 more hours.)

Collaborators will be **coauthors** on the resulting publication, and will also be **paid a small honorarium** (pending the results of funding applications). In principle, all audio recordings will be published using a CC BY-NC 4.0 non-commercial open access license, but exceptions can be discussed on a case-by-case basis (e.g., if this conflicts with taboos or policies regarding indigenous data sovereignty). We seek collaborators aged **18 and over** who are speakers of diverse 1st/heritage languages.

Once you have finished the recordings and created the segmented text files, please:

- **email us your text files (but NOT your audio recordings) to psavage@sfc.keio.ac.jp and yozaki@sfc.keio.ac.jp.**
- email your **audio recordings to globalsongspeech@gmail.com**, where they will be securely monitored and checked by our RA, Tomoko Tanaka, who is not a coauthor on the manuscript.

This folder shows an example template of one full set of recordings and text files**:**
https://drive.google.com/drive/folders/1qbYpv_gxy-gQTBpATA3WwtPHkj14-lSU?usp=sharing

If you have any questions about the protocol, please email:

- Dr. Patrick Savage (psavage@sfc.keio.ac.jp), Associate Professor, Keio University
- Yuto Ozaki (yozaki@sfc.keio.ac.jp), PhD student, Keio University

---

[Recording content]

- Please choose one traditional song to record. This should be a song you know how to sing that is one of the oldest/ most "traditional" (loosely defined)/ most familiar to your cultural background. This might be a song sung to you as a child by your parents/relatives /teachers, learned from old recordings, etc**.** (we plan to include other genres in future stages). Since there is no universally accepted definition of "song" (which is an issue we hope to address in this study), you are free to interpret "song" however feels appropriate in your language/culture. Please contact us if you would like to discuss any complexities of how to define/choose a "traditional song".

- Please choose a song that you can record yourself singing for a **minimum of 30 seconds**. However, we encourage you to record yourself for as long as makes sense for your song to enable more in-depth future studies without having to go back and re-record yourself (though we request you keep within a maximum of 5 minutes if possible). Note that it is fine if it takes less than 30 seconds to recite the same lyrics when spoken, but please ensure that your free spoken description also lasts a minimum of 30 seconds.

- Please use your **1st/heritage language for every recording** (except for the instrumental track). If you speak multiple languages, please choose one language (and let us know which one ahead of time) and avoid combining multiple languages in singing, recitation and spoken description.

- Please record song, lyric recitation, spoken description and instrumental in the order that you feel natural.

    - **Song:** When you sing, please sing solo without instrumental accompaniment, in a pitch range that is comfortable to you. You do not need to follow the same pitch range sung by others. Feel free to sing while reading lyrics/notation if it is helpful.
    - **Lyric recitation:** When you recite the lyrics, please speak in a way you feel is natural. Feel free to read directly from written lyrics if it is helpful.
    - **Spoken description**: Please describe the song you chose (why you chose it, what you like about it, what the song is about, etc.). However, please avoid quoting the lyrics irn your description. Again, aim for **minimum 30 seconds.**
    - **Instrumental version:** Please also record yourself playing the melody of your chosen song(s). We would be delighted for you to play with a traditional instrument in your culture or country. Continuous-pitch instruments (e.g., violin, trombone, erhu) are especially helpful, but fixed-pitch instruments (e.g., piano, marimba, koto) are fine, too. Please do not use electronic instruments (e.g. electric keyboard). Choose whatever pitch/key is comfortable for you to play (this need not be the same pitch/key as the sung version). Please contact us if you want to discuss any complexities involved in trying to play your song's melody on an instrument.
        - ➢ If you do not play a melodic instrument, it is also acceptable to just record the song's rhythm using tapping sounds or other percussive sounds (e.g., drums). In this case, this "instrumental" recording will only be used to analyze rhythmic features. In this case, you can tap the rhythm while singing in your head, but please do not sing out loud.

[Recording method]

- Please record in a quiet place with minimal background noise.

- Please record each description/recitation/song/instrumental separately as different files. The file name should be "[Given name]_[Surname]_[Language]_Traditional_[Song title]_[YYYYMMDD of the time you record]_[song|recit|desc|inst].[file format]". For example,

    - Yuto_Ozaki_Japanese_Traditional_Sakura_20220207_song.wav
    - Yuto_Ozaki_Japanese_Traditional_Sakura_20220207_recit.wav
    - Yuto_Ozaki_Japanese_Traditional_Sakura_20220207_desc.wav
    - Yuto_Ozaki_Japanese_Traditional_Sakura_20220207_inst.wav

- **Please ensure that your mouth (or instrument) is the same distance from your recording device for each recording, and please make all recordings during one session (to avoid differences in recording environment and/or your vocal condition on that day).**

- Regarding the recording device, a high-quality microphone would be great, but a smartphone or personal computer built-in microphone is also fine. Preferred formats are: .mp4, .MOV, .wav, with sampling rate: 44.1kHz or higher / bit rate: 16bit or higher for .wav and lossless codecs (e.g. Apple Lossless Audio Codec) and 128kbps or higher for .MOV and .mp4 with lossy compression codecs. If you are an iPhone user and considering using the Voice Memos app, please set the "Audio Quality" configuration to "Lossless".

- Note: although we only require and will only publish audio data for the main study, we have found that default audio quality can be higher when recording video via smartphone than when recording audio. Also, when it comes time to publish the findings with accompanying press releases, we plan to ask for volunteers who want to share videos of their own singing/speaking. So if you want to make your initial recordings using video, it may save time if you decide you want to volunteer video materials later on.

[Segmented texts]

- After the recording of spoken description, lyric recitation or song, please create a Word file or Rich Text Format file per recording that segments your utterance based on the onset of acoustic units (e.g., syllable, note) that you feel natural. It is up to you how you divide song/speech into what kind of sound unit.
    - Technically, we would like you to focus on the perceptual center or "P-center" (Morton, Marcus, & Frankish, 1976), which is "the specific moment at which a sound is perceived to occur" (Danielsen et al., 2019).
    - Segmentation by the acoustic unit of language (e.g. syllable, mora), by the acoustic unit of music (e.g. note, 節 fushi), and by the P-center are not necessarily the same. For example, one syllable may sometimes be sung across multiple notes (and vice versa).
- **Please use a vertical bar ("|") to segment recordings (see examples below).**
- Please use romanization when writing and also write it based on the phoneme in your native script if it doesn't use Roman characters. You may use IPA (International Phonetic Alphabet) instead of romanization if you prefer.
- Please start a new line in the segmented text at the position where your utterance has a pause for breathing
- When there are successive sound units that keep the same vowels (e.g. "melisma" in Western music, "kobushi" in Japanese music, etc.) and you feel have separate onsets, then you can segment the text by repeating vowels (e.g. A|men → A|a|a|a|men).
- Please include a written English translation of the text of the spoken description and the sung lyrics.

- Example (Japanese)
    - Singing of Omori Jinku
      (Segmented texts with romanization)
      Ton|Bi|Da|Ko|Na|Ra|Yo|O|O|O
      I|To|Me|Wo|O|Tsu|Ke|E|Te
      Ta|Gu|Ri|Yo|Se|Ma|Su|Yo|O|O
      I|To|Me|Wo|O|Tsu|Ke|E|Te

      Hi|Za|Mo|To|Ni|I|Yo|O
      Ki|Ta|Ko|Ra|Yoi|Sho|Na

      (Original lyrics)
      鳶凧ならヨ　糸目をつけて

(コイコイ)
手繰り寄せますヨ　膝元にヨ
(キタコラヨイショナ)

(English translation of the lyrics)
Tie the bridle of a kite kite (Tonbi-dako), pull it in to your knees.
(Kita-ko-ra Yoi-sho-na)

- Lyrics recitation of Omori Jinku
  (Segmented texts with romanization)
  Ton|Bi|Da|Ko|Na|Ra|Yo
  I|To|Me|Wo|Tsu|Ke|Te
  Ta|Gu|Ri|Yo|Se|Ma|Su|Yo
  Hi|Za|Mo|To|Ni|I|Yo
  Ki|Ta|Ko|Ra|Yoi|Sho|Na

- Spoken description of Omori Jinku
  (Segmented texts with romanization)
  E-|Wa|Ta|Shi|Ga|E|Ran|Da|No|Ha, |Oo|Mo|Ri|Jin|Ku, |To|Iu, |E-, |Tou|Kyou|No|Min|You|De|Su.
  Oo|Mo|Ri|To|Iu|No|Ha|Tou|Kyou|No|Ti|Mei|De,
  I|Ma|Wa|Son|Na|O|Mo|Ka|Ge|Ha|Na|In|Desu|Ke|Re|Do|Mo
  Ko|No|U|Ta|Ga|U|Ta|Wa|Re|Te|I|Ta|To|Ki|Ha,|Sono,|No|Ri|Ga,|Ni|Hon|De|I|Ti|Ban|To|Re|Ru|Ba
  |Sho|To|Iu|Ko|To|De,
  Maa|Wa|Ri|To|So|No,|Kai|San|Bu|Tsu|De|Nan|Ka|Yuu|Mei|Na, |Ti|I|Ki|Dat|Ta|Mi|Ta|I|De|Su.
  Kyo|Ku|No|Ka|Shi|Mo,
  E-, |Sou|Des|Ne, |Ho|Shi|Za|Ka|Na, |To|Ka, |Sou|Iu|Ki-|Wa-|Do|Ga|De|Te|Ki|Ma|Su.

  (Original spoken description)
  えー、私が選んだのは、大森甚句、という、えー、東京の民謡です。
  大森というのは東京の地名で、
  今はそんな面影はないんですけれども
  この歌が歌われていたときは、その、海苔が、日本で一番取れる場所ということで、
  まぁ割とその、海産物でなんか有名な、地域だったみたいです。
  曲の歌詞も、
  えー、そうですね、干し魚、とか、そういうキーワードが出てきます。

  (English translation of the spoken description)
  Ah, the song I chose is entitled Omori-Jinku, ah, a Minyo song from Tokyo. Omori is the name of a place in Tokyo, and it has changed a lot these days, but in those days when this song was sung, the place was known for producing the largest amount of nori (seaweed) in Japan, and it also seemed popular due to seafood. Speaking of the lyrics of the song, ah, yeah, like dried fishes, such keywords appear.

- Example (English)

  - Singing of Scarborough Fair
    (Segmented texts with romanization)
    Are |you |go|ing |to |Scar|bo|rough |Fair
    Pars|ley, |sage, |rose|ma|ry |and |thyme
    Re|mem|ber |me |to |one |who |lives |the|ere
    She |once |was |a |true |love |of |mine
    Tell |her |to |make |me |a |cam|b|ric |shirt
    Pars|ley |sage, |rose|ma|ry |and |thyme
    With|out |no |seam |or |nee|dle|wo|ork
    Then |she'll |be |a |true |love |of |mine

  - Lyrics recitation of Scarborough Fair
    (Segmented texts with romanization)

Are |you |go|ing |to |Scar|bo|rough |Fair
Pars|ley, |sage, |rose|ma|ry |and |thyme
Re|mem|ber |me |to |one |who |lives |there
She |once |was |a |true |love |of |mine
Tell |her |to |make |me |a |cam|bric |shirt
Pars|ley |sage, |rose|ma|ry |and |thyme
With|out |no |seams |nor |nee|dle|work
Then |she'll |be |a |true |love |of |mine

- Spoken description of Scarborough Fair
  (Segmented texts with romanization)
  For |my |tra|di|tio|nal |song |I'm |gon|na |sing |Scar|bo|rough |Fair,|
  um, |be|cause |it |is |one |of |the |ol|dest|
  songs |that |is, |uh, |quite |well |known |be|cause |it |was, |ah, |made |po|pu|lar |by, |ah, |Paul |Si|mon |and |Art |Gar|fun|kle.|
  Um,
  and |it |al|so |has |this |nice |kind |of |haun|ting,|
  beau|ti|ful |me|lo|dy |with |this, |uh, |nice |Do|ri|an |scale |that |gives |it |this |kind |of |old |fa|shioned |feel |that |I |quite |like.|
  And |then |the, |the |mea|ning |is |quite |um, |ah, |In|t'res|ting,|
  has |this |kind |of |strange,|
  um, |im |pos|si|ble |rid|dle |kind |of |theme |where |the,|
  ah, |cha|rach|ter |keeps |as|king |the, |um,|
  o|thers |to |do |these |im|pos|si|ble |things, |so |it's |kind |of |this|
  cryp|tic, |old|fa|shioned |song |that |I, |ah, |I |quite |like.

- Please save the segmented texts of each description/recitation/song separately as different files. The file name should be "[Given name]_[Surname]_[Language]_Traditional_[Song title]_[YYYYMMDD of the time you record]_[song|recit|desc].[file format]". For example,
  - Yuto_Ozaki_Japanese_Traditional_Sakura_20220207_song.docx
  - Yuto_Ozaki_Japanese_Traditional_Sakura_20220207_recit.docx
  - Yuto_Ozaki_Japanese_Traditional_Sakura_20220207_desc.docx
    - ➢ Therefore, you will upload 7 files in total as your deliverables (i.e. 4 audio files and 3 Word/RTF files) in the end.

Appendix 2 Collaboration agreement form[4]

Collaboration agreement form for "Similarities and differences in a global sample of song and speech recordings"

This project uses an unusual model in which collaborators act as both coauthors and participants. All recorded audio data analyzed will come from coauthors, and conversely all coauthors will provide recorded audio data for analysis. Collaborators will be expected to provide data within 2 months of when these are requested. Please do NOT send data now - we are following a Registered Report model where data must not be collected until the initial research protocol has been peer-reviewed and received In Principle Acceptance. We estimate this will be in early 2023, and ask that you provide your audio recordings and accompanying text within 2 months of In Principle Acceptance. We estimate this recording/annotation will take approximately 1-2 hours to complete. This will be followed by an additional 1-2 hours to check/correct the final files we prepare at a later date.

All collaborators reserve the right to withdraw their coauthorship and data at any time, for any reason, until the manuscript has passed peer review and been accepted for publication. In such cases, their data will be immediately deleted from all computers and servers, public and private (though be aware that if this happens after posting to recognized preprint/data servers such as PsyArXiv or Open Science Framework some data may remain accessible). The corresponding authors (Patrick Savage and Yuto Ozaki) also reserve the right to cancel this collaboration agreement and publish without a given collaborator's data and coauthorship if necessary (e.g., if data are not provided according to the agreed timeline, or if an insurmountable disagreement about manuscript wording arises). In such a case, any contributions made will be acknowledged in the manuscript.

Collaborators will be coauthors on the resulting publication, and will also be paid a small honorarium (pending the results of funding applications) unless they choose to waive the honorarium. In principle, all audio recordings will be published as supplementary data with this manuscript and permanently archived via recognized preprint/data servers (e.g., PsyArXiv, Open Science Framework, Zenodo) using a CC BY-NC 4.0 non-commercial open access license, but exceptions can be discussed on a case-by-case basis (e.g., if this conflicts with taboos or policies regarding indigenous data sovereignty). We seek collaborators aged 18 and over who speak a diverse range of 1st/heritage languages.

For analysis, we plan to collect and publish demographic information about each collaborator along with their recordings (language name, city language was learned, biological sex [optional], birth year

---

[4] **NB: This agreement had a different timeline from that eventually adopted, because after beginning the process of scheduled review and discussing the issue of confirmation bias with our editor, we concluded that we needed to modify our planned level of bias control from Level 6** ***("No part of the data that will be used to answer the research question yet exists and no part will be generated until after IPA [In Principle Accepantce] (so-called 'primary RR')")*** **to Level 2** ***("At least some data/evidence that will be used to answer the research question has been accessed and partially observed by the authors, but the authors certify that they have not yet sufficiently observed the key variables within the data to be able to answer the research question AND they have taken additional steps to maximise bias control and rigour (e.g., conservative statistical threshold, recruitment of a blinded analyst, robustness testing, the use of a broad multiverse/specification analysis, or other approaches for controlling risk of bias)"*****; cf. "Registered Reports with existing data").**

We thus had to ask collaborators to record themselves several months earlier than they had originally agreed. Most of them managed to do this, but some did not. Because the number of collaborators who could not meet the revised timeline was small enough not to affect our planned power analyses or robustness analyses, we shared the manuscript with all authors and will incorporate those who had not yet made their recordings in the robustness analyses, along with the other authors who made their recordings after knowing the hypotheses.

[optional]). Providing your biological sex or birth year are optional - if you opt not to include these, we will simply exclude your audio data from exploratory analyses that use these variables. (Though please note that biological sex and age may be guessed from your recordings even if you opt not to answer these questions.)

For compliance purposes, CompMusic Lab ("we" or "us") is the data controller of demographic data and audio recordings we hold about you, and you have a right to request information about that data from us (including to access and verify that data). We would like your informed consent to hold and publish demographic data and recordings that you provide to us. All such data will be treated by us under agreed license terms. Please tick the appropriate boxes if you agree and then sign this form:

- ☐ I agree for my data (audio recordings, written transcriptions, and demographic information [language, city language learned, and biological sex and birth year if provided]) to be used as part of research.
- ☐ I agree to provide my audio recordings and text annotations within 2 months of the Stage 1 protocol's In Principle Acceptance, and to check/correct the final annotated files within 2 months of their preparation.
- ☐ I agree to publish my data under a CC BY-NC 4.0 non-commercial open access license.
    a. (If you do not agree to publish your data under CC BY-NC 4.0 [e.g., for reasons relating to Indigenous data sovereignty]) please state your conditions for sharing your audio recording data.:_____
- ☐ I agree to be a coauthor of the manuscript.
- ☐ I agree for a preprint of the manuscript and accompanying data to be posted to recognized preprint/data servers (e.g., PsyArXiv, Open Science Framework, Zenodo).

If you would like to waive the honorarium, you can also tick this box. If you do not waive the honorarium, we will contact you separately to provide bank account details for the wire transfer after you have provided all data.

- ☐ I choose to waive the honorarium

Name: _____
Affiliation (e.g., Department, University, Country): _____
1st/heritage language(s) spoken: _____
Primary city/town/village(s) where language(s) were learned: _____
[Optional] Biological sex (e.g., male, female, non-binary, etc.):_____
[Optional] Birth year: _____

Appendix 3: Open call for collaboration to the International Council for Traditional Music (ICTM) email list. Adapted versions of this email were also used later in tandem with in-person recruitment at the conferences described in the main text). Note that in later meetings we decided to relax the restriction of one collaborator per language, in part due to difficulties of defining the boundaries separating languages and the desire to maximize inclusion.

From: Patrick Savage <psavage@sfc.keio.ac.jp>
Subject: Call for collaboration on global speech-song comparison
Date: July 15, 2022 9:49:57 JST
To: "ictm-l@ictmusic.org" <ictm-l@ictmusic.org>

Dear ICTM-L members,

I am emailing to inquire if any of you are interested in collaborating on a project comparing speech and song in diverse languages around the world to determine what, if any, cross-culturally consistent relationships exist.

I mentioned this project briefly back in January in response to the discussion about Don Niles' post to this list entitled "What is song?". Since then, we have recruited several dozen collaborators speaking diverse languages (see attached rough map), but would like

to open up the call to recruit more. As you can see from the map, our current recruitment is quite unbalanced, particularly lacking speakers of indigenous languages of the Americas, Oceania, and Southeast Asia. We hope you can help us correct that!

Collaborators will be expected to make short (~30 second) audio recordings of themselves in four ways:
1) singing a traditional song in their native language
2) reciting the lyrics of this song in spoken form
3) describing the meaning of the song in their native language
4) performing an instrumental version of the song's melody on an instrument of their choice (negotiable)
They will also provide written transcriptions of these recordings, segmented into acoustic units (e.g., syllables, notes) and English translations. Later, they will check/correct versions of these recordings created by others with click sounds added to the start of each acoustic unit. Finally, they will help us interpret the results of acoustic comparisons of these recordings/annotations. Our pilot studies suggest that this should all take 2-4 hours for one set of 4 recordings.

Collaborators will be coauthors on the resulting publication, and will also be paid a small honorarium (pending the results of funding applications). In principle, all audio recordings will be published using a CC BY-NC non-commercial open access license, but exceptions can be discussed on a case-by-case basis (e.g., if this conflicts with taboos or policies regarding indigenous data sovereignty).

We seek collaborators aged 18 and over who are native speakers of diverse languages, but we are open to collaborators who are non-native speakers in cases of endangered/threatened languages where there are few native speaker researchers available. During this first stage, we only plan to recruit one collaborator per language, on a first-come first-served basis in principle (in future stages we will recruit multiple speakers per language).

More details and caveats (e.g., how to interpret "traditional" or "song") can be found in a draft protocol here: https://docs.google.com/document/d/1qICFXwew7OEj06dkSoR59TIF7HCmVGcudkenMwHREmM/edit

We actually are not quite ready to begin the formal recording/analysis process yet as we are still working out some methodological and conceptual issues (for which we would also welcome your contributions). The reason I am putting out this call now is that I will be presenting at ICTM in Lisbon next week and I know many of you will also be there, so I wanted to use this chance to reach out in case any of you want to meet and discuss in person in Lisbon.

I'll be mentioning more details about this project briefly during a joint ICTM presentation on "Building Sustainable Global Collaborative Networks" at 9am on July 26th (Session VIA01), and would be delighted to meet anyone interested in collaboration following this session or at any other time during the week of the conference.

Please email me (mentioning your native language[s]) if you're interested in collaborating or in meeting in Lisbon to discuss possibilities!

Cheers,
Pat

Dr. Patrick Savage (he/him)
Associate Professor
Faculty of Environment and Information Studies
Keio University SFC (Shonan Fujisawa Campus)
http://compmusic.info

Stage 2 Supplementary Materials

S8 Break annotation

Break is defined as the end of a continuous sequence of sounds before relatively long pauses. Breaks are used to avoid creating inter-onset intervals that do not include sounds. For vocal recordings, that would typically constitute when the participant would inhale. In the case of instrumental recordings, how to determine break points between instrumental phrases is up to the person who made the recording, but it is expected to indicate pauses during sound production.

S9 Exploratory analysis figures

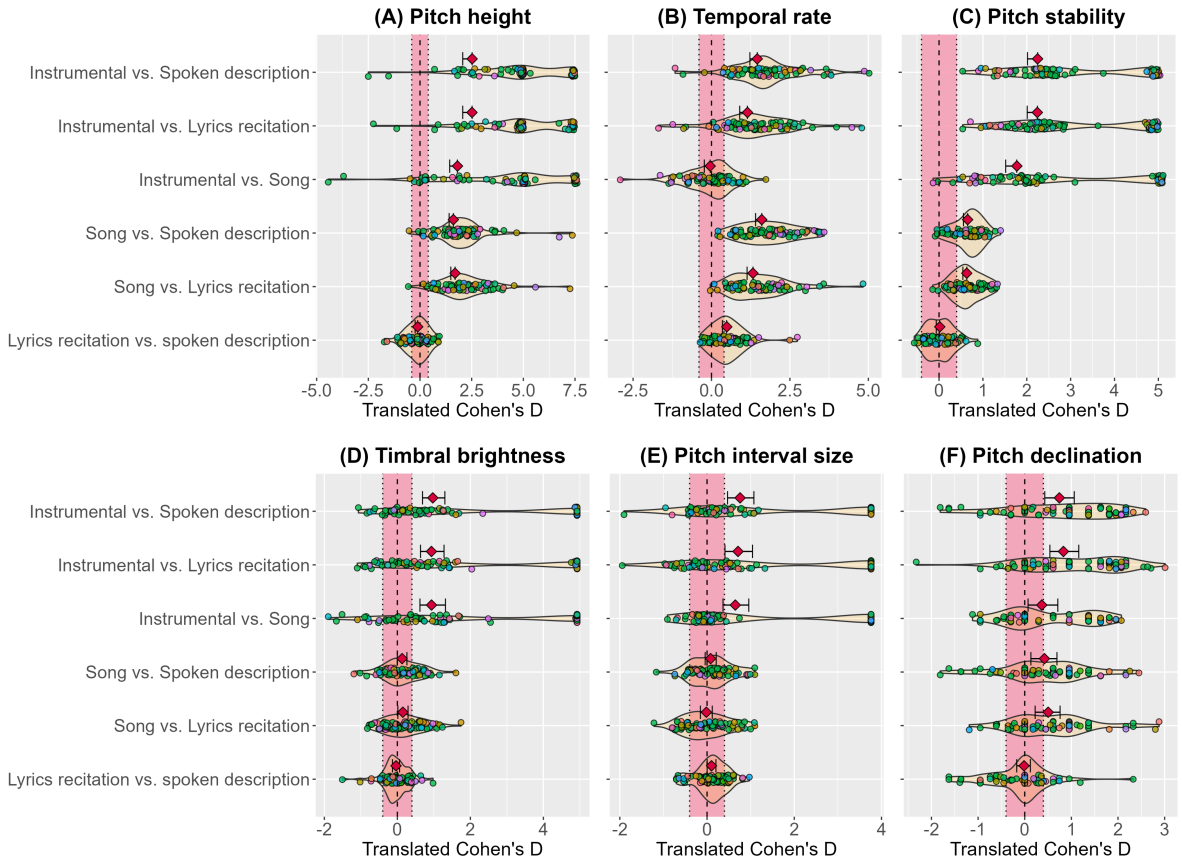


Figure S13. Effect sizes of each feature using the same data as in Figure 5 but with exploratory comparisons with recitation and instrumental recording types.

Table S2. Nonparametric trend test (Jonckheere-Terpstra test) for the shift of mean values of features across different acoustic forms. The category is ordered as 1 = instrumental, 2 = song, 3 = lyrics recitation, and 4 = spoken description. Note that the Jonckheere-Terpstra test assumes observations in each category to be independent of the other categories (e.g., between-subjects design), but our data are collected in a within-subjects design. Therefore, the p-values can be somewhat inaccurate in testing the null hypothesis (i.e., H_0: $\theta_1 = \theta_2 = \theta_3 = \theta_4$) if there is a strong correlation within subjects. The p-values were calculated by a Monte Carlo permutation procedure.

Feature	JT statistics	P-value
Pitch height	6752	1.2×10^{-4}
Temporal rate	27672	1.2×10^{-4}
Pitch stability	3569	1.2×10^{-4}
Timbral brightness	16864	1.2×10^{-4}
Pitch interval size	13340	0.30
Pitch declination	10288	1.2×10^{-4}
Phrase length	10876	1.2×10^{-4}
Intensity	13787	3.7×10^{-4}
TImbral noisiness	22998	1.2×10^{-4}
Rhythmic regularity	23484	1.2×10^{-4}
Pitch interval regularity	20329	1.2×10^{-4}
Pulse clarity	9911	1.2×10^{-4}
Pitch range	13114.5	0.20

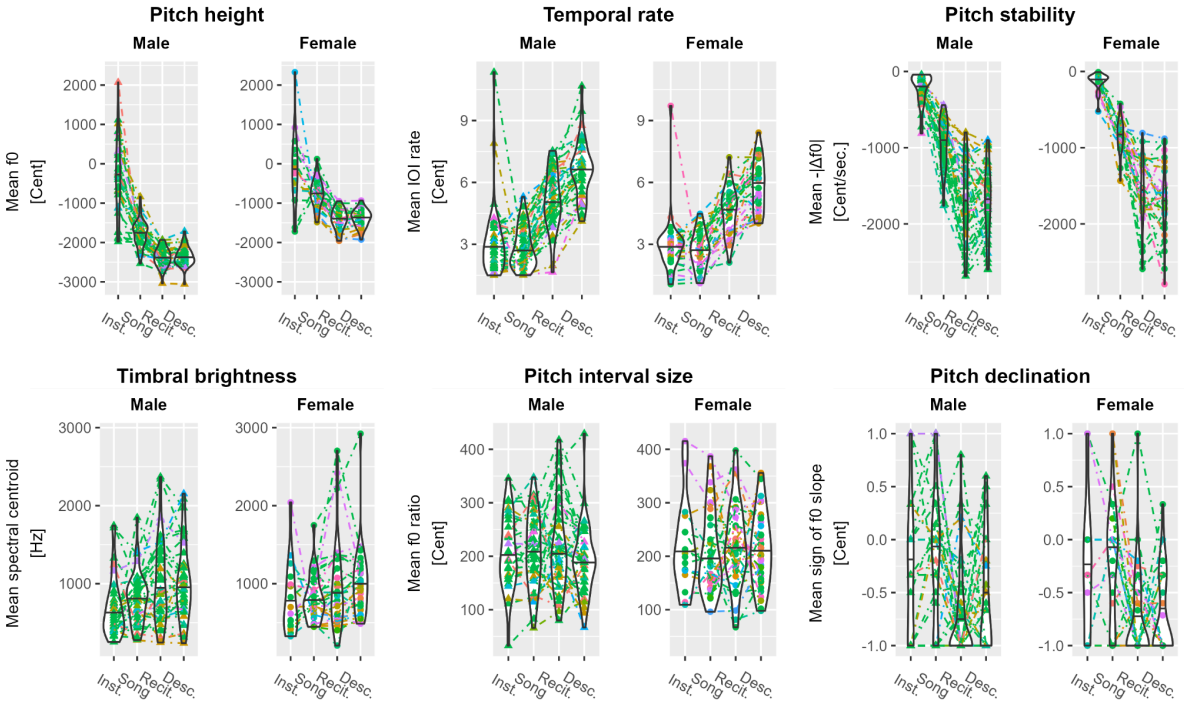


Figure S14. Alternative visualization of Figure 9 showing mean values of each feature by biological sex and focusing on the features subject to the main confirmatory analysis. Note that the colors of data points indicate language families, which are coded the same as in Figure 3

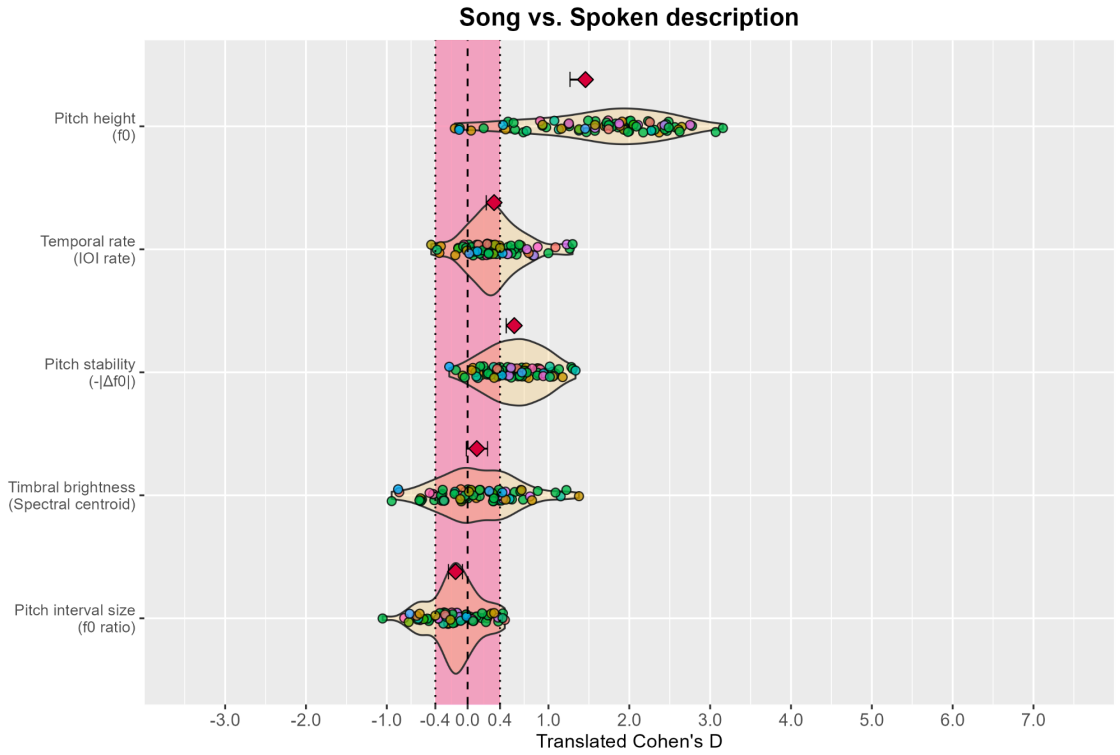


Figure S15. Re-running of the analysis on our full data with automated feature extraction. pYIN (Mauch & Dixon, 2014) was used for f0 extraction and de Jong & Wemp's (2009) Praat script was used for onset timing extraction. Break annotation was not automated so pitch declination was not measured.

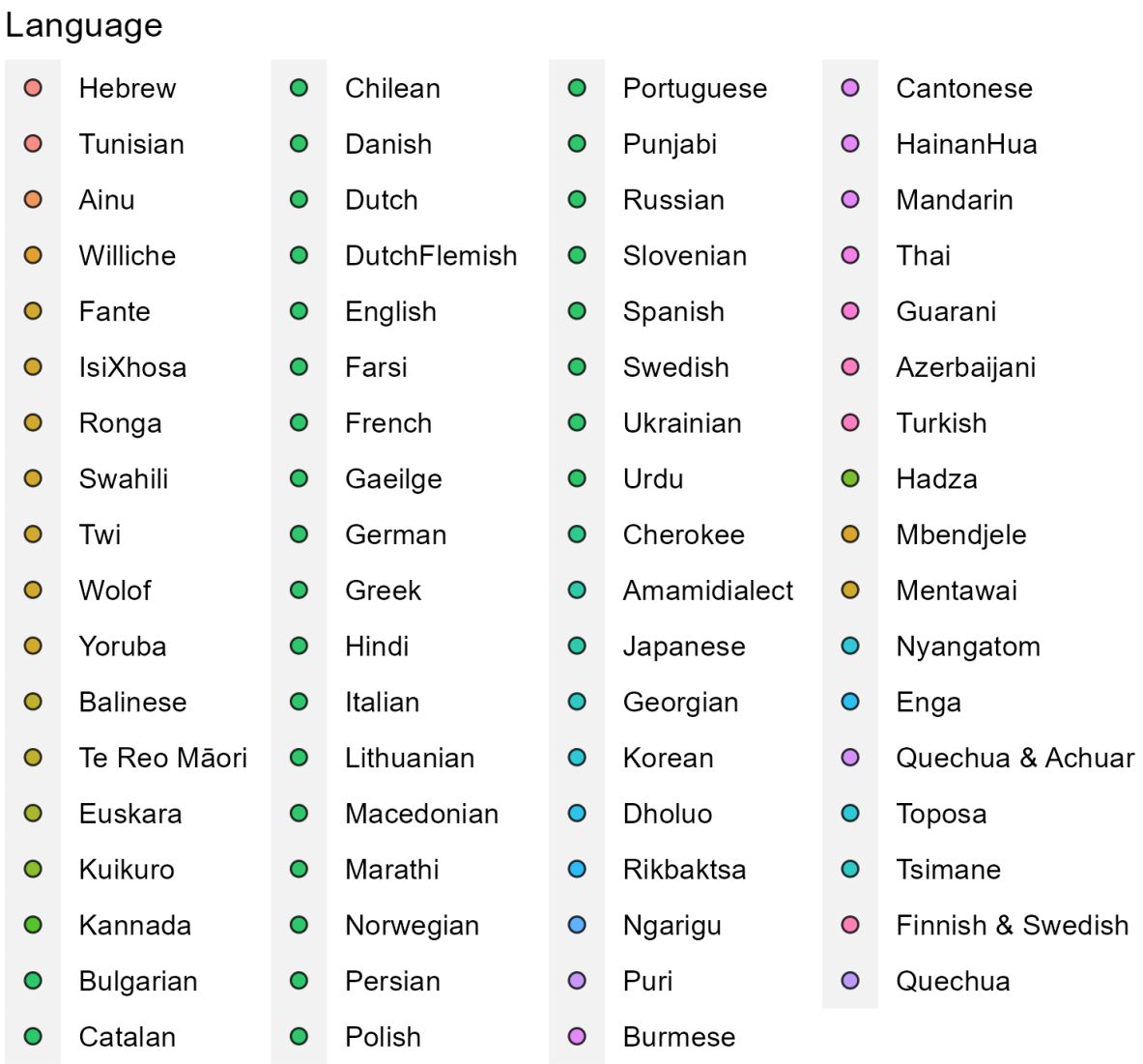


Figure S16. Color mapping of Figure 12.

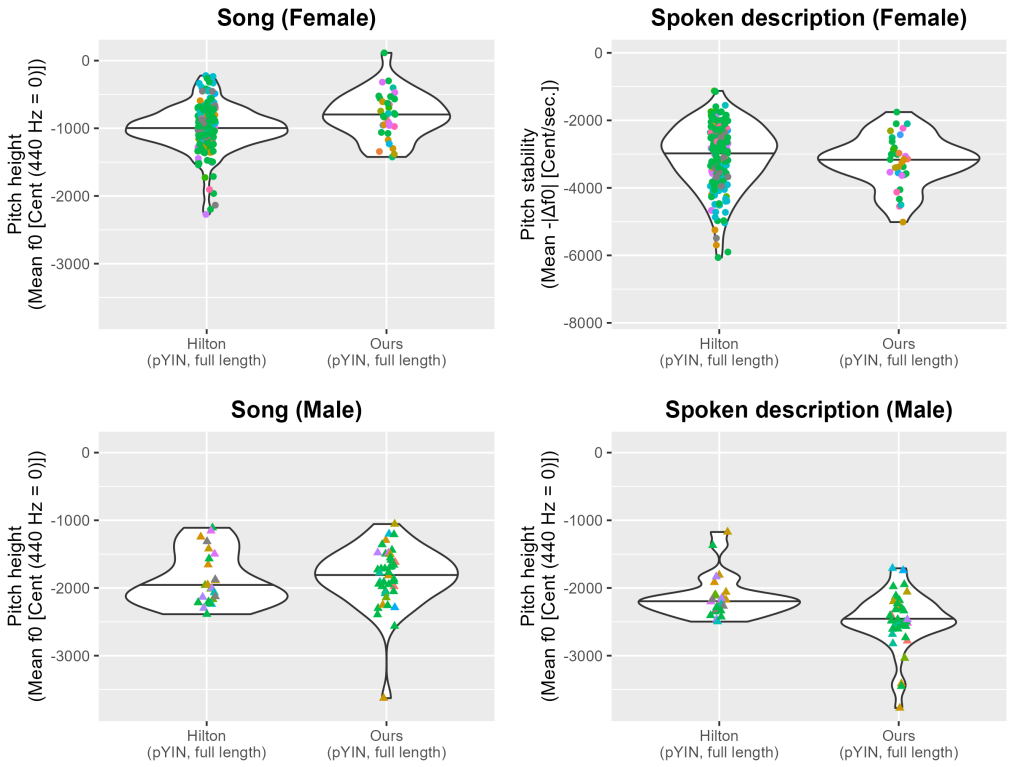


Figure S17. Supplementary information for Fig. 10. Mean values of pitch height of each recording are displayed. f_0s were extracted by pYIN (Mauch & Dixon, 2014). The horizontal lines in the violin plots are median.

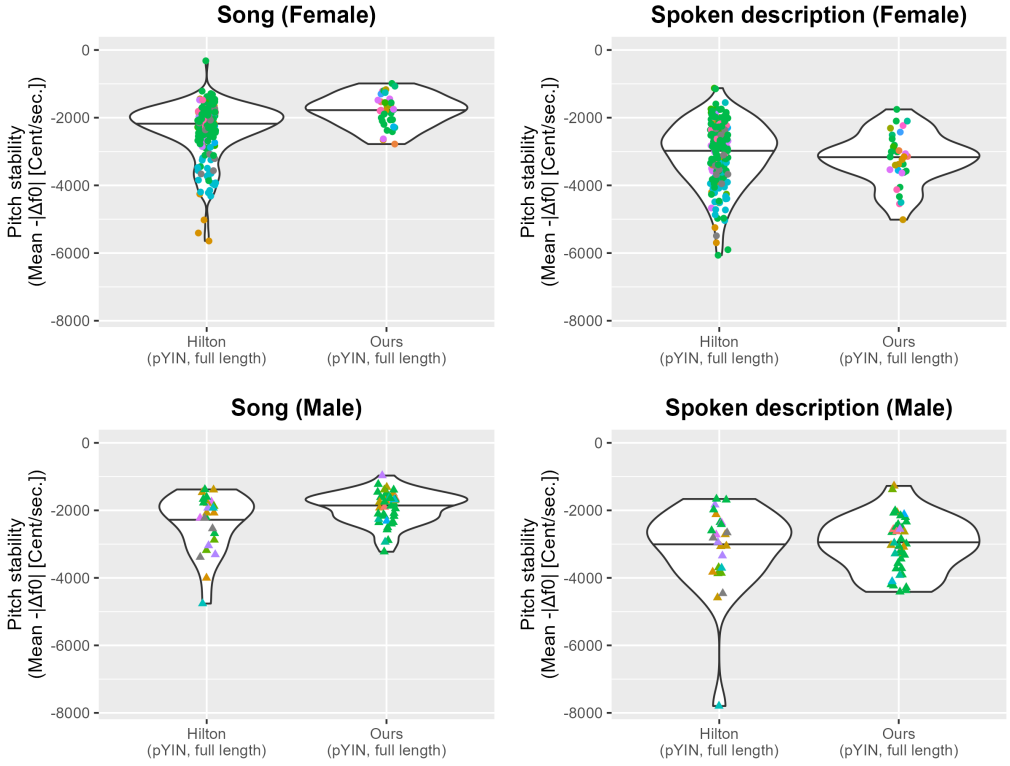


Figure S18. Supplementary information for Fig. 10. Mean values of pitch stability of each recording are displayed. f_0s were extracted by pYIN (Mauch & Dixon, 2014). The horizontal lines in the violin plots are median.

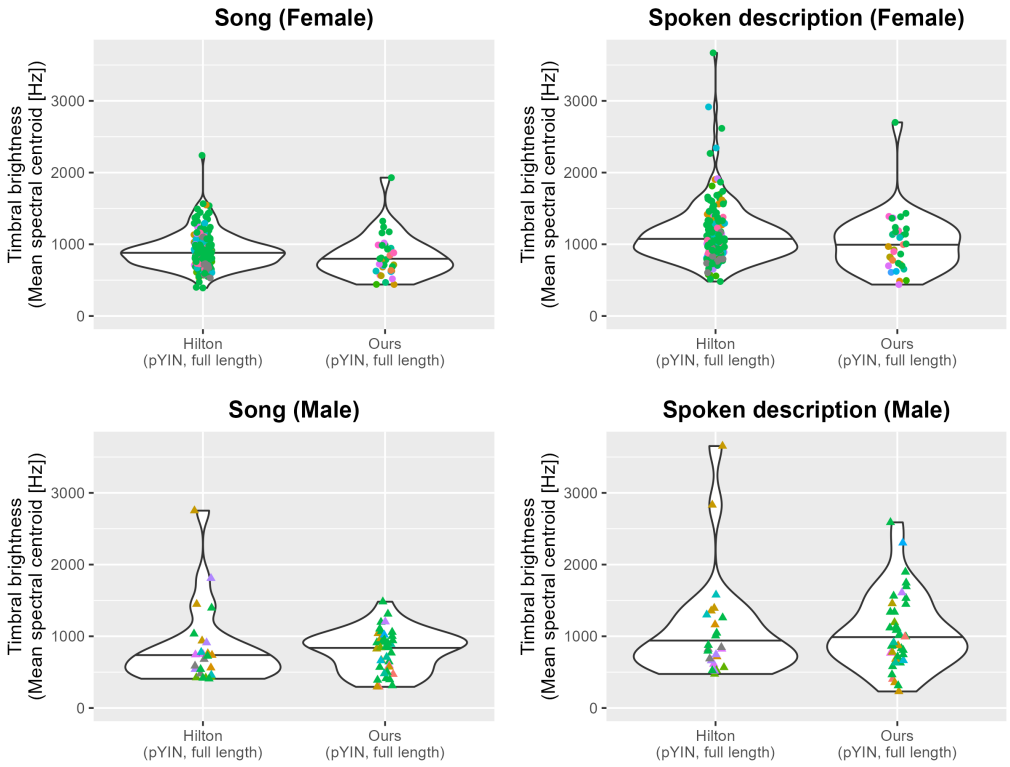


Figure S19. Supplementary information for Fig. 10. Mean values of timbral brightness of each recording are displayed. f_0s were extracted by pYIN (Mauch & Dixon, 2014). The horizontal lines in the violin plots are median.

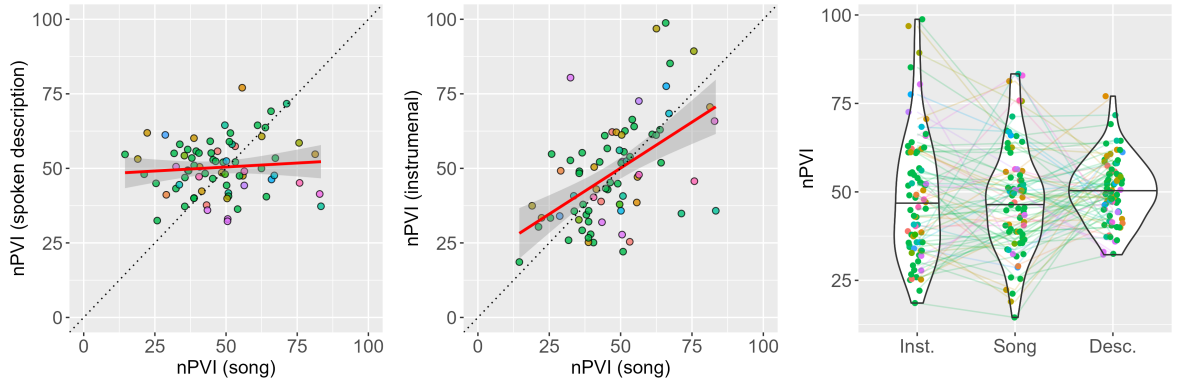


Figure S20. Mapping data by nPVIs of song and spoken description by each collaborator and its song-instrumental version, and the density plot of nPVIs of each . The red lines are linear fitting of nPVIs of spoken description and nPVIs of song, and the dotted line is y = x which can be used to grasp if nPVIs of spoken description is larger than that of song and vice versa.

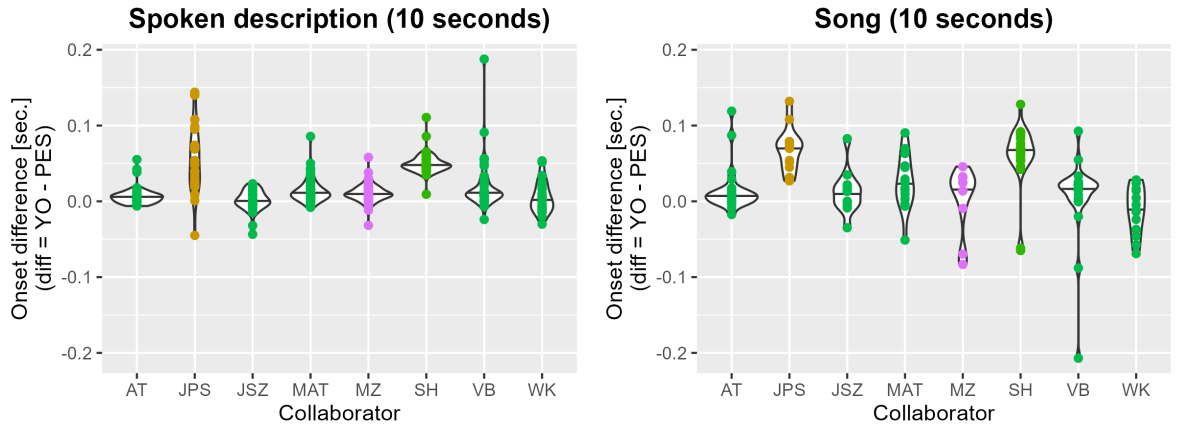


Figure S21. Difference between onset times annotated by Ozaki (YO) and onset times annotated by Savage (PES) per recording for the 8 codings re-annotated by Savage to assess inter-rater reliability. The horizontal lines in the violin plots indicate the median. Color is coded as the same in Fig. 3.

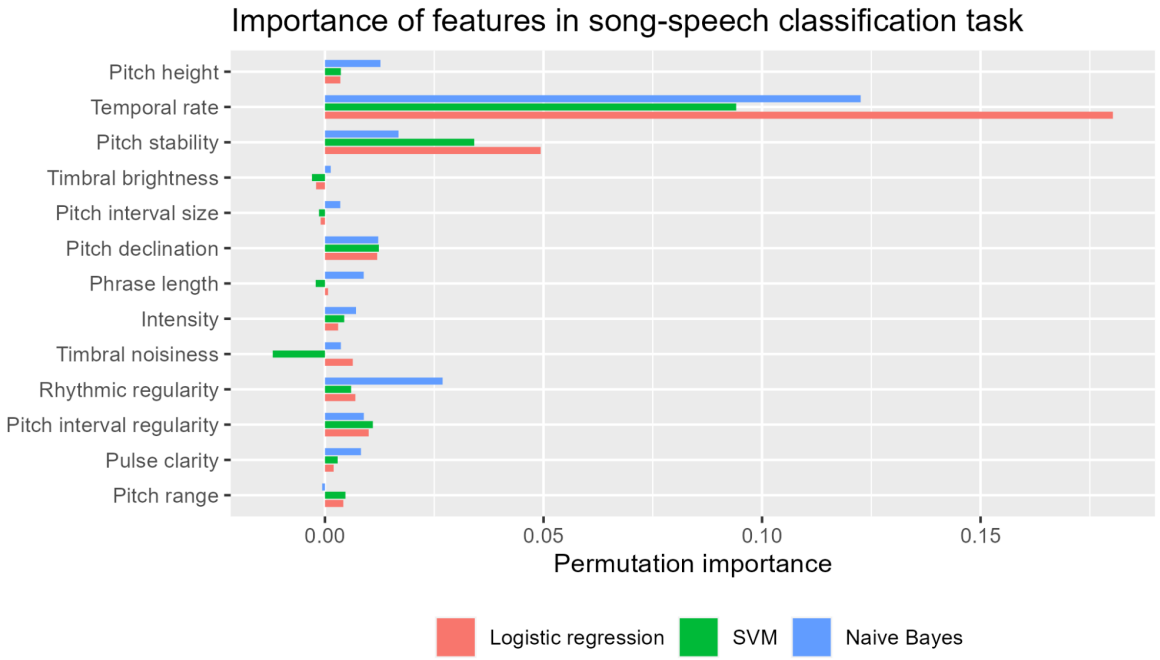


Figure S22. Permutation importance of the features in three binary classifiers.

Table S3. Average over performance metrics measured by randomly splitting recording sets into training and test sets 1024 times.

		Logistic regression	SVM	Naive Bayes
Accuracy		95.78%	93.75%	92.94%
Song	Precision	96.66	92.68	92.81
	Recall	95.25	95.70	93.98
	F1 score	95.72	93.92	93.03
Spoken description	Precision	95.74	95.89	94.45
	Recall	96.31	91.80	91.91
	F1 score	95.80	93.50	92.76

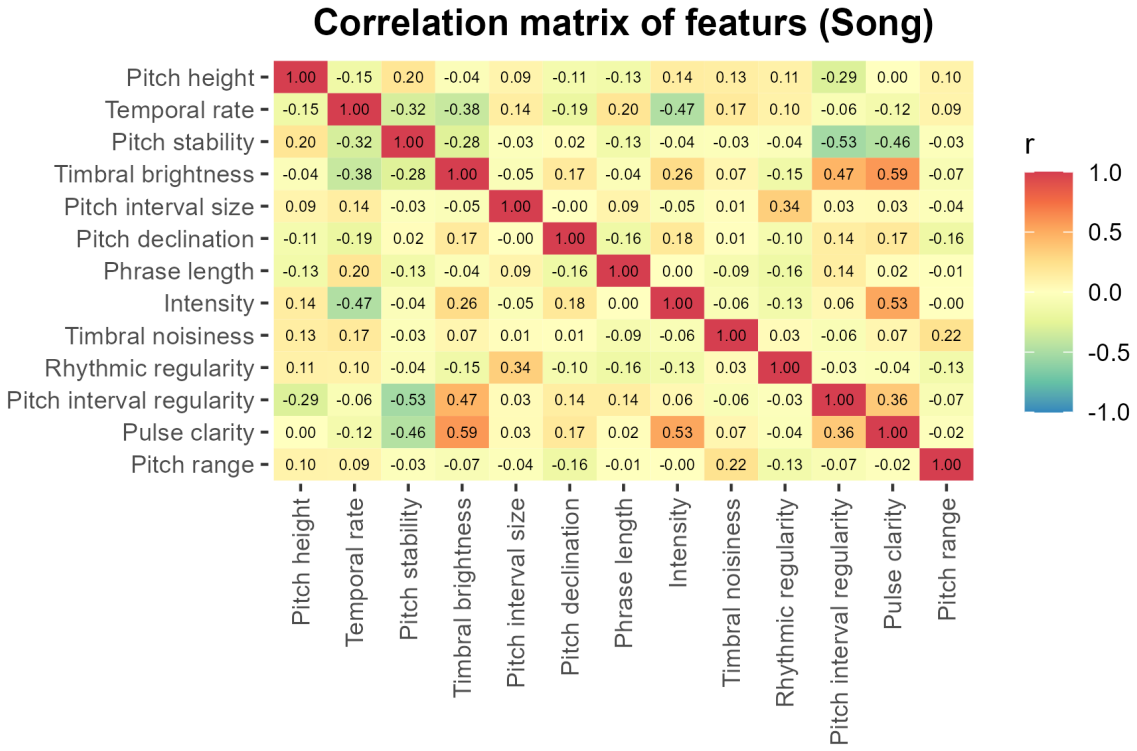


Figure S23. Correlation matrix of the features within song recordings. The data are the mean values of the features, which are plotted in Figure 6.

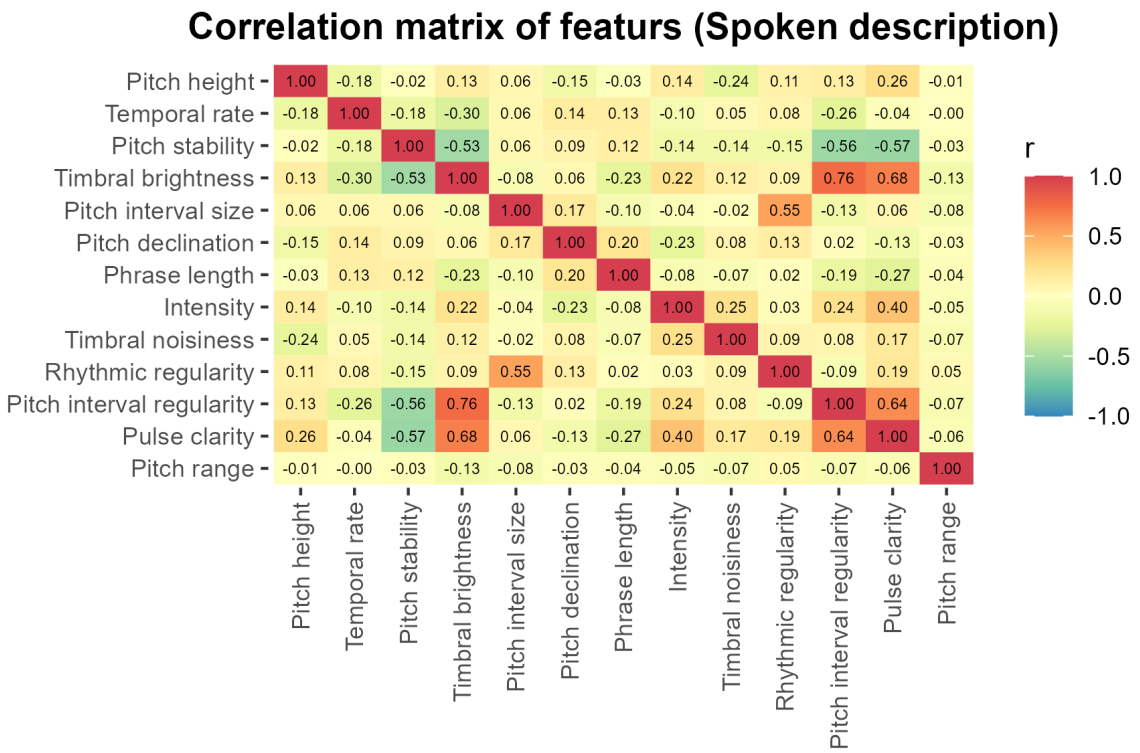


Figure S24. Correlation matrix of the features within spoken description recordings. The data are the mean values of the features, which are plotted in Figure 6.

Appendix 3 List of songs

#	Name	Song title (Romanization)	Language	Instrument
1	Nori Jacoby	Laila Laila	Modern Hebrew [Jerusalem]	Whistle
2	Limor Raviv	ירושלים של זהב (Yerushalayim ShelZahav)	Modern Hebrew [Tel Aviv]	Tapping
3	Iyadh El Kahla	لاموني اللي غاروا مني	Tunisian Arabic	Aerophone
4	Utae Ehara	イタサン (Itasan)	Aynu (Hokkaido Ainu)	Tapping
5	Neddiel Elcie Muñoz Millalonco	Ñaumen pu llauken	Tsesungún (Huilliche)	Clapping
6	Nozuko Nguqu	Ulele	IsiXhosa (Xhosa)	Piano
7	Mark Lenini Parselelo	Lala Mtoto Lala	Kiswahili (Swahili)	Tapping
8	Cristiano Tsope	Hiya Tlanguela xinwanana xinga pswaliwa namuntla	Ronga	Clapping
9	Florence Nweke	Pat omo o	Yoruba	Piano
10	Adwoa Arhine	Yɛyɛ Eguafo	Fante (Akan)	Clapping
11	Jehoshaphat Philip Sarbah	Daa na se	Twi (Akan)	Piano
12	Latyr Sy	Mbeuguel	Wolof	Clapping
13	I Putu Gede Setiawan	Putriceningayu	Balinese	Suling
14	Suzanne Purdy	Pōkarekare Ana	Te Reo Māori (Māori) [Auckland]	Tapping
15	Rob Thorne	Ko Te Pū	Te Reo Māori (Māori) [Wellington]	Kōauau rākau
16	Nerea Bello Sagarzazu	Xoxo Beltza	Euskara (Basque) [Hondarribia]	Aerophone
17	Urise Kuikuro	Toló	Língua Kuikuro (Kuikúro-Kalapálo)	Clapping
18	Shantala Hegde	Moodala Maneya	Kannada	Clapping
19	Rytis Ambrazevičius	Sėjau rugelius	Lithuanian	Idiophone
20	Tadhg Ó Meachair	Éiníní	Gaeilge (Irish)	Piano Accordion
21	Niels Chr. Hansen	I Skovens Dybe Stille Ro	Danish	Piano
22	Mark van Tongeren	Hoor De Wind waait	Dutch [Heemstede]	Piano
23	Kayla Kolff	Dikkertje Dap	Dutch [Nairobi]	Membranophone
24	Adam Tierney	Simple Gifts	English [Indiana]	Electric Piano
25	Christina Vanden	Sleep Now Rest Now	English [Michigan]	Cello

	Bosch der Nederlanden			
26	Patrick Savage	Scarborough Fair	English [Nevada]	Piano
27	John McBride	Arthur McBride	English [Newry]	Flute
28	William Tecumseh Fitch	Rovin’ Gambler	English [Pennsylvania]	Guitar
29	Peter Pfordresher	America the Beautiful	English [Washington D.C.]	Piano
30	Yannick Jadoul	VandaagIs't Sinte Maarten	Flemish (Dutch)	Piano
31	Felix Haiduk	Die Gedanken Sind Frei	German	Melodica
32	Ulvhild Færøvik	Nordmannen	Norwegian	Clapping
33	Daniel Fredriksson	Ho Maja	Svenska (Swedish)	Offerdalspipa
34	Emmanouil Benetos	Saranta Palikaria	Greek	Clapping
35	Dhwani P. Sadaphal	Saraswatee maateshwaree	Hindi	Harmonium
36	Parimal M. Sadaphal	Sukhakartaa	Marathi	Sitar
37	Meyha Chhatwal	ਬਾਜਰੇ ਦਾ ਸਿੱਟਾ (Bajre Da Sitta)	Punjabi (Eastern Panjabi)	Harmonium
38	Ryan Mark David	Dil Dil Pakistan	Urdu	Acoustic guitar
39	Shahaboddin Dabaghi Varnosfaderani	Morgh e Sahar	Western Farsi [Isfahan]	Clapping
40	Shafagh Hadavi	Mah Pishanoo	Western Farsi [Tehran]	Piano
41	Manuel Anglada-Tort	La Presó de Lleida	Catalan	Piano
42	Pauline Larrouy-Maestri	À la claire fontaine	French	Piano
43	Andrea Ravignani	Bella Ciao	Italian	Saxophone
44	Violeta Magalhães	O milho da nossa terra	Portuguese [Porto]	Tapping
45	Camila Bruder	A Canoa Virou	Portuguese [São Paulo]	Tambourine
46	Marco Antonio Correa Varella	Suite do Pescador	Portuguese [São Paulo]	Nose flute
47	Juan Sebastián Gómez-Cañón	El pescador	Spanish [Bogotá]	Guitar
48	Martín Rocamora	Aquello	Spanish [Montevideo]	Guitar
49	Javier Silva-Zurita	Un gorro de lana	Spanish [Santiago]	Guitar
50	Ignacio Soto-Silva	El Lobo Chilote	Spanish [Osorno]	Clapping
51	Dilyana Kurdova	Zarad tebe, mome, mori	Bulgarian	Clapping
52	Aleksandar Arabadjiev	Jovano	Macedonian	Kaval

53	Wojciech Krzyżanowski	Wlazł Kotek Na Płotek	Polish	Guitar
54	Polina Proutskova	Dusha moia pregreshnaia	Russian	Violin
55	Vanessa Nina Borsan	En Hribček Bom Kupil	Slovenian	Tapping
56	Olena Shcherbakova	Podolyanochka	Ukrainian	Piano
57	Diana Hereld	ᎤᏁᎳᏅᎯ ᎤᏪᏥ (unelanvhi uwetsi)	Cherokee	Tapping
58	Gakuto Chiba	津軽よされ節 (Tsugaru-yosarebushi)	Japanese [Hokkaido]	Tsugaru-shamisen (津軽三味線)
59	Shinya Fujii	デカンショ節 (Dekansho-bushi)	Japanese [Hyogo]	Clapping
60	Yuto Ozaki	大森甚句 (Omori-Jinku)	Japanese [Tokyo]	Guitar
61	Naruse Marin	朝花節 (Asabana-bushi)	Northern Amami-Oshima	Sanshin (三線)
62	Teona Lomsadze	Nana (Lullaby)	Georgian	Chonguri
63	Sangbuem Choo	아리랑 (Arirang)	Korean	Guitar
64	Patricia Opondo	Ero Okech Nyawana	Luo (dholuo) (Luo (Kenya and Tanzania))	Whistle
65	Rogerdison Natsitsabui	Jakara Wata	Rikbaktsa	Clapping
66	Jakelin Troy	Gundji gawalgu yuri	Ngarigu	Percussion
67	Tutushamum Puri Righi	Petara	Puri Kwaytikindo (Puri)	Terara (bamboo flute)
68	Su Zar Zar	Mya Man Giri	Myanmar (Burmese)	Saung-gauk
69	Psyche Loui	梁祝 (Butterfly Lovers)	Cantonese (Yue Chinese)	Violin
70	Minyu Zeng	五指山歌 (The Song of the Five-Fingers Mountain)	HainanHua (Min Nan Chinese)	Idiophone
71	Fang Liu	送别 (Farewell)	Mandarin Chinese	Clapping
72	Great Lekakul	ลาวดวงเดือน (Lao Doung Duan)	Thai	"Klui"(ขลุ่ย) (a Thai flute)
73	Brenda Suyanne Barbosa	Apykaxu	Mbyá-Guaraní	Clapping
74	Polina Dessiatnitchenko	Ay Lachin	North Azerbaijani	Tar
75	Olcay Muslu	Uzun Ince Bir Yoldayim	Turkish	Tapping